JCTC Journal of Chemical Theory and Computation

Article

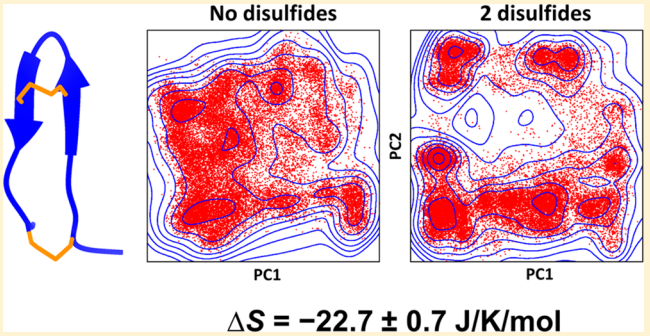
pubs.acs.org/JCTC

Calculation of Configurational Entropy Differences from Conformational Ensembles Using Gaussian Mixtures

Gergely Gyimesi,[‡] Péter Závodszky, and András Szilágyi*

Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Magyar tudósok krt. 2, H-1117 Budapest, Hungary

ABSTRACT: We present a novel, conceptually simple approach to calculate the configurational entropy difference between two conformational ensembles of a molecular system. The method estimates the full-dimensional probability density function of the system by a Gaussian mixture, using an efficient greedy learning algorithm with a cross-validation-based stopping criterion. An evaluation of the method on conformational ensembles corresponding to substates of five small peptide systems shows that excellent agreement is found with the exact entropy differences obtained from a full enumeration of conformations. Compared with the quasiharmonic method and two other, more recently developed methods, the Gaussian mixture method yields more accurate results at smaller sample sizes. We illustrate the power of the method by calculating the backbone torsion angle entropy difference between disulfide-bonded and nondisulfide-bonded states of tachyplesin, a 17-residue antimicrobial peptide, and between two substates in the native ensemble of the 58-residue bovine pancreatic trypsin inhibitor.



INTRODUCTION

Entropy and free energy are fundamental thermodynamic quantities that play an essential role in determining the macroscopic behavior of peptides, proteins, and other biomolecular systems. Spontaneous processes at constant temperature (and pressure) tend to reduce the free energy of the system. The behavior of entropy during the process provides insight into the mechanism and the main forces driving the changes at the molecular level. A main driving force of protein folding, for example, is the hydrophobic interaction, partly driven by an increase in the entropy of water.[1] To calculate entropies from theoretical models and molecular simulations is an important challenge.

However, theoretical efforts in the past decades have mostly been devoted to calculating free energies rather than entropies.[2] The most popular methods to calculate free energy changes are thermodynamic integration and the related free energy perturbation and histogram analysis methods. These methods work best when the free energy difference between two similar states of a biomolecular system is to be calculated. A large structural difference between the two states makes the calculation increasingly difficult, and sometimes requires simulations that are unfeasible. A promising solution to this problem is to calculate the absolute free energies of the two states and simply take their difference, thus entirely avoiding the need to find or set up an integration or transition path between the two states. For a solvated system, the internal energy may be calculated using a molecular mechanics force field, and the solvation free energy may be derived from implicit solvent models, such as in the popular MM/PBSA approach.[3] To obtain the total free energy, however, the conformational entropy of the solute must also be estimated, requiring integration over the configurational space. This is typically neglected in the MM/PBSA method. Consequently, there is a need for methods that can efficiently calculate configurational entropies from conformational ensembles obtained from experiments or generated by molecular dynamics (MD) or Monte Carlo (MC) simulations.

Calculating entropies is a challenging problem because, in principle, the whole configurational space has to be evaluated. A wide range of methods for calculating entropies from conformational ensembles have been developed over the past decades; we direct the reader to in-depth reviews of the subject.[2,4] The available methods differ from each other in several aspects: whether they calculate quantum mechanical or classical entropies; the type of coordinates they use; how they deal with the problem of dimensionality; and how they estimate probability densities.

Many methods aim to calculate classical entropy, typically on internal coordinates or only torsion angles.[5−19] These entropies are relative because in classical mechanics, entropy is undetermined up to an arbitrary additive constant depending on the chosen phase space cell volume; this is not a problem as the laws of thermodynamics are only concerned with changes in entropy. Another group of methods calculates quantum mechanical entropy, typically from Cartesian coordinates.[20−25] The quantum mechanical entropy is considered absolute

Received: August 23, 2016
Published: November 29, 2016

DOI: 10.1021/acs.jctc.6b00837
J. Chem. Theory Comput. 2017, 13, 29−41

because the phase space cell volume is tied with Planck's constant to account for the uncertainty principle. Several methods combine the two approaches by treating the fast degrees of freedom quantum mechanically and the slow degrees of freedom classically.[26−28]

With systems larger than a small molecule with a few degrees of freedom, the high number of dimensions of the configurational space introduces a computational challenge. Only a few methods aim to calculate a full-dimensional entropy, and they are still limited to a few dozen dimensions.[14,27] Some methods simply neglect correlations between coordinates and calculate the entropy as a sum of one-dimensional entropies.[7,8,16,29,30] More commonly, principal component analysis is applied to obtain linearly independent degrees of freedom,[5,6,9,12] or vibrational modes are determined and treated as independent.[20−25] A number of methods employ mutual information expansion (MIE) to approximate the full-dimensional entropy using marginal distributions (usually up to two-dimensional marginals).[13,15,26,28,31,32] Other solutions include clustering the coordinates into minimally coupled subspaces (MCSA) by full correlation analysis (FCA),[27,28] and methods based on series expansions such as maximum information spanning tree (MIST)[33] and multibody local approximation (MLA).[18,19]

Entropy calculation involves estimating the probability distribution of the system in phase space, or (in the case of configurational entropy) in configurational space. Several methods discretize the data into bins and calculate the entropy from the resulting histogram,[7,8,13,17−19,31,33] although this is only practical for a small number of dimensions. To estimate the probability density beyond histograms, various parametric and nonparametric methods have been used. The classical quasiharmonic method[5] is parametric; it approximates the probability distribution as a single multidimensional Gaussian; this can be improved by a cubic correction.[9] Equivalently, the quantum quasiharmonic method approximates the system as a set of harmonic oscillators.[20−25] While the quasiharmonic method is easy and fast to calculate, it has obvious limitations as it assumes that the system has a single harmonic energy well; even simple molecules have a significantly more complex energy landscape. The inaccuracy of the quasiharmonic method has been demonstrated in multiple studies.[4,34,35] Thus, a number of methods have been developed to obtain more accurate probability densities and entropies. One solution involves decomposing entropy into vibrational and conformational contributions;[36,37] this leads to an approximation of the energy landscape by a number of distinct local energy minima (corresponding to distinct conformations), each of which is approximated by the quasiharmonic approximation.[20,35] Another group of methods uses the quasiharmonic approximation with correction terms for anharmonicity and supralinear correlations.[26,38] An approach based on Fourier expansions is also parametric.[10,11] Nonparametric density estimation methods include the k-nearest-neighbor approximation,[14,15,26] and kernel density estimation based on von Mises kernels,[16,29,30] Gaussian kernels,[12] and adaptive anisotropic kernels.[27,28] A completely different approach is used by Meirovitch and coworkers,[39,40] who use a complex reconstruction algorithm to establish the probability of each configuration.

As expected, the currently existing methods still have limitations, for example, many of them use low-dimensional approximations to the full-dimensional entropy, or have a high computational burden. Here, we report a novel method to calculate the (classical) configurational entropy of a system from a conformational ensemble. As a natural extension of the quasiharmonic method, our method uses a Gaussian mixture model to estimate the full-dimensional probability density function of the molecular system. Gaussian mixture functions are weighted sums of individual multivariate Gaussians, and can approximate any smooth function to arbitrary accuracy. Gaussian mixtures can be efficiently estimated by a greedy expectation maximization method,[41] and the entropy can be easily calculated. This offers us a conceptually simple way to calculate the configurational entropy of molecular systems having an arbitrarily complex energy landscape. We have tested the accuracy of the method on five small peptides and found that it provides more accurate results at smaller sample sizes than several existing methods. The method scales well to larger molecules; we demonstrate this by using it to calculate full-dimensional entropies for a 17-residue peptide and a 58-residue protein.

THEORY

In a classical system with M degrees of freedom, the entropy is defined as[5,7]

$$S = -k_B \int P(\mathbf{p}, \mathbf{q}) \ln(P(\mathbf{p}, \mathbf{q})h^M)\, d\mathbf{p}\, d\mathbf{q}$$

where k_B is Boltzmann's constant (to be replaced by the ideal gas constant R when calculating molar entropies), $P(\mathbf{p}, \mathbf{q})$ is the probability density in phase space with $\mathbf{p}$ and $\mathbf{q}$ representing momentum and position variables, respectively, and h^M is the cell volume in phase space, with h often chosen to be Planck's constant to reproduce quantum mechanical entropies. As noted in the Introduction, classical entropy is undetermined up to an arbitrary additive constant, therefore h^M in the above formula can be omitted. In a conservative system, the probability density in Cartesian phase space factorizes into a kinetic and a configurational part, and the entropy becomes

$$S = S_p + S_c$$

where S_p and S_c are the kinetic and the configurational entropy, respectively. The latter one is

$$S_c = -k_B \int P(\mathbf{x}) \ln P(\mathbf{x})\, d\mathbf{x}$$

where $P(\mathbf{x})$ is the probability density in the space of Cartesian coordinates $\mathbf{x}$. It should be noted that while the total entropy S is unit-independent, both S_p and S_c depend on the length unit, but this does not pose any problems because changing the unit only adds an additive constant to the calculated entropies, which cancels out when calculating entropy differences.

