JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Vibrational Entropy of a Protein: Large Differences between Distinct Conformations

Martin Goethe,*[,†] Ignacio Fita,[‡] and J. Miguel Rubi[†]

[†]Departament de Física Fonamental, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain

[‡]Institut de Biologia Molecular de Barcelona, (CSIC), Baldiri Reixac 10, 08028 Barcelona, Spain

Supporting Information

ABSTRACT: In this article, it is investigated whether vibrational entropy (VE) is an important contribution to the free energy of globular proteins at ambient conditions. VE represents the major configurational-entropy contribution of these proteins. By definition, it is an average of the configurational entropies of the protein within single minima of the energy landscape, weighted by their occupation probabilities. Its large part originates from thermal motion of flexible torsion angles giving rise to the finite peak widths observed in torsion angle distributions. While VE may affect the equilibrium properties of proteins, it is usually neglected in numerical calculations as its consideration is difficult. Moreover, it is sometimes believed that all well-packed conformations of a globular protein have similar VE anyway. Here, we measure explicitly the VE for six different conformations from simulation data of a test protein. Estimates are obtained using the quasi-harmonic approximation for three coordinate sets, Cartesian, bond-angle-torsion (BAT), and a new set termed rotamer-degeneracy lifted BAT coordinates by us. The new set gives improved estimates as it overcomes a known shortcoming of the quasi-harmonic approximation caused by multiply populated rotamer states, and it may serve for VE estimation of macromolecules in a very general context. The obtained VE values depend considerably on the type of coordinates used. However, for all coordinate sets we find large entropy differences between the conformations, of the order of the overall stability of the protein. This result may have important implications on the choice of free energy expressions used in software for protein structure prediction, protein design, and NMR refinement.

INTRODUCTION

In principle, the equilibrium state of a protein in solution can be identified from the minimum of a suitable free energy function F. In practice, however, finding an accurate expression for F is a difficult task, not to mention the subsequent challenge of computing its minimum. To a good approximation, F can be decomposed into five contributions,[1] namely, (i) the average intramolecular energy, (ii) the average solvation free energy, (iii) the configurational entropy (times $(-T)$) and two conformation independent contributions, (iv) the free energy of global translation and rotation of the protein, and (v) the free energy of the pure solvent. Much work has been done in developing good approximations for the first two contributions which has yielded a variety of effective energy expressions (or potentials of mean force)[1] involving implicit solvation models of different complexity and precision.[2,3]

Much less is known about configurational entropy which is a measure of the thermal motion of the protein atoms (except for roto-translation). It can further be decomposed into two contributions, namely, conformational entropy and vibrational entropy (VE).[4,5] This decomposition is based on the multivalley structure of the (effective) energy landscape where the protein adopts different conformations in the valleys.[6,7] Gradually, many valleys are visited by the protein, which causes conformational entropy $S_{\mathrm{Cfm}} = -\kappa_{\mathrm{B}}\sum_{i=1}^{M} p_i \log p_i$, a sum over all M energy minima where p_i is the occupation probability of minimum i. The occurring conformational changes can roughly be divided into backbone and side-chain rearrangements, and hence, conformational entropy is often also partitioned into backbone conformational entropy and side-chain conformational entropy.[8]

VE is the weighted average $S_{\mathrm{Vib}} = \sum_{i=1}^{M} p_i S_{\mathrm{vib}}^i$ of the entropies $\{S_{\mathrm{vib}}^i\}$ in the different minima, formally defined as $S_{\mathrm{vib}}^i = -\kappa_{\mathrm{B}} \int_{v_i} \mathrm{d}^n\boldsymbol{x}\, \rho(\boldsymbol{x}) p_i^{-1} \log(\rho(\boldsymbol{x}) p_i^{-1})$ where $\rho(\boldsymbol{x})$ is the Boltzmann weight and the integral is restricted to minimum i.[4,5] Vibrational motion inside the minima causes S_{Vib}. To avoid confusion, let us emphasize that the term VE is not uniquely used in the literature. As outlined above, we use the terminology introduced by Karplus et al.[4] which was adopted by many authors afterward (for reviews see refs 5 and 9). According to this definition, also thermal motion of (soft) torsion angles inside a given minimum of the energy landscape contribute to S_{Vib}. In contrast, some authors use the term VE to denote only the entropy caused by vibrations of bond lengths and bond angles.[10]

In what follows, we restrict our considerations to proteins whose native states have a very dominant backbone conformation such as globular proteins with well-defined X-ray and NMR structures. For those proteins, backbone

