# Free-Energy Landscape of RNA Hairpins Constructed via Dihedral Angle Principal Component Analysis

**Laura Riccardi, Phuong H. Nguyen, and Gerhard Stock*,†**

*Institute of Physical and Theoretical Chemistry, Goethe University, Max-von-Laue-Str. 7, D-60438 Frankfurt, Germany*

*Received: August 6, 2009; Revised Manuscript Received: October 13, 2009*

To systematically construct a low-dimensional free-energy landscape of RNA systems from a classical molecular dynamics simulation, various versions of the principal component analysis (PCA) are compared: the cPCA using the Cartesian coordinates of all atoms, the dPCA using the sine/cosine-transformed six backbone dihedral angles as well as the glycosidic torsional angle $\chi$ and the pseudorotational angle $P$, the aPCA which ignores the circularity of the $6 + 2$ dihedral angles of the RNA, and the dPCA$_{\eta\theta}$, which approximates the 6 backbone dihedral angles by 2 pseudotorsional angles $\eta$ and $\theta$. As representative examples, a 10-nucleotide UUCG hairpin and the 36-nucleotide segment SL1 of the $\Psi$ site of HIV-1 are studied by classical molecular dynamics simulation, using the Amber all-atom force field and explicit solvent. It is shown that the conformational heterogeneity of the RNA hairpins can only be resolved by an angular PCA such as the dPCA but not by the cPCA using Cartesian coordinates. Apart from possible artifacts due to the coupling of overall and internal motion, this is because the details of hydrogen bonding and stacking interactions but also of global structural rearrangements of the RNA are better discriminated by dihedral angles. In line with recent experiments, it is found that the free energy landscape of RNA hairpins is quite rugged and contains various metastable conformational states which may serve as an intermediate for unfolding.

## I. Introduction

Much of our recent understanding of biomolecular processes such as molecular recognition, folding, and aggregation has been promoted by the concept of the free energy landscape[1−5]

$$\Delta G(r) = -k_B T[\ln P(r) - \ln P_{max}] \quad (1)$$

Here $P$ is the probability distribution of the molecular system along some (in general multidimensional) coordinate $r$ and $P_{max}$ denotes its maximum, which is subtracted to ensure that $\Delta G = 0$ for the lowest free energy minimum. Popular choices for the coordinate $r$ include the fraction of native contacts, the radius of gyration, and the root-mean-square deviation of the molecule with respect to the native state. Characterized by its minima (which represent the metastable conformational states of the systems) and its barriers (which connect these states), the energy landscape allows us to account for the pathways and their kinetics occurring in a biomolecular process.

Compared to the vast number of studies on peptides and proteins, there are relatively few experimental[6−11] and theoretical[12−18] works that consider the energy landscape of nucleic acids. This situation is likely to change, however, due to the discovery that RNA may fulfill a considerable diversity of biological tasks, including, apart from its encoding and translational activity, also enzymatic and regulatory functions.[19] Despite the limited number of ribonucleotide residues, RNA molecules are capable to fold into a wide variety of secondary and tertiary structures.[20,21] Besides structural characteristics, also dynamic properties appear to play an important role in maintaining the functional diversity of RNA.[22,23]

The goal of this paper is to systematically construct a low-dimensional free-energy landscape of RNA from a classical molecular dynamics (MD) simulation. To this end, principal component analysis (PCA), also called quasiharmonic analysis or essential dynamics method, has been found valuable.[24−28] By diagonalizing the covariance matrix of some system variables, it has been shown that a large part of the system's fluctuations can be described in terms of only a few PCA eigenvectors. Apart from the common usage of Cartesian coordinates, it has recently been shown that a PCA based on the backbone dihedral angles, referred to as dPCA, may be advantageous.[29] To avoid problems arising from the circularity of these variables, a transformation from the space of dihedral angles $\{\varphi_n\}$ to a linear metric coordinate space (i.e., a vector space with the usual Euclidean distance) was built up by the trigonometric functions sin $\varphi_n$ and cos $\varphi_n$. The dPCA is appealing, because other internal coordinates such as bond lengths and bond angles usually do not undergo changes of large amplitudes. Hence the analysis already starts with the relevant part of the dynamics, thus avoiding unnecessary noise. Moreover, problems associated with the mixing of internal and overall motion are naturally avoided by using internal coordinates.[30] Recently, the theoretical foundations of the dPCA was established,[31] the method was implemented in the GROMACS MD program,[32] and various application have been considered.[33−36]

On the basis of the dPCA, a systematic characterization of the free energy landscape can be performed as follows:[33] (i) By requiring that the energy landscape reproduces the correct number, energy, and location of the system's metastable states and barriers, the dimensionality of the free energy landscape (i.e., the number of essential components) is obtained. This dimensionality can be determined from the components' distribution (number of non-Gaussian components) and the autocorrelation (number of "slow" components), resulting typically

* To whom correspondence should be addressed. E-mail: stock@physik.uni-freiburg.de.
† Present address: Institute of Physics, Albert Ludwigs University, Hermann-Herder-Strasse 3, 79104 Freiburg, Germany.

10.1021/jp9076036 
Published on Web 11/30/2009

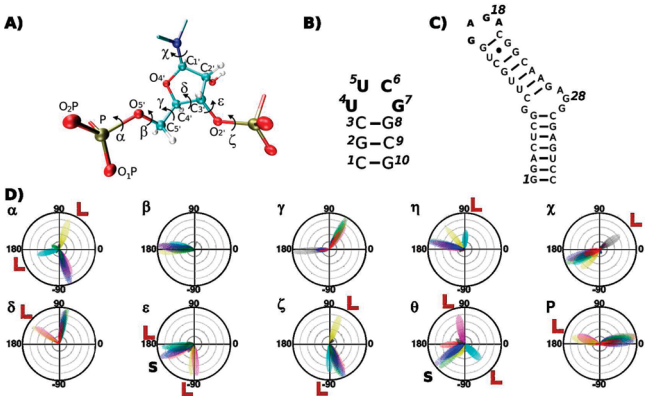


**Figure 1.** (A) Definition of the dihedral angles in RNA, following the IUPAC nomenclature.[54] Secondary structures of (B) the UUCG tetraloop and (C) the SL1m segment of HIV-1. (D) Distribution of RNA dihedral angles as obtained from a MD simulation of the UUCG hairpin. Apart from the backbone torsional angles α, $\beta$, $\gamma$, $\delta$, $\epsilon$, and $\zeta$, and the glycosidic torsional angle $\chi$, also the pseudorotational phase angle $P$ and the pseudotorsional backbone angles $\eta$ and $\theta$ (definition see text) are considered. The polar diagrams are generated by distributing the angular data into 360 bins on the circle.[55] The data are normalized such that all distributions show the same value at the main maximum. Different colors represent different residues of the hairpin, "L" and "S" label loop and stem residues, respectively. The absence of a label means that the conformational state is visited by both loop and stem residues.