For nondiffusive systems where overall translation and rotation can be ignored, it is often advantageous to switch to internal coordinates $\mathbf{q}$ (typically bond lengths, bond angles, and torsions), which introduces the Jacobian $J(\mathbf{q})$ so that $d\mathbf{x} = J(\mathbf{q})\, d\mathbf{q}$, and the configurational entropy becomes

$$S_c = -k_B \int P'(\mathbf{q}) \ln P'(\mathbf{q})\, d\mathbf{q} - k_B \int P'(\mathbf{q}) \ln J(\mathbf{q})\, d\mathbf{q}$$

where $P'(\mathbf{q})$ is the probability density in internal coordinate space. Internal coordinates are often divided into "hard" ones ($\mathbf{q}_{hard}$) that only vary in very narrow ranges (typically bond lengths and bond angles) and "soft" ones ($\mathbf{q}_{soft}$) that vary widely (typically torsions). It has been shown[42] that the Jacobian $J(\mathbf{q})$ only depends on the bond lengths and bond angles, and is independent of the torsions, that is, $J(\mathbf{q})$ =

DOI: 10.1021/acs.jctc.6b00837
J. Chem. Theory Comput. 2017, 13, 29−41

Table 1. Details of the Test Systems

system	degrees of freedom (list of torsions)	lattice unit φ_0 (degrees)	number of conformations	subset B[a]
ala3	4 (ψ_1, φ_2, ψ_2, φ_3)	4	65 610 000	$-120° \le \psi_1 < 0°$
ala-val-ala	5 (ψ_1, φ_2, χ, ψ_2, φ_3)	10	60 466 176	$-120° \le \chi < 120°$
ile	4 (φ, ψ, χ_1, χ_2)	4	65 610 000	$-120° \le \psi < 0°$
val	3 (φ, ψ, χ)	1	46 656 000	$-120° \le \psi < 120°$
val2	4 (ψ_1, φ_2, χ_1, χ_2)	4	65 610 000	$-120° \le \psi_1 < 0°$

[a]Subset A is the complement of subset B.

$J(\mathbf{q}_{\text{hard}})$. Assuming that the "hard" variables can be replaced by their equilibrium values (an approximation which may not always be satisfactory[5,7,42] but is routinely used), the probability density also factorizes into $P'(\mathbf{q}) = P'(\mathbf{q}_{\text{hard}})P'(\mathbf{q}_{\text{soft}})$, and after performing the integration over the hard coordinates, the configurational entropy becomes

$$S_c = -k_B \int P'(\mathbf{q}_{\text{soft}}) \ln P'(\mathbf{q}_{\text{soft}}) \, d\mathbf{q}_{\text{soft}} + \text{const}$$

where the additive constant can be omitted. Again, the entropy calculated this way is dependent on the angle unit ($\mathbf{q}_{\text{soft}}$ representing torsion angles), but only via another additive constant, therefore entropy differences will be unit-independent. Note that the above treatment assumes "hard" but still flexible bond lengths and bond angles; assuming completely rigid bonds and bond angles results in a more complex formula containing the mass-metric tensor.[42] With the approach presented above, calculating the classical configurational entropy from an ensemble of conformations involves estimating the probability density in the space of torsion angles and computing the integral shown. This basic approach has been used, in many variations, in a number of entropy calculation methods.[5−19,29,30,33]

In this paper, we use the terms "configurational entropy" and "conformational entropy" interchangeably because the conformations of a flexible molecule are its configurations in a statistical mechanics sense. We treat conformational space as a continuous space, unlike approaches for which configurational space is viewed as a set of discrete conformations with harmonic vibrations around them.[20,35,36]

METHODS

Peptide Test Systems. Five small peptides were used to test the accuracy of our entropy calculation method. The initial atomic models of peptide molecules were generated with SYBYL 7.3. The geometry was optimized by energy minimization with the G53a6 force field using the L-BFGS method in GROMACS 4.5[43] with all nonbonding interactions turned off. This was done in order to optimize bond lengths and angles for the G53a6 united atom force field, which was used for all subsequent energy evaluations.

The selected molecules were Ala3, Ala−Val−Ala, Ace−Ile−Nme, Ace−Val−Nme, Val2, and the resulting ensembles were named "ala3", "ala-val-ala", "ile", "val" and "val2", respectively. Amino and carboxyl termini, if present, were left uncharged.

Generation of the Full Set of Conformations for the Peptide Test Systems. To calculate the exact configurational entropy, a set of conformations was generated for each peptide test system that uniformly samples the phase space of internal coordinates. This was done by taking all conformations along a lattice $\varphi = (k_1\varphi_0, k_2\varphi_0, ..., k_d\varphi_0)$ in internal coordinate space, where k_i are integers and φ_0 is a fixed sampling interval. Only values of k_i were considered where $k_i\varphi_0$ lies in the $[-180°, 180°)$ interval. The conformations were generated by rigid-body rotation along the torsion angles representing the degrees of freedom of the molecule. The value of φ_0 was adjusted for each system to yield ∼40−60 millions of conformations. The parameters for each system are summarized in Table 1. The potential energy according to the G53a6 force field was associated with each conformation.

Calculation of Exact Entropies for the Peptide Test Systems. The calculation of the exact configurational entropy of the systems was carried out based on the classical statistical physical definition of entropy. The configurational entropy is

$$S_{\text{conf}} = k_B(\ln Z_{\text{conf}} + \beta\langle E_{\text{conf}}\rangle)$$

where $\beta = 1/(k_BT)$ is the temperature factor, Z_{conf} is the configuration integral, and $\langle E_{\text{conf}}\rangle$ is the average potential energy. These were approximated as sums based on the full set of conformations generated for each molecule. The formulas used were

$$Z_{\text{conf}} = \int e^{-\beta U(\mathbf{q})} \, d\mathbf{q} \approx \sum_k e^{-\beta\varepsilon_k}\varphi_0^d = Z_{\text{disc}}\varphi_0^d$$

$$E_{\text{conf}} = \frac{1}{Z_{\text{conf}}} \int U(\mathbf{q}) e^{-\beta U(\mathbf{q})} \, d\mathbf{q} \approx \frac{1}{Z_{\text{conf}}} \sum_k \varepsilon_k e^{-\beta\varepsilon_k}\varphi_0^d$$

$$= \frac{1}{Z_{\text{disc}}} \sum_k e_k e^{-\beta\varepsilon_k} = \langle E_{\text{disc}}\rangle$$

where $U(q)$ represents the potential energy as a function of the torsion angles q, the summation runs over the discrete conformations and ε_k is the potential energy of the kth conformation, d is the number of dimensions, and Z_{disc} and E_{disc} denote the discrete partition function and the discrete average energy, respectively. From these, the discrete entropy is obtained as

$$S_{\text{disc}} = k_B(\ln Z_{\text{disc}} + \beta\langle E_{\text{disc}}\rangle) + k_B \, d \ln \varphi_0$$

Generation of Monte Carlo (MC) Ensembles for the Peptide Test Systems. For the creation of canonical ensembles for the peptide test systems, a Metropolis Monte Carlo algorithm was implemented that samples the full set of conformations at a fixed temperature of T = 300 K. After an initial sampling, the configurational space of each system was examined as scatterplots of the torsion angles, and was divided into two subsets (denoted by A and B) for the purpose of entropy difference calculations between the two subsets. The definitions of subsets B are shown in Table 1; subsets A are the complements of subsets B. The exact entropy of each subset was calculated as described in the previous section. For each subset, 100 000 conformations were generated by MC sampling.

Because the points in the samples generated this way are on lattice points, many of them coincide (i.e., they are repeated). We found that the k-nearest neighbor method for entropy

DOI: 10.1021/acs.jctc.6b00837
J. Chem. Theory Comput. 2017, 13, 29−41

estimation (which we used for comparisons with our method) does not handle repeated points well; this is understandable as the method uses the distance of a point from its kth nearest neighbor, which is often zero when the points are repeated. To eliminate the repeated sample points, we applied a random shift to the points according to a uniform distribution with a maximum absolute value of half the lattice unit along each dimension. This "smearing" of the sample does not introduce any distortions into the underlying probability density and should conserve the entropy.

Generation of Molecular Dynamics (MD) Ensembles for Tachyplesin. We used tachyplesin,[44] a 17-residue antimicrobial peptide to test our entropy calculation method. This peptide is stabilized by two intramolecular disulfide bonds between residues 7—12 and 3—16. Starting with the solution structure (PDB ID: 1MA2[45]), we generated two modified molecules by cutting one (3—16) or both disulfide bridges. MD simulations were run on the wild-type and the two disulfide-cut molecules using GROMACS 5.0[43] in GBSA implicit solvent[46] with the stochastic dynamics integrator at 300 K with LINCS[47] bond constraints and a 2 fs time step; infinite cutoffs were used for the nonbonded interactions. Thirty-two independent simulations of 50 ns were run for each molecule for better sampling; the first 3 ns were considered equilibration and discarded. A total of 125 344 conformations were used for entropy estimations for each molecule.

Molecular Dynamics Samples for BPTI. Bovine pancreatic trypsin inhibitor (BPTI) is a 58-residue protein whose native-state dynamics has been characterized in detail by a 1 ms MD simulation.[48] We have obtained this 1 ms trajectory (courtesy of D. E. Shaw Research), and used every seventh frame in the trajectory to obtain an ensemble of 589 281 conformations for our analyses. Subsets for entropy difference calculations were defined as described in the Results section.

