CONFORMATIONAL ENTROPY OF BIOMOLECULES: BEYOND THE QUASI-HARMONIC APPROXIMATION

JORGE NUMATA[1]
numata@chemie.fu-berlin.de

MICHAEL WAN[1, 2]
mwan@fas.harvard.edu

ERNST-WALTER KNAPP[1]
knapp@chemie.fu-berlin.de

[1]*Macromolecular Modeling Group, Dept. of Chemistry and Biochemistry, Freie Universitaet Berlin, Takustr. 6, Berlin 14195 Germany*
[2] *Aspuru-Guzik Research Group, Harvard University, Dept. of Chemistry and Chemical Biology, 12 Oxford Street, Cambridge, MA 02138, USA*

A method is presented to calculate thermodynamic conformational entropy of a biomolecule from molecular dynamics simulation. Principal component analysis (the quasi-harmonic approximation) provides the first decomposition of the correlations in particle motion. Entropy is calculated analytically as a sum of independent quantum harmonic oscillators. The largest classical eigenvalues tend to be more anharmonic and show statistical dependence beyond correlation. Their entropy is corrected using a numerical method from information theory: the k-nearest neighbor algorithm. The method calculates a tighter upper limit to entropy than the quasi-harmonic approximation and is likewise applicable to large solutes, such as peptides and proteins. Together with an estimate of solute enthalpy and solvent free energy from methods such as MMPB/SA, it can be used to calculate the free energy of protein folding as well as receptor-ligand binding constants.

Keywords: conformational entropy; vibrational entropy; principal component analysis; quasi-harmonic approximation; anharmonicity; k-nearest neighbor entropy.

1. Introduction: thermodynamics of biological macromolecules

1.1. *Entropy of protein folding and ligand binding*

Protein folding and receptor-ligand binding occur in a spontaneous and specific way because the folded and bound states have a lower free energy than their unfolded and unbound counterparts, respectively. The Helmholtz free energy change ΔF or the Gibbs free energy change $\Delta G=\Delta F+P\Delta V$ predict the equilibrium constant (K_{eq}) for folding and binding. For incompressible fluids like water, the volume term $P\Delta V$ is negligible:

$$\Delta G \approx \Delta F = \Delta U - T\Delta S = -k_B T \ln K_{eq} \tag{1}$$

The net enthalpic (ΔH) and entropic ($T\Delta S$) contributions from all particles (solute and solvent) almost cancel out in natural or properly engineered proteins [1]. Stability against unfolding is typically around ΔG= 5 to 15 kcal/mol (K_{eq}= 10^{-4} to 10^{-11}). Upon folding, the solute becomes more rigid and loses conformational entropy. This unfavorable contribution is typically $T\Delta S_{\text{conformational}}$= 10 to 100 kcal/mol. Any estimation of free energy lacking this contribution will grossly overestimate the stability against unfolding.

A physically realistic combination of models to estimate the contributions to free energy can be found in the MM/PBSA method [2,3]:

$$\Delta G = \Delta U^{solute} + \Delta G^{solvent}_{electrostatic} + \Delta G^{solvent}_{hydrophobic} + \Delta U^{solu-solv}_{vdW} - T\Delta S^{solute}_{conformational} \quad (2)$$

The solute internal energy is evaluated by the molecular mechanics (MM) force field. Electrostatic solvation free energy can be obtained from the Poisson-Boltzmann (PB) equation, or its approximations. Hydrophobic free energy may be estimated as proportional to the solvent accessible surface (SA). An often neglected, but important term is the change in solute-solvent van der Waals interactions. Finally, the **solute conformational entropy** can be estimated with the method presented here.

Kollman, Case et al. say in a seminal article on MMPB/SA that "it would clearly be of interest to have better ways to estimate solute entropy, but there can be difficulties when the dynamics at room-temperature jump between basins." [3] The present method can deal with the multimodal probability distributions resulting from those jumps.

1.2. *Entropy quantifies conformational freedom*

Entropy is both a measure of disorder and of correlation. More formally, for a set of particles, entropy is a measure of the phase-space accessibility. Maximum entropy and accessibility in phase space for the non-interacting particles of an ideal gas are reached in a state of complete and random occupation of the container volume. For interacting particles, like the atoms in an organic molecule, entropy is still a measure of disorder (spread of the coordinates). But now the interactions between the atoms have to be included in the calculation by estimating the correlation between the atomic displacements. If we just add the entropy due to the conformational freedom of individual atoms, severe overestimation will occur. For macromolecules, correlation and dependence manifest themselves as concerted, delocalized motions spanning many atoms. [4] This phenomenon is incorporated in the form of Shannon information redundancy to provide a better estimate of entropy.

