A quantitative measure for protein conformational heterogeneity

ARTICLE *in* THE JOURNAL OF CHEMICAL PHYSICS · SEPTEMBER 2013

Impact Factor: 2.95 · DOI: 10.1063/1.4812791 · Source: PubMed

3 AUTHORS:

Nicholas James Lyle
Washington University in St. Louis
19 PUBLICATIONS **212** CITATIONS
SEE PROFILE

Rahul K. Das
Washington University in St. Louis
12 PUBLICATIONS **241** CITATIONS
SEE PROFILE

Rohit V Pappu
Washington University in St. Louis
108 PUBLICATIONS **2,789** CITATIONS
SEE PROFILE

AIP | The Journal of Chemical Physics

A quantitative measure for protein conformational heterogeneity

Nicholas Lyle, Rahul K. Das, and Rohit V. Pappu

Citation: J. Chem. Phys. **139**, 121907 (2013); doi: 10.1063/1.4812791
View online: http://dx.doi.org/10.1063/1.4812791
View Table of Contents: http://jcp.aip.org/resource/1/JCPSA6/v139/i12
Published by the AIP Publishing LLC.

A quantitative measure for protein conformational heterogeneity

Nicholas Lyle,[1,a)] Rahul K. Das,[2,b)] and Rohit V. Pappu[2,c)]
[1]*Computational and Systems Biology Program, Division of Biology and Biomedical Sciences, Washington University in St. Louis, One Brookings Drive, Campus Box 1097, St. Louis, Missouri 63130, USA*
[2]*Department of Biomedical Engineering and Center for Biological Systems Engineering, Washington University in St. Louis, One Brookings Drive, Campus Box 1097, St. Louis, Missouri 63130, USA*

(Received 29 April 2013; accepted 18 June 2013; published online 11 July 2013)

Conformational heterogeneity is a defining characteristic of proteins. Intrinsically disordered proteins (IDPs) and denatured state ensembles are extreme manifestations of this heterogeneity. Inferences regarding globule versus coil formation can be drawn from analysis of polymeric properties such as average size, shape, and density fluctuations. Here we introduce a new parameter to quantify the degree of conformational heterogeneity within an ensemble to complement polymeric descriptors. The design of this parameter is guided by the need to distinguish between systems that couple their unfolding-folding transitions with coil-to-globule transitions and those systems that undergo coil-to-globule transitions with no evidence of acquiring a homogeneous ensemble of conformations upon collapse. The approach is as follows: Each conformation in an ensemble is converted into a conformational vector where the elements are inter-residue distances. Similarity between pairs of conformations is quantified using the projection between the corresponding conformational vectors. An ensemble of conformations yields a distribution of pairwise projections, which is converted into a distribution of pairwise conformational dissimilarities. The first moment of this dissimilarity distribution is normalized against the first moment of the distribution obtained by comparing conformations from the ensemble of interest to conformations drawn from a Flory random coil model. The latter sets an upper bound on conformational heterogeneity thus ensuring that the proposed measure for intra-ensemble heterogeneity is properly calibrated and can be used to compare ensembles for different sequences and across different temperatures. The new measure of conformational heterogeneity will be useful in quantitative studies of coupled folding and binding of IDPs and in *de novo* sequence design efforts that are geared toward controlling the degree of heterogeneity in unbound forms of IDPs. © *2013 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4812791]

I. INTRODUCTION

Proteins undergo disorder-to-order transitions either as units that fold autonomously[1] or as intrinsically disordered proteins (IDPs)[2] that couple their folding to binding[3] or self-assembly.[4] The driving forces for and mechanisms of disorder-to-order transitions are governed by the degree of conformational heterogeneity within disordered states and the extent of overlap between conformational ensembles of disordered and ordered states. Therefore, there is growing interest in quantitative studies of disordered states of proteins.[5–8]

Studies of disorder in protein folding are focused on characterizing the ensemble of non-native conformations under denaturing as well as native conditions.[9–11] Of interest are questions pertaining to the degree of conformational heterogeneity,[12,13] the balance between intrachain and chain-solvent interactions that define polymeric properties,[14–16] effects of macromolecular crowding,[17,18] intermolecular interactions that lead to protein aggregation,[19–21] and the timescales for conversion between distinct conformations that contribute to internal friction.[22] Recent interest has also focused on the topic of IDPs. Their sequences encode preferences for heterogeneous ensembles of conformations as the thermodynamic ground state under standard physiological conditions (aqueous solutions, 150 mM monovalent salt, low concentrations of divalent ions, pH 7.0, and temperature in the 25 °C–37 °C range).[23,24] Conformational heterogeneity of IDPs in their unbound forms influences their ability to adopt different folds in the context of binary and multimolecular complexes.[25,26] In IDPs, disorder-to-order transitions are realized by coupling the folding process to either binding or self-assembly providing the heterotypic or homotypic interactions in *trans* can stabilize the IDP in a specific fold. The stabilities of complexes are thermodynamically linked to the ensemble of conformations that IDPs sample as autonomous units.

Thermodynamic descriptions of disorder-to-order transitions require the use of a suitable order parameter. A *bona fide* order parameter has to quantify the symmetry that is broken as a result of the disorder-to-order transition. Proteins are polymers and can expand to form low-density conformations that have large interfaces with the surrounding solvent; alternatively, they can collapse to form high-density conformations that minimize the chain-solvent interface. It is well established that $s^2 = \frac{\langle R_g^2 \rangle}{N}$ is a *bona fide* order parameter for quantifying density changes that accompany coil-to-globule transitions.[27,28] Here, R_g denotes the radius of gyration and

[a)] Electronic mail: lylenj@gmail.com
[b)] Electronic mail: chemrahul82@gmail.com
[c)] Electronic mail: pappu@wustl.edu; Telephone: (314) 935-7958.

N is the chain length. In protein folding, changes in density are also associated with the acquisition of a homogeneous ensemble of conformations. However, s^2 can be used as the sole parameter to monitor folding if and only if proteins follow two-state behavior.[29] Theories, simulations, and experiments have established that while s^2 is extremely important for understanding the convolution of coil-to-globule transitions with protein folding, it is inadequate for providing a complete description of transitions between unfolded and folded states.[30–37] Recent simulations and experiments have also shown that several IDPs undergo collapse to form globules under standard physiological conditions.[38–44] This preference for globules can be reversed through increases in temperature,[45] net charge per residue,[46,47] or concentrations of chemical denaturants.[39] Collapse in globule-forming IDPs does not have to imply the acquisition of a homogeneous ensemble of conformations. These results highlight the need for additional parameters that report on overall conformational heterogeneity.

Figure 1 summarizes the temperature dependence of s^2 and densities that are obtained from atomistic simulation results for five archetypal systems. Details of the simulations that were used to generate the temperature dependent profiles are discussed in Sec. II. The N-terminal domain of the ribosomal L9 protein (NTL9) and the B1 domain of protein G (GB1) undergo unfolding transitions as temperature increases. This unfolding is also associated with chain expansion as shown in Figure 1. We compare these profiles to the temperature dependence of s^2 and density (ρ) for three homopolymeric systems. These are polyproline (P_{56}), which is intrinsically stiff, polyarginine (R_{56}), which is a highly charged, rod-like polyelectrolyte, and polyglutamine (Q_{56}), which is an intrinsically disordered polar tract. P_{56} shows weak chain contraction as temperature increases. This feature is consistent with the so-called inverse transition temperature[48] that has been observed for poly-L-proline polymers and derives, partially, from the increased fraction of *cis* peptide bonds at higher temperatures.[49] R_{56} and Q_{56} show distinct limiting behaviors; the former maintains its rod-like behavior across all temperatures whereas the latter undergoes a globule-to-coil transition as temperature increases. Despite undergoing reversible coil-to-globule transitions, previous simulations and experimental studies demonstrate that collapse does not imply the acquisition of an ensemble of a homogeneous ensemble of conformations, i.e., collapse does not imply folding.[50,51] This in turn implies that while the temperature dependence of s^2 provides information regarding coil-to-globule transitions, it fails to discriminate between systems such as NTL9 and GB1 on the one hand and Q_{56} on the other. Analysis of the temperature dependence of the specific heat capacities (Figure 2), which reports on the temperature dependence of the energy variance, does not provide any additional information that cannot be obtained by analyzing the temperature dependence of density fluctuations.

In the parlance of energy landscape theory,[52–54] a system such as polyglutamine has a rugged landscape below its collapse transition temperature.[55] Indeed, such a scenario has been predicted for IDPs[56,57] and random polypeptide sequences.[58,59] This ruggedness is not registered in measures such as estimates of density or energy fluctuations because distinct conformations of equivalent compactness have negligible energy differences and hence equivalent likelihoods of being accessed. In this scenario, both the energy and density fluctuations will be small and the sharpness of the change in energy and density fluctuations masks the fact that the globule-to-coil transition in a system like polyglutamine might actually be a "disorder-to-disorder" transition where the transition is between distinct classes of heterogeneous conformational ensembles.