Implementation of Other Entropy Estimation Methods. To compare our method with other, published entropy estimation methods, we implemented the following methods. The classical quasiharmonic method[5] consists of fitting a single Gaussian onto the sample, and thus is identical with the first step of our Gaussian mixture method. The k-nearest neighbor (kNN) method is used in several published entropy estimation methods,[14,15,26] and it is based on estimating the probability density at a point based on its distance from its kth nearest neighbor. Here, k is an arbitrary parameter; we tested values from 1 to 4 in accordance with literature recommendations.[26] We implemented the nearest-neighbor search with k–d trees.[49] Unlike some implementations in the literature,[14,15] we did not use any extrapolation technique with our kNN tests as this would require a range of large sample sizes and we were interested in how the methods handle smaller sample sizes. The "2D entropy" method of Wang & Brüschweiler[12] treats torsion angles as complex numbers, performs principal component analysis of the sample, and uses kernel density estimation to estimate the probability densities along each principal axis, followed by numerical integration to obtain the entropies; the total entropy is obtained as the sum of entropies calculated for the individual eigenmodes. Here, the bandwidth σ of the kernel is a free parameter; we tested multiple values between 0.05 and 0.5 in accordance with literature recommendations.[50] We also tested low-dimensional approximations to the entropy based on mutual information expansion (MIE).[15]

Measuring Correlations between Coordinates. Linear associations between coordinates were measured by the Pearson correlation coefficient r. For a more general measure of association between coordinates q_i and q_j, we used the mutual information defined as

$$I(q_i, q_j) = S(q_i) + S(q_j) - S(q_i, q_j)$$

where $S(q_i)$ and $S(q_j)$ denote the marginal entropies and $S(q_i, q_j)$ is the joint entropy of q_i and q_j. If the correlation coefficient between q_i and q_j is zero (as is the case when the coordinates have been transformed by principal component analysis) then the mutual information describes the nonlinear or supralinear association between the coordinates.[26] To handle cases in which the correlation coefficient r is nonzero (as when using the original, nontransformed coordinates), we introduce the quasiharmonic mutual information as

$$I^{(\mathrm{qh})} = \frac{1}{2}\sum_{i=1}^{d} \log \sigma_{ii} - \frac{1}{2}\log\det(\boldsymbol{\sigma})$$

where $\boldsymbol{\sigma}$ is the covariance matrix of the coordinates as calculated from a sample. $I^{(\mathrm{qh})}$ is the difference between the sum of one-dimensional marginal quasiharmonic entropies and the full-dimensional quasiharmonic entropy; thus it is the mutual information taken into account by the quasiharmonic approximation. It essentially captures the linear correlations between the coordinates; a similar approach has been described in the literature.[51] The difference between the full mutual information I (obtained by summing the pairwise mutual information values) and the quasiharmonic mutual information $I^{(\mathrm{qh})}$ is thus a measure of supralinear correlations in the system. When used as contributions to thermodynamic entropies, the mutual information values are multiplied by the gas constant R.

RESULTS

Entropy Estimation by Gaussian Mixtures. Any continuous multivariate probability density can be arbitrarily closely approximated by a Gaussian mixture function of the form

$$p_k(\mathbf{q}) = \sum_{i=1}^{k} w_i N(\mathbf{q}; \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)$$

where the vector $\mathbf{q}$ represents the variables (in our case, torsion angles), k is the number of Gaussian components, w_i is the weight of the ith component, and $N(\mathbf{q}; \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)$ is the multivariate normal distribution with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\sigma}_i$. Given a sample of n input points $\mathbf{Q} = (\mathbf{q}_1, ..., \mathbf{q}_n)$, in our case an ensemble of n molecular conformations described by sets of torsion angles, the underlying probability density can be estimated as a Gaussian mixture, which involves estimating the number of components, the weights, and the means and covariance matrices. This is a computationally highly intensive task if all the parameters (including the number of components) are estimated simultaneously as the parameter space to be searched is high-dimensional. Therefore, we use the greedy learning method developed by Verbeek et al.,[41] which first fits a single Gaussian onto the sample, and then adds new components one by one, choosing the new component in each step from a number of locally optimal candidate components and then applies expectation-maximization (EM) to the resulting mixture. The number of candidate components tested in each step and the convergence threshold for the EM step are free parameters that can be set by the user; we set them to 30 and 10^{-5}, respectively, but we found that the results were rather

DOI: 10.1021/acs.jctc.6b00837
J. Chem. Theory Comput. 2017, 13, 29–41

insensitive to these settings. New components are added until a predefined stopping criterion is met (see next subsection). To implement the algorithm, we adapted the Matlab code downloaded from http://lear.inrialpes.fr/~verbeek/software to Python/NumPy, with some modifications.

Once we obtain the estimated Gaussian mixture p_k, the entropy could be obtained by integration according to the Shannon formula. However, this extra calculation is not necessary. As the greedy expectation-maximization algorithm outputs the maximized log-likelihood

$$L = L(\mathbf{Q}, p_k) = \sum_{i=1}^{n} \log p_k(\mathbf{q}_i)$$

the estimated information entropy of the distribution is simply obtained as the negative average log-likelihood

$$H = -L/n$$

which is a consequence of the weak law of large numbers, that is, the fact that the sample average of a quantity converges to its expected value; entropy is by definition the negative of the expected value of the logarithm of the probability density. To obtain the molar entropy, we multiply the information entropy by the universal gas constant R:

$$S = RH$$

Stopping Criterion. The greedy expectation-maximization algorithm as described by Verbeek et al.[41] does not specify a particular criterion determining when we should stop adding new components to the Gaussian mixture. We tested a number of stopping criteria including a simple threshold for the decrease in entropy upon adding a new component, the Bayesian Information Criterion, and the Akaike Information Criterion, but none of these criteria worked consistently across sample sizes and dimensionalities. Therefore, we introduced a cross validation based stopping criterion. Briefly, the input sample $\mathbf{Q}$ is randomly divided into two equal parts, $\mathbf{X}$ (as the training set) and $\mathbf{Y}$ (as the testing set), and the greedy learning method is applied to $\mathbf{X}$, yielding a k-component Gaussian mixture $p_{\mathbf{X},k}$. To decide whether to keep the kth component, we calculate the log-likelihood of $\mathbf{Y}$ using $p_{\mathbf{X},k}$:

$$L(\mathbf{Y}, p_{\mathbf{X},k}) = \sum_{\mathbf{q}_i \in \mathbf{Y}} \log p_{\mathbf{X},k}(\mathbf{q}_i)$$

and we stop the algorithm and discard the kth component when the log-likelihood decreased relative to the previous step, that is, if $L(\mathbf{Y}, p_{\mathbf{X},k}) < L(\mathbf{Y}, p_{\mathbf{X},k-1})$. Otherwise, the new component is added and the algorithm is continued. We found that this stopping criterion is rather robust and reliably prevents both underfitting and overfitting with all sample sizes and dimensionalities, and has the advantage that it does not require any arbitrary parameters. When the algorithm has converged, the entropy can be estimated from the log-likelihood on either $\mathbf{X}$ or $\mathbf{Y}$ (or even the full sample $\mathbf{Q}$), although the log-likelihood on $\mathbf{X}$ is always higher, this difference usually disappears when calculating entropy differences. In our tests with small peptide systems (see later), we found that the entropy estimated from the training set $\mathbf{X}$ still tends to slightly overestimate the exact entropy, indicating that our method is unlikely to be affected by the bias found in histogram-based methods,[31] as it would result in underestimated entropies.

As a consequence of the random division of the sample into two parts, repeating the calculation several times usually gives slightly different results as the sample is divided differently each time. This variation decreases with increasing sample size. To obtain more accurate entropy estimates, the calculation can be repeated a number of times and the average of the results can be taken; the standard error can also be calculated. When calculating entropy differences, it is desirable that the two samples are about the same size as this ensures similar accuracy of their calculated entropies.

Scaling of the Input Data to Prevent Arithmetic Errors. The formula for the individual Gaussian components contains the determinant of the covariant matrix in the normalizing factor, and this determinant scales exponentially with the number of dimensions d. Therefore, with higher dimensionalities, the normalizing factor can easily result in an arithmetic overflow or underflow during the computation. To eliminate this problem, before starting the greedy learning algorithm, we fit a single Gaussian to the whole sample and check whether the determinant D of the covariance matrix is larger than 10^{20} or is smaller than 1. In either case, we scale the input data by the scaling factor $f = (10^{20}/D)^{1/d}$. This will ensure that no underflows or overflows occur during the greedy learning. At the end of the calculation, the effect of the scaling is removed by adding $d \log f$ to the obtained log-likelihood.

Distribution Centering. As torsion angles are circular variables usually mapped to the −180° to 180° interval, fitting a probability distribution to such samples may cause artifacts if significant portions of the probability density fall near the periodic boundary; for example, a peak near 180° gets split in half and erroneously appears as two peaks. To alleviate these problems, centering of the data was performed. This was done by creating a histogram of values independently for each torsion angle and finding the longest contiguous interval of angles with minimum frequency. A transformation was then applied to rotate the torsion angle values to move the periodic boundary (180°) to the middle of the interval. This way, we make sure that no peaks fall on the periodic boundary. This procedure was repeated for each internal coordinate. The procedure shifts possible peaks at the periodic boundary but otherwise does not change the information entropy of the distribution.

Accuracy of the Gaussian Mixture Entropy Estimation Method. *Peptide Test Systems.* The accuracy of our entropy estimation method was tested on five small peptide systems with 3 to 5 rotatable torsion angles. The molecules are shown in Figure 1, and the details of the systems are listed in Table 1. To be able to calculate entropy differences, the configuration space of each molecule was divided into two parts (subsets A and B; see table for their definition) as described in the Methods section.