1.3. *Thermodynamics and information theory*

The information entropy of Claude Shannon [5] and the thermodynamic entropy of Rudolf Clausius [6] have the same functional form. This similarity alone makes it plausible that these quantities are the same. The formula for entropy (Eq. 3) contains a constant k and the natural or Naperian logarithm of the probability of the microstates of a molecular system. When dealing with statistics of coin tosses or genes, it is simpler to choose k=1 and other bases for the logarithm. As a fossil of the parallel historical development, the symbol for entropy is different in the Shannon context of communications theory (*H*) and thermodynamics (*S*). We will use *S*:

$$S = -k_B \sum_{i=1}^{t} p_i \ln p_i \quad (3)$$

The temptation is great to equate thermodynamic and information theoretical entropy. This temptation has been resisted in the present work, and each kind of entropy (quantum, classical and statistical) treated in its own mathematical framework.

1.4. *Novelty of the method*

The established protocol for estimating conformational entropy from molecular dynamics trajectories, the quasi-harmonic approximation [7, 8], provides a strict upper limit but is known to overestimate it considerably [9]. The method presented here takes the quasi-harmonic entropy and corrects it using a numerical method from information theory, the k-nearest neighbor entropy algorithm [10, 11]. The result is a method to estimate the entropy of solute molecules which accounts for anharmonicity (beyond Gaussians) and supralinear motion correlation (beyond covariance). This is achieved:

- Without omission of the mass-metric tensor [12, 13] to avoid deforming phase space. This is in contrast to other applications of numerical methods to calculate molecular entropy using dihedral angles, which will not yield the proper thermodynamic entropy without consideration of the Jacobian of the transformation from Cartesian to internal coordinates.
- Realizing that classical force-fields are fitted on quantum mechanical data. Quasi-harmonic frequencies higher than $k_B T / \hbar$ in the simulation correspond to quantum mechanical behavior, whose entropy should be calculated using all accessible states of a harmonic oscillator, and not just the ground state. Quantum entropy has the additional advantage of yielding an absolute value.
- By making use of an unbiased numerical method (the k-nearest neighbor entropy), which can give more precise results than simple binning of the trajectory.

2. Method for calculating conformational entropy of biomolecules

2.1. *Calculate principal components from covariance of a molecular dynamics trajectory*

Our starting point is a reasonably long and correctly set up molecular dynamics (MD) trajectory in the NVT ensemble. All particles except the solute atoms are deleted. Center-of-mass rotation and translation should be removed to calculate only the conformational entropy. If relevant, the rotational and translational components of entropy may be added later as ideal-gas S_{rot} and S_{trans} (See Chap 10 Ref. [14], Chap. 5 Ref. [15]).

The processed MD coordinate trajectory matrix **X** has dimensions $(3N_a,n_f)$ where N_a is the number of atoms and n_f the number of simulation frames selected for analysis.

$$X = (x_1,\ x_2, x_3,\ ..\ ,\ x_{nf}) \tag{4}$$

We wish to apply principal component analysis (PCA) on this trajectory. It is desirable to calculate the principal components of variance directly in Cartesian space,

without the matrix being singular [8]. Correct application of the mass-metric tensor is essential to diagonalization and uncoupling of the Hamiltonian in Cartesian coordinates. We thus apply mass weighting to each frame:

$$y_i = M^{1/2} x_i \qquad \text{for i=1,..,n}_f \tag{5}$$

Where **M** is a ($3N_a$,$3N_a$) matrix with the atomic masses repeated three times in the diagonal elements m_{ii} and $m_{ij}=0$ for $i \neq j$. We thus obtain the ($3N_a$,n_f) mass weighted trajectory matrix **Y**. To apply PCA on **Y**, we calculate its covariance matrix. The individual elements of the ($3N_a$,$3N_a$) covariance matrix $\boldsymbol{\sigma}^2_{m.w.}$ are:

$$(\sigma^2_{m.w.})_{i,j} = \left\langle (y_i - < y_i >)(y_j - < y_j >) \right\rangle \tag{6}$$

with <...> denoting average across the considered simulation frames or trajectory. $<y_i>$ is the precalculated average coordinate. Now $\boldsymbol{\sigma}^2_{m.w.}$ is diagonalized by an orthogonal transformation. After diagonalization, we obtain a new PCA coordinate system matrix with $3N_a$-6 eigenvalues or variances:

$$\lambda_{ii} = (m_{eff} \sigma^2_{PCA})_i \tag{7}$$

m_{eff} means that the PCA masses are combinations of the original ones. Of the total of $3N_a$ eigenvalues, 6 will have very high frequencies and may be discarded if translation and rotation were properly removed.