in 3−10 essential components.[37] (ii) By use of this low-dimensional space, one may employ a clustering algorithm such as *k*-means[38] that facilitates the identification of the system's metastable states and connecting barriers. (iii) Finally, the conformational dynamics of the low-dimensional model can be studied via a Markov state model[28,39−42] (if the dynamics in state space is Markovian), a Langevin simulation,[43−46] or a nonlinear dynamic model.[37,47−49]

In this work, we apply the dPCA to construct the free energy landscape of RNA molecules. While protein conformations can be well characterized through two backbone dihedral angles $\phi$ and $\psi$, RNA has six backbone torsional angles α, $\beta$, $\gamma$, $\delta$, $\epsilon$, and $\zeta$ (Figure 1A). Of interest are, moreover, the glycosidic torsional angle $\chi$ of the bond between the ribose ring and the base as well as the pseudorotational phase angle $P$[50] accounting for the conformation of the sugar.[51] Since these dihedral angles are commonly employed to characterize the structure of the RNA,[20,21,52] they also appear to be prime candidates for the coordinates used in a PCA. Furthermore, we also investigate the performance of various other variants of the PCA, including a PCA on the Cartesian coordinates of all atoms (cPCA), an approximate version of the dPCA which ignores the circularity of the dihedral angles (aPCA), and a PCA using the pseudotorsional angles $\eta$ and $\theta$.[53]

As a well-established example of conformational heterogeneity and dynamics, we first adopt a short RNA hairpin with a UUCG tetraloop (see Figure 1B), which has been the subject of various experimental[56,57] and molecular dynamics (MD) studies.[58−61] The hairpin motif is one of most common secondary structure elements in RNA and plays important roles in both RNA structure and function, e.g., hairpins are thought to provide nucleation sites for RNA folding[62] and tertiary recognition sites for both proteins and nucleic acids.[63] To warrant sufficient conformational sampling of the 10-mer hairpin under consideration, we employ replica-exchange MD (REMD) simulations,[64−66] using the Amber all-atom force field[67,68] and explicit solvent. As a second example, we consider the 36-nucleotide segment SL1 of the Ψ site of the human immunodeficiency virus type-1 (HIV-1), which is known to play an important role both during virus assembly and infectivity.[69] Besides the loop, the hairpin consists of two stems separated by a bulge (see Figure 1C). A 100-ns MD simulation is performed of a mutated form of this region, called SL1m, in which the GC-rich loop is substituted with a GAGA tetraloop to avoid dimerization.[69]

Interestingly, we find for both RNA hairpins that only an angular PCA such as the dPCA is able to resolve the conformational heterogeneity of these systems. In line with recent experiments,[6−11] we demonstrate that the free energy landscape of the RNA hairpins is quite rugged and contains various conformational states which may serve as an intermediate for unfolding. It is shown that these states are well characterized by specific hydrogen bonding and stacking interactions and can therefore, at least in principle, be discriminated by nuclear magnetic resonance (NMR), infrared, or fluorescence experiments.[6,8−11] Furthermore, we find that the dPCA is also sensitive to global structural rearrangements of the RNA which, for example, can be observed in electronic paramagnetic resonance (EPR) experiments that provide the distance distribution between various parts of the molecule.[70,71]

## II. Theory and Methods

**A. MD Simulations.** We used the AMBER force field (parm98)[67,68] to model the UUCG hairpin and the TIP3P model[72] to describe the solvent water. We note that recent MD simulations of a UUCG loop using the AMBER parametrizations parm94,[68] parm98,[67] and parm99[73] gave quite similar structures and virtually the same agreement with experimental data.[74] In all simulations, the GROMACS program suite[32] was employed. The 10-mer cgcUUCGgcg was placed in a cubic box containing 1770 water molecules and 9 $Na^+$ ions. The equation of motion was integrated by using a leapfrog algorithm with a time step of 2 fs. Covalent bond lengths were constrained by the procedure SHAKE[75] with a relative geometric tolerance of 0.0001. We employed the particle-mesh Ewald method to treat the long-range electrostatic interactions.[76] The nonbonded interaction pair-list were updated every 5 fs, using a cutoff of 1.2 nm.

Starting from an NMR structure,[56] the hairpin was minimized using the steepest descent method. Subsequently, the solvated system was equilibrated for 100 ps at constant pressure (1 atm) and temperature ($T$ = 350 K), respectively, using the Berendsen coupling method.[77] The systems were then equilibrated further at constant temperature ($T$ = 350 K) and constant volume for 10 ns. For the REMD simulations, we used 40 replicas covering a temperature range from 295 to 453 K to simulate each system. The temperature gap of 4−6 K between adjacent replicas ensures that the acceptance ration of the exchanges of 20−30%. By employing the last structures of the 10 ns trajectories as starting structures for all 40 replicas, each replica was run independently at its own temperature for 200 ps. Then, the exchange procedure between the replicas was turned on, using a time step of 1.5 ps between two attempts of exchange. This time interval should be large enough compared to the coupling time of the heat bath (0.1 ps), such that the trajectory is roughly equilibrated before the next exchange is attempted. Each replica was run for 10 ns, yielding a total sampling time of 400 ns. The data were collected every 0.1 ps for subsequent analysis. For the detailed descriptions of the REMD algorithm, see refs.[34,64−66]

The second system under consideration was the 36-nucleotide segment SL1m known as the Ψ site of HIV-1, which is essential for HIV-1 to selectively recognize and package its genomes. The engineered structure SL1m (PDB code: 1N8X) was determined by NMR experiment[69] and consists of two stems

separated by a bulge and ending in a tetraloop (see Figure 1C). This system was simulated using the AMBER force field parm99.[73] The molecule was placed in an octahedron box containing 10459 water molecules and 35 $Na^+$ ions. The whole system was minimized using the steepest descent method. Subsequently, the solvated system was equilibrated for 500 ps at constant pressure (1 atm) and temperature ($T = 320$ K), respectively, using the Berendsen coupling method.[77] The system was then run at constant temperature and volume for 100 ns and data were saved every 5 ps for analysis.