Received: August 1, 2014
Published: December 1, 2014

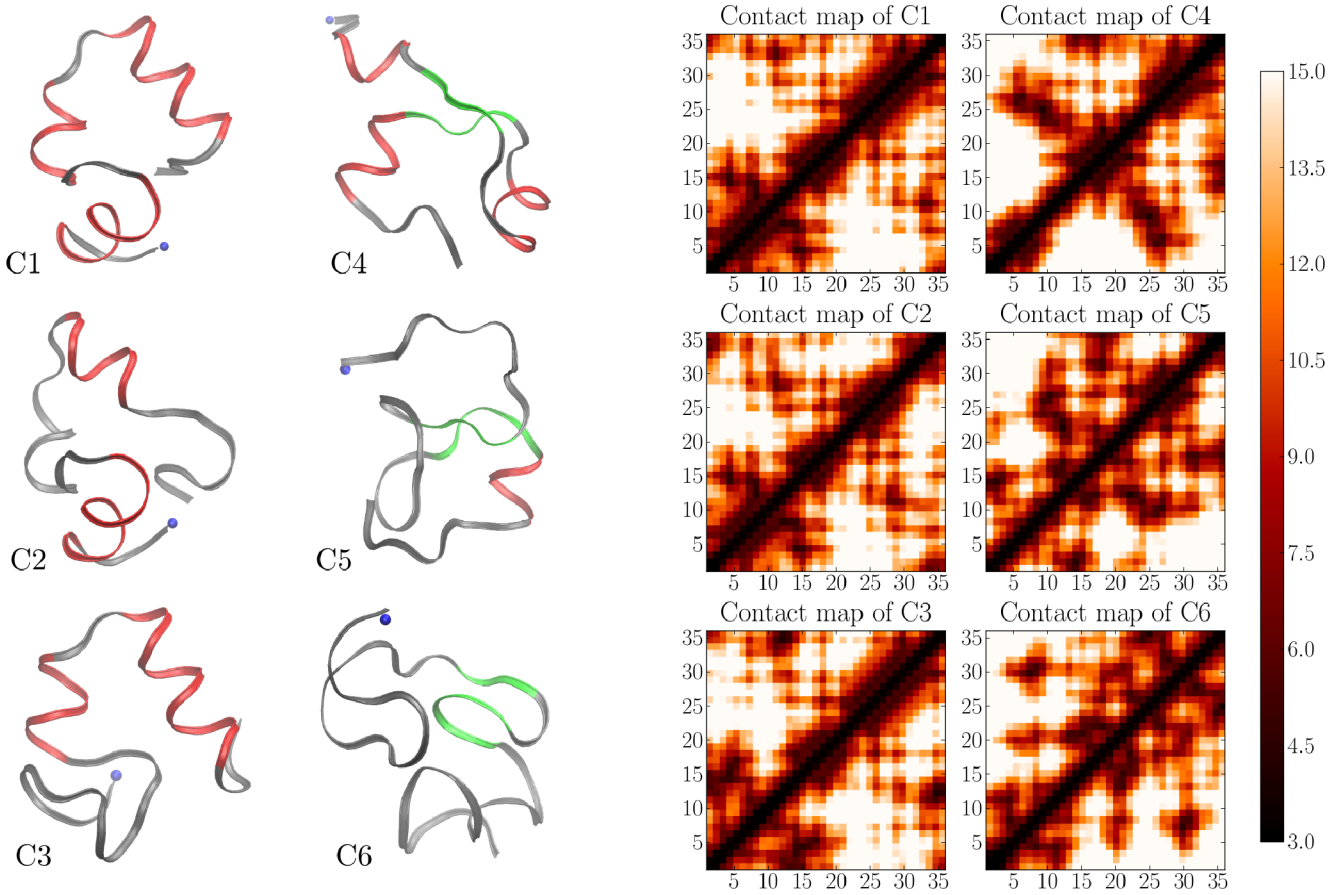

Figure 1. Structural diversity of the conformations. Left: Backbone traces of the six conformations C1−C6. Helices and beta-strands are shown in red and green, respectively. The blue dot indicates the N-terminus. Right: Contact maps of C1−C6 showing C^{α}−C^{α} distances in the range of 3−15 Å. Structures of C4, C5, and C6 are very different from each other and also different from those of C1, C2, and C3.

conformational entropy is negligible and computing the native state by comparing the free energies of different backbone conformations involves the calculation of side-chain conformational entropies and VEs.

Strong evidence has been obtained for the importance of side-chain conformational entropy[11−13] on the native state of proteins, and hence, it is sometimes considered in free energy expressions.[14] In contrast, VE is often assumed to be approximately equal for all well-packed conformations and, hence, not to affect the native state selection. Consequently, basically all free energy expressions used in software for protein structure analysis neglect S_{Vib}. This is supported by two studies[4,15] where VE was found to be similar in different well-packed conformations. However, these studies also found that TS_{Vib} usually represents a large contribution to the free energy at ambient conditions, of the order of several hundreds of $\kappa_B T$, which is very large compared to typical stabilities of proteins (about 8−25 $\kappa_B T$[16]). Therefore, one may only conclude that S_{Vib} is irrelevant if S_{Vib} differs by less than a percent between different conformations, an accuracy not achieved in those studies.

Whether VE can indeed be disregarded or not is currently a matter of debate because also plenty of evidence for the thermodynamical importance of VE has been obtained in closely related studies (e.g., on the insulin dimerization process[17] and the molten globule state of α-lactalbumin,[18] from a normal-mode analysis of high-energy random-coil states of various peptides,[19] from secondary-structure stability analysis in vacuum,[19−22] and from studies on ligand binding using inelastic neutron scattering,[23,24] NMR relaxation experiments,[25] and X-ray crystallography,[26,27] as well as numerics[28,29]), although it is not evident whether these findings can be generalized to well-packed protein conformations in solution.

In this work, we investigate VE differences between well-packed protein conformations by measuring explicitly the VE of a test protein in six different metastable backbone conformations. To our knowledge, no such direct exploration has been reported. We find considerable VE differences between the conformations studied, which is strong evidence for the thermodynamical importance of VE.

RESULTS

Conformation Diversity. We analyzed molecular-dynamics data[30] of a thermostable subdomain of the chicken villin headpiece which consists of 36 amino acids or N = 295 heavy (non-hydrogen) atoms. From a total of five times 100 ns simulation data, we selected six time windows of 26−44 ns (for details see Methods). During each time window, the system is in a different metastable backbone conformation.

The six conformations are sketched in Figure 1 and referred to as C1, C2, ..., C6. They all are well-packed with similar radii of gyration ranging from 9.3(2) Å to 10.1(3) Å. The conformations C1, C2, and C3 are quite closely related to each other, as indicated by the corresponding contact maps (see Figure 1) and the root-mean-square deviation (RMSD) values for the mutual superpositions (ranging from 3.5 to 4.0 Å). In

dx.doi.org/10.1021/ct500696p | J. Chem. Theory Comput. 2015, 11, 351−359

contrast, C1, C2, C3 and C4, C5, C6 are very different, with RMSDs between 5 and 8 Å and dissimilar contact maps (see Figure 1). Moreover, the secondary-structure elements differ between the conformations as indicated in Figure 1. Conformation C1 contains three helices, while C2 and C3 have only two, and none is found in C6, which exhibits a beta-sheet instead. Finally, C4 and C5 contain both helices and beta-strands.

Entropy Estimation. We measured the VE of the six conformations using the well-known quasi-harmonic (QH) approximation which basically relies on approximating the probability density function (pdf) of the protein by a multivariate Gaussian distribution. As the accuracy of this approach depends strongly on the chosen coordinate representation, we used three different coordinate sets and compared the distinct estimates. For details on the QH approximation and a survey of alternative methods for entropy estimation, we refer to the section Methods.

Entropy Estimation in Cartesian Coordinates. First, we used Cartesian coordinates to measure the VE of the conformations which yielded the VE estimate S_{Cart}. To this end, we subtracted the global translation and rotation of the protein, measured the covariance matrix of the $3N$ Cartesian coordinates, and applied the semiclassical version[31,32] of the QH approximation given in eq 2. Figure 2 shows $S_{\mathrm{Cart}}(t)$ for the six conformations as a function of the simulation time t, where t is such that $t = 0$ corresponds for each conformation to the time when the protein adopts the conformation. Mainly as a consequence of the number of available snapshots (see Methods), the calculated entropy values have not converged yet. Nevertheless, it is clear from Figure 2 that the entropy differences between the conformations *grow* with t. Hence, the entropies should be different for $t \to \infty$. The function $a[1 - \exp(-bt)]$ fits the data well (see shaded areas in Figure 2),

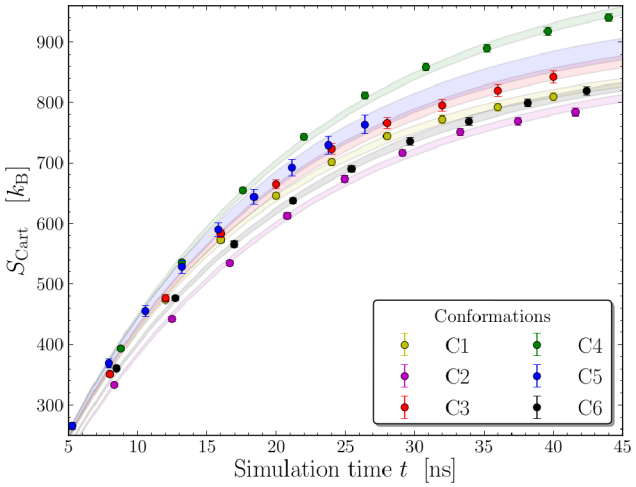

Figure 2. Entropy estimation in Cartesian coordinates. Convergence plot of $S_{\mathrm{Cart}}(t)$ as a function of the simulation time t, starting from the time when the simulation adopts each conformation. The data are far from convergence; however, the differences grow with t between the conformations C1−C6. The data were fitted to the function $a[1 - \exp(-bt)]$. Each shaded area represents this function for all values of a inside its obtained error interval (and for the best fit value for b). From the fits, we obtained the extrapolated values collected in Table 1. As $S_{\mathrm{Cart}}(t)$ is imprecise, the magnitude of the observed values may not be realistic (see text).

which allowed us to obtain the extrapolated entropy values $S_{\mathrm{Cart}} = \lim_{t\to\infty} S_{\mathrm{Cart}}(t)$ from the fitting parameter a. These values are given in Table 1. Indeed, as previously found,[4,15] all values of S_{Cart} for C1−C6 are of the same order of magnitude, with differences of less than 16%. However, since the values are large, some absolute differences are also large. In particular, the entropies of C3 and C5 are clearly larger than the ones of C1, C2, and C6, and C4 has clearly larger entropy than all other conformations. Differences inside these groups can hardly be resolved within the achieved precision.

