Evaluation of Configurational Entropy Methods from Peptide Folding−Unfolding Simulation

Da-Wei Li,[†] Mina Khanlarzadeh,[‡] Jinbu Wang,[†,⊥] Shuanghong Huo,*[,†] and Rafael Brüschweiler*[,§]

Carlson School of Chemistry and Biochemistry, Clark University, Worcester, Massachusetts 01610, Department of Physics, Clark University, Worcester, Massachusetts 01610, Department of Chemistry and Biochemistry & National High Magnetic Field Laboratory, Florida State University, Tallahassee, Florida 32306

Received: July 4, 2007; In Final Form: August 30, 2007

A 4-μs molecular dynamics simulation of the second β-hairpin of the B1 domain of streptococcal protein G is used to characterize the free energy surface and to evaluate different configurational entropy estimators. From the equilibrium folding−unfolding trajectory, 200 000 conformers are clustered according to their root-mean-square deviation (RMSD). The height of the free energy barrier between pairs of clusters is found to be significantly correlated with their pairwise RMSD. Relative free energies and relative configurational entropies of the clusters are determined by explicit evaluation of the partition functions of the different clusters. These entropies are used to evaluate different entropy estimators for the largest 20 clusters as well as a subensemble comprising exclusively extended conformers. It is found that the quasi-harmonic entropy estimator operating in dihedral angle space performs better than the one using Cartesian coordinates. A recent generalization of the quasi-harmonic approach that computes Shannon entropies of probability distributions obtained by projecting the conformers along the eigenvectors of the covariance matrix performs similarly well. For the best entropy estimators, a linear correlation coefficient between 0.92 and 0.97 is found. Unexpectedly, when correlations between dihedral angles are neglected, the agreement with the reference entropies improved.

1. Introduction

The populations of the states of a macromolecular system, such as open and closed or folded and unfolded, are determined by their respective free energies, which have contributions from both energy and configurational entropy. While energies can be readily computed from macromolecular simulations, determination of the free energy is generally much harder.[1] It is therefore not uncommon to employ effective energies, rather than free energies, to predict and assess the stability of molecular states such as native protein structures.[2] Assuming equivalence between energy and free energy is tantamount to assuming that the entropies of different protein states are nearly equal and that their contributions to the free energy difference cancels. However, it has been shown that force fields that are optimized to maximize the effective energy difference between native and non-native states may produce incorrect results when applied to protein folding simulations,[3] which suggests that the entropy differences of different states are not negligible. Furthermore, the influence of entropic barriers on protein folding has been illustrated by a simple model,[4] and the side-chain entropy has been demonstrated to be an important determinant of folding kinetics.[5]

Monte Carlo and molecular dynamics (MD) simulations are potentially powerful tools for estimating entropies as they provide information about the relevant configuration space sampled by the molecular system. Calculation of the ensemble average of the energy $\langle E \rangle$ is fairly straightforward using these simulation techniques. However, the entropy S, and thereby also the Helmholtz free energy $F = \langle \mathrm{E} \rangle - TS$, is harder to deduce directly from simulations.[6] Reversible thermodynamic integration is widely used to obtain free energy differences between two predefined molecular states. However, when the conformational change is large, involving, for example, an unfolding process, it can be prohibitively expensive.[6] Therefore, it is desirable to develop robust and efficient methods to estimate the entropy difference between different states from simulation data.[7−21] This work focuses on the estimation of the configurational or conformational entropy (which in the following is often simply referred to as entropy) of finite molecular ensembles.

The quasi-harmonic method by Karplus and Kushick provides an entropy estimate from MD trajectories.[8] In this approach, the distribution of the various degrees of freedom is assumed to have a multivariate Gaussian form. Cartesian coordinates of all atoms or internal coordinates, such as the bond lengths, bond angles, and dihedral angles can be chosen. Because of the singularity of the covariance matrix in Cartesian coordinates, internal coordinates are often employed instead of Cartesian coordinates. Correlated motions in dihedral angle space have

* Corresponding author. (S.H.) Tel.: 508-793-7533. Fax: 508-793-8861. E-mail: shuo@clarku.edu. (R.B.) Tel.: 850-644-1768. Fax: 850-644-8281. E-mail: bruschweiler@magnet.fsu.edu.

[†] Carlson School of Chemistry and Biochemistry, Clark University.

[‡] Department of Physics, Clark University.

[§] Florida State University.

[⊥] Current address: Structural Biophysics Laboratory, Center for Cancer Research, National Cancer Institute−Frederick, National Institutes of Health, Frederick, MD 27102.

10.1021/jp075220e CCC: $37.00
Published on Web 11/20/2007

been reported by several groups.[22−24] To solve the singularity problem, Schlitter[11] introduced an approximation to the entropy based on a quantum-mechanical treatment of a simple harmonic oscillator, which was reexamined by van Gunsteren and coworkers.[12,25] The quasi-harmonic method in Cartesian space was subsequently reformulated by Andricioaei and Karplus.[13] Although quasi-harmonic analysis has been extensively used over the years, its accuracy has only recently been evaluated for small molecules[26] and idealized model systems[20] but not for entire peptides or proteins.