Also obtained are eigenvectors describing the PCA modes. Each PCA mode is a linear combination of the original coordinates. These combinations come from the weightings in the eigenvector matrix **W** of size (3Na,3Na-6).

PCA assumes that the particles have a Gaussian (normal) distribution with variance λ_{ii} in each mode i. PCA on the mass weighted MD-trajectory is also the called **quasi-harmonic approximation** because it implies fitting effective harmonic potentials on the observed coordinate covariance, smoothing over any anharmonicity. We may thus regard the method as producing the eigenvalues corresponding to a series of uncorrelated simple harmonic oscillators. We will later make use of this to calculate thermodynamic variables. Within the harmonic oscillator model we may connect the eigenvalues $\lambda_{ii}=m_{eff}\sigma^2_{PCA}$ to frequency ω through the **equipartition theorem.** Kinetic and potential energy are equal in the time average [15]:

$$\omega_i^2 \left(m_{eff} \sigma^2_{PCA} \right)_i = k_B T \tag{8}$$

PCA frequencies:

$$\omega_i = \sqrt{\frac{k_B T}{m_{eff} \sigma^2_{PCA}}} = \sqrt{\frac{k_B T}{\lambda_{ii}}} \tag{9}$$

The equipartition theorem is only valid in the classical limit $\omega << k_BT/\hbar$. Fortunately, high frequency quantum vibrations, where this approximation breaks down contribute less to molecular entropy. They also tend to be very close to Gaussian, thus not requiring corrections for anharmonicity and supralinear dependence beyond linear correlation.

We may now sort the $3N_a$-6 PCA frequencies. Large eigenvalues correspond to low frequency correlated motion. The lowest frequency modes are the most interesting collective motions, deemed important for enzymatic catalysis. [4] They are also the highest contributors to molecular conformational entropy. Still, it is not advisable to throw away higher frequency modes, as they may together make a sizeable contribution to entropy.

Using all the eigenvectors **W**, we may project the mass-weighted coordinates into the PCA collective coordinates **Z**:

All PCA modes: $$Z_{all} = W^T Y \tag{10}$$

Alternatively, we may select a subset of d eigenvectors, thus using a reduced eigenvector matrix of size (d,n_f):

Subset of d PCA modes: $$Z_d = W_d^T Y \tag{11}$$

2.2 Absolute entropy of a harmonic oscillator as an upper limit

Our objective is to provide an upper limit estimate of entropy. Assuming a Gaussian distribution is the safest bet because it has the highest entropy among all statistical distributions with the same given variance [16]. A statistical mechanical model that produces Gaussian-distributed coordinate displacements is the harmonic oscillator. For the quantum mechanical harmonic oscillator (including zero-point energy), the partition function is:

$$Q = \sum_{j=0}^{\infty} e^{-\alpha(j+1/2)} = e^{-\alpha/2} \sum_{j=0}^{\infty} \left(e^{-\alpha}\right)^j \quad \text{with} \quad \alpha = \beta\hbar\omega = \hbar\omega/(k_B T) \tag{12}$$

From it, we may directly calculate thermodynamic functions [14, 15] in the NVT ensemble, such as the Helmholtz free energy F and the internal energy U. Our main interest, however, is the quasi-harmonic entropy:

$$S_{Quantum,i} = \frac{F-U}{T} = k_B \frac{\alpha_i}{e^{\alpha_i}-1} - k_B \ln\left(1-e^{-\alpha_i}\right) \tag{13}$$

We now wish to use this result to calculate the entropy of all PCA modes. For many coordinates, we make a multivariate generalization of the formula. In other words, we create an approximation to molecular conformational entropy as a series of uncorrelated simple harmonic oscillators (SHO). Each SHO has frequency ω_i estimated from the classical mass weighted variance, and gives a total entropy [7]:

$$S_{Quantum} = k_B \sum_{i=1}^{3N-6} \left[\frac{\alpha_i}{e^{\alpha_i} - 1} - \ln(1 - e^{-\alpha_i}) \right] \tag{14}$$

with $\alpha_i = \frac{\hbar \omega_i}{k_B T} = \frac{\hbar}{\sqrt{k_B T}} \frac{1}{\sqrt{\lambda_{ii}}}$ in terms of the m.w. PCA eigenvalues λ_{ii}

This absolute entropy excludes the kinetic contribution (which is independent of conformation and may be added analytically from the equilibrated momenta [12]) and assumes distinguishable particles (Boltzmann statistics, as usual for systems of covalently bound atoms). Furthermore, zero-point energy does not add to the entropy because its contribution $\hbar\omega/2$ in F and U mathematically cancels out in Eq. 13.