Analysis of the trajectories was performed with tools from the GROMACS package and with modified versions of them. To define the presence of a hydrogen bond, an acceptor−donor distance smaller than 0.35 nm and a donor-hydrogen-acceptor angle greater than 150 degree was requested. A stacking interaction between a pair of bases is defined by two criteria that need to be satisfied simultaneously. First, the distance between the bases (defined as the distance between the average positions of all base atoms except for the hydrogen atoms) has to be smaller than 0.45 nm. Second, the angle between the base planes (defined by atoms C2, C4, and C6) has to be less than 30° or more than 150°. Figures showing molecular structures were generated using the graphical package VMD.[78]

**B. PCA.** PCA is a well-established method to reduce the dimensionality of a high-dimensional data set.[79] Considering the dynamics of $M$ atoms, the basic idea is that the correlated internal motions are represented by the covariance matrix

$$\sigma_{ij} = \langle (q_i - \langle q_i \rangle)(q_j - \langle q_j \rangle) \rangle \tag{2}$$

where $q_1$, ..., $q_{3M}$ are the mass-weighted Cartesian coordinates of the molecule and $\langle \cdots \rangle$ denotes the average over all sampled conformations.[24−28] By diagonalizing the covariance matrix, we obtain $3M$ eigenvectors $\mathbf{v}^{(i)}$ and eigenvalues $\lambda_i$, which are rank ordered such that $\lambda_1$ represents the largest eigenvalue. The eigenvectors and eigenvalues of $\sigma$ yield the modes of collective motion and their amplitudes, respectively. The first few principal components $V_i = \mathbf{v}^{(i)}\mathbf{q}$ of the data $\mathbf{q} = (q_1, \cdots, q_{3M})^T$ can then be used, for example, to represent the free-energy surface (eq 1) of the system. As discussed previously,[33] we include only principal components with multimodal probability distribution, as their peaks correspond to distinct metastable conformational states, whereas unimodal probability distributions account for fluctuations rather than for conformational transitions.

The basic idea of the dPCA[29] is to perform the PCA on sin- and cos-transformed dihedral angles

$$\begin{aligned} q_{2n-1} &= \cos \varphi_n \\ q_{2n} &= \sin \varphi_n \end{aligned} \tag{3}$$

where $n = 1$, ..., $N$ and $N$ is the total number of peptide backbone and side-chain dihedral angles used in the analysis. Equation 3 represents a transformation from the space of dihedral angles $\{\varphi_n\}$ to a linear *metric* coordinate space (i.e., a vector space with the usual Euclidean distance), thus avoiding problems arising from the circularity of angular variables.[80] Although this procedure results in $2N$ variables $q_n$ for $N$ angles, it should be stressed that the transformation 3 amounts to a one-to-one representation of the original angle distribution.[31]

Four variants of the PCA were considered. To perform a dPCA of the UUCG hairpin, we used the six backbone dihedral angles $\alpha$, $\beta$, $\gamma$, $\delta$, $\epsilon$, and $\zeta$ (see Figure 1A) of the RNA as well as the glycosidic torsional angle $\chi$ and the pseudorotational angle $P$. [By ignoring the pseudorotational angle $P$ in the analysis, the resulting energy landscapes look quite similar (data not shown), due to high correlation of $P$ and $\delta$.] This yields in total 152 sin/cos variables $q_i$. About 9 of the 152 principal components exhibited a non-Gaussian distribution. These 9 components contained 69% of the total fluctuations of the system. To perform the PCA on the $3M$ ($M = 319$) mass-weighted Cartesian coordinates of the RNA (cPCA), we first removed the translational and rotational motion of the trajectory. Of the 957 components, only 5 showed a non-Gaussian distribution and yielded 75% of the fluctuations.

We also considered a simplified model of the RNA backbone suggested by Duarte and Pyle,[53] who introduced two pseudotorsional angles $\eta$ and $\theta$ which are obtained by imagining a connection between the P and $C_{4'}$ backbone atoms. That is, $\theta$ is defined as $P_{(n)}-C_{4'(n)}-P_{(n+1)}-C_{4'(n-1)}$ and $\eta$ is defined as $C_{4'(n-1)}-P_{(n)}-C_{4'(n)}-P_{(n+1)}$, where $n$ labels the reference nucleotide, and $n - 1$ and $n + 1$ the previous and the following nucleotide, respectively. Proceeding this way, the conformation of RNA can be described by only two backbone dihedral angles per residue, just as in the case of proteins. By performing a dPCA on the angles $\eta$ and $\theta$ (for the backbone) as well as on $\chi$ and $P$ (for the side chain), we obtain an approximate description of energy landscape referred to as $dPCA_{\eta\theta}$. In this case, 7 out of the 72 principal components exhibited a non-Gaussian distribution, yielding 68% of the fluctuations.

Finally, we have performed a PCA directly on the $6 + 2$ dihedral angles of the RNA, referred to as aPCA. Ignoring the circularity of the dihedral angles, the aPCA represents an approximation of the sin/cos dPCA described above. This is because given a range larger than 180°, standard statistical quantities such as mean or difference cannot generally be calculated as in the case of linear data. (For example, given two angles with values of 20 and 350°, their distance is 30° instead of 330° because the maximum distance between two circular points is half of the total range, i.e., 180°.) The angular average $\langle\varphi\rangle$ over $n$ values $\varphi_i$ can be calculated via[80]

$$\langle\varphi\rangle = \arctan S/C + \delta \tag{4}$$

where $S = \Sigma_{i=1}^{n} \sin(\varphi_i)$ and $C = \Sigma_{i=1}^{n} \cos(\varphi_i)$. Moreover, $\delta = 0$ for $C > 0$, $\delta = \pi$ for $C < 0$ and $S > 0$, and $\delta = \pi$ for $C < 0$ and $S < 0$. Different methods have been suggested to calculate the covariance of angular data.[31,81] Here, we center all values of $\varphi$ around its mean by the transformation

$$\varphi \rightarrow \begin{cases} \varphi - 2\pi & \text{if } \varphi - \langle\varphi\rangle > \pi \\ \varphi + 2\pi & \text{if } \varphi - \langle\varphi\rangle < -\pi \\ \varphi & \text{otherwise} \end{cases} \tag{5}$$

This ensures that $|\varphi - \langle\varphi\rangle| < \pi$ and that the directionality of the data is preserved. That is, positive if the data point is on the left-hand-side of the mean value (in the clockwise half part of the circle) and negative otherwise. In the case of the aPCA, 6 out of the 76 principal components exhibited a non-Gaussian distribution, yielding 60% of the fluctuations.