Hence, when using Cartesian coordinates we observe entropy differences ranging from 41(13) κ_{B} up to 160(10) κ_{B}. These differences are so large compared to the typical stability of proteins (about 8−25 $\kappa_{\mathrm{B}}T$[16]) that we wonder if they are correct. In fact, we believe that these values suffer from a fundamental problem of the QH approximation in Cartesian coordinates.[33] In general, only constraints, which are described by linear functions of the coordinates, can be accounted for when the pdf is approximated by a multivariate Gaussian distribution. However, typical constraints in proteins are rigid bond lengths and angles which are highly nonlinear functions in Cartesian coordinates. Therefore, these constraints are poorly captured and the QH approximation in Cartesian coordinates overestimates entropies.[33] This effect is particularly severe for the outer atoms of large and rigid side chains such as for example the C^{η} atom of tryptophan. Its true entropy contribution is small (as it is part of the indole rings) while χ^1 and χ^2 torsion angle fluctuations cause a large QH estimate in Cartesian coordinates instead.

Entropy Estimation in BAT Coordinates. We then measured the VE estimate S_{BAT} using bond-angle-torsion (BAT) coordinates (standard tree-like definition). We employed the classical version of the QH approximation[34] given in eq 1, as the semiclassical one applies only to Cartesian coordinates. In this approach, entropies are calculated for a subset $\mathcal{K}$ of so-called "important" coordinates for which the classical entropy expression holds. We used three possible choices, namely, $\mathcal{K} = \{\psi, \phi\}$, $\mathcal{K} = \{\psi, \phi, \chi\}$, and $\mathcal{K}$ = {all torsion angles}. Here, ψ,ϕ denote the 70 backbone torsion angles around the N−C^{α} and C^{α}−C bonds, respectively, χ is the set of 76 flexible side-chain torsion angles, and the last set incorporates all $N - 3$ torsion angles, i.e., also those that are strongly constrained due to the covalent nature of the protein.

Figure 3 shows $S_{\mathrm{BAT}}(t)$ for $\mathcal{K} = \{\psi, \phi, \chi\}$ as a function of t. We shifted all data such that C1 has zero entropy, as only differences are meaningful. The data for C1, C3, and C4 are close to convergence and show entropy differences far beyond the error bars. In the same way as before, we obtained the extrapolated values S_{BAT} given in Table 1. These reveal mutual entropy differences for the conformations C1−C5 ranging from 5(4) κ_{B} to 49(2) κ_{B}. No reliable extrapolation could be obtained for C6; however, as data grow with t, its entropy should be larger than that of C1 and C2. The extrapolated values S_{BAT} for $\mathcal{K} = \{\psi, \phi\}$ and $\mathcal{K}$ = {all torsion angles} are also given in Table 1, while the corresponding convergence plots are shown in Supporting Information. Qualitatively, the results are similar. However, the numerical values depend on $\mathcal{K}$, and more precisely, the entropy differences increase with the size of $\mathcal{K}$.

Also S_{BAT} overestimates the true VE mainly because of multiply populated rotamer states of side chains.[33] Torsion angles associated to side chains, which occasionally jump

Table 1. Extrapolated Entropy Estimates[a]

estimate	$\mathcal{K}$	C1	C2	C3	C4	C5	C6
S_{Cart}	N/A	0(7)	−7(7)	46(11)	153(8)	68(19)	5(7)
S_{BAT}	$\{\psi, \phi\}$	0(1)	4(1)	21(2)	28(1)	14(2)	≥16(2)[c]
S_{BAT}	$\{\psi, \phi, \chi\}$	0(1)	−8(1)	23(2)	41(2)	18(3)	≥13(3)[c]
S_{BAT}	{all torsion angles}	0(9)	−36(9)	30(10)	85(11)	N/A[b]	22(8)
S_{rdl}	$\{\psi, \phi\}$	0(1)	6(1)	14(2)	25(1)	19(1)	≥18(2)[c]
S_{rdl}	**$\{\psi, \phi, \chi\}$**	**0(7)**	**-4(4)**	**12(8)**	**36(6)**	**10(6)**	**≥18(7)**[c]
S_{rdl}	{all torsion angles}	0(12)	−13(13)	25(13)	81(9)	N/A[b]	37(14)

[a]Extrapolated vibrational entropy estimates S_{Cart}, S_{BAT}, and S_{rdl} measured in Cartesian, BAT, and rdl-BAT coordinates, respectively. All data are shifted such that conformation C1 has zero entropy as only entropy differences are meaningful for S_{BAT} and S_{rdl} (S_{Cart} of C1 equals 879(7) κ_B). All values are given in units of $\kappa_B \approx 0.6$ kcal/(302 K·mol). The number in parentheses represents the uncertainty of the last significant figure(s) (concise notation). For S_{BAT} and S_{rdl} we used three sets of coordinates $\mathcal{K}$. All entropy estimates show significant differences between specific conformations. Moreover, sorting the conformations according to their entropies (from small to large entropy) gives very similar results for all estimates. However, the precise values vary since the quality of the QH approximation depends on the choice of coordinates. We conclude that the entropy differences are at least as large as the ones for S_{rdl} with $\mathcal{K} = \{\psi, \phi, \chi\}$ (shown in bold). [b]Insufficient data available. [c]No reliable extrapolation possible.