The equipartition theorem (Eq. 8) is only strictly valid in the classical limit $\hbar\omega$ << $k_B T$. The limiting frequency corresponds to $k_B T/\hbar$ = 4.06E13 Hz or f<< 216 cm^{-1} at T=310K.

Thermodynamic entropy of a classical harmonic oscillator

Using the classical partition function Q=1/α, we may calculate the classical thermodynamic entropy of a harmonic oscillator. This will become important to compute a correction to the entropy values within the classical region:

$$S_{Classical,thermo,i} = -k_B \left[\ln(\alpha_i) + 1 \right] \tag{15}$$

Classical entropy is a lower limit for quantum entropy at all points. At values α<0.6, quantum and classical entropy agree better than 99%. At α=1, the classical entropy still agrees to 96% with the quantum one. At α>1 it incorrectly diverges towards -∞ (See Fig. 1). Classical entropy calculations should be avoided in the quantum region, which is nonetheless relevant for molecular motion.

Statistical entropy of a Gaussian distribution

The classical statistical entropy for each eigenvalue can be calculated from the definition of differential entropy and the Gaussian probability density function. With the eigenvalues (variances) in terms of our dimensionless α:

$$S_{Class,Gaussian,i} = -k_B \left[\ln(\alpha_i) + \ln\left(\sqrt{2\pi e}\right) \right] \tag{16}$$

Comparing Eq. 16 to Eq. 15, one realizes that the additive constant is different for thermodynamic and statistical entropy. The important point is that Eq. 16 should be used to calculate the correction, because the k-nearest neighbors statistical method used to calculate non-Gaussian entropy in Sec. 2.4 produces comparable values.

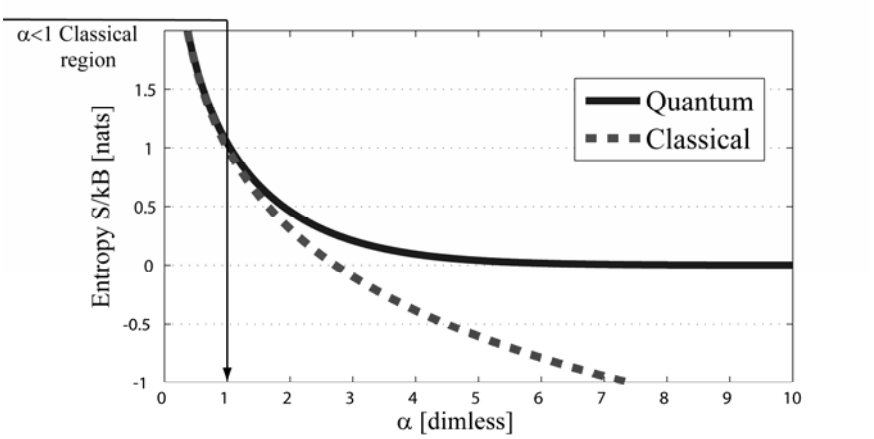

Fig. 1: Plot of thermodynamic entropy as a function of $\alpha = \hbar\,\omega/(k_B T)$, where ω is the harmonic oscillator frequency. Shown are the quantum (Eq. 14) and classical (Eq. 15) results. For large frequencies, quantum entropy displays the expected limiting behavior, such that $S_{Quantum} \rightarrow 0$. Classical entropy diverges $S_{Classical} \rightarrow -\infty$ as ω grows. The classical approx. breaks down in the quantum region because the spacings between energy levels become too small, and higher energy levels become significantly occupied.

2.3 Pairwise non-Gaussian corrections for anharmonicity and supralinear dependence

We now wish to provide a tighter upper limit to entropy than the quasi-harmonic approximation can afford. PCA decomposes the linear (Gaussian) correlation. If histograms are made of the principal components **Z**, most will be actual Gaussians. PCA modes with the largest classical eigenvalues are major contributors to entropy. They also tend to deviate from the Gaussian distribution.

An important property of Gaussians is that they have the largest entropy possible for a given variance [17]. This means that each principal component that displays **anharmonicity** will overestimate the entropy [9]. Furthermore, deviations from the Gaussian distribution also imply that higher order dependence in the trajectory data beyond linear correlation was missed. This supralinear dependence is quantified by **mutual information** (**M.I.**). Missing higher order correlations causes a more severe overestimation of entropy than anharmonicity [12]. For instance, a kind of dependence that is systematically missed by PCA is that of 90° phase shifted atoms moving in parallel lines [18]. This is captured by the correction for the classical region.