The quasi-harmonic approximation does not rigorously hold in cases where the distribution of degrees of freedom is significantly non-Gaussian. To address this problem, various methods have been proposed. In the method by Edholm and Berendsen, the configurational entropy is separately determined for each internal coordinate from the probability distribution of the ensemble using histograms with variable bin width.[9,10] More rigorous and computationally quite expensive alternatives are Meirovitch's hypothetical scanning and local states methods[15−17,19] that are capable of including correlation effects beyond second order and the method by Demchuk and coworkers[14,18,21] that was applied to systems with only one or two torsional angles. Recently, Wang and Brüschweiler[20] proposed an entropy estimator, S_{2D}, in which principle component analysis (PCA) was applied to the covariance matrix of dihedral angles represented in the complex plane. Each mode was then subjected to an entropy analysis using a smoothed form of the original distribution. While it includes second-order correlation effects, this model is relatively simple and has high computational efficiency. For a recent review on computational methods of entropy and free energy estimation, see ref 1.

In the present work, several entropy estimators are tested on the basis of a 4-μs MD simulation of the 16-amino-acid second β-hairpin of the B1 domain of streptococcal protein G (GB1). The use of an implicit solvation model allows the efficient simulation of an equilibrium folding−unfolding trajectory, during which numerous peptide folding and unfolding events are observed, as has been demonstrated previously.[27] The thorough sampling of the important portions of the free energy surface warrants a direct comparison between the free energies estimated according to the relative populations of the substates and those calculated by the different entropy estimators.

2. Method

2.1. MD Simulation. A 4-μs MD simulation was performed on the β-hairpin peptide with sequence GLY-GLU-TRP-THR-TYR-ASP-ASP-ALA-THR-LYS-THR-PHE-THR-VAL-THR-GLU, with no blocking groups at terminal residues. The simulations were carried out using the program CHARMM[28] and the CHARMM19 force field using a protocol that is essentially identical to the one described by Krivov and Karplus.[27] Briefly, to enhance sampling and speed up the simulation, Langevin dynamics with a friction constant of 0.1 ps^{-1} was employed, and the temperature of the system was set to 360 K. The implicit solvent model EEF1[29] as implemented in CHARMM is employed to describe the aqueous environment. EEF1 has previously been successfully applied to simulate protein unfolding[30] and β-hairpin folding.[31,32] Nonbonded interactions were truncated at a 9-Å cutoff, as required by the EEF1 implicit solvent model. A fully extended backbone structure was chosen as the initial conformation. A time step of 2 fs was used, and all bond lengths involving hydrogen atoms were constrained. Snapshots were saved every 20 ps, yielding a total of 200 000 conformers for subsequent analysis.

The snapshots were clustered according to several criteria detailed below and the relative free energy, energy, and entropy were determined for each cluster, allowing a direct comparison between the different approaches. In principle, the MD simulation should give the correct entropy difference between clusters, provided that the MD simulation is well equilibrated. On the other hand, if, during the simulation, the system is kinetically trapped in a state, the artificially long residence time in this state will give a lower free energy than the true value. In the present simulation, our trajectory represents a reasonably good equilibrium state because the peptide undergoes about 10 transitions between folded and unfolded states, consistent with the observations reported in ref 27.

2.2. Clustering. All conformations were clustered according to their geometric similarity in terms of the all-atom root-mean-square deviation (RMSD), with a RMSD threshold of 2 Å. The same clustering criterion as that in ref 27 was used, but with a different clustering algorithm. Clustering was performed by generating for each conformer a neighbor list based on the RMSD. Then the conformer with the largest number of neighbors was identified and assigned along with all its neighbors to the first cluster. All the conformers of this cluster were then eliminated from the ensemble of conformations, and the neighbor list of all remaining conformers was updated. This process was repeated until the ensemble was empty. In this way, a series of nonoverlapping clusters of conformations is obtained.[33,34] For each cluster, the conformer with the most neighbors is used as the representative of the cluster and is referred to as the cluster center.

Because any single conformer does not have a meaningful free energy without making additional assumptions, a free energy (F) must be defined on a finite region of the configuration space according to

$$F_i = -k_B T \ln Z_i = -k_B T \ln \int_{\text{cluster}(i)} \exp(-E(\Omega)/k_B T)\mathrm{d}\Omega \quad (1)$$

where Z_i is the partition function of cluster i, and k_B is the Boltzmann constant. The number of conformers, N_i, of a given cluster is related to the partition function Z_i and the Helmholtz free energy F_i of the cluster through

$$F_i = -k_B T \ln Z_i = -k_B T \ln(N_i) + c \quad (2)$$

where c denotes a constant that is related to the total number of MD snapshots analyzed and that is inconsequential when considering free energy differences. This method for obtaining free energies and entropies is sometimes referred to as the "counting method".[1] The effective energy of all conformers that belong to a given cluster was calculated, and the average value was taken to be the effective energy of the cluster (U_i) (note that, because an implicit solvent model is employed, the effective energy here is the potential energy of the protein plus the solvation free energy). From the free energies F_i and the effective energies U_i, the configurational entropies S_i can be calculated for each cluster according to

$$TS_i = U_i - F_i \quad (3)$$

where $T = 360$ K is the temperature of the MD simulation. In what follows, these entropies represent reference entropies and are referred to as S_{MD}. Following Krivov and Karplus,[27] the connected graph and the minimal cut tree were constructed, which preserve the information about the free energy minima and the free energy barriers between each pair of minima. The