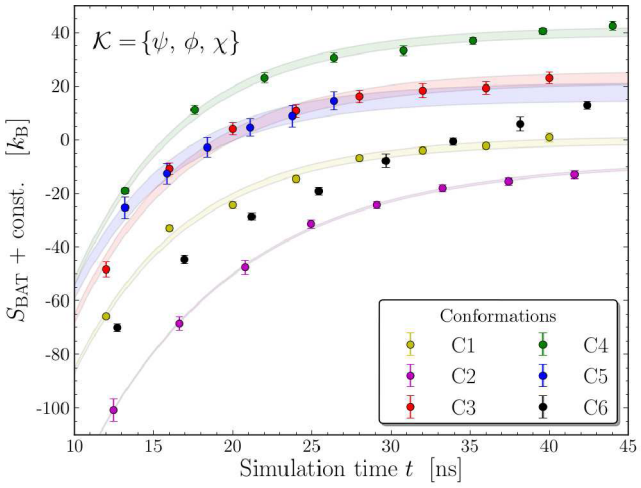

Figure 3. Entropy estimation in BAT coordinates. Convergence plot of $S_{BAT}(t)$ for the coordinate set $\mathcal{K} = \{\psi, \phi, \chi\}$. All data are shifted as only differences are meaningful. Most curves are close to convergence, allowing for reliable extrapolation (see Table 1). This reveals clear entropy differences. An exception is C6 for which only a lower bound on S_{BAT} can be found. Note that the estimate S_{BAT} is imprecise due to multiply populated rotamer states (see text).

between different rotamer states, are not unimodal but multimodal distributed. Hence, they can hardly be represented by a Gaussian. In particular, the standard deviation of the Gaussian fit will be of the order of the peak separation, which is irrelevant and usually much larger than the actual widths of the individual peaks. These pathological fits generate severe systematic errors in the calculation of $S_{BAT}(t)$ since all studied backbone conformations contain many side-chain conformations. Similarly, also multimodal distributed backbone angles (as they can occur in flexible tails and protruding loops) generate some imprecision; however, the major problem is caused by χ angles as they are far more frequently multimodal distributed than ψ and ϕ.

Entropy Estimation in Rotamer-Degeneracy Lifted BAT Coordinates. To overcome this problem, we introduced a new set of coordinates which we termed "rotamer-degeneracy lifted BAT (rdl-BAT) coordinates". This set consists of the unimodal distributed BAT coordinates and a degeneracy lifted version of the multimodal distributed ones. The transformation from BAT to rdl-BAT is described in detail in the section Methods. In short, we identified the torsion angles whose marginal distributions contain more than a single peak and shifted the peaks on top of each other. In this way, the standard deviations of all rdl-BAT coordinates represent the typical peak widths of the coordinates which is essential for the entropy estimation. More precisely, the VE of a multimodal distributed torsion angle becomes (up to a constant) $\kappa_B \log[\sum_i^M p_i \sigma_i^2]/2$ using the QH approximation in rdl-BAT where $\{\sigma_i^2\}$ are the variances of the peaks. As shown in Supporting Information, this is a good approximation for the true expression $\kappa_B \log[\prod_i^M (\sigma_i^2)^{p_i}]/2$, obtained for M Gaussian peaks, because the weighted geometric and arithmetic means in the logarithms are of the same order of magnitude.

In the Supporting Information, we report a benchmark test for the proposed rdl-BAT approach performed on "simplified butane", a system for which the VE can be calculated exactly. The test confirms that the use of rdl-BAT coordinates overcomes the shortcoming of BAT coordinates discussed in the previous paragraph leading to a tremendous increase of the achieved accuracy. Furthermore, in the benchmark test, our method performs similarly well as the second-generation mining minima algorithm of Gilson and co-workers.[35] This is remarkable as our method can be applied to all proteins of arbitrary size, while the M2 method is limited to studies of small peptides.

For the six conformations, we measured the VE estimate $S_{rdl}(t)$ in rdl-BAT coordinates using the classical version of the QH-approximation (eq 1) and the same coordinate sets $\mathcal{K}$ as before. Data for $\mathcal{K} = \{\psi, \phi, \chi\}$ are shown in Figure 4. The error-bars on $S_{rdl}(t)$ are considerably larger than for $S_{BAT}(t)$ as they also account for the uncertainty in the peak detection. The extrapolations for C1−C5 are given in Table 1. As before, S_{rdl} is similar for C1 and C2, larger for C3 and C5, and even larger for C4. Between these groups we can resolve entropy differences of about 13(9) κ_B, 25(10) κ_B, and 38(9) κ_B. The table also shows S_{rdl} for $\mathcal{K} = \{\psi, \phi\}$ and $\mathcal{K}$ = {all torsion angles}, while the corresponding plots are given in Supporting Information. As before, the numerical values depend on $\mathcal{K}$. Within the precision, the entropy differences increase with the size of $\mathcal{K}$.

In summary, we found large VE differences between the conformations *for all coordinate sets* used. We cannot give precise values for the differences as our results were obtained in

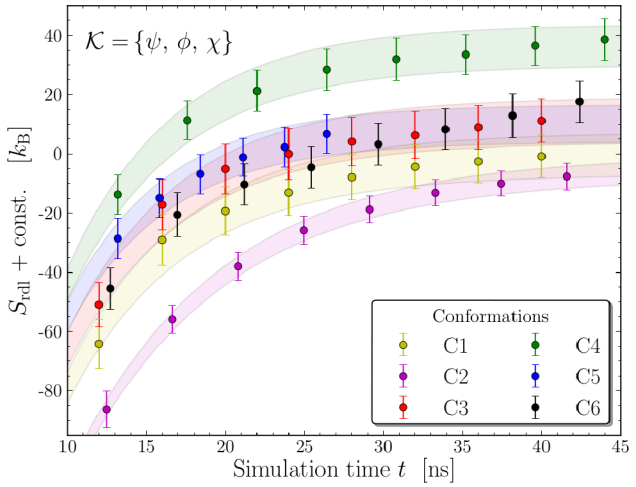

Figure 4. Entropy estimation in rotamer-degeneracy lifted BAT coordinates. Similar plot to Figure 3 but we show $S_{rdl}(t)$ computed using rdl-BAT coordinates instead of ordinary BAT coordinates. The pdf in this new coordinate set is practically unimodal as multimodality (mainly arising from multiply populated rotamer states) is removed. This yields more accurate VE estimates (see text). The error bars are considerably larger than in Figure 3 as they also account for the uncertainty in rotamer detection. The extrapolated values S_{rdl} (see Table 1) reveal clear entropy differences.

QH approximation and we are not able to estimate the systematic error coming from this approximation. However, we are confident that the VE differences are *at least* as large as the ones found for S_{rdl} with $\mathcal{K} = \{\psi, \phi, \chi\}$ (shown in bold in Table 1) because first, the entropy differences grow with increasing number of coordinates in the $\mathcal{K}$ sets used, and second, because the pdf in rdl-BAT can well be approximated by a multivariate Gaussian, as (i) the typical protein constraints are captured, (ii) the pdf is practically unimodal,[36] and (iii) the remaining uncertainties due to anharmonicities are small.[37,38]

DISCUSSION

In this study, we measured the VE of six well-packed backbone conformations of a test protein. We found that the VEs differ substantially between the conformations. More precisely, we observed entropy differences between specific conformations of *at least* 13(9) κ_B, 25(10) κ_B, and 38(9) κ_B, which are of the order of typical protein stabilities (about 8−25 $\kappa_B T$[16]). We cannot rule out that the differences may even be larger.

We measured VEs in quasi-harmonic approximation whose precision depends strongly on the chosen coordinate system. In addition to Cartesian and BAT coordinates, usually used in the literature, we proposed a new set of coordinates (referred to as rotamer-degeneracy lifted BAT coordinates) which overcomes a known shortcoming of the QH approximation in BAT coordinates for multiply populated rotamer states.[33] Using rdl-BAT coordinates gives improved QH estimates, as (i) all typical protein constraints (i.e., stiff bond-length and angles) are captured, (ii) the pdf is practically unimodal,[36] and (iii) remaining anharmonicities are known to have only small effect.[37,38]

Our result is in strong contradiction to a common belief that VE differs only negligibly between well-packed conformations of a protein. We show that, although the VEs of different conformations are of the same order of magnitude,[4,15] they still vary considerably between conformations. Therefore, our result does not support the usual practice of neglecting VE in equilibrium calculations of proteins.