Calculation of Exact Entropies. The configuration space of each peptide was enumerated on a lattice in the space of torsion angles. Each conformation was assigned a potential energy, and the exact entropies for subsets A and B were calculated using the partition function and the average energy (see Methods for details).

Correlations between Coordinates. When selecting the molecules for testing, we tried to ensure that there is a sufficient amount of correlation between the degrees of freedom to pose a challenge to the entropy estimation methods. Indeed, each test peptide has one or more pairs of highly correlated torsion angles: the correlation coefficients with the highest absolute value are −0.60, −0.40, −0.62, −0.59, and −0.70 for the ala3, ala-val-ala, ile, val, and val2 peptides, respectively. We also

DOI: 10.1021/acs.jctc.6b00837
J. Chem. Theory Comput. 2017, 13, 29−41

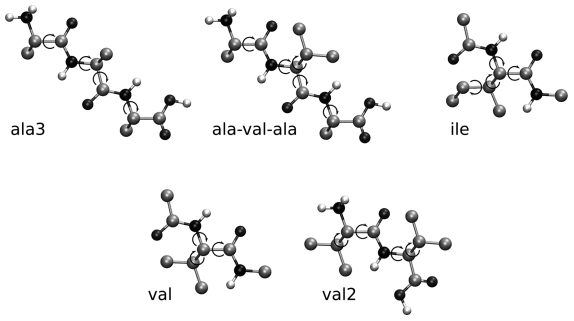


Figure 1. Molecular systems of small amino acids or derivatives selected for ensemble generation. Selected molecules were Ala_3 ("ala3"), Ala-Val-Ala ("ala-val-ala"), Ace-Ile-Nme ("ile"), Ace-Val-Nme ("val") and Val_2 ("val2"). Termini were left uncharged. United carbon atoms are displayed in gray, oxygen and nitrogen are black, and hydrogen atoms are white. Rotatable bonds, which represent the degrees of freedom, are indicated.

calculated the total mutual information for each molecule from the discrete marginal and joint probability distributions, as well as the quasiharmonic mutual information (see Methods). We find that correlations not captured by the quasiharmonic model account for 26%, 63%, 54%, 53%, and 54% of the total mutual information for the five molecules, respectively. Thus, the degrees of freedom of our test systems have a considerable amount of both linear and supralinear associations.

Monte Carlo (MC) Sampling. To obtain suitable samples for testing the entropy estimation algorithms, the set of conformations generated in the previous step was sampled by MC sampling to generate 100 000 conformations for both subsets A and B of each peptide.

Accuracy of Estimated Entropies. The Gaussian mixture method and three other, widely used methods, namely the classical quasiharmonic method,[5] the *k*-nearest neighbor method (kNN),[14,15,26] and the "2D entropy" method[12] were used to calculate entropy differences between substates A and B of each test peptide from the generated MC samples. Various sample sizes were tested in order to compare how the results depend on sample size. Table 2 shows the exact entropy differences ΔS_{AB} in comparison with the estimates obtained with each algorithm at a sample size of 10 000 conformations, and the average deviation from the exact entropy difference. Because the entropy estimated by the Gaussian mixture method can be different at each run due to the random selection of subsets for the cross-validation stopping criterion, we performed the calculation 20 times, and show the mean and the standard deviation of the results. The Gaussian mixture method has proven to be by far the most accurate, with an average deviation of only 0.54 ± 0.33 J/K/mol from the exact entropy difference. The quasiharmonic method is the least accurate with a deviation of 7.42 J/K/mol, while the 2D entropy and the kNN methods are about equally accurate with a deviation of 1.41 and 2.73 J/K/mol, respectively. It should be noted that we tested the kNN method with k values from 1 to 4, and the 2D entropy method with several values of the σ parameter between 0.05 and 0.5; we obtained the most accurate results with $k = 4$ and $\sigma = 0.5$; only these are shown in the table.

To see how each method performs at various sample sizes, we performed the calculations with several sample sizes from 2500 to 100 000. The results are shown in Figure 2 for each molecule, and the bottom right graph shows the averaged accuracy over all five peptides. As expected, all methods are inaccurate at very small sample sizes, and get more accurate as sample size increases, although the quasiharmonic method remains inaccurate even at large sample sizes. The accuracy of the Gaussian mixture method increases significantly faster with sample size than does the accuracy of the kNN method, and the Gaussian mixture method is about as accurate at a sample size of 5000 as the kNN method is at a sample size of 100 000. The accuracy of the 2D entropy method is less dependent on the sample size than the kNN method, but it is less accurate overall than the Gaussian mixture method. Overall, the Gaussian mixture method provides more accurate entropies at smaller sample sizes than the other methods.

Application to Tachyplesin. *MD Simulations.* To probe the behavior of our algorithm on more complex systems, we applied it to larger molecules. Tachyplesin[44] is a 17-residue antimicrobial peptide whose structure is stabilized by two disulfide bridges between residues 3—16 and 7—12. We used our Gaussian mixture entropy estimation method to estimate the entropic effect of cutting one (the 3—16) or both disulfide bridges in this molecule. Large MD samples (32 parallel simulations of 50 ns to obtain a total of 1600 ns simulation time for each molecule) were generated as described in the Methods section; we chose to perform many shorter simulations instead of a single long one as this approach is known to provide better sampling.[52] Only the backbone torsion angles φ and ψ were used in the entropy calculation; the degrees of freedom are thus 32.

Convergence of the Entropy during the Algorithm. Figure 3 shows the convergence of the entropies during the running of the algorithm as a function of the number of Gaussian components added, both for the native tachyplesin with two disulfide bridges and the variant with no disulfide bridges. The entropies calculated for the training set (denoted by **X**, see subsection "Stopping Criterion") and the testing set (denoted by **Y**) are separately shown. The entropy $S_{\mathbf{X}}$ calculated for the testing set **X** is always a bit lower than that ($S_{\mathbf{Y}}$) of the training

Table 2. Accuracy of Entropy Calculation Methods on Small Samples[a]

molecule	exact	Gaussian mixture	quasiharmonic	nearest neighbor ($k = 4$)	2D entropy ($\sigma = 0.5$)
ala3	4.20	4.01 ± 0.27	3.82	3.45	1.17
ala-val-ala	−5.00	−3.6 ± 0.56	−15.31	−0.86	−3.95
ile	−1.30	−1.45 ± 0.26	5.32	−5.63	0.00
val	−0.87	−0.75 ± 0.21	−11.22	0.63	−1.60
val2	1.10	1.92 ± 0.35	10.54	−1.81	2.00
average deviation	0	0.54 ± 0.33	7.42	2.73	1.41

[a]Entropy differences (in J/K/mol) between the substates of five molecules as defined in Table 1, calculated with various methods, from ensembles of 10 000 conformations. The last row shows the absolute deviation from the exact value averaged over the five molecules. Results with larger samples are presented in Figure 2. For the Gaussian mixture method, the mean ± standard deviation is shown based on running the calculation 20 times.

DOI: 10.1021/acs.jctc.6b00837
J. Chem. Theory Comput. 2017, 13, 29−41

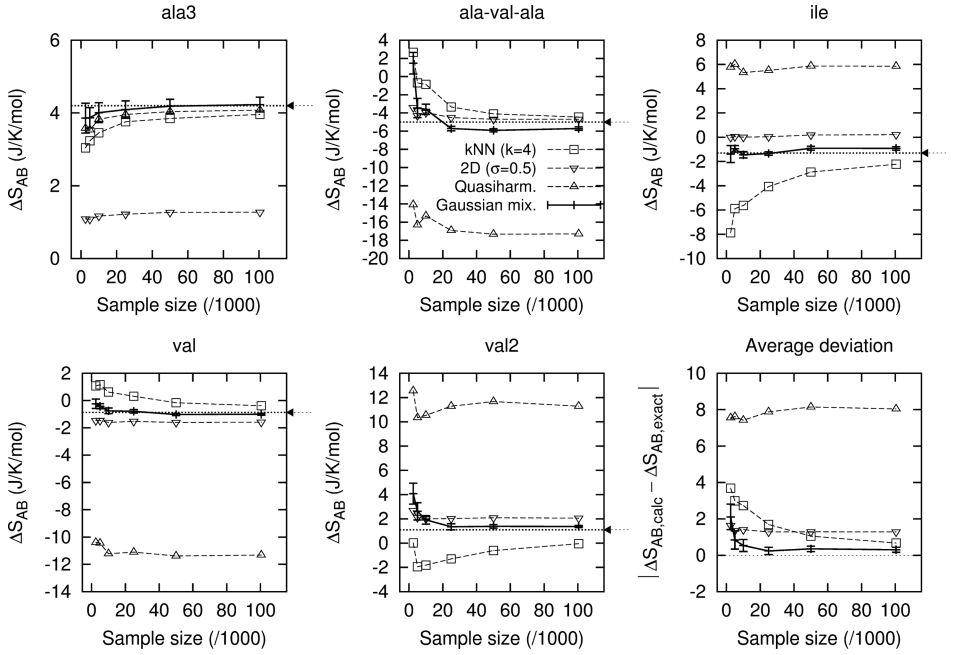


Figure 2. Entropy differences between the A and B substates of five molecules as defined in Table 1, calculated with various methods, plotted against the sample size. The legend is shown in the top center plot. The exact entropy difference is indicated as a horizontal dotted line with an arrow at the right side of each plot. The bottom right plot shows the absolute deviation from the exact value averaged over the five molecules. Error bars represent the standard deviation of the results from the Gaussian mixture method.