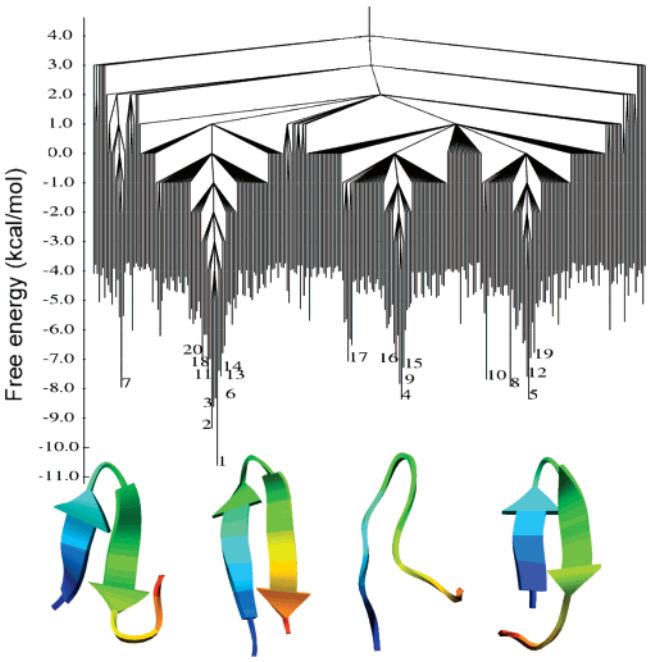

Figure 1. Minimal cut tree that includes the 300 lowest free energy minima. The four representative conformations are the cluster centers of clusters 1, 4, 5, and 7. Clusters 1−20 are labeled. The same clustering criteria as in ref 27 were used, but with a different clustering algorithm. The overall shape of the minimal cut tree is similar to that of Figure 2 of ref 27.

resulting minimal cut tree for the 300 lowest minima is depicted in Figure 1. The free energy barrier between two clusters i and j is

$$F_{ij} = -k_B T \ln Z_{ij} \tag{4}$$

where Z_{ij} is the partition function of the barrier and is related to the minimum cut value (n_{ij}) by[27,34]

$$Z_{ij} = \frac{1}{2} n_{ij} \times \frac{h}{k_B T} \times \frac{1}{\Delta t} \tag{5}$$

where h is Planck's constant, $T = 360$ K, and $\Delta t = 20$ ps is the sampling interval.

2.3. Quasi-harmonic Approach. In the quasi-harmonic (QH) approach, the coordinate distribution is fitted to a multivariate Gaussian distribution, and the entropy is then calculated as

$$S_{QH} = k_B \sum_{j=1}^{3N-6} \frac{\hbar\omega_j/k_B T}{\exp(\hbar\omega_j/k_B T) - 1} - \ln(1 - \exp(-\hbar\omega_j/k_B T)) \tag{6}$$

$$\omega_j = \sqrt{k_B T/\lambda_j} \tag{7}$$

where λ_j is the jth eigenvalue of the mass-weighted covariance matrix, and ω_j is the "frequency" of the eigenmode mode j. To compute S_{QH} of eq 6, we employ the vibration module of CHARMM. Computation of the covariance matrix from the Cartesian coordinates is straightforward. Alternatively, one can compute the covariance matrix in internal coordinates, such as mobile dihedral angles, which are major contributors to the configurational entropy.

Note that when internal coordinates are used, the Jacobian is typically assumed to be a constant,[7,9] which makes these computations feasible. Entropy estimators that do not rely on this assumption are available for small systems.[14,18,21]

Equation 6 includes vibrational zero-point quantum effects, which are important for absolute entropies. When considering entropy differences only, a classical treatment is normally justified provided that vibrations with high frequencies involving protons do not change. The entropy difference between free energy minima a and b is then given by[8]

$$\Delta S = S(b) - S(a) = k_B \ln[\prod_{j=1}^{3N-6} \omega_j(a) / \prod_{j=1}^{3N-6} \omega_j(b)] = \frac{1}{2} k_B \ln[\prod_{j=1}^{3N-6} \lambda_j(b) / \prod_{j=1}^{3N-6} \lambda_j(a)] \tag{8}$$

Quasi-harmonic entropy estimates from dihedral angles are referred to as $S_{QH,dihe}$. Because of the 2π periodicity of dihedral angles, the computation of the covariance matrix in dihedral angles can be ambiguous. To address this problem, we choose the 2π angle for each dihedral angle as $[x - \pi, x + \pi]$ so that its midpoint x coincides with the center of the dihedral angle distribution. The 2π periodicity problem can also be alleviated by using a complex notation as described below.

2.4. Extended Principal Component Analysis Approach. The quasi-harmonic approach can be extended as follows:[20] after eigenvalues and eigenvectors are determined in a quasi-harmonic analysis, all conformers are projected along a given real or complex eigenvector (mode) q. This leads to a probability distribution P_q along this mode from which the entropy is calculated according to

$$S = -k_B \int P_q \ln(P_q) dq \tag{9}$$

The total entropy is determined by summing over the entropy contributions of all modes. It should be pointed out that quantum effects are not included in this estimator, so that only relative entropies can be determined. The approach can accommodate probability distributions other than Gaussian distributions, including multi-modal distributions. For an ensemble with a finite number of conformers, the distribution along each mode takes a discrete form, which may be unsuitable for integration, and a smoothing procedure prior to integration is applied. Following Wang and Brüschweiler,[20] each discrete coordinate q along each mode is converted into a (pseudo-)continuous distribution by convolution with a Gaussian function:

$$P_q dq = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{q^2}{2\sigma^2}\right) dq \tag{10}$$

where σ is a smoothing parameter whose effect is discussed below. We refer to this entropy estimator as S_{1D}.