**C. Clustering of Data.** To identify and illustrate geometric clusters of the free energy landscape, the $k$-means algorithm[38] as implemented in the $R$ program suite[82] was employed. After choosing the desired number of clusters, $k$, the algorithm initially randomly selects $k$ data points as cluster centers and assigns the remaining data points to the closest center. Then $k$ new centroids are calculated as the average over all data points in their corresponding groups, and this is repeated until the data points no longer switch clusters. As this method is sensitive to the initial conditions, the algorithm was run 400 times and the partition with the minimal root-mean-square distances (rmsd) between the elements and their corresponding cluster centroid was used. As input data, all principal components with a non-

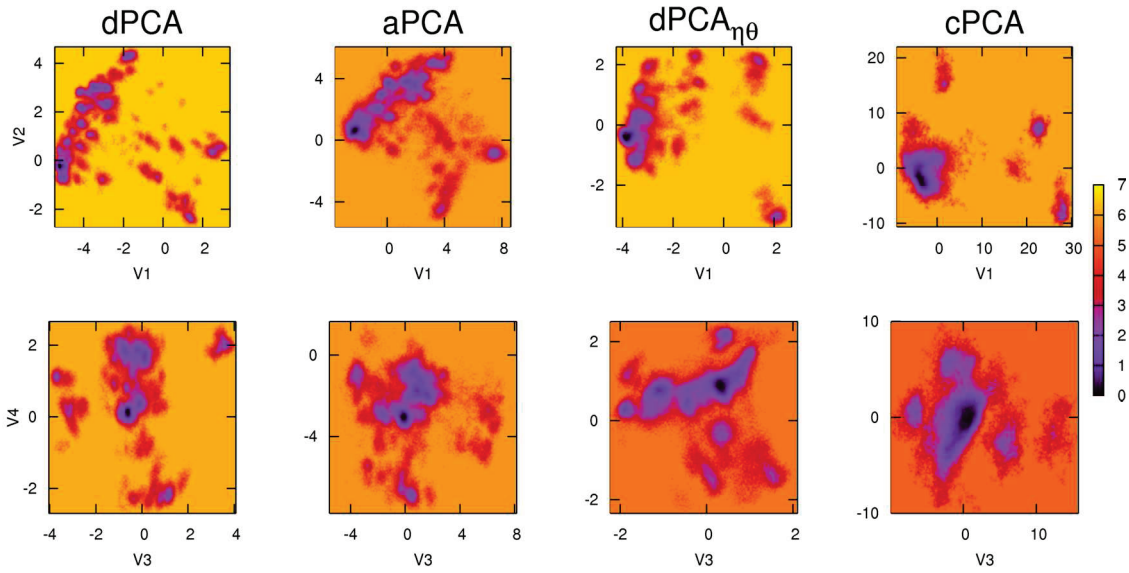


**Figure 2.** Two-dimensional representations of the free energy landscapes $\Delta G$ (in kcal/mol) as obtained by dPCA, aPCA, dPCA$_{\eta\theta}$, and cPCA. Shown are $\Delta G(V_1, V_2)$ (upper panels) and $\Delta G(V_3, V_4)$ (lower panels).

Gaussian distribution (see above) were used. As discussed previously,[33] this procedure yields the correct number of minima and barriers of the free energy landscape, while being much faster than performing the clustering on the complete data set.

To determine an appropriate value for the number of clusters $k$, we have employed the following procedure: By construction, the $k$-means algorithm minimizes the rmsd of the structures within each cluster. To estimate the quality of the clustering, it is therefore helpful to introduce the sum $S$ of these $k$ intracluster RMSDs. In general, $S$ decreases when the number of clusters $k$ is increased. It ranges from 0, when $k$ is equal to the total number of points, to a maximum, when $k = 1$ and all points are grouped together. Plotting $S$ as a function of $k$, we obtain curves as shown in Figure S1 of Supporting Information. Initially we observe a rapid drop of $S$, meaning that some clusters still contain points far from each other and a higher number of clusters is required to better partition the data. After this initial phase, $S$ decreases only slowly. Exploiting this behavior, we determine $k$ as the minimal number of clusters which gives a sum of intracluster RMSDs less than a given threshold.

By dividing the conformational space of the molecule into $k$ clusters, we obtain conformational states for which a $k \times k$ transition matrix $\mathbf{T}(\tau)$ was calculated from the REMD trajectories.[39] Its elements $T_{ij}(\tau)$ denote the probability of observing the system in state $j$ at time $t + \tau$ given that it is in state $i$ at time $t$. As the lag time needs to be shorter than the time between two replica exchanges (1.5 ps), we choose $\tau = 1$ ps. If the process under consideration can be described by a Markov chain,[39] a master equation using transition matrix $\mathbf{T}(\tau)$ provides the complete information of the time evolution of the system.[28,40−42]

## III. Results

**A. Conformational Distribution of the UUCG Hairpin.** In what follows, we restrict the analysis to the REMD results at room temperature, combining the data of three replicas around 300 K for better statistics. The discussion of the folding and unfolding of the RNA hairpin at increased temperature will be the subject of a subsequent paper.[83] At room temperature, all simulations yielded stable RNA structures with a mean rmsd of 0.22 nm with respect to the initial structure. To obtain a first impression of the conformational heterogeneity of the RNA hairpin, Figure 1D shows the probability distribution of the backbone dihedral angles. Interestingly, we find that, except for the angle $\beta$, all distributions are of multimodal character and typically show two peaks. As indicated by the labels "L" and "S", it is usually the loop residues that are found in several conformations. Comprising the information of the six backbone dihedral angles into two variables, the pseudotorsional angles $\eta$ and $\theta$ show multipeak distributions. Moreover, the pseudorotational angle $P$ reflects considerable conformational heterogeneity of the sugar puckers, which coexist in the C2′-endo state (corresponding to $P \approx 150°$) and in the C3′-endo state (corresponding to $P \approx 20°$). These conformational states are also represented in the distribution of the dihedral angle $\delta$.

For comparison, we have also considered the angular distribution obtained from REMD simulations of the UUCG loop around 450 K. At these high temperatures, the hairpin may unfold and fold reversible and therefore exhibits a large variety of structures including numerous single-strand conformations of the RNA. As shown in Figure S2 of the Supporting Information, the resulting angular distributions were nevertheless quite similar to the low-temperature results in Figure 1D, apart from the fact that the distributions become somewhat wider at elevated temperatures. Moreover, Figure S2 of Supporting Information shows very similar results for the 36-nucleotide segment SL1m discussed below. Hence, the angular distributions shown in Figure 1D appear to be quite representative for RNA molecules.

**B. Comparison of Various PCAs.** It is interesting to study to what extent the various versions of the PCA are able to resolve the conformational heterogeneity of the RNA hairpin. To this end, Figure 2 shows two-dimensional free energy landscapes as obtained from the dPCA, aPCA, dPCA$_{\eta\theta}$, and cPCA (see Methods for definitions). As suggested by recent experiments,[6−11] the RNA energy landscape of the RNA hairpins is quite rugged, i.e., there exist numerous conformational states which are separated by free energy barriers $\gtrsim$2 kcal/mol. Although all representations use the identical MD trajectories, the general appearance of the energy landscape is seen to significantly depend on the variables used in the PCA. Generally speaking, we find that the dPCA yields the best and the cPCA

the lowest resolution of conformational states. As the minima of the dPCA energy landscape correspond to well-defined conformational structures (Figure 4), it is interesting to investigate the origin of the lower resolution of the other representations.

