JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

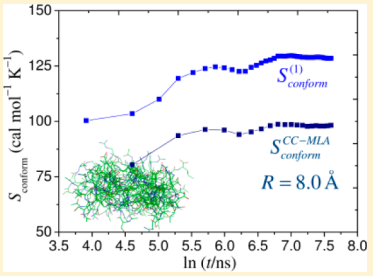
Sampling Assessment for Molecular Simulations Using Conformational Entropy Calculations

Dimas Suárez* and Natalia Díaz

Departamento de Química Física y Analítica, Universidad de Oviedo, Julián Clavería 8, 33006 Oviedo, Asturias, Spain

Supporting Information

ABSTRACT: The extent and significance of conformational sampling is a major factor determining the reliability of long-scale molecular simulations of large and flexible biomolecules. Although several methods have been proposed to quantify the effective sample size of molecular simulations by transforming root mean squared distances between pairs of configurations into statistical/probabilistic quantities, there is still no standard technique for measuring the size of sampling. In this work, we study conformational entropy ($S_{conform}$) as a purely informational and probabilistic measure of sampling that does not require the adoption of any clustering protocol or distance metric between configurations. In addition $S_{conform}$, which is calculated from the probability mass functions associated with discretized dihedral angles, offers other potential advantages for sampling assessment (e.g., global character, thermodynamic significance, relationship with internal degrees of freedom, etc.). The utility of $S_{conform}$ in sampling assessment is illustrated by carrying out test calculations on configurations produced by two extended molecular dynamics simulations, namely, a 2.0 μs trajectory of a highly flexible 17-residue peptide and the trajectory data set of the 1.0 ms bovine pancreatic trypsin inhibitor simulation provided by the D. E. Shaw research group.



INTRODUCTION

During the last years we are witnessing impressive developments in the scope and predictive value of computational biochemistry fueled by continuous improvements in theoretical methods, computational algorithms, and hardware technologies. However, the accurate prediction of free energies and molecular properties of flexible biomolecules in solution is still a challenging problem. Assuming that an all-atom representation of the system of interest is considered and that other choices like the treatment of long-range electrostatic effects and interaction cutoff distances are properly selected, the reliability of molecular simulations depends mainly on two factors: (i) the accuracy and precision of the molecular mechanics (MM) and/or quantum mechanical (QM) methods employed to calculate the potential energy and forces; (ii) the extent and quality of the sampling of the configuration space performed by either Molecular Dynamics or Monte Carlo approaches. Consequently, both error assessment in potential energy models and sampling assessment are of undeniable importance for the application of computational methods to the study of biomolecular systems.

Error Assessment in Energy and Free Energy Calculations. The systematic and random errors present in MM and QM methods have been studied by Merz and co-workers relying on statistical error analyses that yield the total error in the energy of a large biomolecule through the propagation of fragment-based error estimates.[1,2] Similarly, they have investigated how energy function errors depend on system size and how they propagate when computing statistical mechanics-derived thermodynamic quantities.[3] As noticed by the authors, knowledge gained from these studies could pave the way for designing *posthoc* corrections that improve the reliability of energy calculations. On the other hand, poor sampling of important but rare events is commonly recognized as the major source of error in free energy calculations, which usually start from a presumably equilibrated reference system and perturb it to a target system.[4] In fact errors on the ΔG values connecting the two end states that may arise from the choice of the energy methods are expected to cancel out to a large extent. More emphasis is therefore placed on sampling assessment to understand and characterize the uncertainty of ΔG values.[5] Thus, it has been shown that errors in the computed ΔG values can be reduced by increasing the sample size or by using more efficient free energy or sampling methods.

Structural Assessments of Sampling. General sampling assessment has received little attention as noticed by Zuckerman and co-workers,[6] who have emphasized the lack of a standard for quantifying the amount of effective sampling, although several proposals can be found in the literature.[7] Following Zuckerman's view, the effective sample size of dynamical simulations can be expressed as the number N_{eff} of statistically independent configurations generated, which, in turn, can be derived from a simple ratio: $N_{eff} \approx t_{sim}/t^*_{corr}$, where t_{sim} is the simulation time and t^*_{corr} represents the correlation time for the observable physical property that relaxes most slowly along the simulation. However, the choice of the slowest observable is not obvious, and the costly computation of t^*_{corr} can be very sensitive to numerical noise. Therefore, most of the previously proposed measures for sampling assessment rely on structural and dynamical analyses of the configuration space explored by the

Received: May 15, 2014

dx.doi.org/10.1021/ct500700d | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

simulations using clustering techniques, principal component analyses (PCA), and all-to-all root-mean-square deviation (RMSD) analyses. However, these measures of sampling depend on the arbitrary choice of some external parameter (e.g., intercluster distance).[7] To remove such dependency, Zuckerman and co-workers have designed an automated and parameter-free algorithm based on hierarchical RMSD clustering that derives t^*_{corr} from the overall distribution in configuration space, estimating the time that must elapse between snapshots included in a sample for that sample to exhibit the statistical properties expected for independent and identically distributed configurations. The resulting effective sampling size, N_{eff}, is a relative measure of how much sampling is produced by a given simulation, what seems actually more compatible with the statistical nature of the sampling problem than absolute measures attempting to provide a yes/no answer to the question of whether or not convergence has been achieved.[6] However, the applicability of this statistical technique for sampling assessment has been rather limited to date, although test applications on peptide molecules have resulted in t^*_{corr} estimations that lie in the nanosecond time scale.[8,9]

Very recently, a combined probabilistic and structural measure of convergence has been proposed by Koukos and Glykos that employs the Good−Turing statistical formalism to estimate the probability of unobserved configurations based on the observed frequencies of the configurations actually generated by a given molecular simulation.[10] More specifically, this technique builds a probability distribution, p_{unobs}(RMSD), as a function of the RMSD distance between a hypothetical unobserved molecular configuration and all the observed configurations (other metrics than RMSD can be used). From the shape of the p_{unobs}(RMSD) function and its dependence on the extent of sampling, various assessments about the quality and convergence of sampling can be formulated.

Entropy as a Sampling Measurement. The classical configurational entropy (S_{conf}) constitutes a probabilistic measure of the size of the phase space that is accessible to the system of interest through thermal fluctuations. For a single-molecule, S_{conf} is usually defined as the Gibbs−Shannon entropy associated with the probability distribution function of its internal coordinates Q.[11,12]

$$S_{conf}(Q) = -k_B \int \rho(Q) \ln \rho(Q) dQ$$

An alternative definition of S_{conf} can be formulated[13,14] if we assume that the potential energy surface determining the molecular motions can be described as a collection of disjoint harmonic wells, which leads to the following additive approximation for estimating S_{conf}:

$$S_{conf} = \overline{S}_{vib} + S_{conform}$$

where $\overline{S}_{vib}$ is the average vibrational entropy over the set of energy wells and $S_{conform}$ is the conformational contribution that arises from the discrete probability distribution $\{P_\alpha\}$ associated with the population of the different minima.

$$S_{conform} = k_B \sum_\alpha -P_\alpha \ln P_\alpha$$

In principle, entropy methods that directly estimate either S_{conf} or $S_{conform}$ from the configurations generated by a molecular simulation can measure the sampling size. Furthermore, convergence of the computed $S_{conf}/S_{conform}$ values with respect to the sampling size can be indicative of convergence of the simulation. Thus, Genheden and Ryde have questioned whether or not molecular dynamics (MD) of proteins can reach convergence,[15] by computing the marginal contributions of internal rotations to S_{conf} using a direct dihedral histogram (DDH) method. The resulting entropies for three proteins ranging from small to medium-sized systems show relatively poor convergence properties ($\pm$10−70 cal mol^{-1} K^{-1}) after 380−500 ns MD simulations. Similarly, these authors have also found that the DDH entropy along the extra-long 1.0 ms simulation of the bovine pancreatic trypsin inhibitor (BPTI)[16] results in an entropy difference between the first and second halves of the trajectory of 25 cal mol^{-1} K^{-1}. According to Genheden and Ryde, previous entropy calculations reported in the literature for other proteins or peptide molecules exhibit a similar lack of convergence. In particular, a former 1.1 μs simulation of the so-called cc-β peptide,[17] which is a highly flexible 17-residue peptide, has been considered as a representative example of a nonequilibrated trajectory. Overall, the authors of this stimulating paper conclude that it is practically impossible to fully equilibrate MD simulations of proteins given that the configurational entropy in their analyses turns out to be rather dependent on the simulation time. This would suggest the convenience of replacing entropic sampling assessments by specific analyses adapted to the molecular properties of interest in each particular protein simulation.

Goals of the Present Work. Certainly, the analyses and results reported in the literature raise several questions about the actual feasibility of entropy-based measurements of sampling that may be better analyzed in the case of simulations of small, but flexible, peptide molecules. In fact, we have found in previous work that conformational entropy calculations for various peptide systems reach reasonable convergence after 1.0−2.0 μs[14,18,19] and, thereby, we wonder whether or not microsecond-scale simulations of other highly flexible peptide molecules can reach entropic convergence. Moreover, we are also interested in analyzing the role of correlation among the torsional degrees of freedom to assess the quality of sampling. To this end, we reexamine in this work the dynamics of the cc-β peptide (Ace-SIRELEARIRELELRI), whose sequence has been designed to study the transition from a complex formed by a coiled-coil of three α-helices at ambient conditions into β-sheet-rich amyloid fibrils at higher temperatures.[20] Computational methods have been previously employed to investigate the structural stability of the cc-β coiled-coil at various temperatures with the help of MD simulations and quasi-harmonic (QHA) configurational entropy calculations.[21] The structure of the cc-β monomer has also been investigated by Baron et al. in their theoretical work aimed to study the effects of anharmonicity and supralinear correlations on absolute entropy estimations based on the QHA method.[17] Herein, we are only concerned with the cc-β monomer for which we report the results of an all-atom MD simulation in explicit solvent that extended up to 2.0 μs. The large flexibility of cc-β is described with the help of clustering analyses and by the fluctuations of the molecular mechanics Poisson−Boltzmann (MM-PB) energy components of the solute molecule. Then the quality of the 2.0 μs sampling for the cc-β peptide is assessed basing on the N_{eff} method of Zuckerman, the Good−Turing test of Koukos and Glykos as well as on the results of conformational entropy calculations. The same techniques were used to analyze the convergence properties of the 1.0 ms trajectory of the small BPTI protein,[16] which exhibits a rather stable secondary structure that exchanges reversibly among various conformational states on the micro-

dx.doi.org/10.1021/ct500700d | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

second time scale. On the basis of these results, we discuss the ability of the direct $S_{conform}$ calculations to capture the entropic contributions arising from the molecular degrees of freedom and to provide an objective measure of the conformational sampling achieved by molecular simulations.