To our knowledge, no direct evidence for the importance of VE on the native state of proteins at ambient conditions has been obtained so far, while plenty of evidence for its importance was reported in studies on related systems involving proteins and peptides.[17−29] However, it stayed unclear to which extent these results carry over to proteins since these systems (albeit similar) have different physical properties. In addition to this indirect support, our results are also plausible by means of the following simple argument: Distinct conformations contain different secondary-structure motifs with different internal properties (such as interactions, packing fractions, etc.), allowing for different thermal motion. For example, atoms are usually stronger correlated in helices than in random coil regions and fluctuate stronger in turns than in beta-sheets.[39] This makes it likely that VE depends on the secondary-structure composition[19−22,40] and, hence, varies between conformations.

Our result has important practical implications for numerical computations on proteins. Basically all software based on free-energy minimization and designed for protein structure prediction, drug design, and NMR refinement neglect VE. However, as VE differs between conformations, a deep "narrow" minimum of a used free energy approximation neglecting VE may have higher free energy than a shallow "broad" one. Therefore, neglecting VE may lead to the incorrect selection of a metastable state instead of the native state and, hence, incorrect output. To our knowledge, only ref 41 considers VE for the prediction of mutant stabilities. The method relies on the *CONCOORD* algorithm[42] producing random perturbations of a given structure based on geometric restrictions. While the method is fast and gives slightly improved predictions,[43] the physical legitimization of *CONCOORD* is not evident. As an alternative, we currently devise a knowledge-based method for VE prediction[44] where we use information on "typical" thermal motion of proteins, similarly to the way knowledge-based rotamer potentials have been developed.

METHODS

In this section we discuss the simulations, the conformation selection process, the quasi-harmonic approximation, the structure alignment, the covariance estimation, the rotamer-degeneracy lifted BAT coordinates, and the error estimation.

Simulations. All simulations were performed by Colombo and co-workers.[35,45,46] The authors described in detail their accurate simulations which we summarize briefly in this paragraph. Five molecular dynamics (MD) simulations were performed starting from different initial structures, namely, one NMR structure and four completely artificial structures, previously generated using Monte Carlo dynamics of a coarse-grained representation of the protein. The structures were accurately translated to an all-atom representation and equilibrated carefully to the simulation conditions. The production runs were performed under NPT conditions at 300 K and 1 bar with explicit (SPC) water, using *GROMOS96*[47] with an integration time step of 2 fs. Differences between the NPT and NVT ensembles could safely be neglected.[5] Each simulation ran for 100 ns. Snapshots were taken every 80 ps and made available through the *Decoys 'R' Us* server[48] after slight energy minimization. The chicken villin headpiece protein has been extensively investigated exper-

dx.doi.org/10.1021/ct500696p | J. Chem. Theory Comput. 2015, 11, 351−359

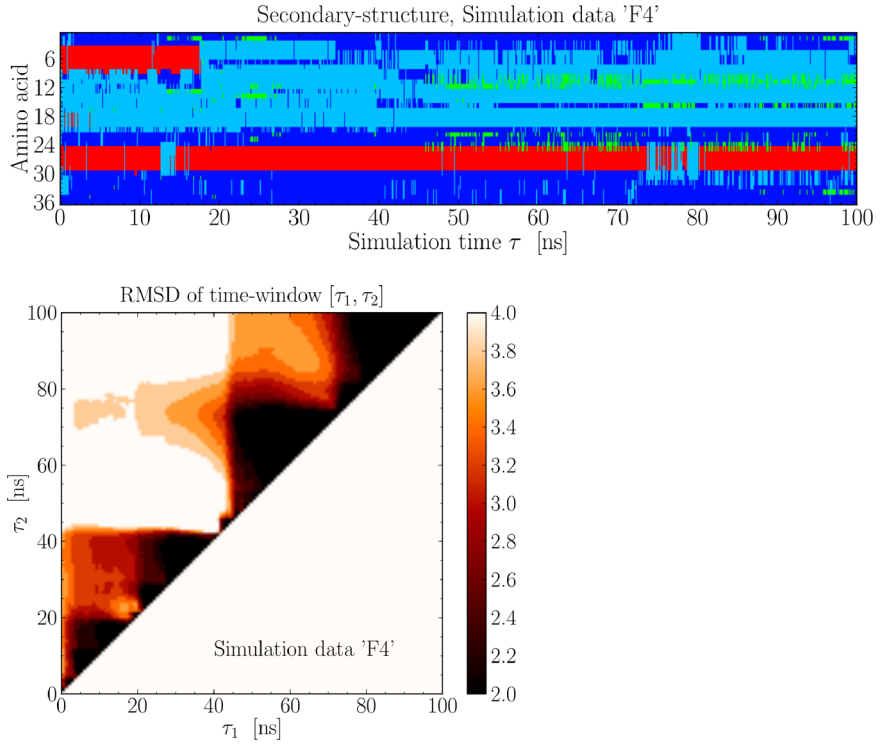

Figure 5. Conformation selection via secondary-structure composition and backbone similarity. Top: The secondary structure (as assigned by *STRIDE*) of one simulation (originally called "F4"[30]) is shown as a function of the simulation time τ and the residue number. Helices, beta-strands, turns, and random-coil regions are represented as red, green, light-blue, and dark-blue areas, respectively. This representation reveals conformational changes occurring during the simulation. Bottom: Average RMSD of the aligned snapshots of a time window $[\tau_1, \tau_2]$ as a function of the limiting times τ_1 and τ_2. Data for the same simulation as in the other panel are shown. Time windows consisting of similar snapshots appear as dark areas of triangular shape.

imentally and also as an ideal benchmark for theoretical studies mainly due to its small size, with no cysteines, which folds autonomously and rapidly.[46,49]

Conformation Selection. From the MD trajectories, we selected six time windows in which the system is in a single conformation. This was necessary as we measured VE in metastable states and the simulations occasionally jumped from one to another metastable state. In this work, we selected conformations regarding two criteria, secondary-structure composition and backbone similarity.

Selection via Secondary-Structure Composition. We used the secondary-structure assignment tool *STRIDE*[50] to extract the secondary-structure composition of all snapshots. Figure 5 (top panel) shows exemplarily the assigned secondary structure of one simulation (originally called "F4"[30]). In it, assigned helices, beta-strands, turns, and random-coil regions are represented by red, green, light-blue, and dark-blue areas, respectively, as a function of the simulation time τ (x-axis) and the residue number (y-axis). This representation shows clearly the presence of conformational changes. While the MD simulation was initialized with a two-helices structure, one of the helices (residues 4−9) disappeared at simulation time τ = 17.5 ns. The other helix was mostly stable throughout the simulation, but for 12.6 ns < τ < 14.5 ns, it also melted completely, and for 72.8 ns < τ < 80 ns it was marginally stable. Further, a beta-sheet (residues 10−12 and residues 22−25) arose at $\tau \approx 46$ ns, disappeared after some time, and recurred. From these observations, we established a first criterion to define a conformation: a conformation consists of the snapshots of a time window in which no such secondary-structure changes occur.