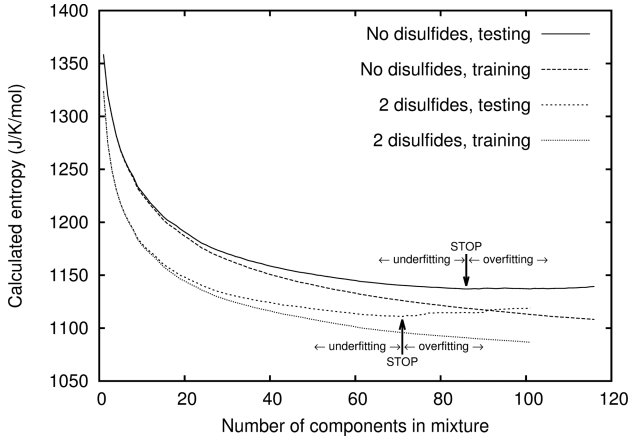


Figure 3. Entropies calculated by Gaussian mixture fitting to backbone torsion angle ensembles for tachyplesin with no disulfide bridges and two disulfide bridges from both the training and testing ensemble subsets, plotted against the number of components in the mixture. Ensembles from the first 21 ns of the simulation were used (~80 000 conformations per molecule). The calculation was continued beyond the point (indicated by arrows with the "STOP" label) where the stopping criterion was met to illustrate underfitting and overfitting behavior.

set **Y**; while S_X keeps decreasing as more components are added, S_Y stops decreasing at a certain number of components (50 to 60 components in this case), and that is where the algorithm is stopped. Figure 3 also shows the behavior of the entropies when the algorithm is not stopped: the training set entropy keeps decreasing while the testing set entropy increases, indicating overfitting.

Convergence of the Entropy with Simulation Time. Reliable estimation of the entropy requires adequate sampling; thus, it is a good practice to observe how the estimated entropy changes as the simulation progresses. As tachyplesin is a 17-residue peptide, a full sampling of its entire configurational space may require milliseconds, which is not feasible by conventional MD simulations. However, by performing 32 parallel simulations starting from the same initial structure, we can at least explore the vicinity of the starting structure and obtain an adequate sample of it. Figure 4 shows the variation of

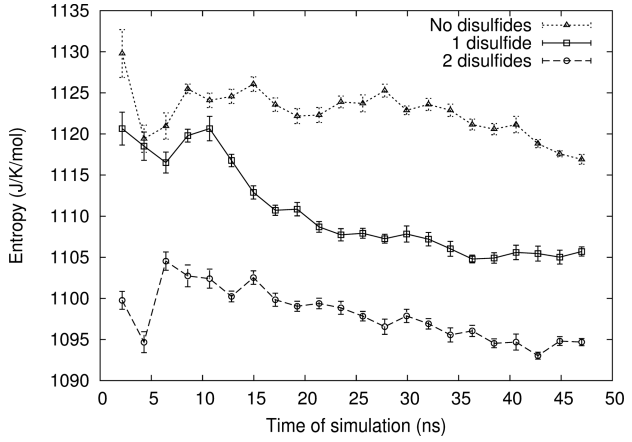


Figure 4. Entropies calculated by Gaussian mixture fitting to backbone torsion angle ensembles from MD simulations of tachyplesin with 0, 1, or 2 disulfide bridges as a function of simulation time. Ensembles were generated by merging the trajectories from 32 parallel independent simulations. Each data point is an average of 10 entropy calculations, with the standard errors shown as error bars.

entropy with simulation time for tachyplesin with 2, 1, and 0 disulfide bridges. The first 3 ns of each simulation was discarded; thus we used a maximum simulation time of 47 ns for the calculations. The trajectories from the 32 independent simulations were merged, and each entropy calculation was performed 10 times to allow us to calculate the averages and

DOI: 10.1021/acs.jctc.6b00837
J. Chem. Theory Comput. 2017, 13, 29−41

the standard errors. As expected, there are large variations in the entropy initially, which diminish as time progresses. After ~30 ns, the entropies of the disulfide-bonded species appear sufficiently converged, while there is still a slightly decreasing trend in the entropy of the molecule without disulfide bridges. This is understandable as the lack of disulfide bonds makes this molecule very flexible and it explores a significantly larger region in conformational space than the disulfide-bonded species. For this reason, its estimated entropy should be considered less reliable than that of the disulfide-bonded species.

Calculated Entropy Differences. The greedy learning method fitted 119 to 145, 96 to 121, and 93 to 106 Gaussian components to the ensembles of backbone torsion angles obtained from MD simulations of tachyplesin with 0, 1, and 2 disulfide bonds, respectively. The higher number of components needed to fit the ensembles of the 0-disulfide species indicates the significantly larger complexity of its energy landscape. Figure 5 shows scatterplots of the ensembles for

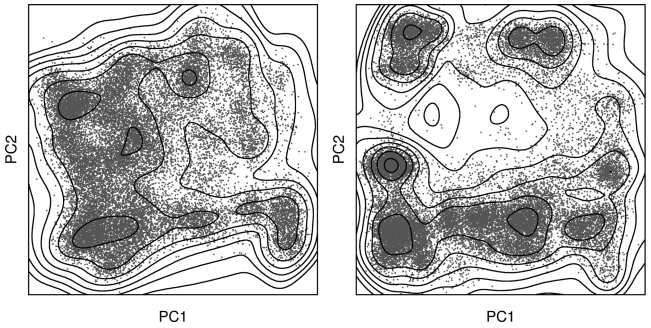


Figure 5. Projection of torsion angles from MD simulations of tachyplesin onto the plane of the first two principal components. Left, tachyplesin with no disulfide bonds; right, tachyplesin with two disulfide bonds. The gray dots represent conformations; the black contour lines represent the Gaussian mixtures fitted onto the samples.

the species with 2 disulfides and no disulfides, projected onto the plane of the first two principal components, and the fitted Gaussian mixtures are represented as contour lines. The graphs show that the probability density has narrower peaks in the disulfide-bonded state than without disulfide bridges. The entropy differences obtained from the calculations are $S_2 - S_0 = -22.2 \pm 0.7$ J/K/mol, $S_1 - S_0 = -11.2 \pm 0.8$ J/K/mol, and $S_2 - S_1 = -11.0 \pm 0.7$ J/K/mol where S_m denotes the entropy of the species with m disulfide bridges. We can compare the $S_1 - S_0$ value with that calculated with the formula $\Delta S = -8.7822 - 1.5R \ln n$ from polymer theory[53] for the entropy reduction of a random coil due to introducing a cross-link with n residues between the linked residues. From this theoretical formula, $S_1 - S_0$ should be about −26 J/K/mol. Considering the fact that the theoretical formula is for a noninteracting random coil, our result of −11.2 J/K/mol is in reasonable agreement with this, and the remaining difference is to be expected due to the intrachain interactions in tachyplesin which reduce its entropy in the non-cross-linked state. Our estimated entropies also come with the caveat of being derived from sampling a portion of conformational space reachable within 50 ns from the initial structure.

Application to BPTI. *Definition of Substates.* To test whether our Gaussian mixture method can be used for even larger systems, we applied it to calculate the entropy difference between two native substates of the 58-residue bovine pancreatic trypsin inhibitor. The native-state dynamics has been characterized in detail by a 1 ms MD simulation,[48] and we used this 1 ms trajectory (courtesy of D. E. Shaw Research) for our calculations as described in the Methods section. To visualize the substates of the native state, we performed principal component analysis on an ensemble of 589 281 frames described by 221 rotatable torsion angles (including backbone and side-chain torsion angles). Figure 6 shows these

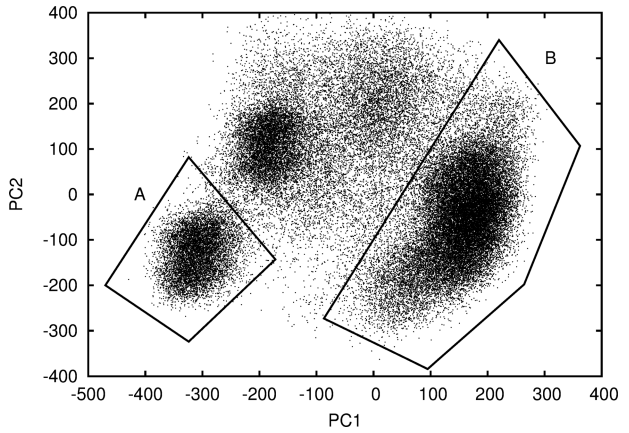


Figure 6. Definition of A and B substates for BPTI from an ensemble generated by a long MD simulation. All 221 torsion angles were subjected to principal component analysis and projected to the plane of the first two principal components. The subsets were defined as the prominent clusters A and B as shown in the graph.

torsion angles projected onto the plane of the first two principal components. We selected the two distinct sets of points shown in the figure as substates A and B, and used 85 000 frames from each subset for entropy calculations.

Calculated Entropy Difference. Entropy calculation by our Gaussian mixture method was applied to the 116 backbone torsion angles of the 85 000 frames in each subset selected as described above. Repeating the calculations 7 times for each subset, the greedy learning procedure fitted 10 to 14 Gaussian components onto subset A, and 9 to 12 components onto subset B. The entropy difference $S_A - S_B$ was estimated to be 31.8 ± 0.5 J/K/mol. Although we have no way to independently verify this result, it demonstrates the applicability of the method to similar systems.

Probing the Limits of the Method. We used the 1 ms BPTI trajectory to test how our method can cope with large samples of high dimensionality. Thus, we applied the greedy learning procedure to all 589 281 frames of all 221 torsion angles of the molecule. The program completed in 57 h of CPU time on an Intel Xeon E5430 CPU, and fitted 51 Gaussian components onto the data. The calculated entropy value is not shown as only entropy differences are physically meaningful.