METHODOLOGY

Simulation Details. Initial coordinates of the cc-β peptide were generated using the LMOD algorithm included in the Amber10 package[22,23] and the ff99SB force field[24] coupled with the Hawkins−Cramer−Truhlar pairwise generalized Born solvent model.[25] A total of 1000 LMOD iterations were computed by exploring a one-frequency vibrational mode, where all the peptide residues were allowed to move, recalculating the eigenvectors every 25 LMOD iterations. The structure with the lowest LMOD energy was then surrounded by a truncated octahedral periodic box of TIP3P water molecules that extended 18 Å beyond the peptide atoms. In addition, a counterion (Na^+) was placed by the *LEaP* program included in Amber10 to neutralize the system, resulting in a total of 298 peptide atoms being solvated by 6579 water molecules under periodic boundary conditions.

Both energy minimization and MD calculations were performed with the *NAMD* 2.7b1 program.[26] The full system was initially relaxed by means of 5000 energy minimization steps and heated gradually to 300 K during 200 ps of MD. The *SHAKE* algorithm was used to constrain all R−H bonds, and periodic boundary conditions were applied to simulate a continuous system. A nonbonded cutoff of 10.0 Å was used, whereas the particle−mesh Ewald method with a grid spacing of ∼1 Å was employed to include the contributions of long-range interactions. Langevin dynamics was employed to control the temperature (300 K) using a damping factor of 2 ps^{-1}, whereas pressure control (1 atm) employed the Berendsen bath coupling. Starting from the thermalized system, a 2.0 μs trajectory was computed with a time step of 2 fs, and coordinates were saved every 1 ps.

The 1.0 ms BPTI trajectory analyzed in this work corresponds to that previously reported by Shaw and co-workers using similar settings to those employed in the cc-β simulation in explicit solvent.[16] However, we only analyzed 4.0 million MD snapshots of the BPTI solute, which was modeled using a variant of the AMBER ff99SB force field.[27]

Trajectory Analyses. Determination of structural properties, clustering of MD configurations, and structurally based sampling assessment were performed using various software tools for the analysis of MD trajectories. Thus, we employed the *PTRAJ* module of Amber10[23] and some other specific software developed locally for RMSD, radius of gyration, and H-bond analyses. Secondary structure assignment was done using the 2002 CMBI version of the *DSSP* program.[28]

Coordinates of the backbone atoms in the cc-β and BPTI trajectories extracted from every 10th snapshot were clustered with a *k*-centers/*k*-medoids algorithm as implemented in the *MSMBuilder*2 package.[29] We adopted the RMSD distance metric and applied the *k*-centers clustering until the intercluster distance had a value of 3.0 Å for the cc-β backbone atoms, or 1.5 Å for the BPTI backbone, followed by 200 iterations of hybrid *k*-medoids refinement. The remaining 90% of the data of each trajectory was assigned to the corresponding clusters. The kinetic clustering resulted in microstate models containing 253 and 532 states for the cc-β and BPTI systems, respectively, so that the solute conformations that interconvert rapidly are grouped into the same microstate. The eigenvalues of the transition probability matrices for changes among the microstates in a certain lag time τ are the key quantities to validate the models as Markov State models (MSMs).[30,31] From such eigenvalues, it is possible to estimate the so-called implied time scales k for interstate transitions. For a perfect Markovian model, the graphical representation of its k values as a function of τ must converge to constant values for a lag time greater than the so-called Markov time. However, the top $k(\tau)$ curves corresponding to the cc-β/BPTI MSM models tend to increase smoothly in the explored lag time intervals (10−500 ps and 1−100 ns for cc-β and BPTI, respectively). This trend, which seems slightly more accentuated in the case of the cc-β trajectory, may be indicative of some structural and kinetic heterogeneity in the microstates. We attempted to use lower values of the intercluster distances to generate less coarse microstate models, but this strategy produced poorer plots of the implicit time scales. We used then the improved Perron−Cluster analyses algorithm[32] to construct macrostate MSM models with a minimal number of states starting from the selected microstate models. After some trial and error, we generated macrostate models containing 16/15 states for the cc-β/BPTI trajectories that exhibit similar implied time scales to those of the parent microstates and partition the phase space into structurally meaningful regions as shown by the time evolution of the radius of gyration/RMSD for cc-β/BPTI.

The number N_{eff} of statistically independent configurations in the cc-β/BPTI simulations was estimated using the latest version of the Zuckerman's methodology[8] as implemented in the *LOOS* package.[33] The determination of N_{eff} as a sampling measure is based on a hierarchical clustering approach that is fully automatized and does not depend on arbitrary parameters. Since this procedure starts by picking up one MD snapshot at random, the N_{eff} analyses for a series of trajectory cuts were repeated 10 times.

Both the calculation of the PCA modes in Cartesian coordinates and the estimation of the Good−Turing p_{unobs}(RMSD) distributions were carried out with the help of the *grcarma* interface[34] to the *carma* program for analysis of large-scale MD trajectories.[35]

Energetic Analyses. MM-PB calculations[36] were performed over 10 000/20 000 MD snapshots extracted from the cc-β/BPTI simulations every 200/50 000 ps, respectively. The cc-β snapshots were postprocessed through the removal of all the water molecules and the Na^+ counterion. The MM-PB energy was computed according to the following expression:

$$G_{\text{MM-PB}} = E_{\text{MM}} + \Delta G_{\text{solv}}^{\text{PB}} + \Delta G_{\text{solv}}^{\text{nonpolar}} - TS_{\text{RRHO}}$$

where E_{MM} is the molecular mechanics energy, ΔG_{solv}^{PB} is the electrostatic solvation energy obtained from Poisson−Boltzmann (PB) calculations,[37] $\Delta G_{solv}^{nonpolar}$ is the nonpolar part of solvation energy[36] and S_{RRHO} is the absolute entropy derived from normal mode calculations. For the cc-β trajectory, $\Delta G_{solv}^{nonpolar}$ is determined as the sum of $\Delta G_{solute-solvent}^{vdW}$, which stands for the vdW interaction energy between the solute and a 12.0 Å shell of explicit water molecules, and the cavitation energy, which is determined by a molecular surface area dependent term (γA). For the BPTI trajectory lacking the coordinates of explicit water molecules, the nonpolar contribution is simply estimated by a solvent-accessible surface area (SASA) dependent term (i.e., $\Delta G_{solv}^{nonpolar} = a\text{SASA} + b$).

The *SANDER* program in Amber10 was used to compute (no cutoff) the molecular mechanics energy terms, while the

dx.doi.org/10.1021/ct500700d | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

electrostatic contributions to the solvation free energy were determined using the *PBSA* program with the atomic charges and radii from the ff99SB force field. We solved the linearized PB equation on a cubic lattice with a grid spacing of 0.33 Å applying the Debye−Hückel potentials at the boundary of the grid defined by the contact surface between the radii of the solute and the radius (1.4 Å) of a water probe molecule.

The van der Waals interaction energies between solute and solvent atoms were determined for a water shell of 12 Å thickness around the solute with no cutoff using *SANDER*. To estimate the cavitation energy, the surface tension proportionality constant γ was set to 69 cal mol^{-1} Å^{-2}, and the molecular surface area was determined using the *MOLSURF* program included in Amber10, applying Bondi radii for the solute atoms and a water probe radius of 1.4 Å.

The RRHO entropy was estimated by molecular mechanics normal mode calculations carried out with the *NAB* package.[38] Prior to the normal mode calculations, the geometries of the systems were minimized until the RMSD of the elements in the gradient vector was less than 10^{-5} kcal mol^{-1} Å^{-1}. These minimizations and the subsequent normal mode calculations were carried out using the Hawkins−Cramer−Truhlar version of the generalized Born model[25] for representing a solvent environment.

Error estimates in the mean values of the MM-PB components were calculated using the *g_analyze* program included in the GROMACS 4.5 suite.[39]

Conformational Entropy Calculations. Conformational entropies were calculated using the *CENCALC* program,[19] which uses both trajectory coordinates and topology information to select the appropriate dihedral angles for describing each rotatable bond and to characterize the conformational states of the molecule of interest. To this end, *CENCALC* discretizes the time evolution of the selected dihedral angles and calculates the probability mass functions of the discretized dihedral angles and those of groups of angles (pairs, triads, ...) to estimate an upper limit to S_{conform} using various methodologies.

In this work, we calculated the marginal (first-order) entropies of all the rotatable bonds in the cc-β and BPTI solutes. We also employed the multibody local approximation (MLA)[18] as implemented in *CENCALC* to include high-order correlation effects among groups of dihedral angles that are presumably more correlated based on a predefined distance-based cutoff R. To this end, the distance between two different dihedral angles θ_A and θ_B defined by the atoms A1−A2−A3−A4 and B1−B2−B3−B4 is computed as $1/4\Sigma_{i=2,3;\ j=2,3}\overline{r}(A_i,B_j)$, where $\overline{r}(A_i,B_j)$ is the average interatomic distance between the atoms A_i and B_j along the considered trajectory length. Assuming that the sum of first-order marginal entropies is reasonably converged, the bias of the MLA entropy due to finite sampling can be removed by means of the correlation-consistent MLA (CC-MLA) entropy estimator.[18] The calculation of the $S_{\mathrm{conform}}^{\mathrm{CC\text{-}MLA}}$ entropy as a function of the cutoff R allows us to determine the best cutoff R, which results in the lower entropy bound and extracts the maximum amount of genuine correlation from the available sampling.