Selection via Backbone Similarity. It is not sufficient to define a conformation only via its secondary structure as spatial rearrangements of those structures are not accounted for. Therefore, we further defined a similarity criterion: All snapshots of a conformation must have a RMSD of less than 3.2 Å. To this end, we superimposed the alpha carbons and phosphorus atoms of all possible time windows using the structural-alignment software *THESEUS*.[51] Then we measured the RMSD for the superposition, more precisely, the average RMSD for all pairwise combinations of snapshots in the ensemble. Figure 5 (bottom panel) shows this value as a function of all possible time windows exemplarily for the simulation discussed in the previous paragraph. Darker (lighter) colors represent smaller (larger) RMSD values. In this representation, time windows of similar snapshots emerge as dark regions of triangular shape.

By applying these two criteria to all available MD data, we selected six conformations sufficiently large to allow for entropy measurements. Finally, we confirmed the selection by visual inspection of the simulation trajectories. Precisely, we defined the conformation C1 (C2, C3, C4, C5, C6) as the time window $\tau \in$ [3.2 ns, 43.2 ns] ($\tau \in$ [57.6 ns, 99.2 ns], $\tau \in$ [20.8 ns, 60.8 ns], $\tau \in$ [55.2 ns, 99.2 ns], $\tau \in$ [46.4 ns, 72.8 ns], and $\tau \in$ [4.8 ns, 47.2 ns]) of the simulation originally[30] termed "NATIVE" ("NATIVE", "F1", "F3", "F4", "F7"). All conformations are visited only once during the simulations.

Quasi-Harmonic Approximation and Alternative Approaches for Entropy Estimation from MD Simulations.

We estimated VE using the well-known quasi-harmonic approximation[31,32,34] which consists of approximating the pdf of the protein's degrees of freedom in some coordinate representation by a multivariate Gaussian distribution with equal covariance matrix Σ. This simplification enables one to derive explicit expressions for the entropy. Two versions of the approach have been proposed. First, a purely classical calculation by Karplus and Kushick[34] yields the formula

$$S_{\mathrm{cl}} = \frac{k_{\mathrm{B}}}{2} \log(\det \Sigma_{\mathcal{K}}) + \text{const} \tag{1}$$

where $\Sigma_{\mathcal{K}}$ is the covariance matrix Σ restricted to a set $\mathcal{K}$ of "important" coordinates. This set $\mathcal{K}$ is not explicitly defined in ref 34; however, it is the set of coordinates which contains the major part of the thermal motion and can be treated classically due to its magnitude. In other words, $\mathcal{K}$ should be chosen in such a way that addition of new coordinates to the set only yields new eigenvalues of small magnitude which have true (quantum mechanical) entropy close to zero and cannot be described classically. Equation 1 can be used for all coordinates $\{q\}$ in which the Jacobian associated to the transformation from Cartesian coordinates to $\{q\}$ does not depend on coordinates of $\mathcal{K}$. In particular, the formula applies to bond-angle-torsion coordinates. The approach, however, lacks a criterion for selecting the set of coordinates $\mathcal{K}$.

Second, a semiclassical calculation by Schlitter[31] and subsequent refinement[32] yields the expression

$$S_{\mathrm{qm}} = k_{\mathrm{B}} \sum_{i=1}^{3N} \frac{\gamma_i}{e^{\gamma_i} - 1} - \log(1 - e^{-\gamma_i}) \tag{2}$$

with $\gamma_i = \hbar\omega_i/(\kappa_{\mathrm{B}}T)$, $\omega_i = (\kappa_{\mathrm{B}}T/\lambda_i)^{1/2}$, and $\{\lambda_i\}$ being the $3N$ eigenvalues of the mass-weighted correlation matrix in Cartesian coordinates. The advantage of this expression is that no distinction between different types of coordinates has to be made. Unfortunately, the formula applies only to Cartesian coordinates.

As reviewed recently,[9] several alternative approaches have been proposed for estimating configurational entropy from snapshots of a MD or Monte Carlo simulation. For molecules and small peptides, nonparametric methods give precise entropy estimates without assuming a particular form of the pdf. Prominent examples such as the mutual information expansion,[52,53] the maximum information spanning tree technique,[54,55] the nearest neighbor technique,[56,57] the hypothetical scanning method,[58] and combinations of these approaches[59] are based on techniques borrowed from information theory. Another important example is the second-generation mining minima algorithm where minima of the energy landscape are first located and then locally scanned including anharmonicities.[35] While these methods are mathematically very elegant and tempting due to their nonparametric nature, their application to larger biomolecules is hampered by their large computational cost. However, in hybrid entropy methods,[9] nonparametric methods are employed "economically" only for specific entropic contributions whose estimation is difficult with the QH approximation. For example, the precision of S_{Cart} could be improved[60] and very accurate entropy estimates were obtained[61] in this way. Moreover, one can exploit the fact that the correlations in proteins are mostly local and approximate the entropy as a sum of entropies of suitable coordinate subsets which then can be derived using nonparametric methods.[62]

Another class of algorithms was designed to improve the QH approximation without compromising its applicability to large biomolecules. In two such approaches, the marginal distributions of the BAT coordinates are estimated accurately using histograms or von Mises distributions while the correlations are treated on QH level or are neglected.[63,64] The method proposed in this article also belongs to this class. We correct both the fluctuations and the correlations for the adverse effect of multimodality. This greatly reduces the error of the QH estimate for VE while the method is as fast and as widely applicable as the QH approximation itself.

Alignment. To measure $S_{\mathrm{Cart}}(t)$, we subtracted the roto-translational motion of the protein in two ways. First, we used the *THESEUS* algorithm as discussed above. Second, we shifted the center of mass of all snapshots to the origin, calculated the principal axes, sorted and directed them appropriately, and rotated the coordinate system such that these axes fell onto the coordinate axes. We then calculated $S_{\mathrm{Cart}}(t)$, as described in Results. Both methods gave results agreeing in the error-bars.

Covariance Estimation. To estimate the covariance matrix of a conformation, we used the standard sample covariance matrix

$$\Sigma_{i,j} = \frac{1}{n-1} \sum_{t=1}^{n} (q_i(t) - \bar{q}_i)(q_j(t) - \bar{q}_j), \quad \text{for } 1 \le i, j \le k \tag{3}$$

where n is the number of snapshots of the conformation and k is the number of coordinates. The placeholder q_i may stand for a component of a Cartesian coordinate of an atom, a BAT coordinate, or a rdl-BAT coordinate (definition below). The quantity $\bar{q}_i$ is the angle mean[65] when q_i is a torsion angle of BAT, it vanishes for the rdl-BAT coordinates, and it is the geometric mean otherwise.