In order to detect putative disorder-to-disorder transitions that are masked when analyzing s^2 or densities, we need a measure for heterogeneity within a conformational ensemble. For this we introduce a parameter Φ whose design is guided by the need to distinguish between systems that couple their unfolding-folding transitions with coil-to-globule transitions and those systems that undergo coil-to-globule transitions with no evidence of acquiring a homogeneous ensemble of conformations upon collapse. The design of Φ is also

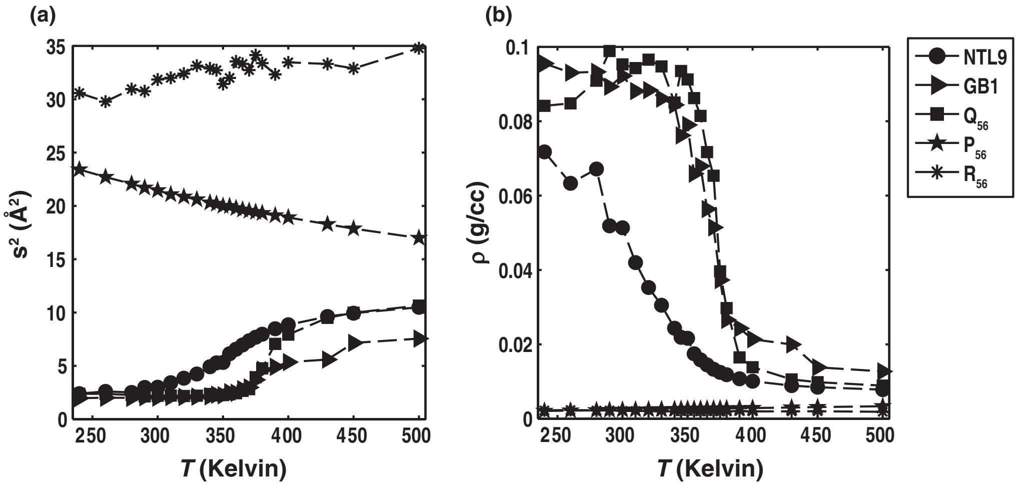

FIG. 1. Temperature dependence of s^2 and density for five archetypal systems. Panel (b) quantifies the temperature dependence of chain density (in units of gm-cm^{-3}), which is calculated as $\langle\rho\rangle = \frac{\mathrm{MW}}{\left\langle R_g^2\right\rangle^{\frac{3}{2}}}$, where MW denotes the molecular weight in gm mol^{-1}.

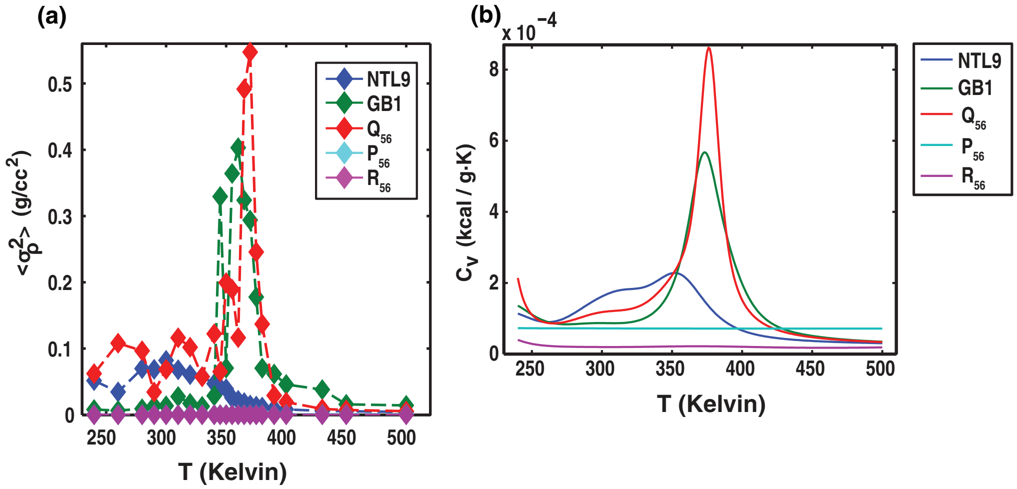

FIG. 2. Temperature dependence of fluctuations in density and energy for five archetypal systems. Panel (a) shows the temperature dependence of the density fluctuations quantified as the variance of the density distribution for a given temperature, i.e., $\sigma_\rho^2 = \langle \rho^2 \rangle - \langle \rho \rangle^2$. Panel (b) shows the temperature dependence of the specific heat capacity. The specific, constant volume heat capacities were calculated as $C_V = \frac{1}{\mathrm{MW}}\left(\frac{\partial \langle E \rangle}{\partial T}\right)_V$, where MW is the molecular weight and $\langle E \rangle$ is the ensemble-averaged potential energy for simulated ensembles at a given temperature. Typically, one expects sharp transitions for well-defined order-to-disorder transitions and yet, interestingly, the Q_{56} system shows the sharpest transition. The relatively broad transitions for NTL9 and GB1 highlight the joint contributions of gradual melting and different degrees of residual local structure in their unfolded states.

intended to accomplish two additional goals for the analysis of results from molecular simulations: (i) To compare the degree of conformational heterogeneity of ensembles obtained for a specific system at different simulation conditions; and (ii) To compare the degree of conformational heterogeneity of ensembles for different polypeptide sequences at equivalent simulation conditions.

The remainder of the narrative is organized as follows: In Sec. II we summarize the simulation approach for generating the conformational ensembles that were used to prototype Φ. Section III is split into two parts. In part 1, we describe the methodological framework for calculating Φ. In doing so, we discuss the choices made in converging upon the overall approach. In part 2, we use Φ to assess recent simulation results that were reported for the basic regions (bRs) of bZIP transcription factors.[60] These results demonstrated the role of sequence contexts in modulating the intrinsic helicities of bZIP-bRs. We show that Φ unmasks the weaknesses inherent to measures of average secondary structure contents as probes for structure and highlight how conformational heterogeneity can prevail in ensembles with high average helicities. We conclude with Sec. IV that summarizes the uses for Φ in analyzing protein disorder and in *de novo* sequence design. The discussion also provides a comparison between Φ and other approaches for quantifying conformational heterogeneity.

II. METHODS

A. Polypeptide systems included in this work

We simulated homopolymers of glutamine (Q_{56}), proline (P_{56}), and arginine (R_{56}) each 56-residues long. In addition, we included two 56-residue polypeptides, NTL9 and GB1 that adopt well-defined folds at low temperatures. Homopolymers were N-terminally acetylated and C-terminally *N*′-methylamidated. Atomistic Markov Chain Metropolis Monte Carlo (MC) simulations[61] were performed in the canonical ensemble using one polypeptide for each construct. Mobile sodium and chloride ions were included for peptides containing charged residues and the ions were represented explicitly. The salt concentration of ion-containing systems was set to 25 mM. The peptides and ions were enclosed in a spherical droplet of radius 200 Å and the droplet boundary was enforced using a stiff harmonic boundary potential. For the basic region leucine zipper transcription factor (bZIP-bR) peptides we analyzed simulation results from the work of Das *et al.*[60]

B. Details of the metropolis Monte Carlo (MC) simulations

The CAMPARI molecular simulation package (http://campari.sourceforge.net/) in conjunction with the ABSINTH implicit solvation model[62] and OPLS-AA/L[63] molecular mechanics force field parameters (abs3.2_opls.prm) were used for four out of five sets of simulations. For the poly-arginine system we used the new ion parameters developed by Mao and Pappu[64] for the mobile Na^+ and Cl^- ions. The spatial cutoffs for Lennard–Jones and electrostatic interactions between net-neutral charge groups were set to 10 Å and 14 Å, respectively. No cutoffs were employed for computing the electrostatic interactions for ions and side chain moieties with a net charge. Sodium and chloride ions were modeled explicitly and polypeptides were modeled in atomic detail. The internal degrees of freedom included the backbone ϕ, ψ, ω and side chain χ dihedral angles. Rigid-body moves simultaneously change rotational and translational degrees of freedom of the protein whereas translational moves were applied to alter the positions of mobile ions. Random cluster moves alter the rigid body coordinates of multiple molecules at once. Pucker moves perturb the ring geometry of proline residues.[49] The frequencies with which different moves were chosen along with parameters specific to each move type are summarized in the decision tree that is similar to that of Mao *et al.*[47] The starting conformation for homopolymers Q_{56}, P_{56}, and R_{56} was generated at random from a pre-equilibrated distribution of atomistic self-avoiding random walks. Starting conformations for NTL9 and GB1 for all simulation temperatures were derived from Protein Data Bank (http://www.rcsb.org) IDs 2HBB and 1GB1, respectively. Additional details regarding the setup of the initial folded conformations are as described in Meng *et al.*[14] The bond lengths and bond angles were fixed at values prescribed by Engh and Huber.[65]