The cPCA landscape exhibits one main minimum containing more than 80% of all structures and about 3−4 smaller minima. To study the origin for this poor resolution, we first repeated the rotational fit of the trajectories (see Methods) using various reference structures, which did not change the cPCA results. Moreover, the rmsd of the majority of the REMD structures show only relatively small fluctuations around their mean of 0.2 nm, which indicates a fairly rigid system. Hence, we may exclude artifacts due to the coupling of overall and internal motion, which have been found a serious problem for the cPCA analysis of folding peptides.[29] A closer analysis of the conformational states resolved by the dPCA but not by the cPCA (see below) suggests that the details of RNA hydrogen bonding and stacking are better discriminated by using dihedral angles. While Cartesian coordinates certainly contain the full information, the noise caused by the high-frequency motions of the RNA seems to obscure the slow conformational dynamics of interest. Containing significantly less coordinates than the Cartesian representation, moreover, the phase space of dihedral angles is likely to be better sampled than the Cartesian phase space.

The overall appearance of the free energy landscape obtained from the dPCA$_{\eta\theta}$ (Figure 2) looks already more similar to the landscape of the full dPCA. As a consequence of the reduction to only two backbone dihedral angles, however, in particular the resolution of the substructures in the main part of the energy landscape is only limited. The free energy landscape obtained from the aPCA (PCA directly on the angles, instead of on the sin/cos-transformed angles) appears quite similar to the results for the full dPCA, at least for the energy along the first two principal components. This reflects the fact that in most cases the dihedral angles of the RNA hairpin stay within a range of 180°, for which the aPCA is a correct representation. Even in these cases, however, it is a general finding that the sin/cos transformation 3 of the dPCA results in a better resolution of the conformational states on the free energy landscape.[33]

**C. Clustering Results.** As detailed in the Methods Section, we employed the *k* means algorithm to identify clusters of the free energy landscape corresponding to metastable conformational states. Guided by visual inspection, we performed the analysis for numerous cluster numbers *k* and choose the cluster number that resulted in an optimal value of the sum of intrastate RMSDs (see Figure S1 of Supporting Information). Reflecting the structural resolution of the various methods, this resulted in total cluster numbers of 11, 9, 6, and 5 for the dPCA, aPCA, dPCA$_{\eta\theta}$, and cPCA, respectively. Figure 3 illustrates the resulting clusters in the ($V_1$, $V_2$) plane. A comparison with Figure 2 reveals that for the most part the clusters coincide with the minima of the corresponding free energy landscape, thus representing metastable conformational states.

Since the various landscapes are different representations of the same data set, it is interesting to find out how corresponding cluster compare to each other. Adopting the dPCA clusters as a reference, Table I reports on the similarity of clusters found in the various representations. Clearly, the highest correspondence between the various landscapes is found for the smallest clusters such as dPCA cluster number **7**, **8**, and **10**. A closer analysis (Table II) reveals that these states are quite different from the rest of the ensemble and can therefore be easily distinguished by all methods. Apart from these similarities, Figure 3 clearly shows the quite different performance of the

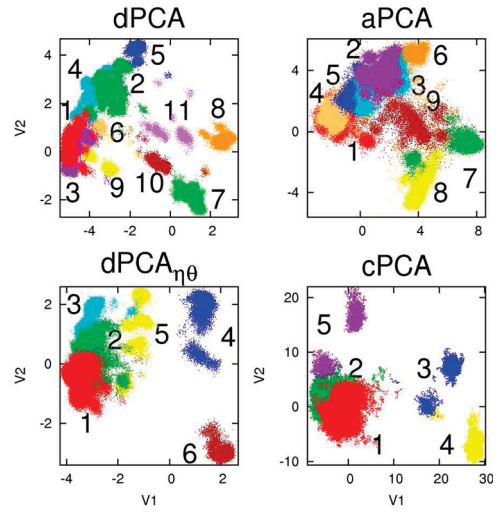


**Figure 3.** Clustering results obtained for the various PCA methods, illustrated by different colors in the $\Delta G(V_1, V_2)$ landscape. In all cases, the cluster numbers are ranked from the most populated to the least populated cluster.

various versions of the PCA. The cPCA combines almost all structures in only two clusters. In particular, the second cPCA cluster is surprisingly mixed, since it contains quite different states such as dPCA clusters **2**, **4**, **5**, **9**, and **11**.

By use of only two backbone dihedral angles, the clusters obtained from the dPCA$_{\eta\theta}$ already represent a clear improvement over the cPCA. However, problems may arise to discriminate conformations with similar backbone structure. For example, the dPCA$_{\eta\theta}$ puts clusters **1** and **9** together, although they look quite different in Figure 3, mainly because of the nucleotide in the 5′ position. This is due to the fact that there is one out of 16 pseudotorsional angles that differs for these two structures in the dPCA$_{\eta\theta}$, while there are 13 out of 56 dihedrals that differ in the full dPCA. The aPCA clusters are quite similar to the results from the full dPCA, which reflects the fact that in most cases the dihedral angles of the UUCG hairpin stay within a range of 180°. However, the aPCA does not resolve all dPCA states. For example, similar to the dPCA$_{\eta\theta}$, the aPCA combines clusters **1** and **9** in a single state.

To verify that the various clusters revealed by the dPCA are not artifacts of the method but correspond to different RNA conformations, we performed a structural analysis of the dPCA clusters. Figure 4 shows representative structures of the UUCG clusters, which exhibit the minimum geometrical distance from the cluster centroid. The hairpin structure does not go into a complete unfolding event: even for significant rearrangements, especially in the less populated clusters, the stem seems to remain quite stable. Table II characterizes the main hydrogen bonds and stacking interactions of the cluster structures. (More details of the hydrogen-bonding are comprised in Table S1 of Supporting Information.) The structure of cluster **1**, which is the most populated state and thus the global minimum of the free energy landscape, shows a pattern of only "native" hydrogen bonds (i.e., bonds that are present in the experimental starting structure). Smaller clusters, whose structure is more disordered (Figure 4) are characterized by a higher number of nonconventional hydrogen bonds. Interestingly, the Watson−Crick base pair between the central nucleotides of the stem 2G:9C are most stable. The strength of this interactions seems to be relevant for the stability of the overall structure. On the other hand, the Watson−Crick base pair between residues 1C:10G is not highly conserved: these nucleotides are at the end of the stem and are