We thus take the calculated absolute quantum entropy and correct it with numerical methods from information theory. This correction is strictly non-positive, because: a) a Gaussian has the maximum entropy for a given variance and b) two or more orthogonal Gaussians (after PCA) have zero Mutual Information. For a total of c modes in the classical regime, a **correction scheme for pairs of modes** (i,j) is proposed (see Fig. 2).

$$\Delta S_{Class,pairwise,correction} = -\left[\sum_{i=1}^{c} S_{anh,(i)} + \sum_{i=1}^{c}\sum_{j=i+1}^{c} M.I._{(i,j)}\right] \quad (17)$$

Anharmonic correction: Let $S_{non\text{-}Gauss,(i)}$ be the marginal (1-dimensional) entropy estimated by a non-Gaussian numerical method for PCA mode i:

$$S_{anh,(i)} = S_{Class,Gaussian,(i)} - S_{non-Gauss,(i)} \quad (18)$$

Supralinear dependence correction: Let $S_{non\text{-}Gauss(i,j)}$ be the joint (2-dimensional) entropy estimated by a non-Gaussian numerical method for PCA modes (i,j). The supralinear dependence correction is given by the mutual information (M.I.) [19]:

$$M.I._{(i,j)} = S_{non-Gauss,(i)} + S_{non-Gauss,(j)} - S_{non-Gauss,(i,j)} \quad (19)$$

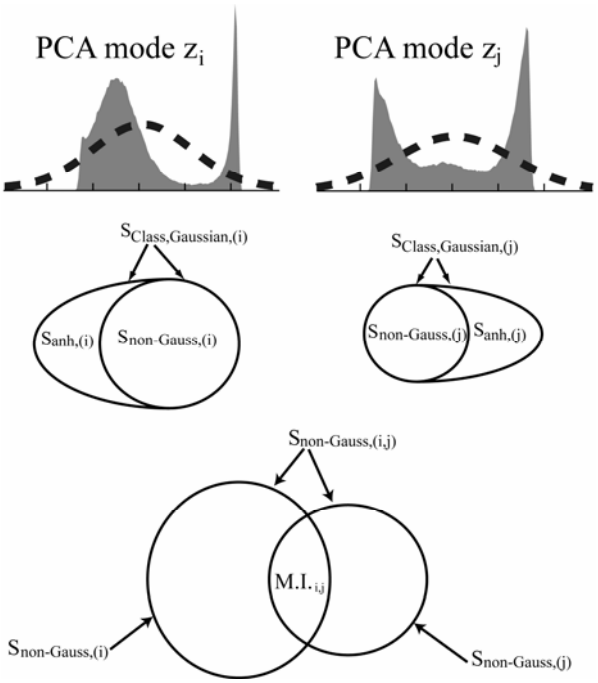

Fig. 2 The pairwise correction is the sum of anharmonicities and supralinear dependence (M.I.). Shown are two strongly anharmonic modes and their Gaussian fits. In the Venn diagrams, entropy is proportional to the area.

Because of numerical inaccuracies, S_{anh} and M.I. can be slightly negative. In the present implementation, a mode is taken as harmonic if $S_{anh}/S_{Quantum}<0.007$. Any M.I.<0 is taken as zero. If a large and negative M.I. is calculated for certain mode pairs, it is likely due to the frequency of data sampling. Both too low and too high resolution cause inaccuracies. For best results, use saved coordinates between every 10fs to 500fs.

In the next section, a practical method called the k-nearest neighbor algorithm ($S_{kNN,k}$) is used to calculate $S_{non\text{-}Gauss}$ for arbitrary dimensions. To guarantee numerical stability, the current implementation is restricted to pairwise corrections (d=1 and d=2).

2.4 Estimation of non-Gaussian entropy

The assumption of a Gaussian distribution is not necessary to estimate entropy. Indeed, one does not need to assume a particular distribution at all. Non-parametric methods avoid forcing a functional form for the distribution. One such method calculates the entropy from the k-nearest neighbor (kNN) points in the sample [10].

Mathematical foundation of k-nearest neighbor entropy

Let matrix $\mathbf{Z}_d$ of dimension (d,n_f) be a random sample of n_f observations of a d-dimensional random vector. In this case, the random vectors are a subset of d square-root-of-mass weighted principal components of the original MD trajectory of a solute molecule for n_f simulation frames:

$$Z_d = (z_1,\ z_2, z_3,\ ..\ ,\ z_{nf}) \tag{20}$$

Now we use a non-parametric estimate of the probability density around each observed PCA conformation. The conformational density for each frame of the simulation is assessed through its distance to the k most similar (nearest neighbor) frames. For the true underlying probability distribution function $f_{c\text{-}true}$ that produced the data, a non-parametric estimate for each d-dimensional sample vector z_i is given by f_c [20]:

$$f_c(z_i) = \frac{k}{n}\frac{1}{V_d\left(R_{i,k}\right)} \quad \text{with} \quad V_d\left(R_{i,k}\right) = \frac{\pi^{d/2}R_{i,k}^d}{\Gamma\left(\frac{d}{2}+1\right)} \tag{21}$$

k is defined below as the nearest-neighbor index and should not to be confused with k_B. V_d is the volume of a d-dimensional hyper-sphere with radius $R_{i,k}$:

$$R_{i,k} = \left\| z_i - z_{i,k} \right\|_2 \tag{22}$$

$R_{i,k}$ is the Euclidean distance between sample point $\mathbf{z}_i$ and its k-th nearest neighbor $\mathbf{z}_{i,k}$. The entropy of the probability distribution may now be estimated in an asymptotically unbiased form. The k-nearest neighbors entropy, following Hnizdo et al is [10, 11]:

$$\frac{S_{kNN,k}}{k_B} = \frac{d}{n}\sum_{i=1}^{n}\ln R_{i,k} + \ln\frac{n\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)} - L_{k-1} + \gamma \tag{23}$$

with $L_0 = 0$; $L_m = \sum_{i=1}^{m} \frac{1}{i}$ for $m \geq 1$; $\gamma = 0.5772157...$ (Euler-Mascheroni constant)

The first term of Eq. 23 may be understood intuitively as an estimate of entropy. The three last terms are an unbiasing correction. They arise to eliminate a known asymptotic bias in this estimator, as detailed by Hnizdo, Demchuk et al [10]. It is an extension to k-neighbors of a similar estimator proposed before by Kozachenko and Leonenko for k=1 [21]. Using higher neighbors, the method becomes less sensitive to the accuracy of the data. In practical tests by others [10] and our own with multidimensional random points shaped by statistical distributions of known entropy, the $k=4^{th}$ nearest neighbor has demonstrated to work well.

The k-nearest neighbors in a d-dimensional sample of n_f-points may be found using the ANN program, a C++ library by David Mount and Sunil Arya [22]. Specifically, the k-d tree nearest neighbors method implemented into ANN was used.

The kNN algorithm produces a relative, statistical and more importantly, a classical entropy. It is thus well suited to compute relative corrections in the classical region.

3. Practical application to small biomolecules

The main focus of the presented method is the calculation of conformational entropy of proteins and other macromolecules. As a first test, however, the combination of methods was applied to smaller molecules; glycine and alanine. Molecular dynamics simulations were done for 0.9 ns (after heating and equilibration for 0.170 ns). The C- and N-termini were charged, but the overall charge is zero. An explicit water box of $16Å^3$ was used together with the Particle Mesh Ewald electrostatic method. The time steps were of 1fs, with freely moving hydrogens except for TIP3P waters (fixed with SETTLE). The Charmm22 force field was used with programs Charmm31a1 and NAMD 2.6b1. For analysis, solvent atoms were deleted. Translation and rotation were removed. Samples were taken every 100 fs for the entropy calculation, so that n_f=9000.

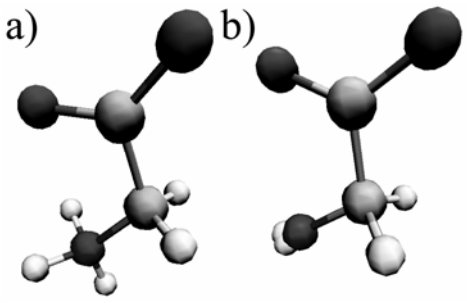

Fig. 3: Glycine molecule in a) its initial conformation and b) its average conformation during a 0.9 ns MD simulation. Notice that the amino group rotates freely, and this is manifested as an average close to the central nitrogen. It is not cause for worry that the average structure looks chemically unsound. In the words of Jaynes, “it is possible to make a sharp distinction in statistical mechanics: the physical and the statistical.” [17] Our physical description is built into the MD simulation. Our statistical analysis happens in “shape and size” space, treating the conformational density with information theory.

The histograms shown in Fig. 4 represent each mass-weighted coordinate in the glycine trajectory. Fig. 4a presents the marginal distributions of the original trajectory. We may already calculate an entropy from the variance of these coordinates. Summing

the marginal Gaussian entropies is tantamount to ignoring both anharmonicity and any correlations. This provides the highest upper limit to entropy (see Table 1). The severe overestimation speaks against proposals for additive entropy contributions, such as an "entropy per side chain" or "per chemical group".