Scaling of the Gaussian Mixture Method with Sample Size and Dimensionality. Being a parametric method, the running time of Gaussian mixture entropy estimation method depends on the input data: more complex distributions require more Gaussian components to fit, and therefore run longer. Therefore, a general formula to estimate the run time for a given sample cannot be given. The greedy learning procedure for a k-component mixture on n samples has a time complexity[41] of $O(k^2n)$, or $O(kmn)$ if m candidate components are used in the algorithm and $k < m$ (we used $m = 30$ in our

DOI: 10.1021/acs.jctc.6b00837
J. Chem. Theory Comput. 2017, 13, 29−41

implementation). The number of components k is in turn determined by the cross-validation stopping criterion in our implementation; thus it depends on the sample size: larger sample sizes will result in more Gaussian components fitted.

To examine the scaling of the method in more detail, we generated samples from 10-component Gaussian mixtures with sample sizes ranging from 1000 to 100,000 and dimensionalities from 1 to 50. The 10 Gaussian components were standard normal multivariate distributions with their means shifted to the position $5n$ for n = 0, ..., 9; thus, they were well separated along one axis.

The Gaussian mixture entropy estimation method was applied to these samples. Figure 7 shows the CPU time

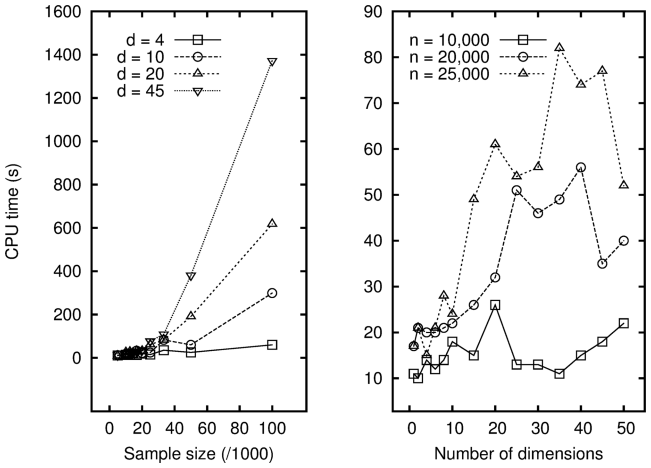


Figure 7. Performance of the Gaussian mixture entropy calculation method as measured by the CPU time needed to evaluate a sample generated from a 10-component Gaussian mixture. The sample size n and the number of variables (dimensions) d were both varied. Left, CPU time vs sample size for various values of d; right, CPU time vs the number of dimensions for various values of n.

needed by the procedure at various sample sizes n and dimensionalities d. At fixed dimensionalities, run time increases roughly linearly with the sample size for small dimensionalities, but faster than linearly for higher dimensionalities (left graph in Figure 7). The dependence on the dimensionality at fixed sample sizes is more complex (right graph in Figure 7), with the run time first increasing but then dropping again at higher dimensionalities. These findings are explained by the fact that the number k of Gaussian components fitted is determined by the cross-validation stopping criterion. If the number of dimensions is too high for the given sample size, the stopping criterion will stop the procedure early; in this case, less than 10 components will be fitted. For example, at a sample size of 10000 in 50 dimensions, only one Gaussian component is fitted because the sample is too small to enable the reliable identification of all 10 components in the probability distribution that was sampled. As sample size increases, more and more Gaussian components will be fitted. Thus, the procedure adjusts itself to the available data and only fits as many components as is possible without overfitting. This ensures that run times are reasonable even for high-dimensional data.

Comparison of the Running Time with Other Methods. We compared the running time of the Gaussian mixture method with that of the kNN and the 2D entropy methods. While run times were similar for small samples and low dimensionalities, both the kNN and the 2D entropy method required much longer times than our method for large sample sizes and high dimensionalities. For example, for 100 000 samples in 50 dimensions, the Gaussian mixture method completed in 1978 s while the kNN method with k = 5 ran for 7790 s, and the 2D entropy method required about 1 day. However, these running times can probably be improved, for example, by using a different algorithm for the kNN method and by less accurate numerical integration for the 2D entropy method. More importantly, as the run time of our method depends on the particular distribution while that of the kNN and 2D entropy methods (being nonparametric methods) does not, it is not possible to provide a universally valid comparison of the run times. In our experience, though, the cross-validation stopping criterion of our method ensures very reasonable running times in all cases as the program will stop when no more Gaussian components can be added without overfitting, the nonparametric methods do not have this favorable property.

Low-Dimensional Approximations. While our method easily works with relatively high-dimensional input data, larger systems with hundreds or thousands of degrees of freedom still pose a difficult problem. Several approximate methods have been developed that make use of low-dimensional marginal distributions of the full probability distribution; these include mutual information expansion (MIE),[13,15,26,28] maximum information spanning tree (MIST),[33] and multibody local approximation (MLA).[18,19] Several studies have suggested that higher-order correlations may be insignificant for the entropy, thus low-dimensional approximations may often be sufficiently accurate.[13,15,16] To test how our Gaussian mixture method works with low-dimensional approximations and to find out whether these are sufficiently accurate, we calculated three approximate entropies for our five peptide test systems. Using the MIE series expansion, the first-order approximation to the full d-dimensional entropy is simply the sum of one-dimensional marginal entropies:

$$S^{(1)} = \sum_{i=1}^{d} S(q_i)$$

where $S(q_i)$ is the marginal entropy calculated for the ith variable q_i; the marginal distribution of each variable was estimated by the Gaussian mixture method. Because the first-order entropy completely ignores correlations between variables, we also calculated a corrected first-order approximation as

$$S^{(1c)} = S^{(1)} - I^{(qh)}$$

where $I^{(qh)}$ is the quasi-harmonic mutual information as defined in the *Methods* section. This corrected first-order entropy fully accounts for anharmonicities in the individual variables but only accounts for correlations in the same way as the quasiharmonic method, that is, it assumes Gaussian joint distributions and it mostly captures the linear correlations between variables.

The second-order approximation is

$$S^{(2)} = S^{(1)} - \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} I(q_i, q_j)$$

where $I(q_i, q_j)$ is the mutual information between variables q_i and q_j (see Methods for its definition). For calculating $I(q_i, q_j)$, the joint entropy $S(q_i, q_j)$ was again calculated by the Gaussian

DOI: 10.1021/acs.jctc.6b00837
J. Chem. Theory Comput. 2017, 13, 29−41

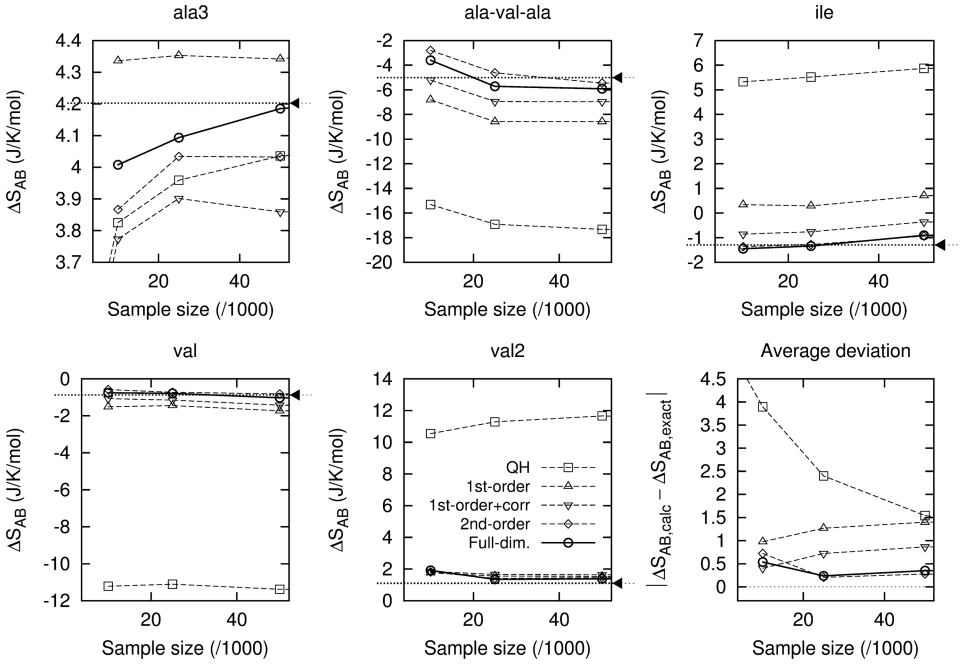


Figure 8. Similar to Figure 2, but the accuracies of the quasi-harmonic, first-order, corrected first-order, and second-order entropy calculations are compared. The indicated values are averages from 20 independent calculations.

mixture method. Thus, obtaining the first-order entropy requires d one-dimensional entropy calculations, and obtaining the second-order entropy requires $d(d-1)/2$ additional two-dimensional entropy calculations.

Figure 8 presents the comparison of quasiharmonic, first-order, corrected first-order, and second-order entropy differences with the full-dimensional calculation for our five peptide test systems, as a function of sample size. The bottom right graph shows the deviations from the exact entropy difference averaged over the five systems. Clearly, the quasiharmonic approximation is the least accurate, followed by the first-order, corrected first-order, and second order entropy differences. This succession shows how entropy estimates become more and more accurate as anharmonicities, linear correlations, and supralinear correlations are accounted. We found no significant difference between the second-order and the full-dimensional entropy differences. Although we were only able to perform these comparisons on small molecules (as we need the exact entropies as a reference), in general we still expect higher accuracy from the full-dimensional calculation, but if the size of the system is prohibitively large, the second-order approximation (when calculated by the Gaussian mixture method) should provide sufficiently accurate results.