**TABLE I: Comparison of Conformational Clusters of the RNA Hairpin as Obtained from Various Methods (dPCA, dPCA$_{\eta\theta}$, aPCA, and cPCA), Where the 11 Clusters of the Full dPCA Are Taken as Reference[a]**

| dPCA | dPCA$_{\eta\theta}$ | dPCA | aPCA | dPCA | cPCA |
|---|---|---|---|---|---|
| 1 + 2(70%) + 4 + 9 | 1 | 1 + 9 | 1 | 1 + 2(65%) + 3 | 1 |
| 2(30%) + 3 + 6 | 2 | 2 | 2(70%) + 3(55%) | 2(35%) + 4 + 5 + 9 + 11 | 2 |
| 5 | 3 | 3 | 4 | 7 | 4 |
| 7 | 6 | 4 | 2(25%) + 5 | 8 | 3 |
| 8 | 4 | 5 | 6 | 6 + 10 | 5 |
| 10 + 11 | 5 | 6 | 3(35%) | | |
| | | 7 | 8 | | |
| | | 8 | 7 | | |
| | | 10 + 11 | 9 | | |

[a] Two clusters are regarded as equivalent, if their structures agree to at least 80%. For example, "9 (aPCA) vs 10 + 11 (dPCA)" means that cluster 9 of the aPCA contains (at least 80% of) the structures comprised in clusters 10 and 11 of the dPCA. In all cases, the cluster numbers are ranked from the most populated to the least populated cluster.

**TABLE II: Main Hydrogen Bonds and Stacking Interactions of the dPCA Conformational States of the UUCG Hairpin[a]**

| | | | | Watson−Crick | | | Wobble |
|---|---|---|---|---|---|---|---|
| cluster id | population (%) | stacking interactions | total no. of HB | 1C-10G | 2G-9C | 3C-8G | 4U-7G |
| 1 | 24.0 | 2−10, 4−6, 8−9 | 11 | $^2/_3$ | y | y | $^1/_2$ |
| 2 | 22.0 | 1−2, 4−6 | 16 | $^2/_3$ | y | y | |
| 3 | 12.0 | 4−6, 8−9 | 9 | | $^2/_3$ | y | |
| 4 | 11.8 | 1−2, 4−6, 8−9 | 14 | y | y | y | $^1/_2$ |
| 5 | 6.5 | 1−2, 4−8, 9−10 | 15 | y | y | y | $^1/_2$ |
| 6 | 5.6 | 8−9, 9−10 | 13 | y | y | $^2/_3$ | |
| 7 | 5.3 | 3−7 | 12 | | y | y | |
| 8 | 4.6 | 4−6 | 16 | y | y | y | |
| 9 | 4.0 | 8−9, 9−10 | 13 | | y | y | $^1/_2$ |
| 10 | 2.3 | 1−10, 2−4, 4−6, 7−9 | 13 | | y | | y |
| 11 | 1.7 | 2−3, 4−6, 7−9, 9−10 | 13 | | y | y | $^1/_2$ |

[a] "y" indicates a stable base pair, fractions indicate the number of stable hydrogen bonds (i.e., $^2/_3$ means two out of three). Stacking interactions are shown by indicating the number of the involved nucleotides.

therefore less constrained. In the loop there are one guanosine and two uridines; of the two, only 4U can interact with 7G via a Wobble base pair.[84] The importance of this interaction is indicated by the observation that the conventional Wobble base pair can be replaced by a hydrogen bond involving atoms N2(7G) and O2(4U) and a hydrogen bond between the base of 7G and the ribose of 4U. Interactions between base atoms only are mainly found in the loop region, because in the stem the formation of Watson−Crick base pairs does not allow for further interactions.

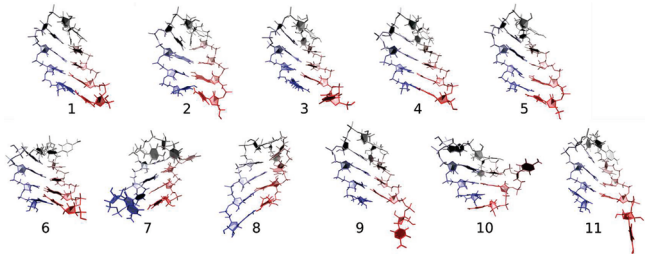


**Figure 4.** Representative structures of the dPCA clusters obtained or the UUCG hairpin. The color scale, from red to blue, indicates the direction of the RNA, from 5′ to 3′.

Besides hydrogen bonding, stacking interactions can play an important role in the stabilization of the RNA structure. Surprisingly, the most conserved stacking interaction is not in the stem but in the loop between residues 4U and 6C, which represents a conserved and stable feature of the UUCG tetraloop.[56] As mentioned above, base pair 1C:10G is the least conserved one, but looking at the structures, it seems that mainly residue 1G is affected by the loss of this interaction. This can

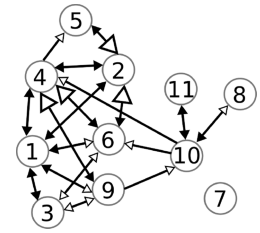


**Figure 5.** Illustration of the transition matrix showing the pathways of the conformational rearrangements between the dPCA clusters of the UUCG hairpin. Big empty, full black, and small empty arrows correspond to ≳100, 10, and 1 transitions within 10 ns, respectively.

be reconnected to the possible alternative stacking interactions that 10C can form with 2G, as in cluster **1**. Indeed, stacking interactions in the 3′ region, 8G-9C and 9C-10G, are generally more stable than those located at the opposite end. Interestingly, cluster **7**, located at the "edge" of all landscapes, quite far from other clusters, is also the one with in total the least number of native interactions: only 6 of 12 hydrogen bonds are native and the only stacking interaction is between 3C-7G. To summarize, we have shown that the dPCA conformational states are well characterized by specific hydrogen bonding and stacking interactions. In principle, these states can therefore be discriminated by NMR, infrared, or fluorescence experiments.[6,8,11]

To reveal the pathways of the conformational rearrangements connecting the various clusters of the UUCG hairpin, we have calculated the transition matrix associated with the dPCA conformational states (see Methods). Because of the short lag time of 1 ps, all clusters are quite stable with metastabilities $T_{ii} \lesssim 0.96$. Figure 5 shows a visual representation of the intercluster transitions, which reveals the preferred pathways of the free energy landscape. Generally speaking, clusters **1**−**6**

and **9** are well connected and exhibit frequent transitions between each other. In particular, cluster **1** representing around 24% of the whole population is highly connected, especially with its direct neighbors. Moreover, it is the only acceptor state of transitions coming from cluster **3**. Cluster **7**, on the other hand, has no connections with the other clusters, because it is located at the very edge of the landscape. Interestingly, clusters **7**, **8**, **10**, and **11** show on average the highest rmsd and are at the same time kinetically well separated from the other clusters. Connections between the left and the right part of the landscape are possible only by passing through cluster **10**. This state might act as a kinetic trap in the folding or unfolding of the RNA.