C. The MC sampling protocol

Simulation results for NTL9, GB1, Q_{56}, P_{56}, and R_{56} were generated using the following protocol: For each system we performed ten independent MC simulations at each of the following temperatures: $T = 240$, 260, 280, 290, 300, 310, 320, 330, 340, 345, 350, 355, 360, 365, 370, 375, 380, 390, 400, 430, 450, and 500 K. Each independent simulation used a different random seed to initialize the MC run. A total of 8×10^7 MC steps were used in each independent simulation and of these, the results from the first 2×10^7 steps were discarded as equilibration. Observables were accumulated every 10^4 MC steps and conformational vectors for the heterogeneity calculation were collected every 10^5 steps. Thus, an ensemble from a single run that was used to calculate Φ contained 6000 members for each temperature. Reproducibility of the simulation results across multiple independent runs negated the need for using enhanced sampling methods.

D. The Flory random coil (FRC) model

The FRC reference state was constructed for each polypeptide.[66] FRC peptides were represented in all atom detail with the same degrees of freedom as used in the MC simulations described above. FRC conformations were generated by random assignment of sterically allowed combinations of backbone ϕ, ψ, ω and side chain χ dihedral angles while ignoring all inter-residue interactions. Each step of FRC sampling consisted of picking a residue at random then assigning all of the torsional degrees of freedom of the residue to a vector of ϕ, ψ, ω, and χ selected at random from a library of size 10^4. These libraries were generated for each residue via MC simulations of the corresponding dipeptides in the excluded volume (EV) limit. EV ensembles were generated using atomistic descriptions of the dipeptide while ignoring all non-bonded interactions excepting steric repulsions. A total of 4×10^7 steps were applied in each FRC simulation and resulting polypeptide conformations were accumulated every 10^5 steps. Ten independent FRC simulations were performed resulting in a total of 4000 reference conformations for each peptide. This pool of conformations is referred to as the FRC ensemble and all members were used in calculating Φ.

III. RESULTS

A. Estimating Φ

Our goal is to quantify the degree of conformational heterogeneity given an ensemble of conformations. This requires a method to quantify the degree of similarity between all distinct pairs of conformations within the ensemble. The resultant distribution of pairwise similarity measures is then used to obtain a value for Φ that reports on the degree of conformational heterogeneity within the ensemble. For a chain of N residues, each conformation c is represented as an $n_\mathrm{d} \times 1$ conformational vector $\mathbf{V}_c$ where $n_\mathrm{d} = \frac{N(N-1)}{2}$,$\mathrm{V}_c = \{d_{12}, d_{13}, \ldots, d_{N-1,N}\}$, and each element d_{ij} in $\mathbf{V}_c$ represents the spatial distance between a unique pair of residues, i and j. For each pair of residues i and j, we calculate $d_{ij} = \frac{1}{Z_{ij}} \cdot \sum_{m \in i} \sum_{n \in j} |\mathbf{r}_m^i - \mathbf{r}_n^j|$. Here, $\mathbf{r}_\mathrm{m}{}^\mathrm{i}$ and $\mathbf{r}_\mathrm{n}{}^\mathrm{j}$ denote the position vectors of atoms m and n within residues i and j, respectively, and Z_{ij} is the number of unique pairwise interatomic distances between the two residues. To compare a pair of conformations k and l, we calculate a pairwise dissimilarity measure $\mathcal{D}_{kl} = 1 - \cos(\Omega_{kl})$ where $\cos(\Omega_{kl}) = \frac{\mathrm{V}_k \cdot \mathrm{V}_l}{|\mathrm{V}_k||\mathrm{V}_l|}$. An ensemble of n_c conformations produces an ensemble of conformational vectors, $\mathbf{V}_1$, $\mathbf{V}_2$,…etc. These vectors are used to calculate a distribution $P(\mathcal{D})$ of $\frac{n_c(n_c-1)}{2}$ conformational dissimilarity values. Examples of these distributions are shown in Figure 4.

For a given simulation temperature T, the first moment of the distribution of dissimilarity values, $\langle \mathcal{D} \rangle$, provides an estimate of the most likely value for the degree of conformational heterogeneity within the ensemble. This measure, however, needs calibration in order for it to be used for comparing ensembles across different simulation temperatures or ensembles for different systems. For a given system, the intrinsic conformational properties of amino acids within the sequence place an upper bound on the degree of dissimilarity that is realizable.[66] These intrinsic biases need to be accounted for and normalized against if we are to compare the degree of conformational heterogeneity between systems. Furthermore, considerations of chain connectivity might render many of the values for inter-residue distances d_{ij} to either be invariant or slowly varying as temperature changes. Accordingly, we use a simulation approximation of the Flory random coil that is based on the rotational isomeric approximation to calibrate the distribution of dissimilarity values obtained for ensembles of a given system. This is accomplished by calculating the pairwise conformational dissimilarity $\mathcal{D}_{kl}$ between each conformation k from the ensemble of interest and conformation l drawn from the ensemble of FRC conformations. The latter ensemble varies depending on the amino acid sequence and remains invariant with temperature. Consequently, for each ensemble corresponding to a given simulation temperature T, we obtain two distributions of dissimilarities, viz., the distribution of $\mathcal{D}$-values for pairs of conformations within an ensemble and a distribution of $\mathcal{D}$-values comparing each conformation to an ensemble of FRC conformations (see Figure 3). Averaging over the former yields $\langle \mathcal{D} \rangle$ and averaging over the latter, which is an ensemble of ensembles yields $\langle\langle \mathcal{D} \rangle\rangle_\mathrm{FRC}$. The values of $\langle \mathcal{D} \rangle$ and $\langle\langle \mathcal{D} \rangle\rangle_\mathrm{FRC}$ lead to an estimate of the degree of heterogeneity Φ within the ensemble at

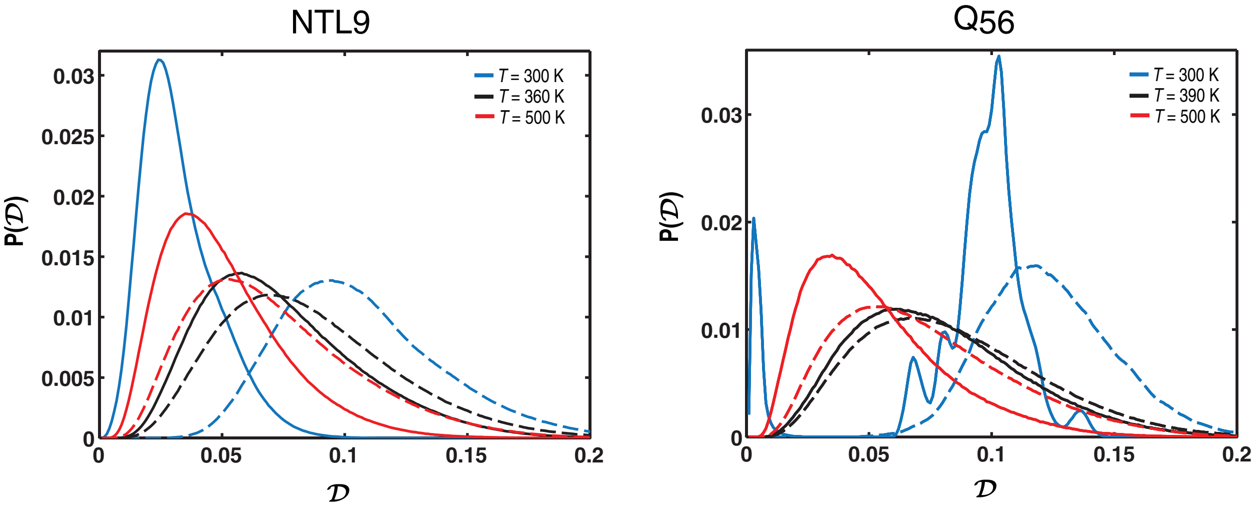

FIG. 3. Sample distributions $P(\mathcal{D})$ for two systems at different temperatures. The panel on the left shows $P(\mathcal{D})$ distributions for NTL9 at three different temperatures and the panel on the right shows these distributions for the Q_{56} system at three different simulation temperatures. In both panels, the solid curves represent intra-ensemble $P(\mathcal{D})$ distributions whereas the dashed curves are for comparisons between conformations within an ensemble at temperature T and conformations drawn from the FRC ensemble.

temperature T. We first compute the ratio $\mathcal{H} = \langle\mathcal{D}\rangle/\langle\langle\mathcal{D}\rangle\rangle_{\mathrm{FRC}}$ and use it to calculate $\Phi = 1 - \mathcal{H}$.