It should be noted that unlike some earlier studies,[26,38] we worked with the original coordinates rather than principal-axis transformed coordinates (i.e., quasiharmonic modes) in our low-dimensional approximations presented above. The reason is that principal-axis transformed coordinates are linear combinations of the original coordinates and therefore their marginal distributions converge to a normal distribution due to the central limit theorem. This effect tends to eliminate anharmonicities, and it can hide much of the complexity of the probability density function, making the Gaussian mixture-based density estimation less accurate.

DISCUSSION

The Gaussian mixture entropy estimation method is a natural and conceptually simple extension of the classical quasiharmonic method as it makes the estimated entropy more accurate by simply adding more Gaussians to the probability density estimate. As any smooth function can be approximated by Gaussian mixtures, the method can naturally model multimodal and anharmonic distributions even in high-dimensional spaces. We have shown that the method provides more accurate results at smaller sample sizes than several other methods described in the literature, while also being faster than other methods in most cases. The success of the Gaussian mixture method can be ascribed to three main causes. First, being a parametric method, it assumes a particular functional form for the probability density (namely, a Gaussian mixture). As Gaussian distributions are associated with harmonic oscillators at constant temperature, it is expected that a Gaussian mixture should be a good description of the probability density function of a molecular system. In contrast, nonparametric methods such as the k-nearest-neighbor method do not assume a particular functional form, and therefore estimate densities that are distant from the reality unless the sample size is sufficiently large. Second, the greedy learning method to estimate Gaussian mixtures is highly efficient, and allows fast and accurate estimation. Third, the cross-validation stopping criterion makes the method robust by ensuring that a sufficient number of Gaussian components are used in the mixture, but not more components than what the available data actually justify.

Another advantage of the Gaussian mixture method is the fact that it is essentially parameter-free. The only user-adjustable parameters are the number of candidate Gaussians to test in each step, and the convergence threshold for the expectation-maximization procedure. However, the algorithm is not sensitive to the values of these parameters, and the default values provide good results regardless of the type or size of the molecular system to be analyzed. In contrast, the value of k in

DOI: 10.1021/acs.jctc.6b00837
J. Chem. Theory Comput. 2017, 13, 29−41

the k-nearest neighbor (kNN) method tends to be arbitrary and its value significantly influences the results (we only showed results obtained with the best k in Table 2). Similarly, the entropies calculated by the 2D entropy method strongly depend on the bandwidth parameter σ, whose value is not transferable between systems,[50] and there is no general rule for its determination (again, we only showed the results obtained with the best σ in Table 2). By introducing the cross-validation stopping criterion, we also eliminated the need for any arbitrarily adjustable convergence-related parameters.

We have shown that the Gaussian mixture method favorably compares to several other, state-of-the-art methods. The methods we used in our comparisons are based on widely used density estimation methods (kNN and kernel density estimation), and are relatively easy to implement. There are many more entropy estimation methods described in the literature, some of which rather complex (see the Introduction). Comparing our method with all of them would be beyond the scope of this article. However, among the methods that are based on probability density estimation, our method should be among the best as Gaussian mixtures can approximate even the most complex probability distributions to arbitrary accuracy.

The Gaussian mixture method bears some similarity to methods based on identifying a set of discrete conformations in local energy minima and assuming harmonic vibrations around them, applying the quasiharmonic approximation to each energy minimum.[20,35] However, our method does not separate vibrational and conformational entropy, and does not assume well-separated harmonic energy basins; the Gaussian mixture can accurately fit densities from arbitrarily shaped, smooth energy landscapes, including overlapping and anharmonic energy basins.

Being a density estimation method, the Gaussian mixture method could be a good alternative to other, less accurate density estimation methods in more complex entropy estimation schemes. This could even allow the incorporation of quantum mechanical entropies. For example, Numata et al.[26] first perform principal component analysis on the mass-weighted Cartesian coordinates to separate the quantum mechanical modes from the classical modes, and then use the kNN method to estimate the density for the classical modes, in the end obtaining an absolute entropy that includes quantum mechanical modes.

There has been some discussion in the literature about whether Cartesian or internal coordinates (or just torsion angles) are best suited for entropy estimation.[4,16,27,33,39] For large systems such as protein molecules, the high dimensionality of the configuration space in Cartesian coordinates poses a serious problem to all density estimation methods. Using internal coordinates, and in particular, exploiting the rigid nature of bond lengths and bond angles to switch to torsion angles only, is an efficient and physically meaningful way to reduce the dimensionality of the system. Although the Gaussian mixture method could in principle deal with Cartesian coordinates, the number of Gaussian components required to fit the distributions would be quite high as atoms tend to move on curved paths which cannot be well approximated with few Gaussians. Thus, the Gaussian mixture method is best for use with ensembles of torsion angles.

We have shown that the Gaussian mixture method is powerful enough to calculate full-dimensional entropies for large samples in relatively high-dimensional spaces (see the example of BPTI). However, for even larger systems such as big proteins, a full-dimensional calculation will not be feasible. We have shown that low-dimensional approximations such as mutual information expansion (MIE) can be sufficiently accurate, and the Gaussian mixture method can be used to calculate the marginal and joint entropies needed for such approximations. The Gaussian mixture method could also be used in combination with other approximation methods based on series expansions such as maximum information spanning tree (MIST),[33] and multibody local approximation (MLA),[18,19] as well as with clustering methods such as identifying minimally coupled subspaces (MCSA) by full correlation analysis (FCA).[27,28]

In addition to providing an entropy estimate, the Gaussian mixture method also provides an analytical function representing the probability density function of the system in torsion angle space. This analytical function only has a limited number of parameters and is easy to evaluate. Potential applications of this function include clustering the conformations, identifying the free energy basins, calculating various ensemble averages, preparing graphical representations of the free energy landscape (as in Figure 5), etc.

Though the Gaussian mixture method yields good results with small sample sizes, we should note that this does not mean that it can calculate accurate entropies from insufficient samples resulting from undersampling. The accuracy of any calculated entropy value critically depends on sampling quality, and no entropy estimation method can compensate for poor sampling.

CONCLUSIONS

The Gaussian mixture method is an accurate and efficient method to estimate classical entropy differences from ensembles of molecular conformations. It can calculate full-dimensional entropies for relatively high dimensionalities, but it can also be used in combination with approximation methods (such as mutual information expansion). Because it is parametric, it is more accurate at smaller sample sizes than several other density estimation methods, and therefore it is a powerful alternative to the k-nearest neighbor and kernel density estimation-based methods. Our implementation of the Gaussian mixture method is available at http://gmentropy.szialab.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: szilagyi.andras@ttk.mta.hu.

ORCID

András Szilágyi: 0000-0002-1773-6861

Present Address

‡Institute of Biochemistry and Molecular Medicine, University of Bern, Bühlstrasse 28, 3012 Bern, Switzerland.

Funding

This work was supported by the Hungarian Scientific Research Fund (OTKA), Grant Nos. K105415 and NK108642. Gergely Gyimesi is currently funded by the Marie Curie Actions International Fellowship Program (IFP) TransCure (www.transcure.org).

Notes

The authors declare no competing financial interest.

DOI: 10.1021/acs.jctc.6b00837
J. Chem. Theory Comput. 2017, 13, 29−41

■ ACKNOWLEDGMENTS

We thank Tamás Hegedüs for his comments on an early version of the manuscript. We thank D. E. Shaw for providing us with the 1 ms BPTI trajectory.