The available data sets are small, i.e., $n \approx 500$ for most conformations (exception C5 with $n = 330$). This limited us in the choice of $\mathcal{K}$. The matrix Σ has $k - n$ zero eigenvalues if $n < k$ (additionally to the zeros due to roto-translation reduction if Cartesian coordinates are used). Schlitter's expression (2) is well-defined, regardless if zero eigenvalues occur or not. In contrast, S_{cl} is only well-defined for $n \ge k$. Therefore, we could not use all BAT coordinates for $\mathcal{K}$. However, we believe that our three choices of $\mathcal{K}$ are sufficient because bond-lengths and angles were confined strongly through the *LINCS* algorithm[66] and the *GROMOS96* force field.[47]

Let us briefly mention an occurring measurement inaccuracy which is in disfavor of adding frozen coordinates to $\mathcal{K}$ anyway. Due to the divergence of the logarithm in S_{cl} in the limit of a zero eigenvalue of $\Sigma_{\mathcal{K}}$, small precision errors in the measure of a tiny eigenvalue lead to a large error on S_{cl}. These errors presumably only cancel in the calculation of entropy differences for very large n.

We also tried alternative estimates for the covariance matrix, usually termed *shrinkage approaches*,[67] which are known to outperform the sample covariance matrix for specific types of correlated data. However, we found that these methods perform poorly in our case, basically (i) because eigenvalues vary on many orders of magnitude and (ii) because not the eigenvalues themselves but their logarithms enter S_{cl}.

Rotamer-Degeneracy Lifted BAT Coordinates. We also measured the entropy S_{rdl} in a new set of coordinates we call

rdl-BAT. The pdf for these coordinates is practically unimodal[36] by construction, and hence, precision problems arising from the fitting of a multimodal distribution to a Gaussian do not occur. The rdl-BAT coordinates are derived from BAT coordinates in the following way:

(1) The marginal distributions of bond-length and angle coordinates in proteins consist of a single narrow peak. We shift these coordinates such that their means are at zero and add them to rdl-BAT.

(2) As discussed in Results, some torsion angles are not unimodal distributed which causes a systematic error of the QH approximation in BAT coordinates.[33] We determined the modality of the marginal distributions $\{\rho(\theta_i)\}$ of the $N - 3$ torsion angles $\{\theta_i\}$ using the *DBSCAN* algorithm[68] with the parameters MinPts = 1, eps $\in [0.11,0.17]$, and the distance function $\text{dist}(\alpha_1, \alpha_2) = \min(|\alpha_1 - \alpha_2|, 2\pi - |\alpha_1 - \alpha_2|)$ of two angles α_1, α_2 in rad. *DBSCAN* groups all n observations $\theta_i(t)$, $t = 1, ..., n$ into N_{peak} clusters such that (i) elements of different clusters have large distance and (ii) all elements of one cluster are connected trough a sequence of elements with pairwise small distance. The clusters were labeled from 1 to N_{peak}. A "membership" function $Z(t)$ was extracted which maps θ_i at time t onto the cluster label it belongs to. By visual inspection of histograms $\rho(\theta_i)$ for a random subset of all torsion angles, we tuned the parameter eps such that the clusters found by *DBSCAN* coincide with the peaks of the histogram. In this way, we obtained the interval for eps given above. The n dependence of this interval could be neglected as C1−C6 have similar n. Alternatively, we could have used other clustering algorithms. *DBSCAN* has the advantage that the total number of clusters does not has to be known beforehand.

(3) For all clusters k $(1 \leq k \leq N_{\text{peak}})$ of θ_i we calculated separately the angle mean $\bar{\theta}_{i,k}$. Then we shifted each observation $\theta_i(t)$ by the mean of its cluster, i.e., $\tilde{\theta}_i(t) = \theta_i(t) - \bar{\theta}_{i,Z(t)}$ (again respecting the 2π periodicity), and added $\tilde{\theta}_i$ to the set of rdl-BAT coordinates.

Uncertainties. The error-bars on $S_{\text{Cart}}(t)$ and $S_{\text{BAT}}(t)$ of Figures 2 and 3 account for the statistical error (obtained with the bootstrap method) and the uncertainty of conformation detection, estimated through shifting all time windows by ±800 ps. The errors of $S_{\text{rdl}}(t)$ additionally reflect the uncertainty of peak detection estimated by varying the parameter eps of *DBSCAN* in the range [0.11, 0.17]. The error on the fit parameter a is the extent in direction of a of the 1-sigma ellipse obtained via synthetic Monte Carlo data sets[69] considering the correlations of individual data points.

■ ASSOCIATED CONTENT

Supporting Information

(1) Additional convergence plots (analogous to Figures 3 and 4), (2) benchmark test for the quasi-harmonic approximation in rdl-BAT coordinates, and (3) error estimation for approximating geometric by arithmetic means. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: martingoethe@ub.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank G. De Mori, C. Micheletti, G. Colombo, F. Fogolari, and S. Tosatto for making available their simulation data. We also thank J. Numata, I. Usón, C. Rovira, A. De Simone, E.-W. Knapp, and H. Grubmüller for helpful discussions.