**D. Conformational Analysis of SL1m.** Our main finding that the conformational states of the UUCG hairpin can only be discriminated by an angular PCA such as the dPCA but not by the cPCA using Cartesian coordinates is somewhat surprising. It raises the question if this result only holds for the special case of a small RNA hairpin or is a general finding that generalizes to larger RNA systems. With this end in mind, we repeated the above analysis for the 36-nucleotide segment SL1m of the Ψ site of HIV-1.[69] Consisting of a GAGA tetraloop and two stems that are separated by a short bulge (Figure 1A), the structure of SL1m is significantly more complex than the small UUCG hairpin considered above. Nevertheless, the results obtained from the various versions of PCA are quite similar to the findings for the UUCG hairpin, see Figures S2 and S3 of Supporting Information. In the following, we therefore restrict the discussion to the free energy landscapes resulting from the dPCA and cPCA, which are shown in Figure 6A.

By employing Cartesian coordinates, only the first out of 3507 principal components shows a multipeak distribution. The two conformational states discriminated along $V_1$ are clearly distinguished by their RMSDs of 0.47 and 0.68 nm, respectively (see Figure S4 of Supporting Information). They correspond to a helical elongated structure occurring for times $t \lesssim 30$ ns and a more compact structure with the upper stem flipped down at longer times; see Figure 6B. This considerable change of the global structure of SL1m may cause non-negligible coupling of overall and internal motion, which is most likely the main reason for the poor resolution of the cPCA energy landscape.[29] In the dPCA, on the other hand, at least 5 out of 568 principal components show a clear non-Gaussian distribution, which facilitates the discrimination of numerous different conformational states. We note in passing that we have also considered the free energy landscapes as obtained by dPCA and aPCA considering only backbone dihedral angles (see Figure S3 of Supporting Information). We found that the aPCA on the six backbone dihedral angles may be regarded as a good compromise between accuracy and computational cost (only 212 instead of 568 (dPCA) or 3507 (cPCA) variables).

Instead of a detailed analysis as done in Table II for the UUCG hairpin, here we wish to focus on the global rearrangements of SL1m as shown by the MD snapshots in Figure 6B at times 0, 13.5, 32.5, and 78 ns, respectively. To this end, we consider the C4′ atoms of residues 1G, 28G, and 18G and introduce the interresidue distances $L$ = 1G−28G describing the length of the lower stem, $U$ = 18G−28G describing the length of the upper stem, and $T$ = 1G−18G accounting for the total length of the RNA. Initially, the whole structure is in an elongated form with values of $L = 2.02$ nm, $U = 4.09$ nm, and $T = 4.27$ nm. After 13.5 ns, we observe a packing of the structure, in which all the distances become smaller by at least 0.5 nm. Moreover, the angle between the three residues decreases by about 20°. At 32.5 ns, the lower stem relaxes to a length $L = 2.58$ nm, which is larger than in the initial structure. At the same time, the upper stem flips down which leads to a short end-to-end distance $T = 1.65$ nm. At 78 ns, the end-to-end distance increases again to ∼3.3 nm. Because of a different folding of upper loop, the end structure is quite far from the starting structure. As indicated by the labeled minima of the dPCA energy landscape, all these conformational states are well resolved in the dPCA. This demonstrates that the dPCA is also sensitive to global structural rearrangements, which may be important for the function of the RNA.[22,23]

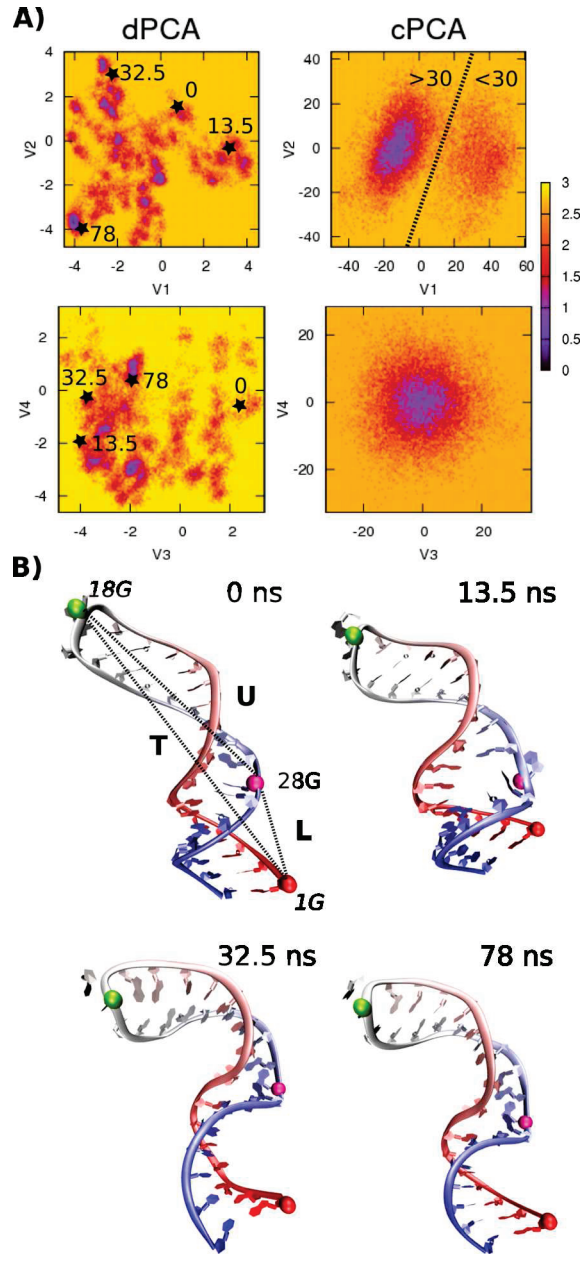


**Figure 6.** (A) Two-dimensional representation of the free energy landscape $\Delta G$ (in kcal/mol) as obtained by the dPCA and the cPCA of the 36-nucleotide segment SL1m of the Ψ site of HIV-1. Shown are $\Delta G(V_1, V_2)$ (upper panels) and $\Delta G(V_3, V_4)$ (lower panels). (B) MD snapshots of SL1m obtained at times 0, 13.5, 32.5, and 78 ns, respectively, which correspond to the labeled minima of the dPCA energy landscape.