RESULTS

Sampling Assessment of the cc-β Trajectory. *Kinetic Clustering and PCA Analyses.* Figure 1 summarizes the structural and dynamical properties of our all-atom simulation of the cc-β peptide. Thus, the cc-β backbone exhibits relatively ample fluctuations of the radius of gyration (r_{gyr} = 7.3 ± 0.8 Å)

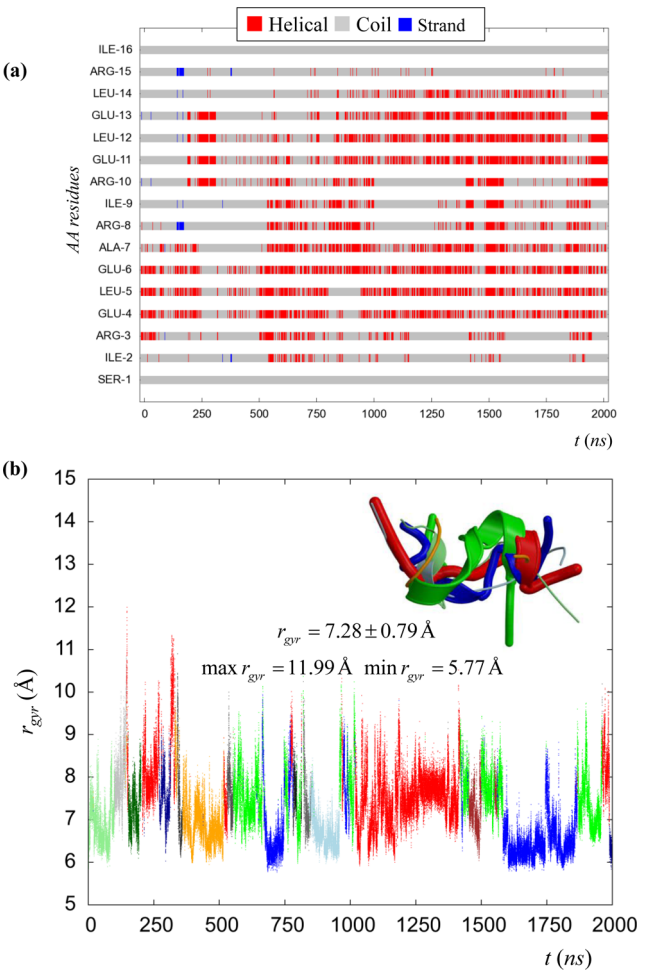


Figure 1. (a) Secondary structure of the cc-β peptide along the 2.0 μs MD simulation. (b) Radius of gyration of the backbone atoms (r_{gyr} in Å). The color of the data points denotes membership to the MSM macrostates. (inset) The superposition of the ribbon models of randomly selected representatives of the macrostates. Thickness of the ribbon models corresponds to population of each macrostate.

that correspond to frequent conformational changes through extended structures with $r_{\mathrm{gyr}} \approx 10-12$ Å. We carried out secondary structure analyses using the DSSP method and found that, although the random coil conformation is the most abundant one, the central residues from Arg_3 to Leu_{14} show helical conformation in ∼20−55% of the analyzed snapshots. This quasi-helical conformation is stabilized by intramolecular H-bond interactions involving backbone groups (e.g., Arg_3−C═O●●●HN−Gln_6, Arg_3−C═O●●●HN−Ala_7, Arg_{10}−C═O●●●HN−Glu_{13}, Arg_{10}−C═O●●●HN−Leu_{14}, Glu_4−C═O●●●HN−Ala_7, etc.), which all have a 30−50% abundance. In addition, other side-chain interactions including two salt-bridge contacts (Arg_3···Glu_6 and Arg_{10}···Glu_{13}) and the Ile_9···Leu_{12} and Leu_{14}···Ile_{16} hydrophobic contacts, also contribute to stabilize the various conformations in ∼30% of the MD snapshots. However, the secondary structure of the peptide changes rapidly and frequently all along the trajectory as seen in Figure 1a, suggesting thus a remarkable structural flexibility of the cc-β monomer at 300 K.

According to our kinetic clustering analyses carried out with the *MSMBuilder* program, the three most important MSM macrostates account for 27%, 20%, and 19% of the MD snapshots and have maximum lifetimes of ∼200, ∼160 and ∼93

dx.doi.org/10.1021/ct500700d | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

ns, respectively (see Figure 1b). These largely populated conformational states exhibit one or two helical turns, while the rest of the states are less populated with abundances ranging from 1 to 8% and present a higher content of residues having a coil conformation. The slowest implied time scales (i.e., relaxation times) associated with the interconversion among the MSM macrostates lie in the range of 30−100 ns. Figure 1b also shows that the trajectory cuts exhibiting stable states are punctuated by series of ∼50−100 ns intervals characterized by frequent exchanges among the MSM macrostates.

To characterize the degree of similarity in the structural fluctuations along the simulation, we computed the PCA overlap Ω_{A-B}, which is defined by Hess as a function of the PCA eigenvalues and eigenvectors computed from the first and second halves of the trajectory (A = 1−1000 ns, B = 1000−2000 ns).[40] The PCA modes were obtained by diagonalizing the mass-weighted covariance matrix in Cartesian coordinates. The resulting Ω_{A-B} values, whose possible values range between 0 (totally dissimilar fluctuations) and 1 (identical structural fluctuations), were 0.58 and 0.64 when considering all-heavy atoms and the backbone atoms, respectively.

Overall, the clustering and PCA analyses point out that although new regions of the conformational space of the cc-β peptide are explored along the MD simulation, the first and second halves of the trajectory still retain an important degree of similarity. This observation may suggest that the 2.0 μs MD simulation of cc-β effectively sampled the most relevant conformational regions available to this flexible molecule, but does not constitute a clear-cut measure of sampling.

Energetic Analyses of the cc-β Trajectory. To complement the structural and dynamical knowledge provided by the kinetic clustering analyses, we characterized the energetic properties of the cc-β trajectory by computing the approximate free energy of the solute molecule according to the MM-PB approach.[36] Figure 2 displays the time evolution of the $G_{MM\text{-}PB}$ values and the various energy components, while the corresponding average values and error estimates are collected in Table 1 for the whole simulation and for the first and second halves.

Inspection of Figure 2 reveals that the MM-PB energy of the cc-β peptide oscillates widely through fast subnanosecond motions as shown by the instantaneous $G_{MM\text{-}PB}$ plot, but also exhibits lower-amplitude oscillations on a longer time scale (∼50−100 ns) as shown by the evolution of the adjacent average along the simulation. Similar but wider oscillations on the short and long time scales can be seen in the anticorrelated plots of the E_{MM} and ΔG_{solv} components that seem in consonance with the winding/unwinding of the central helical twists of the cc−β peptide during the frequent conformational rearrangements.

The S_{RRHO} term accounts only for the entropic effects of the overall translational and rotational motions and mainly of the small amplitude motions of the internal coordinates. We found that the instantaneous values and the adjacent average of the $-TS_{RRHO}$ entropy term exhibit similar fluctuations to those of $E_{MM\text{-}PB}$ and ΔG_{solv}, but their amplitude is considerably lower (e.g., the standard deviation of the $-TS_{RRHO}$ data is ±4.6 kcal mol^{-1} while those for $E_{MM\text{-}PB}$ or ΔG_{solv} are ±40−50 kcal mol^{-1}). Hence, the vibrational entropic contributions seem quite similar for the different conformers of the cc-β molecule.

The long-time oscillations seen in Figure 2 point out that $G_{MM\text{-}PB}$ behaves as a correlated fluctuating quantity in such way that error estimations of its average value should be obtained by means of block averaging.[7] We determined the optimal error

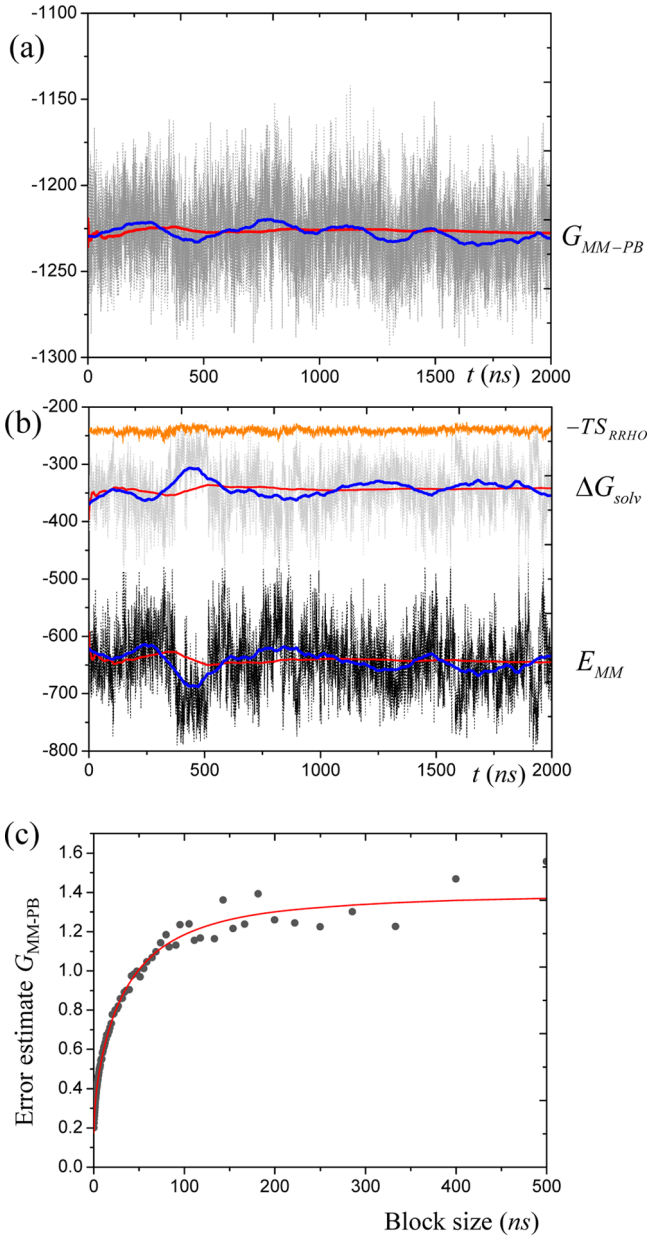


Figure 2. Plots of the time evolution of the MM-PB energy components (in kcal mol^{-1}): (a) $G_{MM\text{-}PB}$ (in dark gray); (b) E_{MM} (in black), solvation energy (in gray) and temperature-weighed RRHO entropy (in orange). The red and blue thick lines represent the evolution of the accumulated and adjacent averages, respectively. (c) Error estimates (●) for $G_{MM\text{-}PB}$ as a function of the block length. The red solid curve is the fit using a double exponential function that includes a short (τ_1) and a long correlation time (τ_2) as fitting parameters.