After PCA is performed (Fig. 4b), the entropy may be calculated from the eigenvalues of the diagonalized mass-weighted covariance matrix. In total we obtain (3N-6)=24 eigenvalues for glycine. Only 3 collective coordinates are in the classical regime (bottom right in Fig. 4b), but together they contribute with 70% of the total entropy. The corresponding estimate can be seen in Table 1 as PCA, also known as the quasi-harmonic approximation [7]. It is lower because of the decomposition of linear correlation.

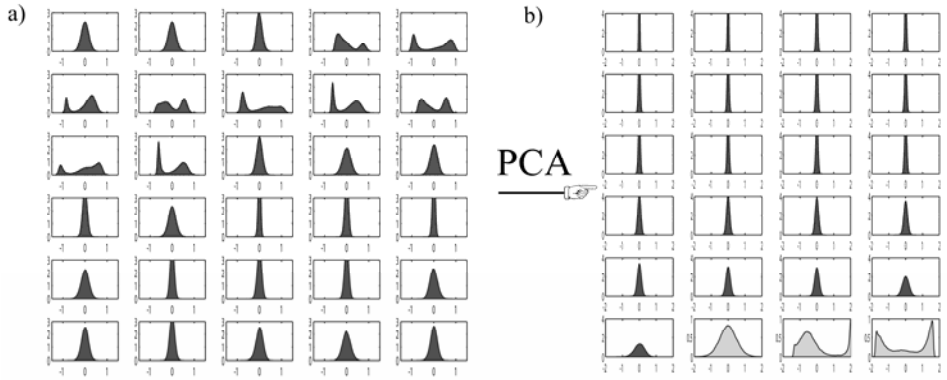

Fig. 4 Histogram from the molecular dynamics trajectory of glycine. a) 30 mass-weighted coordinates before principal component analysis (PCA). Notice how the coordinates 4 to 12 seriously deviate from Gaussians. These correspond to x,y,z fluctuations of the freely rotating hydrogens of the amino group. b) PCA groups motion types by decomposing the linear correlation. Shown are the 24 collective coordinates excluding translation and rotation. Note that although PCA implicitly assumes Gaussian collective coordinates, the actual projections may deviate. This is the case for the last 3 PCA coordinates (light gray), which are also in the classical region. For a) and b): The average of each coordinate has been subtracted for display purposes. Units: Horizontal axis in [a.m.u.^{½}Å] and vertical axis in histogram probability.

Table 1: Absolute molar entropy of two free-form amino acids. All estimates per column are based on the same 0.9 ns trajectory, except for the vacuum $S_{NormalModes}$.

Estimate type	Glycine $S_{conformational}$ cal/(mol K)	Alanine $S_{conformational}$ cal/(mol K)	Comments
Marginal	34.14	55.09	Sum of quantum entropy from variances of the marginal distributions of the original m.w. coordinates.
PCA	15.41	24.44	M.w. PCA, also called the Quasi-Harmonic Approximation (QHA). [7]
PCA-Anh	13.02	22.34	PCA with anharmonic correction.
PCA-Anh-M.I.	**9.44**	**14.88**	**Estimation from present method. PCA with anharmonic and pairwise supralinear (M.I.) corrections.**
Normal Modes	8.56	13.47	NMA After thorough minimization in vacuum. Lower limit benchmark.

In principle, a lower entropy estimate is better. The PCA value was corrected for anharmonicity (PCA-Anh), and then subsequently for anharmonicity and supralinear dependence (PCA-Anh-M.I.). Consistent with a recent study [12], the influence of supralinear dependence (M.I.) was much larger than that of anharmonicity.

For such small molecules, normal mode analysis is still a viable option and can be seen as a benchmark. It is thus encouraging that the corrected values are similar to it. It is not far-fetched to say that the corrected estimates are better gauges of entropy than normal modes, as they account for the entropy of the multimodal rotation of the amino and methyl groups as well as the influence of solvent particles. **For peptides and larger biomolecules, normal mode analysis of single conformations is definitely not appropriate**, as it cannot capture more complex potential energy surfaces. An estimate for the entropy of a protein as a whole can be obtained using the method presented here.

4. Conclusion

Estimation of absolute conformational entropy is important because it allows a detailed understanding of the thermodynamic driving forces at the molecular level [16]. While the quasi-harmonic approximation is known to systematically overestimate entropy [9], it was until recently the only option to calculate absolute entropy from MD simulations. A recent study [12] demonstrated that this overestimation is larger for the folded than for the unfolded state of peptides. Error cancellation is not exhibited because cooperativity and higher order motion correlations are more important in the compact, folded state than in an open, denatured state.