The Flory random coil model helps us dereference three composition-specific contributions to conformational heterogeneity for a given combination of sequence and simulation conditions. These are (i) trivial differences between residue-specific local conformational preferences, (ii) the effects of chain connectivity, and (iii) differences in heterogeneity that arise due to differences in chain length. The use of the FRC model is akin to the use of an ideal fluid prior in calculations of pair correlation functions in atomic and molecular fluids. The Flory random coil model is an informed prior that accommodates maximal conformational heterogeneity by including and accounting for contributions from local biases. Reference models such as the freely jointed or freely rotating chain models entirely ignore residue-specific local conformational biases. Therefore, their usage would be tantamount to pre-multiplying the value of $\langle\mathcal{D}\rangle$ by a constant pre-factor, the value of which depends on chain length and nothing else. The self-avoiding random walk ensemble or conformations drawn from the so-called excluded volume (EV) limit[67] could be used as a reference. However, these conformations are biased in that the ensemble is characterized by correlated fluctuations that afford the unique properties of the EV limit. It is well known that there is a diminution of conformational heterogeneity in the EV limit as compared to the FRC state due to spatial correlations between non-nearest neighbor residues.[68] Hence, in choosing a reference state, we select an ensemble that (a) lacks correlations between non-nearest neighbor residues, (b) retains a minimal degree of sequence specificity, and (c) affords the ability to enable quantitative comparisons between different systems and simulation conditions.

In the FRC model, the conformational partition function for the polypeptide is written as a product of partition functions of independent interaction units. All interactions between non-nearest neighbor residues are ignored while the intrinsic conformational preferences of individual residues are captured in terms of weights for each of the possible rotational isomers. The FRC ensemble therefore represents an intuitive upper bound on conformational heterogeneity and helps ensure that Φ is bounded between the values of 0 and 1, $0 \leq \Phi \leq 1$. This property obtains because $\mathcal{H} \leq 1$ and results from the construction of the reference FRC ensemble, which ensures that $\langle\mathcal{D}\rangle \leq \langle\langle\mathcal{D}\rangle\rangle_{\mathrm{FRC}}$. If the degree of intra-ensemble conformational heterogeneity is akin to the upper bound on heterogeneity expected for an FRC ensemble, then the ratio $\mathcal{H} \rightarrow 1$ and $\Phi \rightarrow 0$, indicating a maximally heterogeneous ensemble. Conversely, for a homogeneous ensemble of conformations it follows that, $\mathcal{D} \rightarrow 0$, $\mathcal{H} \rightarrow 0$, and $\Phi \rightarrow 1$. Therefore, given two values of Φ_{A} and Φ_{B} for two sequences at identical temperatures such that $\Phi_{\mathrm{A}} > \Phi_{\mathrm{B}}$ we infer that when referenced to their respective own Flory random coil states sequence A has higher conformational heterogeneity. It is worth noting that the generation of the FRC reference ensemble adds minimal computational overhead to the overall procedure. Importance sampling methods such as molecular dynamics or Metropolis Monte Carlo sampling are computationally expensive because of the force/energy evaluation that is necessary at each step. In contrast, the FRC reference ensemble is generated using a pre-computed library of rotational isomers for each residue and ignoring all inter-residue interactions. Therefore, the machine time for ensemble generation increases sub-linearly as system size increases. We expect to lower the barrier for generating these reference ensembles by providing a web-based automatic FRC generator at http://pappulab.wustl.edu.

B. Assessment of conformational ensembles using Φ

Figure 4 shows the variation of $\mathcal{D}$ and Φ with temperature for each of the five systems that were introduced in Figure 1. Equivalent inferences can be drawn from the use of either parameter. We focus on the analysis of Φ because

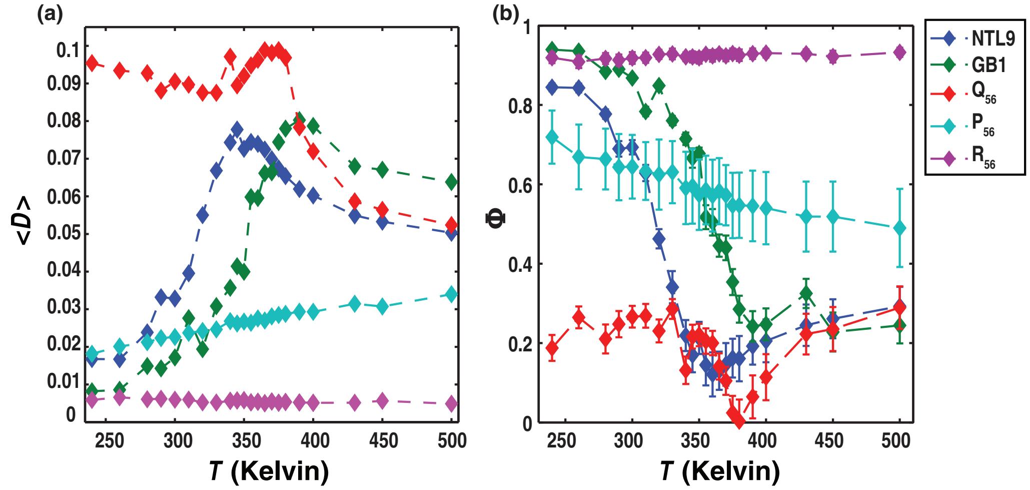

FIG. 4. Temperature dependence of $\langle\mathcal{D}\rangle$ and Φ for the five archetypal systems. Panel (b) includes error bars from a bootstrap analysis whereby 100 distinct bootstrap trials were performed to estimate Φ and the error bars therefore represent standard deviations for the estimate of the mean Φ values.

of the attributes described above, specifically the ability to use it for comparing different systems at similar sets of conditions. For NTL9 and GB1 the unfolding transition is manifest as a transition between a high value for Φ at low temperatures and a low value for Φ at high temperatures and the transition between these two limits is sharp. The slope of the transition region quantifies the "rate" of change in the degree of conformational heterogeneity with temperature. In contrast to NTL9 and GB1, the temperature dependence of Φ for Q_{56} is consistent with equivalent degrees of heterogeneity in the high and low temperature regimes. Previous work on polyglutamine led to an estimate of $T_\theta \approx 390$ K for the theta temperature.[45] At $T \approx T_\theta$, chain-chain and chain-solvent interactions are counterbalanced and statistical properties at the theta temperature resemble that of the FRC model. Accordingly, the temperature dependence of Φ_T for Q_{56} shows a dip and approaches zero near T_θ. Apart from this deviation, the profile of Φ_T for Q_{56} is consistent with the hypothesis of a disorder-to-disorder transition. When combined with the analysis of s^2, it becomes clear that the polyglutamine system transitions between two classes of disorder, viz., a heterogeneous ensemble of compact conformations that maximize the density at low temperatures and a heterogeneous ensemble of expanded conformations that minimize the density at high temperatures.

The Φ profiles for NTL9 and GB1 also show dips at intermediate temperatures, and again these temperatures can formally be shown to correspond to the theta temperatures for these systems. This identification is useful in light of the recent results of Hofmann *et al.*[15] They assessed changes in R_g for unfolded molecules as a function of decreasing denaturant concentration for five different systems and found that for unfolded ensembles under native conditions $R_g \sim N^{0.45\pm\Delta}$ where $\Delta \approx 0.05$. This implies that, on average, intra-chain and chain-solvent interactions are mutually screened because of generic amino acid compositional biases seen in protein sequences giving rise to the property that the statistics of unfolded ensembles mimic those of polymers at theta temperatures. If this proposed equivalence holds up to scrutiny, then the analysis of ensembles generated for temperatures where $\Phi \rightarrow 0$ should lead to insights regarding unfolded states sampled under folding conditions.

The intrinsically stiff P_{56} system shows a linear transition of Φ from a high value to a smaller value. As shown previously,[49,69] the increase in conformational heterogeneity arises mainly due to the increased frequency of generating bends and kinks within polyproline and this is partly due to the increased frequency of sampling *cis* peptide bonds at higher temperatures. Overall, however, the energy scales that encode chain stiffness in polyproline cannot be overcome by increasing temperature, and the transitions of s^2 and Φ reflect this feature.