A more rigorous way to circumvent the 2π periodicity problem of dihedral angles for the computation of the covariance matrix was introduced recently.[20] In this algorithm, dihedral angles φ are represented as points in a complex plane in terms of $\exp(i\varphi)$. The complex covariance matrix still has real eigenvalues, and the entropy, called S_{2D}, is computed along each complex mode according to eq 9 after convolution of the distribution with an axially symmetric bivariate (2D) Gaussian function (same σ along the real and imaginary axes).

3. Results and Discussion

3.1. Clustering and MD Entropies S_{MD}. The 16-amino-acid β-hairpin contains a total of 56 mobile dihedral angles. Within a given cluster, the 30 backbone dihedral angles are dynamically

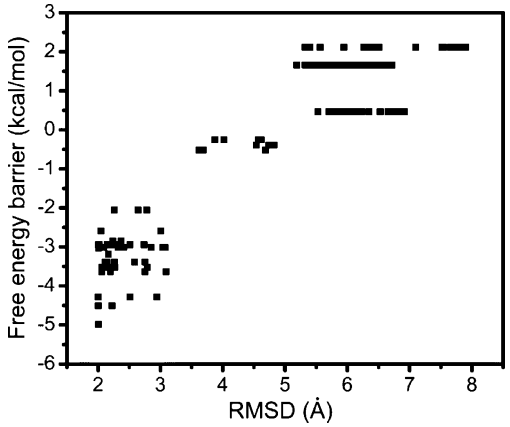

Figure 2. Correlation between the pairwise RMSD and the free energy barrier between clusters 1 and 20.

distributed in a relatively narrow range, while the 26 side-chain dihedral angles generally exhibit a wider distribution. Since most clusters have only a very limited number of conformers (typically well below 1000) that are too small to reliably apply the various entropy estimators, the largest 20 clusters were selected for further analysis and sorted according to cluster size. Cluster 1 contains 40 864 conformers, whereas cluster 20 has 1030 conformers, and the average size of the top 20 clusters is 4791. The minimal cut tree and four representative conformers, which are the cluster centers of clusters 1, 4, 5, and 7, respectively, are shown in Figure 1. Cluster 1 corresponds to the global free energy minimum at 360 K. This cluster, together with its neighboring clusters 2, 3, 6, 11, 13, 14, 18, and 20, which are separated by free energy barriers less than −3.0 kcal/mol, forms a large basin involving 34% (67970) of all conformers, with the center of cluster 1 as its representative. The conformers in the other top 20 clusters are qualitatively similar to the four representative conformers depicted in Figure 1. The centers of cluster 5 and 7 have their turns formed in the wrong positions, while the center of cluster 4 lacks a substantial number of hydrogen bonds. An "extended-state cluster", cluster E, is defined by the conformers whose radius of gyration is larger than 9 Å, irrespective of their mutual RMSD (this cluster does not have a meaningful representative conformer). In Figure 2, the comparison between the pairwise RMSD and the free energy barrier for the first 20 clusters indicates an overall correlation: a large RMSD between two clusters is associated with a high free energy barrier and vice versa.

Clearly, the precise nature of the clusters depends on a number of parameters and choices, such as the nature of the similarity measure (RMSD in our case), the cluster size (cutoff), and the clustering algorithm. RMSD is a common metric for conformer clustering and has also been used to define the native ensemble and an unfolded ensemble to calculate the entropy difference in other peptide systems.[25]

In addition, the free energy barrier between the conformers within a single cluster should be significantly lower than those between the conformers in different clusters. In practice, this property is not always fulfilled because it is hard to find a conformational similarity measure that is strongly correlated with free energy barriers. Figure 2 shows that, for the RMSD measure used here, pairs of conformers with large RMSDs are generally separated by larger free energy barriers, while for RMSDs less than 3 Å, the correlation disappears. To evaluate the effect of the clustering cutoff, clustering was done by using an RMSD threshold of 2.5 Å (instead of 2.0 Å). This increases

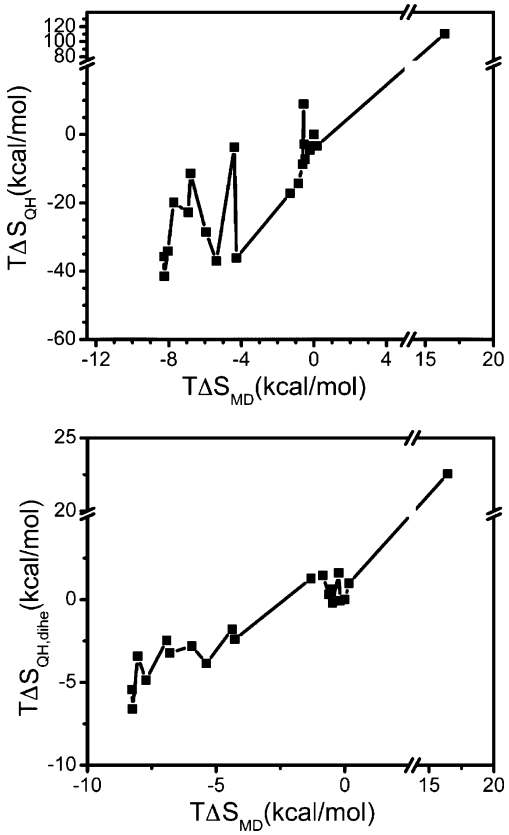

Figure 3. Comparison between $T\Delta S$ (top) from quasi-harmonic analysis using Cartesian coordinates and (bottom) from quasi-harmonic analysis using dihedral angles and $T\Delta S$ from MD. ΔS represents the entropy difference between the entropy of a given cluster and the entropy of the reference cluster (cluster 1). Please note the broken axes. The line connecting the points is intended to guide the eyes (same in Figures 4−6).