estimate (ee) for $G_{MM\text{-}PB}$ from the analytical block average fitting curve proposed by Hess,[41] which also renders two correlation times (τ_1 and τ_2) as fitting parameters. Taking into account their ee in parentheses, the mean value $\bar{G}_{MM-PB}$ for the 2.0 μs simulation, −1227.6 (1.4) kcal mol^{-1}, matches the equivalent $\bar{G}_{MM-PB}$ data for the first and second halves of the trajectory, −1225.9 (1.8) and −1229.3 (1.9) kcal mol^{-1}, respectively. We also found that the longest correlation time present in the $G_{MM\text{-}PB}$ data (∼30 ns) is well below the time interval used in the block-averaging process. Altogether, these observations suggest that the cc-β MD simulation yields reasonably well-converged

dx.doi.org/10.1021/ct500700d | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 1. Average Values for the MM-PB Energy Components (in kcal mol^{-1}) of the cc-β Peptide[a]

	$\bar{E}_{MM}$	$\Delta\bar{G}_{solv}$[b]	$-T\bar{S}^{RRHO}_{MM}$	$\bar{G}_{MM\text{-}PB}$
1−1000 ns (first half)	−639.5 (10.9)	−344.9 (8.6)	−241.4 (0.8)	−1225.9 (1.8)
1000−2000 ns (second half)	−650.3 (4.7)	−338.2 (3.2)	−240.8 (0.7)	−1229.3 (1.9)
full trajectory	−644.9 (5.6)	−341.5 (4.1)	−241.1 (0.5)	−1227.6 (1.4)

[a]Block-average error estimates are indicated in parentheses. [b]$\Delta G_{solv} = \Delta G^{PB}_{solv} + \Delta H^{vdW}_{solute-solvent} + \gamma A$

energetic properties for the peptide molecule in spite of the large flexibility of this system.

N_{eff} Calculations and Good−Turing Test. Table 2 and Figure 3 summarize the results of our effective-sampling size

Table 2. Average Value of the Total Effective Sampling Size[a] ($\bar{N}_{eff}$) and Estimated Range of the Slowest Correlation Time (t^*_{corr} in ns) for Different Cuts of the 2.0 μs cc-β Trajectory

interval, ns	$\bar{N}_{eff}$	t^*_{corr}	$\bar{N}_{eff}$	t^*_{corr}
	all atoms		backbone atoms	
0−250	22.3 ± 3.0	9.8−12.9	23.6 ± 4.5	8.9−13.1
0−500	21.6 ± 1.9	21.3−25.3	24.3 ± 2.5	18.6−22.9
0−750	22.8 ± 1.3	31.0−34.8	27.3 ± 6.2	22.3−35.6
0−1000	27.4 ± 5.6	30.3−45.8	30.8 ± 10.4	24.2−49.1
0−1250	29.4 ± 4.3	37.0−49.5	33.7 ± 8.3	29.8−49.3
0−1500	29.0 ± 3.2	46.5−58.1	36.3 ± 6.6	35.0−50.6
0−1750	28.5 ± 4.4	53.1−72.5	34.8 ± 7.3	41.5−63.6
0−2000	31.8 ± 7.7	50.9−83.6	36.0 ± 11.1	42.4−80.5

[a]The full data set containing 2×10^6 geometries of the cc-β solute was used.

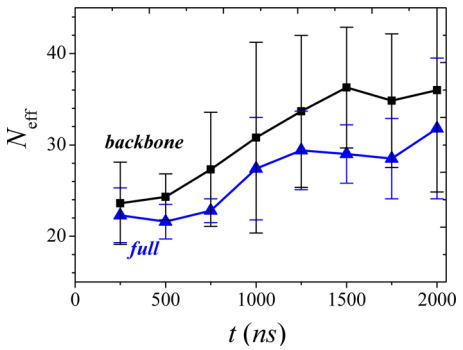


Figure 3. Convergence plot of the average values of N_{eff} for the cc-β trajectory. Error bars account for the standard deviations.

(N_{eff}) calculations on the cc-β trajectory for the backbone and all-heavy atoms. The resulting average values $\bar{N}_{eff}$ tend to increase during the first part of the trajectory, and thereafter they oscillate around $\bar{N}_{eff} \approx 30$ and 35 for the whole cc-β molecule and its backbone chain, respectively. Thus, Figure 3 suggests that the MD simulation reaches structural convergence during the last microsecond and provides a more effective sampling of the backbone chain. In any case, the simulation populates a significant number of effectively independent configurations from which arbitrary properties could be calculated.[6] Concerning the structural correlation times t^*_{corr} determined by the N_{eff} calculations, the resulting values vary significantly for the different subsets of sampling (~40−80 ns), but their order of magnitude resembles those of the long-correlation time (~30 ns) estimated from the $G_{MM\text{-}PB}$ data and of the slowest relaxation times (~30−100 ns) associated with the MSM macrostates. Finally, we note that the large uncertainty of the $\bar{N}_{eff}$ data points, as expressed by their standard deviations, does not allow us to make definitive assessments about improvements in the sampling quality of, for example, the first 1.5 μs ($\bar{N}_{eff}$ = 29.0 ± 3.2 for all-heavy atoms) with respect to the full 2.0 μs trajectory (31.8 ± 7.7).

Figure 4 displays the results of the Good−Turing tests based on the RMSDs for the backbone and all-heavy atoms. As

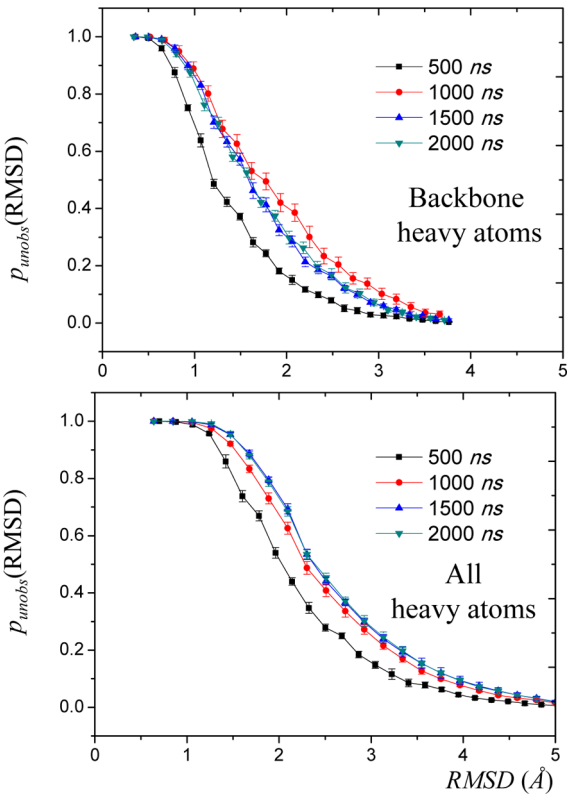


Figure 4. Dependence of the Good−Turing p_{unobs}(RMSD) distributions on the length of subsets of the cc-β trajectory and on the reference atoms for computing the RMSD distance. The analyses were done on 10 000 snapshots extracted every 200 ps from the MD trajectory.

mentioned in the Introduction, this technique estimates the probability p_{unobs}(RMSD) of observing configurations that differ by more than a given RMSD value from the configurations already sampled.[10] By comparing the p_{unobs}(RMSD) curves for different simulation times (0.5, 1.0, 1.5, and 2.0 μs), it turns out that the shape of the probability distribution is hardly changed after 1.5 μs, which might be indicative of a certain structural convergence in the cc-β simulation. It is also seen that the p_{unobs}(RMSD) function for all-heavy atoms adopts consistently higher values and decays to zero more slowly than that for the backbone atoms, suggesting that the simulation provides a more exhaustive sampling of the backbone motions. However, both the backbone and all-heavy atom p_{unobs}(RMSD) curves indicate that the probability of finding new configurations that are relatively close to the explored phase space (e.g., RMSD < 2.0

dx.doi.org/10.1021/ct500700d | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Å) is quite important. This latter feature is probably due to the intrinsic flexibility of the cc-β molecule.

Conformational Entropy Calculations. The simple averaging of the S_{RRHO} term cannot account for the entropy arising from large amplitude motions on the nanosecond time scale that interconnects the distinct conformers of cc-β along the MD simulation. Hence, to assess the quality of the MD sampling in terms of molecular entropy, it is essential to estimate the conformational entropy $S_{conform}$. To this end, the conformational states of cc-β were first characterized by discretizing the time evolution of the internal rotations. From these data, the discrete probability function $P(\theta_j)$ of each rotatable dihedral angle was obtained by applying the maximum likelihood estimator, from which the first-order approximation to the conformational entropy can be computed $S^{(1)}_{conform} = -R\Sigma_j P(\theta_j) \ln P(\theta_j)$.