The current method provides a tighter upper limit to absolute entropy than the quasi-harmonic approximation. As a result, the method stays loyal to the E.T. Jaynes' spirit of maximum entropy as a "method of reasoning which ensures that no unconscious arbitrary assumptions have been introduced" [17].

Future directions include extending the method to more than pairwise supralinear dependencies. Also interesting is the combination with the permutation reduction method to calculate the entropy of diffusive systems (solvation shells and solvent entropy) [23].

A program to calculate the conformational entropy of peptides, proteins or nucleotides from Charmm and NAMD trajectories according to this algorithm will be made available on http://userpage.chemie.fu-berlin.de/~numata

Acknowledgments

This work was supported by the International Research Training Group "Genomics and Systems Biology of Molecular Networks" (GRK1360 of the DFG) and an internship grant for American students from the DAAD RISE program and German SFB498.

References

[1] Loladze, V.V. , Ermolenko, D.N., and Makhatadze, G.I., Thermodynamic consequences of burial of polar and non-polar amino acid residues in the protein interior, *J. Mol. Biol.*, 320(2):343-357, 2002.

[2] Zoete, V., Meuwly, M., and Karplus, M., Study of the insulin dimerization: Binding free energy calculations and per-residue free energy decomposition, *Proteins.*, 61(1):79-93, 2005.

[3] Srinivasan, J., Cheatham III, T.E., Cieplak, P., Kollman, P.A., and Case, D.A., Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices, *J. Am. Chem. Soc.*, 120(37):9401-9409, 1998.

[4] Hammes-Schiffer, S. and Benkovic, S.J., Relating protein motion to catalysis, *Annu. Rev. Biochem.*, 75:519-541, 2006.

[5] Shannon, C.E. and Weaver, W., A mathematical theory of communication, *Bell Syst. Tech. J.*, 1948.

[6] Clausius, R., Über verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie, *Ann. der Physik*, 201:353-400, 1865.

[7] Andricioaei, I. and Karplus, M., On the calculation of entropy from covariance matrices of the atomic fluctuations, *J. Chem. Phys.*, 115:6289-6292, 2001.

[8] Schlitter. J., Estimation of absolute and relative entropies of macromolecules using the covariance matrix, *Chem. Phys. Lett.*, 215:617-621, 1993.

[9] Chang, C-E., Chen, W., and Gilson, M.K., Evaluating the accuracy of the quasiharmonic approximation, *J. Chem. Theory Compu.*, 1 (5):1017-1028, 2005.

[10] Hnizdo, V., Singh, H., Misra, N., Fedorowicz, A., and Demchuk, E., Nearest neighbor estimates of entropy, *Am. J. Math. Manage. Sci.*, 23:301-321, 2003.

[11] Hnizdo, V., Darian, E., Federowicz, A., Demchuk, E., Li, S., and Singh, H., Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules, *J. Comput. Chem.*, 28(3):655-668, 2007.

[12] Baron, R., Biomolecular simulation: Calculation of entropy and free energy, polypeptide and carbopeptoid folding, simplification of the force field for lipid simulations (eth 16584). Zürich: ETH-Zürich; 2006.

[13] Knapp, E.W. and Hoffmann, D., Polypeptide folding with off-lattice monte carlo dynamics: The method, *Eur. Biophysics J.*, 24(6):387-403, 1996.

[14] Cramer, C.J., *Essentials of computational chemistry*, 2nd ed.; 2004.

[15] McQuarrie, D.A., *Statistical mechanics*, Harper & Row; 1973.

[16] Dill, K.A., Bromberg S: *Molecular driving forces* Garland Science; 2003.

[17] Jaynes, E.T., Information theory and statistical mechanics (part 1), *Phys. Rev.*, 106(4):620-630, 1957.

[18] Lange, O.F. and Grubmüller, H., Generalized correlation for biomolecular dynamics, *Proteins.*, 62(4):1053-1061, 2006.

[19] Cover, T.M. and Thomas, J.A, *Elements of information theory*; 1991.

[20] Loftsgaarden, D. and Quesenberry, C., A nonparametric estimate of a multivariate density function, *Ann. Math. Stat.*,1049-1051, 1965.

[21] Kozachenko, L. and Leonenko, N., Sample estimates of entropy of a random vector, *Problems of Information Transmission*, 23:95-101, 1987.

[22] Mount, D., Arya, S., Netanyahu, N., Silverman, R., and Wu, A., An optimal algorithm for approximate nearest neighbor searching fixed dimensions, *JACM*, 45(16):891-923, 1998.
[23] Reinhard, F. and Grubmüller, H., Estimation of absolute solvent and solvation shell entropies via permutation reduction, *J. Chem. Phys.*, 126(1):014102, 2007