The results for poly-arginine are particularly relevant for the study of IDPs. Often one uses web-based assessments of the degree of disorder and as shown previously,[47] these bioinformatics-based servers[70,71] predict that poly-arginine should be highly disordered. This prediction is the result of conflating high charge content and the resultant high s^2 (low density) values to implicitly imply a high degree of conformational heterogeneity. However, we find that $\Phi > 0.9$ across the entire temperature range and this lack of change in Φ with temperature for R_{56} is consistent with the maintenance of a homogeneous ensemble of rod-like conformations. The degree of chain expansion exceeds that of self-avoiding random walks.[47] This expansion results from a combination of long-range electrostatic repulsions and favorable solvation of charged side chains that together give rise to correlated fluctuations and overall rod-like behavior of the chain.[72] Chain compaction and increased conformational heterogeneity can be realized by screening the electrostatic repulsions in the presence of high concentrations of salt as was shown previously.[47]

The preceding analysis shows that it is necessary to use Φ and s^2 jointly to characterize the degree and nature of conformational heterogeneity. In systems such as NTL9 the joint use of s^2 and Φ highlights the positive coupling between increased conformational heterogeneity and chain expansion (see panel (a) in Figure 5). This is consistent with

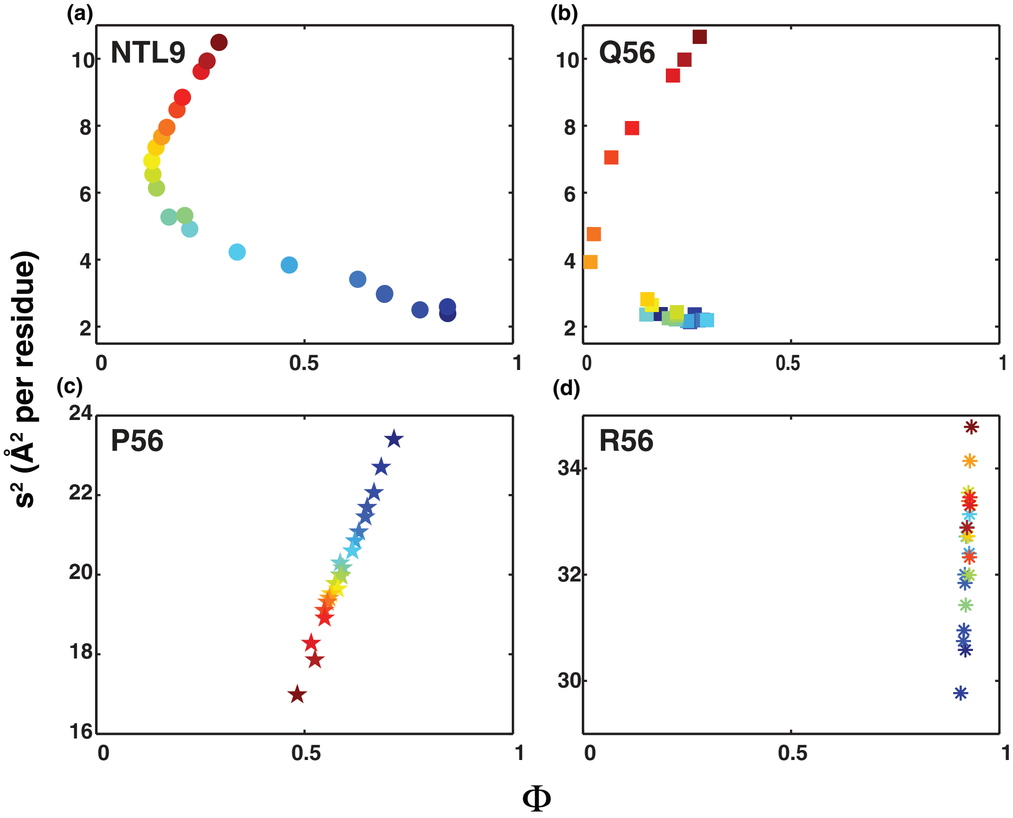

FIG. 5. (a)–(d) Plots to quantify the assessments of conformational properties that derive from the joint analysis s^2 (ordinates) and Φ (abscissae). In each panel, the symbol colors progress from cool to hot as temperature increases.

energy landscape theories that predict strongly funneled landscapes for sequences that fold into well-defined ensembles of self-similar conformations.[54] Conversely, polyglutamine and polyarginine show limiting behaviors (panels (b) and (d) of Figure 5). For polyglutamine, joint use of s^2 and Φ shows that the globule-to-coil transition observed as temperature increases is consistent with ensembles switching from one class of heterogeneity to another. In the parlance of energy landscape theory, temperature modulates the ruggedness of free energy landscapes. For polyarginine, analysis of the temperature dependence of Φ alone might seem confounding in light of bioinformatics-based prediction that this system should be highly disordered.[47,70] Electrostatic repulsions and the favorable solvation of charges side chains give rise to increased electrostatic persistence lengths, long-range correlated fluctuations, and homogeneous ensembles of rod-like conformations for highly charged systems for all temperatures – an observation that is consistent with experimental data[46,47] and polyelectrolyte theories.[73]

C. Application of Φ to assess conformational heterogeneity in IDPs with different secondary structure propensities

In Sec. III B we showed that the joint use of s^2 and Φ provide a more complete picture of conformational heterogeneity. Systems with different degrees of chain compaction can display similar degrees of conformational heterogeneity. Consequently, chain compaction/expansion does not directly imply homogeneous/heterogeneous ensembles. It is also common practice to characterize ensembles in terms of secondary structure propensities because these are accessible to direct inquiry using nuclear magnetic resonance,[74] circular dichroism,[60] and molecular simulations. Such inquiries often show sequence-specific variations in secondary structure propensities and the question is if higher secondary structure content translates to diminished conformational heterogeneity and vice versa? We answer this question by analyzing recent simulation results[60] for the basic regions of bZIP transcription factors.

Basic region leucine zippers (bZIPs) are modular transcription factors that play key roles in eukaryotic gene regulation.[75] The basic regions of bZIPs (bZIP-bRs) adopt regular α-helical conformations when bound to DNA.[76] Bioinformatics predictions and spectroscopic studies suggest that unbound, monomeric bZIP-bRs are uniformly disordered as autonomous units.[77,78] This assumption was recently tested through quantitative characterization of the conformational preferences of fifteen different bZIP-bRs.[60] These were found to have quantifiable preferences for α-helical conformations in their unbound, monomeric forms. This helicity varies from one bZIP-bR to another despite significant sequence similarity of the DNA binding motifs (DBMs). Analysis of the determinants of helicity revealed that intramolecular interactions between DBMs and 8-residue segments directly N-terminal to DBMs are the primary modulators of bZIP-bR helicities. The accuracy of this inference was tested in designed chimeras of bZIP-bRs that have either increased or decreased overall

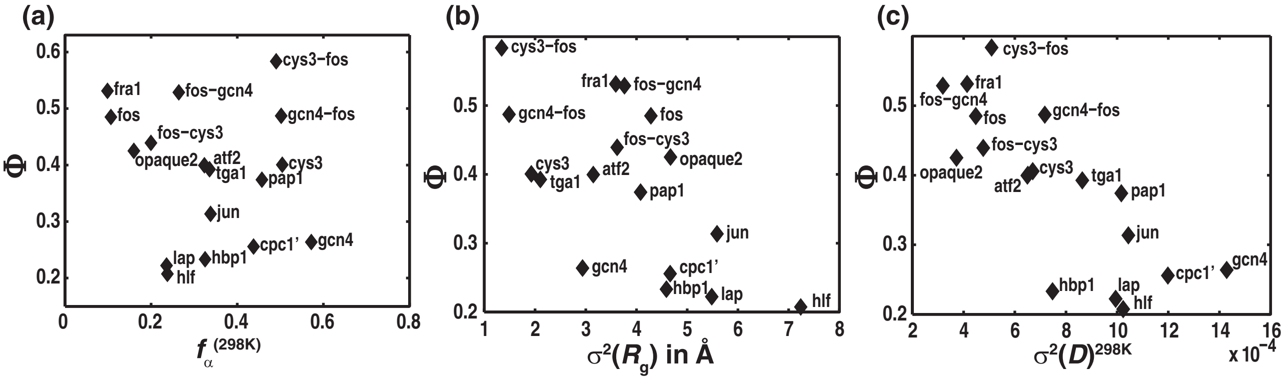

FIG. 6. Assessments conformational heterogeneity in ensembles with different degrees of helical structure. Panel (a) plots Φ against $f_\alpha^{(T)}$ for $T = 298$ K. The results are shown for 17 naturally occurring and designed sequences. Panel (b) plots Φ against $\sigma^2(R_g)$ and panel (c) plots Φ against $\sigma^2(\mathcal{D})$ for each of the 17 bZIP-bRs.