In principle, the $S^{(1)}_{conform}$ values obtained in this work are similar to the DDH configurational entropies[15] in the sense that the two quantities collect the entropy due to the internal rotational motions of the solute molecule neglecting any correlation among the rotational degrees of freedom. However, $S^{(1)}_{conform}$ is a purely informational entropy term that ultimately depends on the labeling of the conformational states accessible to each dihedral angle θ_j, whereas the DDH entropy depends on the marginal probability density functions associated with θ_j.

In Figure 5a, the total $S^{(1)}_{conform}$ entropy and the separate contributions from the backbone and side-chain dihedral angles are plotted as functions of the length of the simulation. The total $S^{(1)}_{conform}$ rises during the first 250 ns, keeps increasing smoothly up to the first microsecond, and evolves as a nearly zero-slope curve afterward. The same behavior is observed in the side-chain and backbone $S^{(1)}_{conform}$ entropies. We note in passing that a similar graphical assessment of entropy convergence can be formulated in terms of logarithmic plots (see Supporting Information). Nevertheless, to minimize any bias when judging the quality of the $S^{(1)}_{conform}$ convergence, we estimated the uncertainty of the limiting entropy value and compared it with reference to the error estimates of the free energy components provided by the MM-PB approach. To this end, we computed the mean value and the standard deviation of the $S^{(1)}_{conform}$data derived from the last 250 ns. The resulting entropy value, 128.6 ± 0.2 cal mol^{-1} K^{-1}, represents a significant fraction (~15%) of the total entropy $\overline{S}_{RRHO}$ + $S^{(1)}_{conform}$, emphasizing thus the importance of the conformational variability of the cc-β peptide. Its standard deviation, ± 0.2 cal mol^{-1} K^{-1}, can be translated into an uncertainty of the temperature-weighted entropy ($-TS^{(1)}_{conform}$) of ±0.06 kcal mol^{-1}, which is nearly negligible compared with the error estimate for the mean values of the potential energy and solvation energy.

From the acceptable convergence properties of the $S^{(1)}_{conform}$ plot displayed in Figure 5a and the small uncertainty of the $-TS^{(1)}_{conform}$ term, it seems quite reasonable to conclude that the 2.0 μs simulation provided a satisfactory sampling of the rich conformational space accessible to the cc-β peptide. Note, however, that this assessment applies only to the *entire* trajectory. In fact, the $S^{(1)}_{conform}$ calculations indicate that the amount of sampling produced by shorter trajectories would not be enough for achieving converged entropy plots. Thus, either the $S^{(1)}_{conform}$ curves for the first half of the trajectory or those obtained using only the snapshots of the second microsecond exhibit poor convergence (see Figure 5a).

The calculation of the first-order entropy terms $S^{(1)}_{conform}$ is the first step for building more accurate approximations to the conformational entropy including correlation effects. Using the

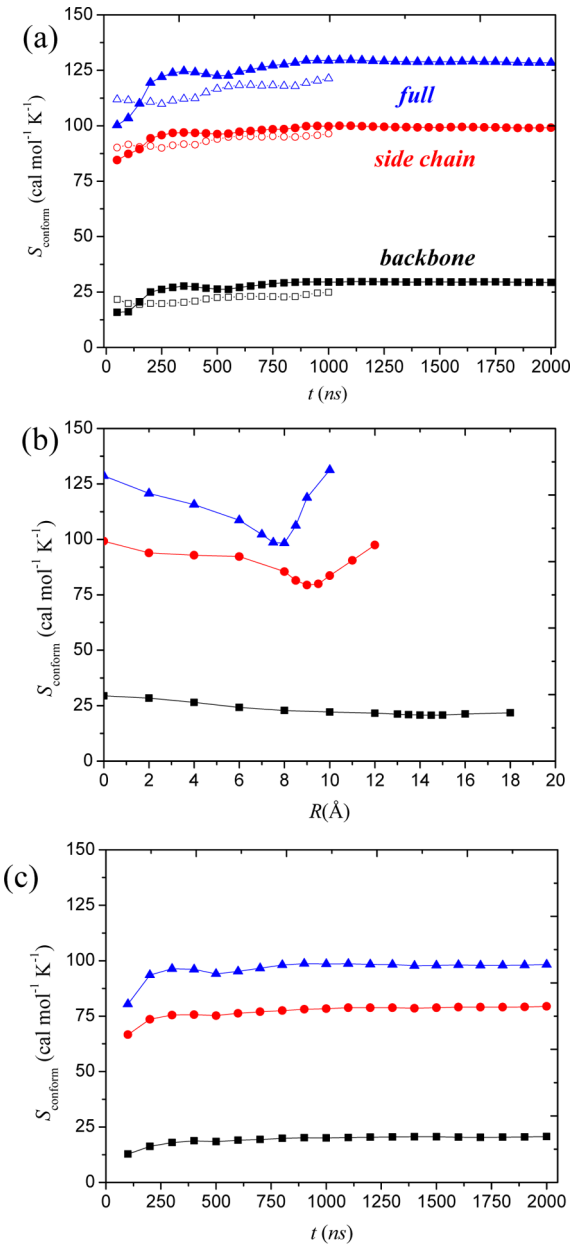


Figure 5. (a) Convergence plots of the first-order contributions to $S_{conform}$ that result from the conformational transitions of 90, 58, and 32 rotatable bonds corresponding to the complete cc-β molecule, its side-chain groups and its backbone chain, respectively. The curves plotted with hollow symbols represent conformational entropy estimations using only the second half of the trajectory. (b) CC-MLA conformational entropies at different cutoff values. (c) Convergence plot of the CC-MLA $S_{conform}$ entropy using a 8.0 Å cutoff.

CENCALC program, we computed the CC-MLA conformational entropy that includes correlation among those groups of dihedral angles that are presumably more correlated basing on a proximity condition expressed by a cutoff distance R. The resulting $S^{CC\text{-}MLA}_{conform}$ values can be taken as upper bounds to the exact entropy so that the optimum value of R is the one giving the lowest entropy estimation.[18] For the cc-β peptide, the corresponding $S^{CC\text{-}MLA}_{conform}$ entropy plot in Figure 5b shows that the global minimum for the total entropy (98.2 cal mol^{-1} K^{-1}) is located at R = 8 Å. By comparing with the marginal entropy, it turns out that inclusion of correlation effects reduces the total

dx.doi.org/10.1021/ct500700d | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

$S_{conform}$ entropy by ~24%. Restricting the $S^{CC\text{-}MLA}_{conform}$ calculations to the separate side-chain and backbone rotations results in optimum cutoffs of 9.0 Å (79.9 cal mol^{-1} K^{-1}) and 14.5 Å (20.6 cal mol^{-1} K^{-1}). The $S^{CC\text{-}MLA}_{conform}(R)$ plots reveal that, unlike the side-chain contribution and the full $S^{CC\text{-}MLA}_{conform}$ entropy, the backbone entropy is smoothly reduced when R increases. Thus, we selected a single optimum cutoff (R = 8 Å) to analyze the convergence properties with respect to the simulation length of the three entropy terms (see Figure 5c) and to estimate the uncertainty of their limiting values (±0.2 cal mol^{-1} K^{-1}), which turned out to be comparable to that of the $S^{(1)}_{conform}$ calculations.

Sampling Assessments of the BPTI Trajectory. To further explore the ability of the structurally based measures of sampling and the $S_{conform}$ calculations to assess the convergence properties of extended MD trajectories, we analyzed the small BPTI protein using 4.0 million configurations extracted from the 1.0 ms MD simulation performed by Shaw and co-workers.[16] The structural, dynamical, and energetic properties of this BPTI simulation have been intensively studied in previous works.[15,16,42] Thus, it has been shown that the secondary structure of BPTI is quite stable along the MD trajectory (e.g., the backbone RMSD fluctuates around ~1.5 Å). However, the analyses of the dynamical content of the trajectory have also revealed that the BPTI backbone exchanged reversibly among various conformational states with lifetimes on the microsecond time scale, while the protein side chains experience large fluctuations on the nanosecond time scale. The computational results have been considered to be in line with experimental NMR observations showing slow internal transitions associated with flipping of aromatic side-chains, bulk-internal water exchange, and disulfide-bridge isomerization.[16]

Kinetic Clustering Analyses. As described in Methodology, we performed kinetic clustering analyses of the BPTI backbone atoms to assign the MD conformations to different macrostates (see Figure 6). The top five implied time scales observed for the

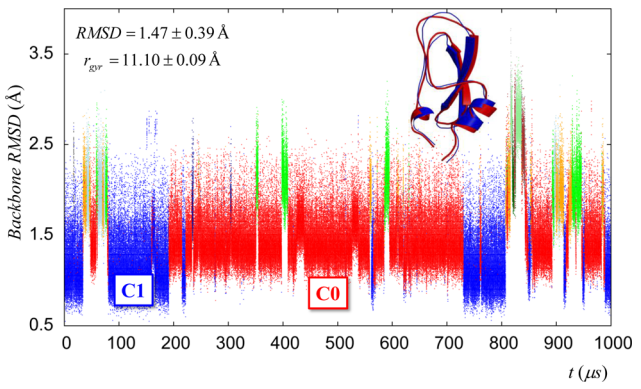


Figure 6. Time evolution of the root-mean-squared deviation (in Å) of the BPTI backbone atoms with respect to the reference X-ray structure. The color of the data points denotes membership to the MSM macrostates generated by *MSMBuilder*. (inset) The superposition of the ribbon models of randomly selected representatives of the **C0** and **C1** macrostates.