the cluster sizes (see Supporting Information), but otherwise does not significantly affect the performance of the entropy estimators.

The definition of the extended state (cluster E) is not straightforward. Naively, one might assign all conformers to this cluster, whose RMSD to the native state exceeds a certain cutoff. However, such a simple definition is not applicable in our case because there are several frequently visited non-native states, in particular, clusters 4, 5, and 7, which possess large RMSDs with respect to the native state. The extended-state cluster is relatively insensitive to the R_g threshold (e.g., for R_g > 9.5 Å instead of >9 Å), and the relative performance of the entropy estimators remains the same.

The entropy difference ΔS_{MD} of each cluster with respect to reference cluster 1 (multiplied by T) is listed in the Supporting Information (Table S1, second column) serving as a reference to evaluate the performance of the different entropy estimators tested here. Since the counting method does not provide absolute entropies (see above), cluster 1 is (arbitrarily) used as the reference state. Similarly, all configurational entropies given for the entropy estimators in the following sections are entropy differences ΔS with respect to the entropy of cluster 1.

3.2. Quasi-harmonic Approach Using Cartesian Coordinates (S_{QH}). The quasi-harmonic method implemented in the vibration module of CHARMM (version c31b1) is applied first. In this approach, ΔS_{QH} is calculated in Cartesian coordinates using eq 6. The comparison between ΔS_{QH} and ΔS_{MD} is shown in Figure 3 (top). When the extended cluster E is excluded, the linear correlation coefficient is 0.78, and the slope of a fitted

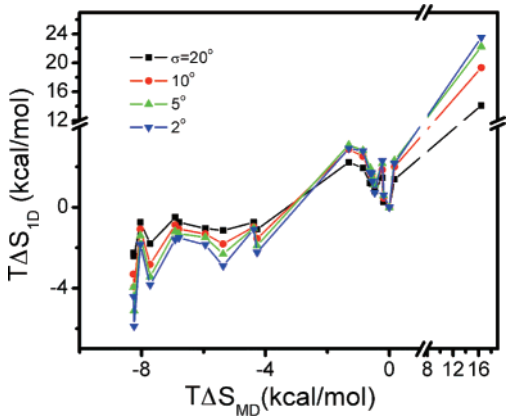

Figure 4. Comparison between $T\Delta S_{1D}$ and $T\Delta S_{MD}$ for 21 MD-derived clusters of the 16-amino-acid β-hairpin peptide.

straight line is 3.5, indicating that this entropy estimator substantially overestimates entropy differences between the different clusters. After the inclusion of cluster E, the correlation coefficient increases to 0.93, but the slope of the linear least-square line increases to 5.3. When calculating the quasi-harmonic entropy without quantum effects (eq 8), the results are similar. All entropies described here and in the following sections are listed in Tables S1−S3 of the Supporting Information.

3.3. Quasi-harmonic Approach Using Internal Coordinates ($S_{QH,\ dihe}$). The covariance matrix is computed using all mobile dihedral angles, both along the main chain and side chains, from which $S_{QH,dihe}$ is determined according to eq 8, whereby the range of each angle is dynamically determined (see Method section). Figure 3 (bottom) shows $\Delta S_{QH,dihe}$ versus ΔS_{MD}. Without the extended cluster E, the linear correlation coefficient is 0.93, and the slope of a fitted straight line is 0.70, which indicates that the entropy estimator generally underestimates the entropy difference of different conformations. When cluster E is included, the linear correlation coefficient is 0.96 with an average slope of 1.02. For cluster E, $\Delta S_{QH,dihe}$ somewhat overestimates entropy ΔS_{MD}. Contributions arising from bond angles and peptide bond torsional angles, treated in the same way as mobile dihedral angles, are found to be negligible as compared to contributions from dihedral angles.

The improvement of the quasi-harmonic method in dihedral angle space over Cartesian space is consistent with a recent analysis of the quasi-harmonic analysis of linear alkanes and a host−guest system.[26] Rationalization of the improvement is not straightforward, as both approaches involve various approximations.[7,9] For example, because of the nonlinear nature of coordinate transformation between dihedral angles and Cartesian coordinates, a Gaussian distribution in one coordinate system leads to a non-Gaussian distribution in the other. Another complication arises from the fact that quasi-harmonic modes can be correlated. If the number of included modes exceeds the actual number of degrees of freedom, some of the modes are not independently populated any longer,[35] which may cause an overestimation of the entropy.