helicities. For a given sequence, the helical propensity $f_\alpha^{(T)}$ at temperature T was calculated using the formula in Eq. (1):

$$f_\alpha^{(T)} = \frac{\sum_{i=1}^{N} \langle p_i^{(\alpha)} \rangle_T}{N},$$

where

$$\langle p_i^{(\alpha)} \rangle_T = \left(\frac{\sum_{k=1}^{n_{\text{conf.}}^{(T)}} \Theta_k^i}{n_{\text{conf.}}^{(T)}} \right) \tag{1}$$

and

$$\Theta_k^i = \begin{cases} 1, & \text{if residue } i \text{ is part of a helical segment in} \\ & \quad \text{conformation } k \\ 0, & \text{otherwise} \end{cases} .$$

In Eq. (1), N denotes the number of residues in a bZIP-bR sequence, $\langle p_\alpha^{(i)} \rangle_i$ is the ensemble-averaged probability of finding residue i as part of a helical segment, $n_{\text{conf.}}^{(T)}$ denotes the number of conformations used for calculating ensemble averages at temperature T, and Θ_k^i is a discrete Heaviside function that determines if residue i is part of an α-helical segment in conformation k. A α-helical segment was identified as a stretch that has *at least* seven consecutive residues with a DSSP (Define Secondary Structure of Proteins)[79] designation of "H", which implies that these residues are part of a regular, hydrogen-bonded α-helix. Panel (a) in Figure 6 shows a plot of Φ against the calculated helicity for seventeen bZIP-bRs that includes 13 naturally occurring bZIP-bRs and four designed chimeric sequences. The results are shown for $T = 298$ K.

Naively one might expect a strong positive correlation between an increase in Φ and an increase in helical propensity. We quantified the linear correlation between Φ and helical propensity using the Pearson product moment correlation coefficient. We find a value of $r = 5 \times 10^{-4}$ when we use the Φ and $f_\alpha^{(T)}$ values for all of the sequences listed in panel (a) of Figure 7. We reasoned that the quantification of helicity, which reports on local structural propensities – especially as calculated in Eq. (1) – masks the degree of conformational heterogeneity that is achievable in the ensemble. The seemingly confounding correlation analysis is impacted by the presence of two types bZIP-bRs that either show high average helicity and low Φ as in the bZIP-bR of gcn4 or those that have low overall helicities on average and higher values of Φ (>0.4). We analyzed the correlation between Φ and the variances of the R_g and $\mathcal{D}$ distributions, i.e., $\sigma^2(R_g)$ and $\sigma^2(\mathcal{D})$ for each of the 17 sequences. Here, we expect a negative correlation between Φ and the σ^2 values because increased conformational heterogeneity should lead to larger fluctuations in chain size and $\mathcal{D}$-values. Panels (b) and (c) in Figure 6 demonstrate these negative correlations. Clearly, locally averaged measures of structure can be misleading because they mask the degree of conformational heterogeneity that can be accommodated within an ensemble despite quantifiable secondary structure content.

Figure 7 shows additional analysis to illustrate the source of the weak correlation between ensemble-averaged helicities and Φ. Panel (a) shows results for the bR of fra1, which has low overall helicity (less than 0.2) and Φ greater than 0.5. The distribution of helical segment lengths, which is narrow, provides an explanation for the higher value of Φ. A similar stretch of 7–10 residues forms helices in roughly 20% of the conformations. Panel (b) shows results for the chimeric bR, cys3-fos, which has a high average helicity and high Φ value (both ≈ 0.6). In this case, fluctuations cause the helical stretch to expand and contract around the central region of the sequence that always spans the DNA binding motif. Panel (c) shows results for the bR of gcn4. Although the ensemble-averaged helicity is high (≈0.6) the ensemble is characterized by a broad distribution of helical segment lengths whereby different sequence stretches fluctuate into and out of helical conformations thus leading to high heterogeneity and a low value (≈0.25) for Φ.

The preceding analysis is important given the previous work Das *et al.*[60] who used *de novo* sequence design to modulate intrinsic helicities of bZIP-bRs. Inasmuch as this effort was geared toward modulating the bias toward or away from α-helical conformations adopted by bZIP-bRs in their bound states, the current analysis highlights the fact that proper

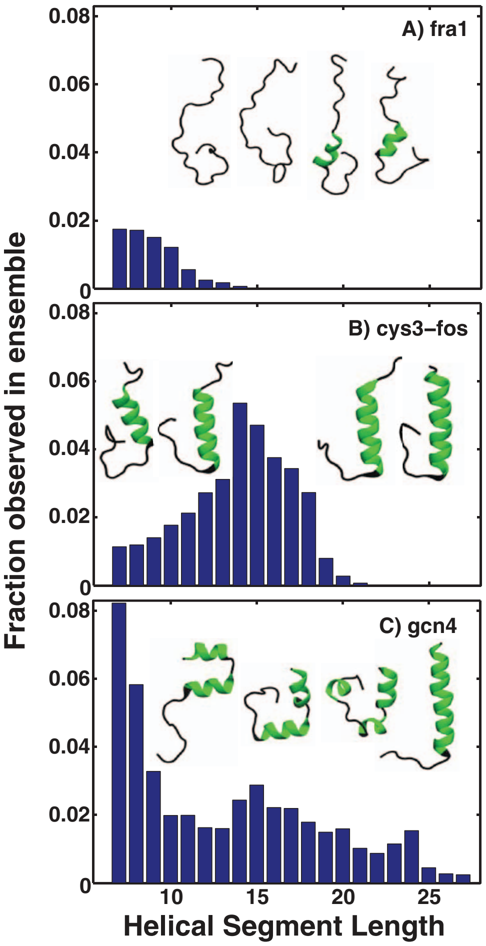

FIG. 7. Analysis of conformational heterogeneity in terms of the distribution of helical segment lengths for three of the bZIP-bRs. The figure shows three panels one each for the bZIP-bR of fra1, the chimeric cys3-fos, and gcn4. Each panel shows a histogram of helical segment lengths within the simulated ensembles. A helical segment corresponds to a consecutive stretch of residues in a conformation with a DSSP "H" designation. The value of Φ is dictated by the width of a segment length distribution as opposed to the ensemble-averaged helicity.

modulation of the degree of disorder in the ensemble requires a joint calculation of helical propensities and Φ. As an illustration of this need for using ensemble heterogeneity as a constraint in *de novo* sequence design we compare the results for the chimeric fos-gcn4 bR to the wild type gcn4-bR. The former was designed to have lower helicity than gcn4, which is indeed the case. Despite this, the heterogeneity is such that the Φ value is higher in the chimera, which has lower helicity than the wild type gcn4-bR. Such results can confound the objectives of *de novo* sequence design especially in a setting where helicities are being modulated to impact the driving forces for and mechanisms of coupled folding to binding reactions.

IV. DISCUSSION

Measures such as s^2 and f_α provide information regarding the global and local conformational preferences, i.e., they help classify the types of local conformational and overall density features of an ensemble of conformations. Conversely, Φ measures the degree of conformational heterogeneity. Hence, combining quantitative analysis of Φ with measures such as s^2 and quantification of local secondary structure preferences provides a complete quantitative summary of the *degree* and *nature* of conformational heterogeneity for the ensemble in question. This approach also facilitates comparisons between ensembles for different sequences and conditions. Diminished conformational heterogeneity will lead to an increase in Φ whereas increased conformational heterogeneity will decrease the value of Φ. Since this parameter is bounded, the result of normalization using the Flory random coil reference state, quantitative changes to Φ imply quantitative changes to the degree of conformational heterogeneity. This parameter should prove useful in comparative assessments of conformational heterogeneity of conformational ensembles generated for a single system at different temperatures and solution conditions as well as for different systems under similar conditions.

To calculate Φ, we used the first moments $\langle \mathcal{D} \rangle$ and $\langle \langle \mathcal{D} \rangle \rangle_{\mathrm{FRC}}$. The first moments provide an assessment of the most likely values for the corresponding distribution of conformational dissimilarity values, i.e., $P(\mathcal{D})$. We have also calculated second moments of the underlying distributions. Figure 8 shows the temperature dependence of the variances, i.e., $\sigma^2(\mathcal{D}) = \langle (\mathcal{D} - \langle \mathcal{D} \rangle)^2 \rangle$ for each of the five archetypal systems. The variance is high when $\langle \mathcal{D} \rangle$ is high and is low when $\langle \mathcal{D} \rangle$ is low implying that the variance and Φ are negatively correlated. Other than this feature, there is no additional insight to be obtained through analysis of the variance and other higher order moments of $P(\mathcal{D})$ distributions, suggesting that Φ proves to be sufficient for a comparative and quantitative assessment of conformational heterogeneity.