MSM macrostates arise in the 1.0−10.0 μs interval. The two most populated macrostates, **C0** and **C1**, which account for 58% and 26% of the MD snapshots, respectively, can be loosely associated with the so-called NMR and crystallographic states.[16] Other conformational states arise during the simulation that deviate more strongly from the reference X-ray structure and tend to be more solvent-accessible, particularly in the 800−850 μs interval (see Figure 6). Notice that the *MSMBuilder* kinetic clustering rendered results very close to those produced by the kinetic clustering protocol employed by Shaw et al.[16]

N_{eff} Calculations and Good−Turing Test. The $\bar{N}_{eff}$ plots shown in Figure 7 reveal that the number of effective structures

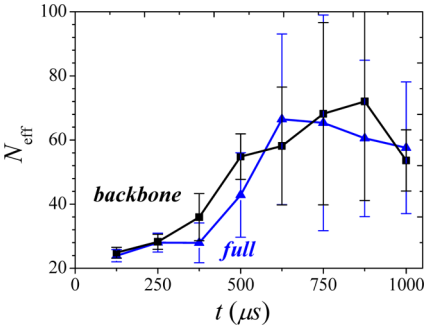


Figure 7. Convergence plot of the average values of N_{eff} for the BPTI trajectory. Error bars account for the standard deviations.

increases rapidly during the first part of the BPTI simulation and then oscillates widely around $N_{eff} \approx 62 \pm 5$ and 63 ± 9 for the all-heavy and the backbone atoms, respectively. These N_{eff} estimations result in approximate values of the slowest correlation time ($t^*_{corr} \approx$ 10−20 μs) that are close to the relaxation times obtained by the kinetic clustering protocol. However, the large magnitude of the error bars in the $\bar{N}_{eff}$ plots as well as the intercrossing of the all-heavy and backbone $\bar{N}_{eff}$ curves impedes us in making a more precise sampling assessment, other than confirming that the BPTI simulation yields a relatively large number of independent structures.

Figure 8 displays the p_{unobs}(RMSD) curves that summarize the results of the Good−Turing tests at various trajectory segments. In contrast with the cc-β simulation, the p_{unobs}(RMSD) distributions for the backbone and all-heavy BPTI at 0.25, 0.5, 0.75, and 1.0 ms atoms are nearly identical to each other. Therefore, it seems that, according to the Good−Turing tests, the BPTI simulation does not access to structurally different regions of the phase space after the first 250 μs. In fact, the p_{unobs}(RMSD) functions decay to zero at low RMSD values, ~1.0 and ~1.5 Å, for the backbone and full protein atoms, which seems in agreement with the global structural stability of the BPTI protein. However, the similar shape of the p_{unobs}(RMSD) curves is somehow inconsistent with the trends shown in the N_{eff} plots.

Conformational Entropy Calculations. Figure 9 displays various $S^{(1)}_{conform}$ plots with respect to the number of frames considered in the conformational entropy calculations using the full set of MD snapshots (4.0×10^6 frames equally spaced every 0.25 ns) or subsets containing the snapshots assigned to the main conformational states **C0** and **C1**. In consonance with the DDH entropy estimated by Genheden and Ryde for the same system,[15] the $S^{(1)}_{conform}$ plot for the whole trajectory in Figure 9 as well as the equivalent logarithmic plot (Supporting Information, Figure S4) show poor convergence properties. During the second and third quarters of the simulation, we observe only small bumps in the $S^{(1)}_{conform}$ curve, which fluctuates around a relatively stable value (~230 cal mol^{-1} K^{-1}). However, a more significant increase of +5 cal K^{-1} mol^{-1} occurs during the last quarter. Thus, the entropy value (234.1 ± 2.1 cal mol^{-1} K^{-1}) estimated by averaging the $S^{(1)}_{conform}$ values from the last 250 μs has a significant uncertainty of ±0.6 kcal mol^{-1} in terms of

dx.doi.org/10.1021/ct500700d | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

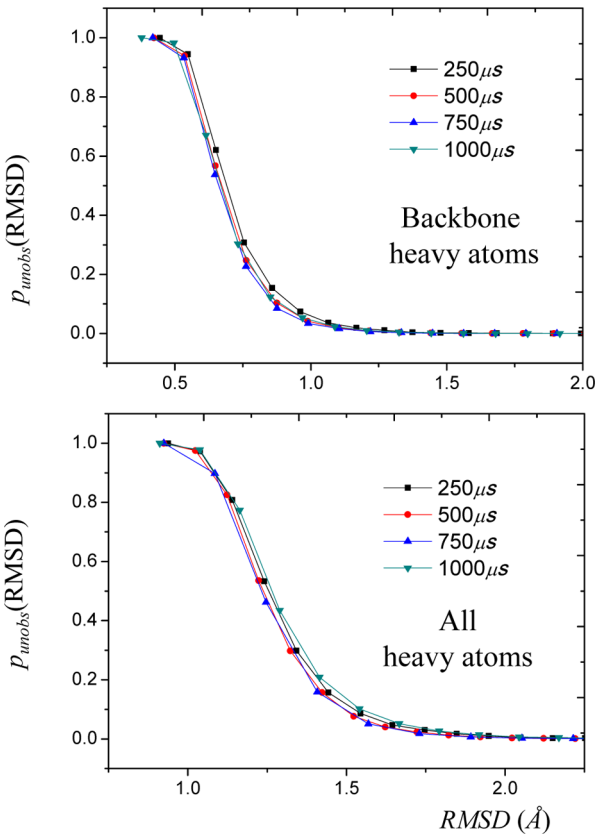


Figure 8. Dependence of the Good–Turing p_{unobs}(RMSD) distributions on the length of subsets of the BPTI trajectory and on the reference atoms for computing the RMSD distance. The analyses were done on 400 000 snapshots extracted every 2500 ps from the MD trajectory.

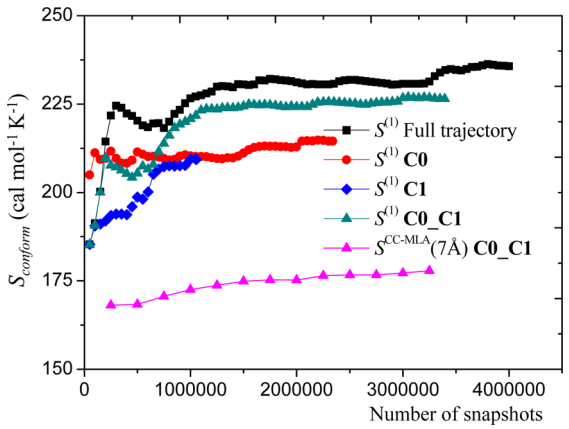


Figure 9. Convergence plots of the first-order contributions to S_{conform} from the 218 rotatable bonds of the BPTI protein. The calculations were performed for the whole set of MD frames and for subsets of frames assigned to the main conformational macrostates (**C0** and **C1**). The CC-MLA entropy at R = 7 Å for the **C0_C1** set is also shown.

temperature-weighted entropy, which lies in between the error estimates for the $\overline{G}_{\text{MM-PB}}$ (ee = 1.2 kcal mol^{-1}) and $-T\overline{S}_{\text{RRHO}}$(ee = 0.2 kcal mol^{-1}) energetic terms calculated with the MM-PB protocol (for the sake of brevity, these MM-PB analyses are summarized in Table S1 and Figure S3 in the Supporting Information).

When comparing with the entropy plots for the cc-β peptide, the worse convergence in the BPTI entropy calculations is not entirely unexpected as BPTI has a larger number of rotatable bonds (218), and the number of frames taken from the 1 ms BPTI simulation (i.e., 4×10^6) only doubles that from the 2.0 μs cc-β trajectory. Given that the previous computational experience on smaller systems[14,18,19] has consistently shown that the estimation of converged probability mass functions for internal rotations demands the accumulation of data from millions of configurations, fast sub-nanosecond conformational motions could be undersampled in the BPTI data set. This has been also noticed in a recent study on the relationship between conformational entropy and NMR order parameters.[43] Hence, the convergence behavior of $S^{(1)}_{\text{conform}}$ could be improved by using a larger number of MD frames extracted at shorter intervals from the same BPTI simulation. Furthermore, the limited sampling of the slow backbone conformational changes occurring on the much longer microsecond time scale also affect the value of $S^{(1)}_{\text{conform}}$. This effect is best illustrated by the time coincidence of the observed $S^{(1)}_{\text{conform}}$ increase during the last quarter of the trajectory with the conformational rearrangements reflected by the RMSD and kinetic clustering analyses (see Figures 6 and 9). In fact, the sudden increase in the $S^{(1)}_{\text{conform}}$ plot at 800–850 μs stresses that the dynamics of BPTI undergoes a qualitative change leading to new conformations in this stage of the trajectory.