■ REFERENCES

(1) Lazaridis, T.; Karplus, M. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 133−152.
(2) Feig, M.; Brooks, C. L., III. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217−224.
(3) Chen, J.; Brooks, C. L., III; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140−148.
(4) Karplus, M.; Ichiye, T.; Pettitt, B. *Biophys. J.* **1987**, *52*, 1083−1085.
(5) Zhou, H.-X.; Gilson, M. K. *Chem. Rev.* **2009**, *109*, 4092−4107.
(6) Frauenfelder, H.; Petsko, G. A.; Tsernoglou, D. *Nature (London)* **1979**, *280*, 558−563.
(7) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598−1603.
(8) Brady, G. P.; Sharp, K. A. *Curr. Opin. Struct. Biol.* **1997**, *7*, 215−221.
(9) Suárez, D.; Díaz, N. *WIREs Comput. Mol. Sci.* **2014**, DOI: 10.1002/wcms.1195.
(10) Ohkubo, Y. Z.; Thorpe, I. F. *J. Chem. Phys.* **2006**, *124*, 024910.
(11) Berezovsky, I. N.; Chen, W. W.; Choi, P. J.; Shakhnovich, E. I. *PLoS Comput. Biol.* **2005**, *1*, 322−332.
(12) Zhang, J.; Liu, J. S. *PLoS Comput. Biol.* **2006**, *2*, 1586−1591.
(13) Sciretti, D.; Bruscolini, P.; Pelizzola, A.; Pretti, M.; Jaramillo, A. *Proteins: Struct., Funct., Bioinf.* **2009**, *74*, 176−191.
(14) Abagyan, R. A. In *Computer Simulations of Biomolecular Systems*; van Gunsteren, W. F., Weiner, P. K., Wilkinson, A. J., Eds.; Springer Netherlands: Dordrecht, 1997; Vol. 3, pp 363−394.
(15) Doig, A. J.; Sternberg, M. J. E. *Protein Sci.* **1995**, *4*, 2247−2251.
(16) Makhatadze, G. I.; Privalov, P. L. *Adv. Protein Chem.* **1995**, *47*, 307−425.
(17) Tidor, B.; Karplus, M. *J. Mol. Biol.* **1994**, *238*, 405−414.
(18) Schäfer, H.; Smith, L. J.; Mark, A. E.; van Gunsteren, W. F. *Proteins: Struct., Funct., Genet.* **2002**, *46*, 215−224.
(19) Ma, B.; Tsai, C.-J.; Nussinov, R. *Biophys. J.* **2000**, *79*, 2739−2753.
(20) Chou, K. *Biophys. Chem.* **1988**, *30*, 3−48.
(21) Ismer, L.; Ireta, J.; Neugebauer, J. *J. Phys. Chem. B* **2008**, *112*, 4109−4112.
(22) Rossi, M.; Scheffler, M.; Blum, V. *J. Phys. Chem. B* **2013**, *117*, 5574−5584.
(23) Balog, E.; Becker, T.; Oettl, M.; Lechner, R.; Daniel, R.; Finney, J.; Smith, J. R. *Phys. Rev. Lett.* **2004**, *93*, 028103.
(24) Balog, E.; Perahia, D.; Smith, J. C.; Merzel, F. *J. Phys. Chem. B* **2011**, *115*, 6811−6817.
(25) Wand, A. J. *Curr. Opin. Struct. Biol.* **2013**, *23*, 75−81.
(26) Williamson, J. R. *Nat. Struct. Mol. Biol.* **2000**, *7*, 834−837.
(27) Gupta, V.; Gupta, R. K.; Khare, G.; Salunke, D. M.; Surolia, A.; Tyagi, A. K. *PLoS One* **2010**, *5*, e9222.
(28) Chang, C.-E.; Chen, W.; Gilson, M. K. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1534−1539.
(29) Killian, B. J.; Kravitz, J. Y.; Somani, S.; Dasgupta, P.; Pang, Y.-P.; Gilson, M. K. *J. Mol. Biol.* **2009**, *389*, 315−335.
(30) Fogolari, F.; Tosatto, S.; Colombo, G. *BMC Bioinformatics* **2005**, *6*, 301.
(31) Schlitter, J. *Chem. Phys. Lett.* **1993**, *215*, 617.
(32) Andricioaei, I.; Karplus, M. *J. Chem. Phys.* **2001**, *115*, 6289−6292.
(33) Chang, C.-E.; Chen, W.; Gilson, M. K. *J. Chem. Theory Comput.* **2005**, *1*, 1017−1028.
(34) Karplus, M.; Kushick, J. N. *Macromolecules* **1981**, *14*, 325−332.
(35) Chang, C.-E.; Potter, M. J.; Gilson, M. K. *J. Phys. Chem. B* **2003**, *107*, 1048−1055.

(36) Strictly speaking, the distribution is unimodal if we superimpose the modes of the individual peaks, not their means. However, this difference is entirely irrelevant for practical purposes as realistic peaks are sufficiently symmetric about their modes.
(37) Baron, R.; van Gunstreren, W. F.; Hünenberger, P. H. *Trends Phys. Chem.* **2006**, *11*, 87−122.
(38) Baron, R.; Hünenberger, P. H.; McCammon, J. A. *J. Chem. Theory Comput.* **2009**, *5*, 3150−3160.
(39) McCammon, J. A.; Harvey, S. C. *Dynamics of proteins and nucleic acids*; Cambridge University Press: Cambridge, 1987.
(40) Bongini, L.; Piazza, F.; Casetti, L.; De Los Rios, P. *Eur. Phys. J. E* **2010**, *33*, 89−96.
(41) Benedix, A.; Becker, C. M.; de Groot, B. L.; Caflisch, A.; Böckmann, R. A. *Nat. Methods* **2009**, *6*, 3−4.
(42) de Groot, B. L.; van Aalten, D. M. F.; Scheek, R. M.; Amadei, A.; Vriend, G.; Berendsen, H. J. C. *Proteins: Struct., Funct., Genet.* **1997**, *29*, 240−251.
(43) Kellogg, E. H.; Leaver-Fay, A.; Baker, D. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 830−838.
(44) Goethe, M.; Fita, I.; Rubi, J. M. Manuscript in preparation.
(45) De Mori, G.; Micheletti, C.; Colombo, G. *J. Phys. Chem. B* **2004**, *108*, 12267−12270.
(46) De Mori, G.; Colombo, G.; Micheletti, C. *Proteins: Struct., Funct., Bioinf.* **2005**, *58*, 459−471.
(47) Scott, W. R. P.; Hünenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krüger, P.; van Gunsteren, W. F. *J. Phys. Chem. A* **1999**, *103*, 3596−3607.
(48) Samudrala, R.; Levitt, M. *Protein Sci.* **2000**, *9*, 1399−1401.
(49) Kuzmanic, A.; Pannu, N. S.; Zagrovic, B. *Nat. Commun.* **2014**, *5*, 3220.
(50) Frishman, D.; Argos, P. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 566−579.
(51) Theobald, D. L.; Steindel, P. A. *Bioinformatics* **2012**, *28*, 1972−1979.
(52) Killian, B. J.; Kravitz, J. Y.; Gilson, M. K. *J. Chem. Phys.* **2007**, *127*, 024107.
(53) Numata, J.; Knapp, E.-W. *J. Chem. Theory Comput.* **2012**, *8*, 1235−1245.
(54) King, B. M.; Tidor, B. *Bioinformatics* **2009**, *25*, 1165−1172.
(55) King, B. M.; Silver, N. W.; Tidor, B. *J. Phys. Chem. B* **2012**, *116*, 2891−2904.
(56) Hnizdo, V.; Darian, E.; Fedorowicz, A.; Demchuk, E.; Li, S.; Singh, H. *J. Chem. Theory* **2007**, *28*, 655−668.
(57) Hensen, U.; Grubmüller, H.; Lange, O. F. *Phys. Rev. E* **2009**, *80*, 011913.
(58) Meirovitch, H. *J. Mol. Recognit.* **2010**, *23*, 153−172.
(59) Hnizdo, V.; Tan, J.; Killian, B. J.; Gilson, M. K. *J. Comput. Chem.* **2008**, *29*, 1605−1614.
(60) Numata, J.; Wan, M.; Knapp, E.-W. *Genome Inform.* **2007**, *18*, 192−205.
(61) Suárez, E.; Díaz, N.; D. Suárez, D. *J. Chem. Theory Comput.* **2011**, *7*, 2638−2653.
(62) Hensen, U.; Grubmüller, H. *PLoS One* **2010**, *5*, e9179.
(63) Edholm, O.; Berendsen, H. J. C. *Mol. Phys.* **1984**, *51*, 1011−1028.
(64) Li, D.-W.; Brüschweiler, R. *Phys. Rev. Lett.* **2009**, *102*, 118108.
(65) The angle mean of a set of angles $\{\alpha(t)\}$ in rad is defined as $\arg[\sum_{t=1}^{n} \exp(i\alpha(t))]$.
(66) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463−1472.
(67) Schäfer, J.; Strimmer, K. *Stat. Appl. Genet. Mol. Biol.* **2005**, *4*, 32.
(68) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*; Simoudis, E., Han, J., Fayyad, U. M., Eds.; AAAI Press: Palo Alto, 1996; pp 226−231.
(69) Press, W. H.; Teukolsky, S.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C, The Art of Scientific Computing*, 2nd ed.; Cambridge University Press: Cambridge, 1992.