Our approach to calculate Φ relied on three distinct choices namely, (i) the use of conformational vectors where the elements are inter-residue distances extracted from a specific conformation; (ii) the use of the distribution of pairwise projections of these vectors to calculate the degree of intra-ensemble dissimilarity; and (iii) the use of the FRC model

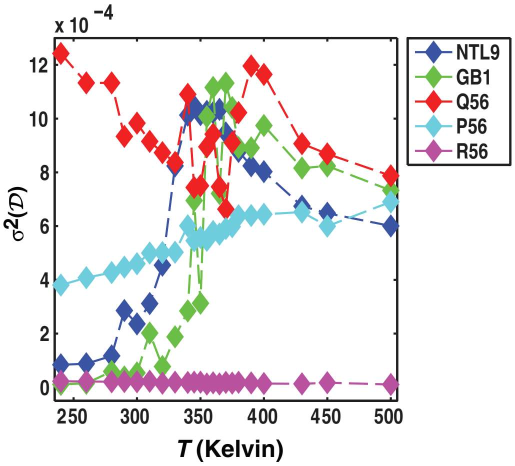

FIG. 8. Temperature dependence of the variance of $\mathcal{D}$ calculated from the distributions of $\mathcal{D}$ values for each of the five archetypal systems. All inferences regarding conformational heterogeneity that are drawn from analysis of the variance are consistent with those drawn from analysis of Φ.

to calibrate the degree of heterogeneity. We discussed the advantages of choice (iii) in the main text. Choices (i) and (ii) lead to an assessment of intra-ensemble conformational dissimilarity. Although these choices are distinct, they are not inherently superior to other methods proposed in the literature. For example, we could have calculated all unique pairwise superpositions of conformations based on least squares optimization and used the resultant distribution of root mean squared deviations (RMSDs)[80,81] as measures of dissimilarities, although the computational expense of these calculations increases substantially with increased number of conformations in an ensemble. One could also use the method of projections to compare conformational vectors comprised of backbone dihedral angles as elements.[82] We do not find any intrinsic advantages with using dihedral angle based conformational vectors and this could be used interchangeably with inter-residue distance based conformational vectors. Recent efforts have focused on the use of the number of inter-residue contacts q.[83] Each conformation within the ensemble is annotated by its q-number and the distributions of q-numbers, viz., $P(q)$ are analyzed to compare different ensembles to each other. This method, which is analogous to methods used in spin glass theories,[84] can be used in conjunction with Φ. It should be noted that the annotation of conformations by q-numbers requires the imposition of an *ad hoc* criterion for defining contacts, which causes an inherent loss of information. This is in contrast to the conformational vectors $\mathbf{V}_c$ used in this work.

Fisher and Stultz[85] introduced an order parameter based on information theory to quantify the degree of conformational heterogeneity. In direct analogy with Φ their order parameter O is bounded, i.e., $0 \leq O \leq 1$; $O \rightarrow 1$ for a homogeneous ensemble whereas $O \rightarrow 0$ for a maximally heterogeneous ensemble. Their procedure for calculating O uses the weights for individual conformations and pairwise conformational similarities that are based on mean square deviations (MSDs). Molecular simulations such as molecular dynamics and Metropolis Monte Carlo methods are classified as importance sampling methods. Accordingly, they yield a set of conformations sampled from the equilibrium distribution but the weights of individual conformations are generally unknown. Assessment of these weights will require *a priori* conformational clustering and the assignment of weights is based on cluster sizes / populations. Alternatively one can use a suitable weighted histogram analysis method such as T-WHAM[86] to assign weights to individual conformations. The procedure for conformational clustering is based on the calculation of MSDs and there is a nonlinear increase in computational expense with sample size. Further, the intrinsic property of MSDs is such that there are more ways to generate higher values of MSDs than lower ones and this bias also shows a nonlinear dependence on the MSD value and must be taken into account. Finally, the assessment of O uses as reference the average pairwise MSD that result from typical thermal fluctuations around a protein structure and hence the assessment of heterogeneity provided by O is intrinsically different from that Φ. According to the former, the degree of conformational heterogeneity in an ensemble is a normalized effective number of conformations in the ensemble given that thermal fluctuation around a protein structure will yield an average pairwise MSD of ≈2.5Å or higher depending on the temperature. The calculation of Φ makes no assumptions regarding the spatial size of thermal fluctuations around specific conformations because this most certainly depends on the density, i.e., the value of s^2 or ρ. Instead the value of Φ quantifies the degree of heterogeneity as the normalized effective number of distinct conformations referenced to a generic, maximally heterogeneous Flory random coil state. It is likely that Φ and O can provide complementary assessments of the degree of disorder in an ensemble, especially in conjunction with assessments of s^2 and other measures of local structure that yield insights regarding the type of conformations sampled. This will require improvements to make the calculation of O more efficient so the ensemble size can be expanded beyond the current limitation of ∼300 conformations.[85]

A. Practical uses for Φ

The calculation of Φ is designed with two practical purposes in mind. As noted in the Introduction, it is important to have measures of conformational heterogeneity that complement the assessments of ensembles that are obtained by quantification of densities, their fluctuations, and variances in energies. In order to understand the mechanisms of coupled folding and binding of IDPs it would be useful to be able to modulate the degree of disorder in the unbound ensemble using *de novo* sequence design. The parameter Φ helps in this regard because it provides a direct measure of conformational heterogeneity and can be used to guide sequence design in a way that heterogeneity is either decreased (increased Φ) or increased (decreased Φ).

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health (5RO1NS056114) and the National Science Foundation (MCB-1121867). We thank Professor Anders Carlsson, Professor Gary Stormo, and two anonymous reviewers for helpful comments and suggestions.

[1] P. G. Wolynes, W. A. Eaton, and A. R. Fersht, Proc. Natl. Acad. Sci. U.S.A. **109**, 17770 (2012).

[2] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic, J. Mol. Graphics Modell. **19**, 26 (2001).

[3] H. J. Dyson and P. E. Wright, Curr. Opin. Struct. Biol. **12**, 54 (2002).

[4] R. Halfmann, S. Alberti, R. Krishnan, N. Lyle, C. W. O'Donnell, O. D. King, B. Berger, R. V. Pappu, and S. Lindquist, Mol. Cell **43**, 72 (2011).

[5] D. Eliezer, Curr. Opin. Struct. Biol. **19**, 23 (2009).

[6] T. R. Sosnick and D. Barrick, Curr. Opin. Struct. Biol. **21**, 12 (2011).

[7] M. Vendruscolo, Curr. Opin. Struct. Biol. **17**, 15 (2007).

[8] A. H. Mao, N. Lyle, and R. V. Pappu, Biochem. J. **449**, 307 (2013).

[9] B. Anil, Y. Li, J. H. Cho, and D. P. Raleigh, Biochemistry **45**, 10110 (2006).

[10] W. Meng, B. Luan, N. Lyle, R. V. Pappu, and D. P. Raleigh, Biochemistry **52**, 2662 (2013).

[11] V. A. Voelz, M. Jaeger, S. Yao, Y. Chen, L. Zhu, S. A. Waldauer, G. R. Bowman, M. Friedrichs, O. Bakajin, L. J. Lapidus, S. Weiss, and V. S. Pande, J. Am. Chem. Soc. **134**, 12565 (2012).

[12] A. Hoffmann, D. Nettels, J. Clark, A. Borgia, S. E. Radford, J. Clarke, and B. Schuler, Phys. Chem. Chem. Phys. **13**, 1857 (2011).

[13] L. J. Lapidus, Curr. Opin. Struct. Biol. **23**, 30 (2013).

[14] W. Meng, N. Lyle, B. Luan, D. P. Raleigh, and R. V. Pappu, Proc. Natl. Acad. Sci. U.S.A. **110**, 2123 (2013).

[15] H. Hofmann, A. Soranno, A. Borgia, K. Gast, D. Nettels, and B. Schuler, Proc. Natl. Acad. Sci. U.S.A. **109**, 16155 (2012).

[16] J. E. Kohn, I. S. Millett, J. Jacob, B. Zagrovic, T. M. Dillon, N. Cingel, R. S. Dothager, S. Seifert, P. Thiyagarajan, T. R. Sosnick, M. Z. Hasan, V. S. Pande, I. Ruczinski, S. Doniach, and K. W. Plaxco, Proc. Natl. Acad. Sci. U.S.A. **101**, 12491 (2004).

[17] H.-X. Zhou, G. Rivas, and A. P. Minton, Annu. Rev. Biophys. Biomol. Struct. **37**, 375 (2008).

[18] A. H. Elcock, Curr. Opin. Struct. Biol. **20**, 196 (2010).

[19] T. R. Jahn and S. E. Radford, Arch. Biochem. Biophys. **469**, 100 (2008).