From the previous considerations, it can be concluded that the BPTI simulation does not provide enough sampling for achieving tight convergence in the S_{conform} contributions arising from the slow backbone conformational changes. Hence, it may be more convenient to restrict the assessment of conformational sampling to each of the long-lived macrostates determined by the kinetic clustering analysis. Basically this approach has already been used by Gilson and co-workers[42] in their energetic and entropic analyses of the BPTI simulation, showing significant differences in the enthalpic and entropic free energy contributions of the three most abundant kinetic clusters. The results of our $S^{(1)}_{\text{conform}}$ calculations for the **C0** and **C1** states using ~1.0 and ~2.3 million frames, respectively, suggest that these two clusters may have conformational entropies of ~214 and ~209 cal mol^{-1} K^{-1}, respectively, but the corresponding entropy plots are clearly far from reaching a stable plateau; that is, the number of frames considered in the $S^{(1)}_{\text{conform}}$ calculation is insufficient for yielding a satisfactorily converged entropy estimate. When the **C0** and **C1** clusters are merged into the **C0_C1** set with ~3.3 million frames, the resulting $S^{(1)}_{\text{conform}}$ entropy plot has slightly better convergence properties than the separate **C0** and **C1** $S^{(1)}_{\text{conform}}$ curves. Thus, the average value for the **C0_C1** $S^{(1)}_{\text{conform}}$ entropy that results from the last quarter of the data set amounts to 226 ± 0.7 cal mol^{-1} K^{-1}, which has a comparable uncertainty to that of $\overline{S}_{\text{RRHO}}$. Assuming that the $S^{(1)}_{\text{conform}}$ entropy of the joint **C0_C1** cluster is converged, the influence of correlation effects can be estimated by means of the CC-MLA technique, which yields a roughly converged $S^{\text{CC-MLA}}_{\text{conform}}$ value of ~177 ± 0.5 cal mol^{-1} K^{-1}. This figure is obtained for an optimum cutoff distance R = 7.0 Å, which corresponds to the global minimum of $S^{\text{CC-MLA}}_{\text{conform}}$ as a function of R (Supporting Information, Figure S5) and is 1.0 Å shorter than the optimum R for the cc-β trajectory. Overall, on the basis of the S_{conform} estimations for the **C0** and **C1** clusters, it may be reasonably expected that well-converged values for the segregated conformational entropy could be calculated provided that a shorter sampling interval is used as aforementioned. However,

dx.doi.org/10.1021/ct500700d | J. Chem. Theory Comput. XXXX, XXX, XXX–XXX

the actual significance of these cluster entropies is not entirely clear due to the global character of $S_{conform}$ that prevents its exact partitioning across disjoint regions of the molecular phase space. For example, when the **C0** and **C1** data sets are merged, the resulting $S^{(1)}_{conform}$ entropy differs appreciably by ∼14 cal mol^{-1} K^{-1} from the population-weighted average value of the separate **C0** and **C1** entropies.

DISCUSSION

The MD simulations analyzed in this work represent two different scenarios of equilibrium sampling of biomolecules. On one hand, the cc-β peptide is a highly flexible molecule that populates a structurally diverse set of conformers that interconvert into one another through structural fluctuations along the picosecond and nanosecond time scales. On the other hand, the dynamics of the highly stable BPTI protein during the 1.0 ms MD simulation encompasses a wider range of time scales: the side chains motions exhibit a liquidlike dynamics, while the solidlike backbone chain undergoes slow transitions on the microsecond time scale.[16] For the two systems, the kinetic clustering calculations and the energetic analyses suggest that the MD simulations may provide sufficient conformational information to estimate their energetic and mechanical properties at 300 K. However, such analyses do not constitute an objective measure of the amount of sampling performed by the MD simulations.

The calculation of either the number of statistically independent configurations (N_{eff}) or the Good−Turing probability distribution (p_{unobs}(RMSD)) yields parameter-free measures of sampling by transforming the RMSD distances between pairs of configurations into statistical/probabilistic quantities. For the cc-β simulation, the application of the two approaches to a series of extending trajectory cuts results in apparently converged sampling measures with regard to the simulation length; that is, both the $\bar{N}_{eff}$ value and the p_{unobs}(RMSD) curve remained approximately constant during the last quarter of the trajectory. These observations suggest that the cc-β configurations generated in the last phase of the trajectory hardly increase the structural and informational content of the simulation. The large fluctuations (±4, ± 7) in the mean values of N_{eff} and the long tail of the p_{unobs}(RMSD) function are probably due to the highly flexible character of the cc-β peptide. However, it remains unclear whether or not these should be considered as indications that more sampling is needed.

For the BPTI protein, the N_{eff} and Good−Turing techniques differ in their account of the sampling size. While the $\bar{N}_{eff}$ plots indicate again that the number of effective configurations tends to increase with the trajectory length, the p_{unobs}(RMSD) curves remain nearly unaltered along the BPTI simulation. Unfortunately, in spite of the rigidity of BPTI, the fluctuations in its N_{eff} values are also larger than those observed for cc-β, further blurring the formulation of convergence assessment based on the N_{eff} data. Furthermore, the comparative assessment of the sampling quality between the cc-β and BPTI simulations is not obvious. Thus, besides the size of the data set, it is clear that the magnitude and fluctuations of the N_{eff} values are determined by the system size and its intrinsic dynamical properties. Similarly, the shape of the probability distribution p_{unobs}(RMSD) also depends on the system properties. Therefore, we conclude that the N_{eff} and Good−Turing analyses introduce some ambiguity when evaluating the overall quality of the MD sampling for the cc-β peptide and BPTI systems.

Assessing the convergence of MD simulations in terms of conformational entropy calculations may have the following advantages with respect to the RMSD-based or other structurally based measurements:

- First, $S_{conform}$ is defined as a purely informational and probabilistic quantity that depends only on the probability mass function associated with the discretized dihedral angles; that is, the estimation of $S_{conform}$ does not require the adoption of any clustering protocol or distance metric between configurations.
- For a given trajectory segment, the calculation of $S_{conform}$ using all the available MD frames yields a unique value because the Shanonn entropy is a global property that cannot be localized to particular configurations. As a consequence, the plotting of $S_{conform}$ as a function of the simulation time constitutes a clear-cut graphical assessment of the extent of conformational sampling.
- The selection of subsets of dihedral angles, together with the use of many-body or cutoff-based expansion techniques for progressively approaching the total $S_{conform}$, deliver additional information concerning both the physical picture of the internal motions in the molecular system and the sampling efficiency.
- The thermodynamic meaning of $S_{conform}$ and $-TS_{conform}$ allows us to compare the convergence degree of $S_{conform}$ with the error estimates of other free energy components and, more particularly, with that of the absolute RRHO entropy.
- Direct $S_{conform}$ calculations enable us to perform a straightforward and balanced comparison of the conformational sampling efficiency among MD trajectories on different molecular systems.

The $S_{conform}$ plots for the cc-β simulation displayed in Figure 5 illustrate the just-aforementioned properties of conformational entropy as a measure of sampling. First, the time evolution of the first-order contributions to $S_{conform}$ derived separately from each dihedral angle approaches a reasonably converged entropy value of ∼128 cal mol^{-1} K^{-1} when all the data gathered from the 2.0 μs simulation are included. Although a similar $S^{(1)}_{conform}$ value is obtained after just 1.0 μs, either the first or second microsecond of the trajectory yields $S^{(1)}_{conform}$ curves with worse convergence properties. Only when the simulation is extended from 1.0 to 2.0 μs and the accumulated $S^{(1)}_{conform}$ estimation is hardly affected by the alternating trajectory phases of fast/slow conformational change, a reliable convergence check of the whole trajectory can be formulated. In terms of thermodynamic properties, the uncertainty in the temperature-weighted entropy $-TS^{(1)}_{conform}$ (<0.1 kcal mol^{-1}) estimated during the latter stages of the simulation is well below the block-averaging error estimates for the MM-PB approximate free energy (∼1.4 kcal mol^{-1}) and the absolute RRHO entropy (∼0.5 kcal mol^{-1}) of the cc-β molecule. Therefore, these observations together with former $S_{conform}$ calculations on other peptide systems[18,19] suggest that MD trajectories on the microsecond time scale are required to sample efficiently the phase space of systems experiencing conformational motions on the sub-nanosecond and nanosecond time scales.

Segregation of the total $S^{(1)}_{conform}$ entropy into the backbone and side chain contributions shows that, in the case of the cc-β peptide, the first-order entropy of the two sets of dihedral angles have the same convergence properties along the simulation. However, when we include the correlation effects among the

dihedral angles in the conformational entropy calculations using the CC-MLA technique, our entropic assessment on the convergence of the 2.0 μs simulation is enhanced based on the optimum cutoffs for the $S_{conform}^{CC\text{-}MLA}$ entropies. On one hand, the conformational space accessible to the backbone chain seems exhaustively explored given that the optimum cutoff (>14 Å) for the backbone $S_{conform}^{CC\text{-}MLA}$ entropy seems long enough to include the most relevant long-range correlation effects. On the other hand, the shorter cutoffs of 8.0−8.5 Å for the $S_{conform}^{CC\text{-}MLA}$ total/side chain entropy reveal that the sampling of the conformational transitions involving side-chain or side-chain/backbone dihedral angles captured short- and medium-range correlation effects. Therefore, the sampling assessment of the cc-β trajectory in terms of $S_{conform}^{(1)}$ and $S_{conform}^{CC\text{-}MLA}$ shows that the simulation succeeded in capturing the essential information associated with the conformational changes of the individual dihedral angles, the correlated motions of the backbone chain atoms, and the medium-range correlation effects among the side-chain and side-chain/backbone dihedral angles. What is missing then is reliable information about the presumably weak long-range correlation effects.

On the basis of the worse convergence of the first-order $S_{conform}$ calculations, we can unambiguously conclude that the conformational sampling achieved by the 4.0 million BPTI configurations is less exhaustive than that performed by the cc-β simulation. Most probably, the sporadic conformational changes from the two most important **C0** and **C1** states to other structurally dissimilar macro states hold the key to explaining the worse entropy convergence, although other factors of the BPTI simulation may contribute as well (i.e., the larger number of degrees of freedom, the potential subsampling of side-chain on the sub-nanosecond and nanosecond time scales). Better convergence of the $S_{conform}$ calculations results when the MD configurations are separated depending on their cluster assignment, leading thus to cluster-based conformational entropies that may be useful to characterize the thermodynamic properties of the main BPTI conformational states. However, this partitioning approach depends on the clustering method and ignores the entropy arising from the conformational fluctuations on the long time scales and/or the interconversion among the various conformational states. Another caveat to the entropic sampling assessment of the BPTI trajectory is that force field errors could have eventually accumulated during the trajectory and induced spurious conformational changes. This possibility seems particularly plausible on the basis of a recent study[44] showing that a series of extended 100 μs long MD simulations have poorly performed in the refinement of protein structure homology models given that many of the MD models drifted away from the initial structures owing to the limitations in the accuracy of the current force fields. Similarly, one or more of the least populated conformational states unveiled by the BPTI simulation might be a consequence of potential force-field artifacts.