3.4. Extended PCA Approach (S_{1D} and S_{2D}). Here we extend the quasi-harmonic approach by projecting all conformers along each given mode and integrate the corresponding distribution over all the modes to obtain the entropy, S_{1D}, using eq 9. The result depends on the width (σ) of the Gaussian with which the original distributions are convoluted. The comparison between S_{1D} and S_{MD} is shown in Figure 4, and the correlation coefficients and average slopes between S_{1D} and S_{MD} are listed in Table 1.

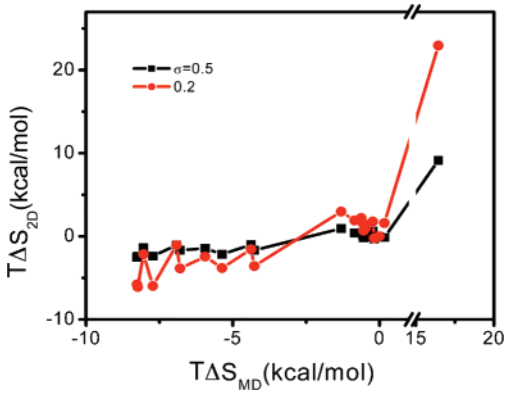

Figure 5. Comparison between $T\Delta S_{2D}$ and $T\Delta S_{MD}$ for 21 MD-derived clusters of β-hairpin peptide.

TABLE 1: Correlation Coefficient and Slope of the Fitted Line for ΔS_{1D} vs ΔS_{MD}

	clusters 1−20		clusters 1−20 and E	
σ (°)	correlation	slope	correlation	slope
20	0.865	0.365	0.943	0.586
10	0.877	0.527	0.947	0.816
5	0.882	0.628	0.949	0.950
2	0.888	0.679	0.952	1.018
1	0.892	0.687	0.952	1.035
0.5	0.887	0.683	0.950	1.050

Parameter σ controls the extent of smoothing introduced by the convolution process and thereby the actual resolution in the projected phase space. For large σ (i.e., low resolution), the fine structure in the distributions with resolution higher than σ is ignored, and therefore the entropy estimates tend to be too large. On the other hand, if σ is chosen too small, undersampling effects start to become an issue. In the case of extreme undersampling and in the absence of degeneracies, that is, when the probability distribution consists of N nonoverlapping Gaussians, the estimated entropy for each mode approaches $k_B \ln(N)$ (where N is the total number of conformers),[20] which bears little relevance to the entropy.

In Table 1, one finds that, for large σ, the slope is small and vice versa. Below $\sigma = 5°$, both the slope and the correlation coefficient change very little, and below 1°, undersampling effects start to become noticeable. Therefore, an acceptable range for σ is between 2° and 5°, for which the linear correlation coefficient is around 0.88 and the slope is between 0.63 and 0.68 for clusters 1−20, and the correlation coefficient is 0.95 and the slope is 0.95−1.02 for all clusters. These results are comparable to the ones obtained for $S_{QH,dihe}$ (see previous section).

In the S_{2D} method, all dihedral angles are expressed in the complex plane and projected along all modes to calculate entropy (S_{2D}) from eq 9. Similar to S_{1D}, S_{2D} also depends on a "smoothing parameter" σ. Figure 5 depicts ΔS_{2D} versus ΔS_{MD}, with the numerical values for correlation coefficients and slopes given in Table 2. While the correlation coefficients are largely independent of σ, the slope varies noticeably, with a value between 0.1 and 0.2 providing the best match. The best results are comparable to the ones of S_{1D} and $S_{QH,dihe}$. Note that there is no one-to-one relationship between these σ values of S_{2D} and those used in S_{1D}.

3.5. Effect of Dihedral Angle Correlations. Because of the dense character of polypeptide folds, it is commonly assumed that different degrees of freedom, such as main-chain and side-chain dihedral angles, are dynamically correlated. To evaluate the importance of these correlations, we set all off-diagonal

TABLE 2: Correlation Coefficient and Slope of the Fitted Line for ΔS_{2D} vs ΔS_{MD}

	clusters 1−20		clusters 1−20 and E	
σ^a	correlation	slope	correlation	slope
0.5	0.892	0.297	0.955	0.426
0.2	0.876	0.772	0.954	1.052
0.1	0.851	1.123	0.949	1.474
0.05	0.848	1.353	0.948	1.722
0.02	0.822	1.435	0.940	1.894

[a] σ in the S_{2D} method is dimensionless.

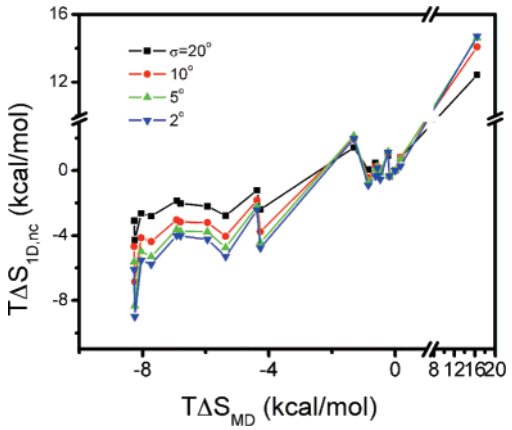

Figure 6. Comparison between $T\Delta S_{1D,nc}$ and $T\Delta S_{MD}$ for 21 MD-derived clusters of β-hairpin peptide.