[20] L. J. Lapidus, Mol. Biosyst. **9**, 29 (2013).

[21] M. M. Dedmon, K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo, and C. M. Dobson, J. Am. Chem. Soc. **127**, 476 (2005).

[22] A. Soranno, B. Buchli, D. Nettels, R. R. Cheng, S. Mueller-Spaeth, S. H. Pfeil, A. Hoffmann, E. A. Lipman, D. E. Makarov, and B. Schuler, Proc. Natl. Acad. Sci. U.S.A. **109**, 17800 (2012).

[23] H. J. Dyson and P. E. Wright, Nat. Rev. Mol. Cell Biol. **6**, 197 (2005).

[24] R. Pancsa and P. Tompa, PLoS ONE **7**, e34687 (2012).

[25] P. Tompa and M. Fuxreiter, Trends Biochem. Sci. **33**, 2 (2008).

[26] T. Mittag, L. E. Kay, and J. D. Forman-Kay, J. Mol. Recognit. **23**, 105 (2010).

[27] A. Y. Grosberg and D. V. Kuznetsov, Macromolecules **25**, 1970 (1992).

[28] I. C. Sanchez, Macromolecules **12**, 980 (1979).

[29] S. Gianni, N. R. Guydosh, F. Khan, T. D. Caldas, U. Mayor, G. W. N. White, M. L. DeMarco, V. Daggett, and A. R. Fersht, Proc. Natl. Acad. Sci. U.S.A. **100**, 13286 (2003).

[30] E. Sherman and G. Haran, Proc. Natl. Acad. Sci. U.S.A. **103**, 11539 (2006).

[31] G. Ziv, D. Thirumalai, and G. Haran, Phys. Chem. Chem. Phys. **11**, 83 (2009).

[32] J. E. Shea and C. L. Brooks, Annu. Rev. Phys. Chem. **52**, 499 (2001).

[33] E. P. O'Brien, B. R. Brooks, and D. Thirumalai, Biochemistry **48**, 3743 (2009).

[34] D. Nettels, I. V. Gopich, A. Hoffmann, and B. Schuler, Proc. Natl. Acad. Sci. U.S.A. **104**, 2655 (2007).

[35] K. K. Sinha and J. B. Udgaonkar, J. Mol. Biol. **353**, 704 (2005).

[36] J. J. Chou and E. I. Shakhnovich, J. Phys. Chem. B **103**, 2535 (1999).

[37] J. B. Udgaonkar, Arch. Biochem. Biophys. **531**, 24 (2013).

[38] S. L. Crick, M. Jayaraman, C. Frieden, R. Wetzel, and R. V. Pappu, Proc. Natl. Acad. Sci. U.S.A. **103**, 16764 (2006).

[39] S. Mukhopadhyay, R. Krishnan, E. A. Lemke, S. Lindquist, and A. A. Deniz, Proc. Natl. Acad. Sci. U.S.A. **104**, 2649 (2007).

[40] D. P. Teufel, C. M. Johnson, J. K. Lum, and H. Neuweiler, J. Mol. Biol. **409**, 250 (2011).

[41] J. A. Marsh and J. D. Forman-Kay, Biophys. J. **98**, 2383 (2010).

[42] N. Jain, M. Bhattacharya, and S. Mukhopadhyay, Biophys. J. **101**, 1720 (2011).

[43] S. Brocca, L. Testa, F. Sobott, M. Samalikova, A. Natalello, E. Papaleo, M. Lotti, L. De Gioia, S. M. Doglia, L. Alberghina, and R. Grandori, Biophys. J. **100**, 2243 (2011).

[44] S. M. Vaiana, R. B. Best, W.-M. Yau, W. A. Eaton, and J. Hofrichter, Biophys. J. **97**, 2948 (2009).

[45] A. Vitalis, N. Lyle, and R. V. Pappu, Biophys. J. **97**, 303 (2009).

[46] S. Muller-Spath, A. Soranno, V. Hirschfeld, H. Hofmann, S. Ruegger, L. Reymond, D. Nettels, and B. Schuler, Proc. Natl. Acad. Sci. U.S.A. **107**, 14609 (2010).

[47] A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine, and R. V. Pappu, Proc. Natl. Acad. Sci. U.S.A. **107**, 8183 (2010).

[48] L. Tooke, L. Duitch, T. J. Measey, and R. Schweitzer-Stenner, Biopolymers **93**, 451 (2010).

[49] A. Radhakrishnan, A. Vitalis, A. H. Mao, A. T. Steffen, and R. V. Pappu, J. Phys. Chem. B **116**, 6862 (2012).

[50] A. Vitalis, X. Wang, and R. V. Pappu, J. Mol. Biol. **384**, 279 (2008).

[51] S. Chen, V. Berthelier, W. Yang, and R. Wetzel, J. Mol. Biol. **311**, 173 (2001).

[52] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins: Struct., Funct., Genet. **21**, 167 (1995).

[53] J. N. Onuchic, Z. LutheySchulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).

[54] J. N. Onuchic and P. G. Wolynes, Curr. Opin. Struct. Biol. **14**, 70 (2004).

[55] A. Vitalis, X. Wang, and R. V. Pappu, Biophys. J. **93**, 1923 (2007).

[56] G. A. Papoian, Proc. Natl. Acad. Sci. U.S.A. **105**, 14237 (2008).

[57] D. A. Potoyan and G. A. Papoian, J. Am. Chem. Soc. **133**, 7405 (2011).

[58] C. J. Camacho and D. Thirumalai, Phys. Rev. Lett. **71**, 2505 (1993).

[59] H. S. Chan and K. A. Dill, J. Chem. Phys. **99**, 2116 (1993).

[60] R. K. Das, S. L. Crick, and R. V. Pappu, J. Mol. Biol. **416**, 287 (2012).

[61] A. R. N. Metropolis, M. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[62] A. Vitalis and R. V. Pappu, J. Comput. Chem. **30**, 673 (2009).

[63] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen, J. Phys. Chem. B **105**, 6474 (2001).

[64] A. H. Mao and R. V. Pappu, J. Chem. Phys. **137**, 064104 (2012).

[65] R. A. Engh and R. Huber, Acta Cryst. **A47**, 392 (1991).

[66] P. J. Flory, *Statistical Mechanics of Chain Molecules* (Oxford University Press, New York, 1969).

[67] H. T. Tran and R. V. Pappu, Biophys. J. **91**, 1868 (2006).

[68] L. Schäfer, *Excluded Volume Effects in Polymer Solutions as Explained by the Renormalization Group* (Springer, Berlin, 1999).

[69] R. B. Best, K. A. Merchant, I. V. Gopich, B. Schuler, A. Bax, and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. **104**, 18964 (2007).

[70] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, and A. K. Dunker, Proteins: Struct., Funct., Bioinf. **61**, 176 (2005).

[71] T. Ishida and K. Kinoshita, Bioinformatics **24**, 1344 (2008).

[72] B. Y. Ha and D. Thirumalai, Phys. Rev. A **46**, R3012 (1992).

[73] B. Y. Ha and D. Thirumalai, Macromolecules **28**, 577 (1995).

[74] T. Mittag and J. D. Forman-Kay, Curr. Opin. Struct. Biol. **17**, 3 (2007).

[75] G. D. Amoutzias, A. S. Veron, J. Weiner, M. Robinson-Rechavi, E. Bornberg-Bauer, S. G. Oliver, and D. L. Robertson, Mol. Biol. Evol. **24**, 827 (2007).

[76] T. E. Ellenberger, C. J. Brandl, K. Struhl, and S. C. Harrison, Cell **71**, 1223 (1992).

[77] J. G. Liu, N. B. Perumal, C. J. Oldfield, E. W. Su, V. N. Uversky, and A. K. Dunker, Biochemistry **45**, 6873 (2006).

[78] K. T. Oneil, R. H. Hoess, and W. F. Degrado, Science **249**, 774 (1990).

[79] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).

[80] G. Vriend and C. Sander, Proteins: Struct., Funct., Genet. **11**, 52 (1991).

[81] E. Lyman and D. M. Zuckerman, Biophys. J. **91**, 164 (2006).

[82] Y. G. Mu, P. H. Nguyen, and G. Stock, Proteins: Struct., Funct., Bioinf. **58**, 45 (2005).

[83] D. A. Potoyan and G. A. Papoian, Proc. Natl. Acad. Sci. U.S.A. **109**, 17857 (2012).

[84] G. Parisi, Phys. Rev. Lett. **50**, 1946 (1983).

[85] C. K. Fisher and C. M. Stultz, J. Am. Chem. Soc. **133**, 10022 (2011).

[86] J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok, and K. A. Dill, J. Chem. Theory Comput. **3**, 26 (2007).