SUMMARY

On the basis of the exhaustive analyses of the cc-β simulation and the BPTI trajectory, we conclude that conformational entropy plots can facilitate the assessment of both the strengths and weaknesses of the sampling performed by MD simulations and complement the information provided by the clustering methods and other structurally based measures of sampling. More specifically, the calculation of the first-order contribution $S_{conform}^{(1)}$, which is computationally straightforward and inexpensive, characterizes to a considerable extent the degree of convergence of the conformational sampling. More detailed assessments can be made when correlation among the dihedral angle motions is included by means of the cutoff-based CC-MLA calculations, although they also demand larger computational resources. Besides evaluating conformational sampling, the estimation of $S_{conform}$ is also useful to complement the results of approximate free energy calculations. Overall, as MD trajectories of biomolecules are rapidly approaching toward the microsecond−millisecond time scales, the conformational entropy measures of the sampling size in combination with clustering methodologies can be really helpful to analyze the successes and challenges faced by molecular simulations.

ASSOCIATED CONTENT

Supporting Information

Implied time scales for the micro- and macrostates generated by kinetic clustering (Figure S1). Convergence plots of the first-order and CC-MLA conformational entropies of cc-β on a logarithmic scale (Figure S2). Data from MM-PB energy analyses of the BPTI system (Table S1 and Figure S3). Average values and correlation time for the N_{eff} analyses on the BPTI trajectory (Table S2). Convergence plots of the first-order and CC-MLA conformational entropies of BPTI on a logarithmic scale (Figure S4). CC-MLA conformational entropy of the main BPTI clusters at different cutoff values (Figure S5). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: dimas@uniovi.es.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We wish to thank the D. E. Shaw research group for having provided us the 1 ms BPTI trajectory. We are also grateful to one of the referees for having suggested the application of kinetic clustering methodologies.

REFERENCES

(1) Faver, J. C.; Merz, K. M. J. Fragment-Based Error Estimation in Biomolecular Modeling. *Drug Discovery Today* **2014**, *19*, 45−50.
(2) Faver, J. C.; Benson, M. L.; He, X.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Kennedy, M. R.; Sherrill, C. D.; Merz, K. M. J. Formal Estimation of Errors in Computed Absolute Interaction Energies of Protein-Ligand Complexes. *J. Chem. Theory Comput.* **2011**, *7*, 790−797.
(3) Faver, J. C.; Yang, W.; Merz, K. M. The Effects of Computational Modeling Errors on the Estimation of Statistical Mechanical Variables. *J. Chem. Theory Comput.* **2012**, *8*, 3769−3776.
(4) *Free Energy Calculations;* Chipot, C., Pohorille, A., Eds.; Springer-Verlag: Berlin, Heidelberg, 2007.
(5) Lu, N.; Woolf, T. B. Understanding and Improving Free Energy Calcualtions in Molecular Simulations: Error Analysis and Reduction Methods. In *Free Energy Calculations*; Chipot, C., Pohorille, A., Eds.; Springer-Verlag, Berlin Heidelberg, 2007; pp 199−247.
(6) Zuckerman, D. M. Equilibrium Sampling in Biomolecular Simulation. *Annu. Rev. Biophys.* **2011**, *40*, 41−62.
(7) Grossfield, A.; Zuckerman, D. M. Quantifying Uncertainty and Sampling Quality in Biomolecular Simulations. In *Annual Reports in Computational Chemistry;* Ralph, A. W., Ed.; Elsevier: The Netherlands, 2009; Vol. 5, pp 23−48.

(8) Zhang, X.; Bhatt, D.; Zuckerman, D. M. Automated Sampling Assessment for Molecular Simulations Using the Effective Sample Size. *J. Chem. Theory Comput.* **2010**, *6*, 3048−3057.
(9) Lyman, E.; Zuckerman, D. M. On the Structural Convergence of Biomolecular Simulations by Determination of the Effective Sample Size. *J. Phys. Chem. B* **2007**, *111*, 12876−12882.
(10) Koukos, P. I.; Glykos, N. M. On the Application of Good−Turing Statistics to Quantify Convergence of Biomolecular Simulations. *J. Chem. Inf. Model.* **2014**, *54*, 209−217.
(11) Zhou, H.-X.; Gilson, M. K. Theory of Free Energy and Entropy in Noncovalent Binding. *Chem. Rev.* **2009**, *109*, 4092−4107.
(12) Suarez, D.; Díaz, N. Direct Methods for Computing Single-Molecule Entropies from Molecular Simulations. *WIREs Comput. Mol. Sci.* **2014**, DOI: 10.1002/WCMS.1195.
(13) Karplus, M.; Ichiye, T.; Pettit, B. M. Configurational Entropy of Native Proteins. *Biophys. J.* **1987**, *52*, 1083−1085.
(14) Suárez, E.; Díaz, N.; Suárez, D. Entropy Calculations of Single Molecules by Combining the Rigid-Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations from Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2011**, *7*, 2638−2653.
(15) Genheden, S.; Ryde, U. Will Molecular Dynamics Simulations of Proteins Ever Reach Equilibrium? *Phys. Chem. Chem. Phys.* **2012**, *14*, 8662−8677.
(16) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; W, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330*, 341−346.
(17) Baron, R.; Hünenberger, P. H.; McCammon, J. A. Absolute Single-Molecule Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction Terms and Convergence Properties. *J. Chem. Theory Comput.* **2009**, *5*, 3150−3160.
(18) Suárez, E.; Suárez, D. Multibody Local Approximation: Application to Conformational Entropy Calculations on Biomolecules. *J. Chem. Phys.* **2012**, *137*, 084115.
(19) Suárez, E.; Díaz, N.; Méndez, J.; Suárez, D. CENCALC: A Computational Tool for Conformational Entropy Calculations from Molecular Simulations. *J. Comput. Chem.* **2013**, *34*, 2041−2054.
(20) Kammerer, R. A.; Kostrewa, D.; Zurdo, J.; A, D.; Garcıa-Echeverrıa, C.; Green, J. D.; Muller, S. A.; Meier, B. H.; Winkler, F. K.; Dobson, C. M.; Steinmetz, M. O. Exploring Amyloid Formation by a De Novo Design. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 4435−4440.
(21) Missimer, J. H.; Steinmetz, M. O.; Baron, R.; Winkler, F. K.; Kammerer, R. A.; Daura, X.; Van Gunsteren, W. F. Configurational Entropy Elucidates the Role of Salt-Bridge Networks in Protein Thermostability. *Protein Sci.* **2007**, *16*, 1349−1359.
(22) Kolossváry, I.; Guida, W. C. Low-Mode Conformational Search Elucidated: Application to $C_{39}H_{80}$ and Flexible Docking of 9-Deazaguanine Inhibitors into Pnp. *J. Comput. Chem.* **1999**, *20*, 1671−1684.
(23) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; R. C..Walker; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; I..Kolossváry; K. F..Wong; Paesani, F.; Vanicek, J.; X..Wu; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A.; *AMBER*, 10th ed.; University of California: San Francisco, CA, 2008.
(24) Hornak, V.; Abel, R.; Okur, A.; B, S.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* **2006**, *65*, 712−725.
(25) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Pairwise Solute Descreening of Solute Charges from a Dielectric Medium. *Chem. Phys. Lett.* **1995**, *246*, 122−129.
(26) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781−1802.
(27) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber ff99sb Protein Force Field. *Proteins* **2010**, *78*, 1950−1958.
(28) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577−2637.
(29) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. Msmbuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **2011**, *7*, 3412−3419.
(30) Singhal, N.; Pande, V. S. Error Analysis and Efficient Sampling in Markovian State Models for Molecular Dynamics. *J. Chem. Phys.* **2005**, *123*, 204909.
(31) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress and Challenges in the Automated Construction of Markov State Models for Full Protein Systems. *J. Chem. Phys.* **2009**, *131*, 124101.
(32) Deuflhard, P.; Weber, M. Robust Perron Cluster Analysis in Conformation Dynamics. *Linear Algebra Appl.* **2005**, *398*, 161−184.
(33) Romo, T. D.; Grossfield, A. Loos: An Extensible Platform for the Structural Analysis of Simulations. *31st Annual International Conference of the IEEE EMBS* **2009**, 2332−2335.
(34) Koukos, P. I.; Glykos, N. M. Grcarma: A Fully Automated Task-Oriented Interface for the Analysis of Molecular Dynamics Trajectories. *J. Comput. Chem.* **2013**, *34*, 2310−2312.
(35) Glykos, N. M. Carma: A Molecular Dynamics Analysis Program. *J. Comput. Chem.* **2006**, *27*, 1765−1768.
(36) Gohlke, H.; Case, D. A. Converging Free Energy Estimates: MM-PB(GB)SA Studies on the Protein−Protein Complex Ras−Raf. *J. Comput. Chem.* **2003**, *25*, 238−250.
(37) Sharp, K.; Honig, B. Electrostatic Interactions in Macromolecules: Theory and Applications. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *19*, 301−332.
(38) *Modeling Unusual Nucleic Acid Structures;* Macke, T., Case, D. A., Eds.; American Chemical Society: Washington, DC, 1998.
(39) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. Gromacs 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* **2013**, *29*, 845−854.
(40) Hess, B. Convergence of Sampling in Protein Simulations. *Phys. Rev. E* **2002**, *65*, 031910.
(41) Hess, B. Determining the Shear Viscosity of Model Liquids from Molecular Dynamics Simulations. *J. Chem. Phys.* **2002**, *116*, 209−217.
(42) Fenley, A. T.; Muddana, H. S.; Gilson, M. K. Entropy-Enthalpy Transduction Caused by Conformational Shifts Can Obscure the Forces Driving Protein-Ligand Binding. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 20006−20011.
(43) Genheden, S.; Akke, M.; Ryde, U. Conformational Entropies and Order Parameters: Convergence, Reproducibility, and Transferability. *J. Chem. Theory Comput.* **2014**, *10*, 432−438.
(44) Raval, A.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Refinement of Protein Structure Homology Models Via Long, All-Atom Molecular Dynamics Simulations. *Proteins* **2012**, *80*, 2070−2079.