TABLE 3: Correlation Coefficient and Slope of the Fitted Line for $\Delta S_{1D,nc}$ vs ΔS_{MD}

	clusters 1−20		clusters 1−20 and E	
σ (°)	correlation	slope	correlation	slope
20	0.920	0.448	0.966	0.595
10	0.924	0.670	0.973	0.750
5	0.924	0.788	0.973	0.822
2	0.924	0.830	0.972	0.851
1	0.922	0.833	0.971	0.860
0.5	0.908	0.831	0.966	0.871

elements of the covariance matrix to zero prior to estimating the entropy using the S_{1D} method. The entropy estimated in this way is called $S_{1D,nc}$ (nc for no correlations). Figure 6 shows $\Delta S_{1D,nc}$ versus ΔS_{MD}. The complete data are listed in Table S4 of the Supporting Information. The linear correlation coefficient and the slope of the fitted line as a function of σ are listed in Table 3, which show a trend similar to that of S_{1D}. Interestingly, $S_{1D,nc}$ performs even better than S_{1D}. S_{2D} was also computed with the off-diagonal elements set to zero, which gives a result similar to that of $S_{1D,nc}$, with the maximal correlation coefficient reaching 0.97 (see Table S5). Neglect of correlation effects in $S_{QH,dihe}$ has only a minor effect: for the 20 clusters (without the extended cluster E), the correlation decreases from 0.93 to 0.90, whereas, for the 21 clusters, the results are essentially unchanged (correlation coefficient 0.97 and average slope 0.99). The results for S_{1D}, S_{2D}, and S_{QH} indicate that the entropy difference between clusters is well captured when considering the distribution of each dihedral angle separately. Meanwhile, statistical noise contained in covariances, caused by finite sampling, are found to be non-negligible and can have an adverse effect on the entropy prediction. This is only one of different possible reasons why the inclusion of correlation effects between dihedral angles (covariances) does not improve (or even worsens) the performance of these entropy predictors in the present case.

3.6. Contribution from Main-Chain and Side-Chain Dihedral Angles. Separate analysis of main-chain and side-chain entropy contributions reveals that they both play an important role. For clusters 1−20, with S_{1D} (σ = 2°), the correlation coefficient drops to 0.82 and 0.72 if only main-chain or only side-chain dihedral angles are included, respectively. If all dihedral angles are included, the correlation coefficient is 0.89.

4. Discussion and Conclusion

For a fully equilibrated MD trajectory, the estimation of configurational entropies of different substates can be achieved by the relatively straightforward counting method. While current computer power and computational protocols afford such trajectories for small molecules and peptides, including the β-hairpin described here, they are not applicable to more complex systems. For complex systems, MD trajectories can probe individual free energy basins, but generally will not display the sufficiently large number of transitions between basins that is required to establish a thermal equilibrium between them.

In such cases, configurational entropy estimators can yield useful thermodynamic information, provided that they are properly calibrated with respect to known reference entropies. The purpose of this work is to evaluate and calibrate a variety of PCA-based entropy estimators for the GB1 β-hairpin for which a well-sampled MD trajectory is feasible. Although the implicit solvent approximation is likely to impact the accuracy of the simulation with respect to reality, the conformational dynamics is sufficiently realistic for the present methodological purposes.

Quasi-harmonic analysis is routinely applied to protein simulations, although the accuracy of quasi-harmonic entropies has rarely been assessed. Recently, the quasi-harmonic approach was tested for small molecules,[26] and it was found that this method, particularly when used in Cartesian coordinates, substantially overestimates the entropies. Consistent with these results, we find the that quasi-harmonic approach in dihedral angle space works clearly better than in Cartesian space, provided that the dynamic range of dihedral angles is selected properly. In this way, good correspondence is achieved between estimated entropies and reference entropies obtained by the counting method, with a linear correlation coefficient of 0.96 and a slope of 1.02.

The recent extension of the quasi-harmonic approach, in which all conformers are projected along each mode (eigenvector of the corresponding covariance matrix) either in real space (S_{1D}) or in complex space (S_{2D}) and the Shannon entropy is computed along each mode, yields results that are comparable to those of $S_{QH,dihe}$. The S_{2D} method was originally calibrated for discrete molecular ensembles for which the analytical entropy is known.[20] For rotameric transitions of 120° and 180°, an optimal value for σ between 0.5 and 0.6 was found. In the case of the GB1 peptide investigated here, the dihedral angle jumps tend to be smaller, which explains why a narrower smoothing function with σ around 0.2 yields the optimal results. For an appropriate choice of the width of the smoothing function, S_{1D} and S_{2D} achieve a correlation and average slope similar to those of $S_{QH,dihe}$. Somewhat unexpectedly, the S_{1D} and S_{2D} methods with all covariances between dihedral angles removed give an even better agreement with S_{MD}. This means that knowledge of individual dihedral angle distributions is sufficient to estimate the configurational entropy changes of this peptide in good approximation. Such a neglect of correlation

effects for entropy estimates is closely related to the simplest treatment by Edholm and Berendsen,[9,10] except that their variable histogram method is replaced here by our convolution procedure that converts potentially rugged probability distributions into smooth functions prior to integration. Interpretation of changes in order parameters derived from NMR relaxation data in proteins in terms of entropy typically assumes uncorrelated motions between different sites and has been shown to correlate well with thermodynamic measurements.[36−38] The findings reported here for a small peptide could explain why such an assumption works better in practice than one might expect on the basis of first-principle considerations. Since the present work is mainly based on β-hairpin-like and extended structures, it will be important to investigate using the computational methods developed here to what extent these conclusions hold for helical structures as well as for larger polypeptides.