■ REFERENCES

(1) Baldwin, R. L. Temperature Dependence of the Hydrophobic Interaction in Protein Folding. *Proc. Natl. Acad. Sci. U. S. A.* **1986**, *83* (21), 8069−8072.
(2) Meirovitch, H.; Cheluvaraja, S.; White, R. P. Methods for Calculating the Entropy and Free Energy and Their Application to Problems Involving Protein Flexibility and Ligand Binding. *Curr. Protein Pept. Sci.* **2009**, *10* (3), 229−243.
(3) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., 3rd Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res.* **2000**, *33* (12), 889−897.
(4) Baron, R.; van Gunsteren, W. F.; Hünenberger, P. H. Estimating the Configurational Entropy from Molecular Dynamics Simulations: Anharmonicity and Correlation Corrections to the Quasi-Harmonic Approximation. *Trends Phys. Chem.* **2006**, *11*, 87−122.
(5) Karplus, M.; Kushick, J. N. Method for Estimating the Configurational Entropy of Macromolecules. *Macromolecules* **1981**, *14* (2), 325−332.
(6) Brady, J.; Karplus, M. Configuration Entropy of the Alanine Dipeptide in Vacuum and in Solution: A Molecular Dynamics Study. *J. Am. Chem. Soc.* **1985**, *107* (21), 6103−6105.
(7) Edholm, O.; Berendsen, H. J. C. Entropy Estimation from Simulations of Non-Diffusive Systems. *Mol. Phys.* **1984**, *51* (4), 1011−1028.
(8) Di Nola, A.; Berendsen, H. J. C.; Edholm, O. Free Energy Determination of Polypeptide Conformations Generated by Molecular Dynamics. *Macromolecules* **1984**, *17* (10), 2044−2050.
(9) Rojas, O. L.; Levy, R. M.; Szabo, A. Corrections to the Quasiharmonic Approximation for Evaluating Molecular Entropies. *J. Chem. Phys.* **1986**, *85* (2), 1037.
(10) Hnizdo, V.; Fedorowicz, A.; Singh, H.; Demchuk, E. Statistical Thermodynamics of Internal Rotation in a Hindering Potential of Mean Force Obtained from Computer Simulations. *J. Comput. Chem.* **2003**, *24* (10), 1172−1183.
(11) Darian, E.; Hnizdo, V.; Fedorowicz, A.; Singh, H.; Demchuk, E. Estimation of the Absolute Internal-Rotation Entropy of Molecules with Two Torsional Degrees of Freedom from Stochastic Simulations. *J. Comput. Chem.* **2005**, *26* (7), 651−660.
(12) Wang, J.; Brüschweiler, R. 2D Entropy of Discrete Molecular Ensembles. *J. Chem. Theory Comput.* **2006**, *2* (1), 18−24.
(13) Killian, B. J.; Yudenfreund Kravitz, J.; Gilson, M. K. Extraction of Configurational Entropy from Molecular Simulations via an Expansion Approximation. *J. Chem. Phys.* **2007**, *127* (2), 24107.
(14) Hnizdo, V.; Darian, E.; Fedorowicz, A.; Demchuk, E.; Li, S.; Singh, H. Nearest-Neighbor Nonparametric Method for Estimating the Configurational Entropy of Complex Molecules. *J. Comput. Chem.* **2007**, *28* (3), 655−668.
(15) Hnizdo, V.; Tan, J.; Killian, B. J.; Gilson, M. K. Efficient Calculation of Configurational Entropy from Molecular Simulations by Combining the Mutual-Information Expansion and Nearest-Neighbor Methods. *J. Comput. Chem.* **2008**, *29* (10), 1605−1614.
(16) Li, D.-W.; Brüschweiler, R. In Silico Relationship between Configurational Entropy and Soft Degrees of Freedom in Proteins and Peptides. *Phys. Rev. Lett.* **2009**, *102* (11), 118108.
(17) Baxa, M. C.; Haddadian, E. J.; Jha, A. K.; Freed, K. F.; Sosnick, T. R. Context and Force Field Dependence of the Loss of Protein Backbone Entropy upon Folding Using Realistic Denatured and Native State Ensembles. *J. Am. Chem. Soc.* **2012**, *134* (38), 15929−15936.
(18) Suárez, E.; Suárez, D. Multibody Local Approximation: Application to Conformational Entropy Calculations on Biomolecules. *J. Chem. Phys.* **2012**, *137* (8), 84115.
(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* (23), 2041−2054.
(20) Karplus, M.; Ichiye, T.; Pettitt, B. M. Configurational Entropy of Native Proteins. *Biophys. J.* **1987**, *52* (6), 1083−1085.
(21) Schlitter, J. Estimation of Absolute and Relative Entropies of Macromolecules Using the Covariance Matrix. *Chem. Phys. Lett.* **1993**, *215* (6), 617−621.
(22) Schäfer, H.; Mark, A. E.; van Gunsteren, W. F. Absolute Entropies from Molecular Dynamics Simulation Trajectories. *J. Chem. Phys.* **2000**, *113* (18), 7809.
(23) Yu, Y. B.; Privalov, P. L.; Hodges, R. S. Contribution of Translational and Rotational Motions to Molecular Association in Aqueous Solution. *Biophys. J.* **2001**, *81* (3), 1632−1642.
(24) Schäfer, H.; Smith, L. J.; Mark, A. E.; van Gunsteren, W. F. Entropy Calculations on the Molten Globule State of a Protein: Side-Chain Entropies of Alpha-Lactalbumin. *Proteins: Struct., Funct., Genet.* **2002**, *46* (2), 215−224.
(25) Andricioaei, I.; Karplus, M. On the Calculation of Entropy from Covariance Matrices of the Atomic Fluctuations. *J. Chem. Phys.* **2001**, *115* (14), 6289.
(26) Numata, J.; Wan, M.; Knapp, E. W. Conformational Entropy of Biomolecules: Beyond the Quasi-Harmonic Approximation. *Genome Inf.* **2007**, *18*, 192−205.
(27) Hensen, U.; Grubmüller, H.; Lange, O. F. Adaptive Anisotropic Kernels for Nonparametric Estimation of Absolute Configurational Entropies in High-Dimensional Configuration Spaces. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **2009**, *80* (1), 11913.
(28) Hensen, U.; Lange, O. F.; Grubmuller, H. Estimating Absolute Configurational Entropies of Macromolecules: The Minimally Coupled Subspace Approach. *PLoS One* **2010**, *5* (2), e9179.
(29) Li, D.-W.; Brüschweiler, R. A Dictionary for Protein Side-Chain Entropies from NMR Order Parameters. *J. Am. Chem. Soc.* **2009**, *131* (21), 7226−7227.
(30) Genheden, S.; Akke, M.; Ryde, U. Conformational Entropies and Order Parameters: Convergence, Reproducibility, and Transferability. *J. Chem. Theory Comput.* **2014**, *10* (1), 432−438.
(31) Numata, J.; Knapp, E.-W. Balanced and Bias-Corrected Computation of Conformational Entropy Differences for Molecular Trajectories. *J. Chem. Theory Comput.* **2012**, *8* (4), 1235−1245.
(32) 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* (8), 2638−2653.
(33) King, B. M.; Silver, N. W.; Tidor, B. Efficient Calculation of Molecular Configurational Entropies Using an Information Theoretic Approximation. *J. Phys. Chem. B* **2012**, *116* (9), 2891−2904.
(34) Chang, C.-E.; Chen, W.; Gilson, M. K. Evaluating the Accuracy of the Quasiharmonic Approximation. *J. Chem. Theory Comput.* **2005**, *1* (5), 1017−1028.
(35) Goethe, M.; Fita, I.; Rubi, J. M. Vibrational Entropy of a Protein: Large Differences between Distinct Conformations. *J. Chem. Theory Comput.* **2015**, *11* (1), 351−359.
(36) Doig, A. J.; Sternberg, M. J. Side-Chain Conformational Entropy in Protein Folding. *Protein Sci.* **1995**, *4* (11), 2247−2251.
(37) Zhou, H.-X.; Gilson, M. K. Theory of Free Energy and Entropy in Noncovalent Binding. *Chem. Rev.* **2009**, *109* (9), 4092−4107.
(38) 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* (12), 3150−3160.
(39) Cheluvaraja, S.; Meirovitch, H. Simulation Method for Calculating the Entropy and Free Energy of Peptides and Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (25), 9241−9246.

DOI: 10.1021/acs.jctc.6b00837
J. Chem. Theory Comput. 2017, 13, 29−41

(40) Cheluvaraja, S.; Meirovitch, H. Calculation of the Entropy and Free Energy of Peptides by Molecular Dynamics Simulations Using the Hypothetical Scanning Molecular Dynamics Method. *J. Chem. Phys.* **2006**, *125* (2), 24905.
(41) Verbeek, J.; Vlassis, N.; Kröse, B. Efficient Greedy Learning of Gaussian Mixture Models. *Neural Comput.* **2003**, *15* (2), 469−485.
(42) Gō, N.; Scheraga, H. A. On the Use of Classical Statistical Mechanics in the Treatment of Polymer Chain Conformation. *Macromolecules* **1976**, *9* (4), 535−542.
(43) 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* (7), 845−854.
(44) Nakamura, T.; Furunaka, H.; Miyata, T.; Tokunaga, F.; Muta, T.; Iwanaga, S.; Niwa, M.; Takao, T.; Shimonishi, Y. Tachyplesin, a Class of Antimicrobial Peptide from the Hemocytes of the Horseshoe Crab (Tachypleus Tridentatus). Isolation and Chemical Structure. *J. Biol. Chem.* **1988**, *263* (32), 16709−16713.
(45) Laederach, A.; Andreotti, A. H.; Fulton, D. B. Solution and Micelle-Bound Structures of Tachyplesin I and Its Active Aromatic Linear Derivatives. *Biochemistry* **2002**, *41* (41), 12359−12368.
(46) Im, W.; Lee, M. S.; Brooks, C. L. Generalized Born Model with a Simple Smoothing Function. *J. Comput. Chem.* **2003**, *24* (14), 1691−1702.
(47) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (1), 116−122.
(48) 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.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330* (6002), 341−346.
(49) Maneewongvatana, S.; Mount, D. M. On the Efficiency of Nearest Neighbor Searching with Data Clustered in Lower Dimensions. In *Computational Science−ICCS 2001*; Alexandrov, V. N., Dongarra, J. J., Juliano, B. A., Renner, R. S., Tan, C. J. K., Eds.; Goos, G., Hartmanis, J., van Leeuwen, J., Series Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2001; Vol. *2073*, pp 842−851.
(50) Li, D.-W.; Khanlarzadeh, M.; Wang, J.; Huo, S.; Brüschweiler, R. Evaluation of Configurational Entropy Methods from Peptide Folding-Unfolding Simulation. *J. Phys. Chem. B* **2007**, *111* (49), 13807−13813.
(51) Polyansky, A. A.; Kuzmanic, A.; Hlevnjak, M.; Zagrovic, B. On the Contribution of Linear Correlations to Quasi-Harmonic Conformational Entropy in Proteins. *J. Chem. Theory Comput.* **2012**, *8* (10), 3820−3829.
(52) Allnér, O.; Foloppe, N.; Nilsson, L. Motions and Entropies in Proteins as Seen in NMR Relaxation Experiments and Molecular Dynamics Simulations. *J. Phys. Chem. B* **2015**, *119* (3), 1114−1128.
(53) Pace, C. N.; Grimsley, G. R.; Thomson, J. A.; Barnett, B. J. Conformational Stability and Activity of Ribonuclease T1 with Zero, One, and Two Intact Disulfide Bonds. *J. Biol. Chem.* **1988**, *263* (24), 11820−11825.

DOI: 10.1021/acs.jctc.6b00837
J. Chem. Theory Comput. 2017, 13, 29−41