Acknowledgment. We are grateful to Prof. David Wales for providing programs for generating disconnectivity graphs. This work was partially supported by the NIH (Grant 1R15 AG025023-01 to S.H.) and by the NSF (Grant 0621482 to R.B. and a Major Research Instrumentation Grant DBI-0320875 to Clark University). The molecular graphics images were generated using the Chimera[39] package from the Computer Graphics Laboratory, University of California, San Francisco, CA (supported by NIH Grant P41 RR-01081). We also acknowledge the National Center of Supercomputing Applications and the Scientific Computing and Visualization group at Boston University for providing part of the computational resources.

Supporting Information Available: Five tables with entropies of clusters using the different entropy estimators described in the text, one table with correlation coefficients and slopes for S_{2D} when correlation effects are ignored, and one table with cluster entropies when a clustering criterion is used that includes both the RMSD and the similarity of the hydrogen-bonding pattern. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Meirovitch, H. *Curr. Opin. Struct. Biol.* **2007**, *17*, 181.
(2) Lazaridis, T.; Karplus, M. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139.
(3) Liwo, A.; Khalili, M.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2362.
(4) Bicout, D. J.; Szabo, A. *Protein Sci.* **2000**, *9*, 452.
(5) Galzitskaya, O. V.; Surin, A. K.; Nakamura, H. *Protein Sci.* **2000**, *9*, 580.
(6) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed.; Academic Press: San Diego, CA, 2002.
(7) Go, N.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 535.
(8) Karplus, M.; Kushick, J. N. *Macromolecules* **1981**, *14*, 325.
(9) Di Nola, A.; Berendsen, H. J. C.; Edholm, O. *Macromolecules* **1984**, *17*, 2044.
(10) Edholm, O.; Berendsen, H. J. C. *Mol. Phys.* **1984**, *51*, 1101.
(11) Schlitter, J. *Chem. Phys. Lett.* **1993**, *215*, 617.
(12) Schafer, H.; Mark, A. E.; van Gunsteren, W. F. *J. Chem. Phys.* **2000**, *113*, 7809.
(13) Andricioaei, L.; Karplus, M. *J. Chem. Phys.* **2001**, *115*, 6289.
(14) Hnizdo, V.; Fedorowicz, A.; Singh, H.; Demchuk, E. *J. Comput. Chem.* **2003**, *24*, 1172.
(15) Cheluvaraja, S.; Meirovitch, H. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9241.
(16) Cheluvaraja, S.; Meirovitch, H. *J. Phys. Chem. B* **2005**, *109*, 21963.
(17) Cheluvaraja, S.; Meirovitch, H. *J. Chem. Phys.* **2005**, *122*, 54903.
(18) Darian, E.; Hnizdo, V.; Fedorowicz, A.; Singh, H.; Demchuk, E. *J. Comput. Chem.* **2005**, *26*, 651.
(19) Cheluvaraja, S.; Meirovitch, H. *J. Chem. Phys.* **2006**, *125*, 24905.
(20) Wang, J.; Brüschweiler, R. *J. Chem. Theory Comput.* **2006**, *2*, 18.
(21) Hnizdo, V.; Darian, E.; Fedorowicz, A.; Demchuk, E.; Li, S.; Singh, H. *J. Comput. Chem.* **2007**, *28*, 655.
(22) van Aalten, D. M. D.; de Groot, B. L.; Findlay, J. B. C.; Berendsen, H. J. C.; Amadei, A. *J. Comput. Chem.* **1997**, *18*, 169.
(23) Abseher, R.; Nilges, M. *J. Mol. Biol.* **1998**, *279*, 911.
(24) Mu, Y.; Nguyen, P. H.; Stock, G. *Proteins* **2005**, *58*, 45.
(25) Schafer, H.; Daura, X.; Mark, A. E.; van Gunsteren, W. F. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 45.
(26) Chang, C. E.; Chen, W.; Gilson, M. K. *J. Chem. Theory Comput.* **2005**, *1*, 1017.
(27) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766.
(28) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
(29) Lazaridis, T.; Karplus, M. *Proteins* **1999**, *35*, 133.
(30) Lazaridis, T.; Karplus, M. *Science* **1997**, *278*, 1928.
(31) Dinner, A. R.; Lazaridis, T.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9068.
(32) Yang, H.; Wu, H.; Li, D.; Han, L.; Huo, S. *J. Chem. Theory Comput.* **2007**, *3*, 17.
(33) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Angew. Chem., Int. Ed.* **1999**, *38*, 236.
(34) Li, D. W.; Han, L.; Huo, S. *J. Phys. Chem. B* **2007**, *111*, 5425.
(35) Prompers, J. J.; Brüschweiler, R. *J. Phys. Chem. B* **2000**, *104*, 11416.
(36) Akke, M.; Brüschweiler, R.; Palmer, A. G. *J. Am. Chem. Soc.* **1993**, *115*, 9832.
(37) Marlow, M. S.; Wand, A. J. *Biochemistry* **2006**, *45*, 8732.
(38) Yang, D.; Kay, L. E. *J. Mol. Biol.* **1996**, *263*, 369.
(39) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *J. Comput. Chem.* **2004**, *25*, 1605.