## IV. Conclusions

By adoption of a 10-nucleotide UUCG hairpin and the 36-nucleotide segment SL1m of the Ψ site of HIV-1 as representa-

tive examples, we have compared the performance of four version of PCA to accurately represent the free energy landscape of the RNA: the cPCA using the Cartesian coordinates of all atoms, the dPCA using the six backbone dihedral angles as well as the glycosidic torsional angle $\chi$ and the pseudorotational angle $P$, the aPCA which ignores the circularity of the $6 + 2$ dihedral angles of the RNA, and the $dPCA_{\eta\theta}$, which approximates the 6 backbone dihedral angles by two pseudotorsional angles $\eta$ and $\theta$. Interestingly, it was found that the conformational heterogeneity of the RNA hairpins can only be discriminated by an angular PCA such as the dPCA but not by the cPCA using Cartesian coordinates. In the case of SL1m which undergoes a considerable structural change, this failure of the cPCA is probably caused by artifacts due to the coupling of overall and internal motion.[29] The short UUCG hairpin, on the other hand, is fairly rigid, and therefore no such coupling occurs. In this case, the poor performance of the cPCA seems to be due to the fact that the details of RNA hydrogen bonding and stacking are better observed by using dihedral angles. Furthermore, the space of dihedral angles is usually better sampled than the much higher dimensional Cartesian space.

The $dPCA_{\eta\theta}$ already presents a significant improvement over the cPCA, although the resolution of the substructures in the main part of the energy landscape is limited. The free-energy landscape obtained from the aPCA appears quite similar to the results for the full dPCA. This reflects the fact that in most cases the dihedral angles of the RNA hairpins stay within a range of 180°, for which the aPCA is a correct representation. By use of the aPCA and exploiting the fact that the backbone dihedral angles contain the main structural information of an RNA molecule, an aPCA on the six backbone dihedral angles may be regarded as a good compromise between accuracy and computational cost (see Figure S3 of Supporting Information). On the other hand, only the full dPCA achieved the complete resolution of the conformational distributions of the RNA hairpins.

Instead of using the first few principal components of a PCA, in principle, a clustering of RNA structures can also be performed in the full Cartesian or angular space of the trajectory. In the case of the UUCG loop, this direct clustering yielded quite similar results to the clustering using the first few principal components (see Table S1 of Supporting Information). However, it should be stressed that even for a very small RNA with $\approx$300 atoms and $\approx 10^6$ structures the $k$-means clustering becomes quite cumbersome. It is the PCA reduction from $\approx 10^3$ to $\lesssim 10$ coordinates that makes the clustering conveniently doable for larger RNAs. Furthermore, the resulting free energy surface represents a reduced dynamic model of the system and can be used, for example, to perform simulations of the molecular dynamics using the Langevin approach[43−46] or a nonlinear dynamic model.[37,47−49] By employing these methods, we are currently working on the interpretation of recent T-jump experiments on RNA hairpins.[6,10,11] As time-resolved infrared and fluorescence experiments are able to discriminate RNA conformations with specific hydrogen bonding and stacking interactions, and since these conformational states are also well resolved in a dPCA, joint experimental/computational studies are a promising approach to further explore the rugged free energy landscape of RNA.

**Acknowledgment.** We thank Alexandros Altis, Jessica Biedinger, Rainer Hegger, and Jens Wöhnert for numerous inspiring and helpful discussions. This work has been supported by the Frankfurt Center for Scientific Computing, the Fonds der Chemischen Industrie, and the Deutsche Forschungsgemeinschaft (via SFB 579 "RNA-Ligand Interactions").

**Supporting Information Available:** This material is available free of charge via the Internet at http://pubs.acs.org.

## References and Notes

(1) Ball, K. D.; Berry, R. S.; Kunz, R. E.; Li, F.-Y.; Proykova, A.; Wales, D. J. *Science* **1996**, *271*, 963–965.
(2) Onuchic, J. N.; Schulten, Z. L.; Wolynes, P. G. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545–600.
(3) Dill, K. A.; Chan, H. S. *Nat. Struct. Bio.* **1997**, *4*, 10–19.
(4) Gruebele, M. *Curr. Opin. Struct. Biol.* **2002**, *12*, 161–168.
(5) Wales, D. J. *Energy Landscapes*; Cambridge University Press: Cambridge, 2003.
(6) Ma, H.; Proctor, D. J.; Kierzek, E.; Kierzek, R.; Bevilacqua, P. C.; Gruebele, M. *J. Am. Chem. Soc.* **2006**, *128*, 1523–1530.
(7) Tinoco, I., Jr. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 363–385.
(8) Fürtig, B.; Richter, C.; Wöhnert, J.; Schwalbe, H. *Chembiochem* **2003**, *4*, 936–962.
(9) Wenter, P.; Fürtig, B.; Hainard, A.; Schwalbe, H.; Pitsch, S. *Angew. Chem., Int. Ed. Engl.* **2005**, *44*, 2600–2603.
(10) Stancik, A. L.; Brauns, E. B. *Biochemistry* **2008**, *47* (41), 10834–10840.
(11) Sarkar, K.; Meister, K.; Sethi, A.; Grübele, M. *Biophys. J.* **2009**, *97*, 1418–1427.
(12) Chen, S.; Dill, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 646–651.
(13) Sorin, E. J.; Rhee, Y. M.; Nakatani, B. J.; Pande, V. S. *Biophys. J.* **2003**, *85*, 790–803.
(14) Hashem, Y.; Westhof, E.; Auffinger, P. *Computational Structural Biology*; Schwede, T., Peitsch, M. C., Eds.; World Scientific: New York, 2008.
(15) Chen, S.-J. *Annu. Rev. Biophys.* **2008**, *37*, 197–214.
(16) Thirumalai, D.; Lee, N.; Woodson, S. A.; Klimov, D. K. *Annu. Rev. Phys. Chem.* **2001**, *52*, 751–762.
(17) Hyeon, C.; Morrison, G.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (28), 9604–9609.
(18) Bowman, G. R.; Huang, X.; Yao, Y.; Sun, J.; Carlsson, G.; Guibas, L. J.; Pande, V. S. *J. Am. Chem. Soc.* **2008**, *130* (30), 9676–9678.
(19) Couzin, J. *Science* **2002**, *298*, 2296–2297.
(20) Murray, L. J. W.; Arendall, W. B.; Richardson, D. C.; Richardson, J. S. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (24), 13904–13909.
(21) Richardson, J. S.; Schneider, B.; Murray, L. W.; Kapral, G. J.; Immormino, R. M.; Headd, J. J.; Richardson, D. C.; Ham, D.; Hershkovits, E.; Williams, L. D.; Keating, K. S.; Pyle, A. M.; Micallef, D.; Westbrook, J.; Berman, H. M.; Consortium, R. N. A. O. *RNA* **2008**, *14* (3), 465–481.
(22) Zhang, Q.; Sun, X.; Watt, E. D.; Al-Hashimi, H. M. *Science* **2006**, *311* (5761), 653–656.
(23) Hall, K. B. *Curr. Opin. Chem. Biol.* **2008**, *12*, 612–618.
(24) Ichiye, T.; Karplus, M. *Proteins* **1991**, *11*, 205–217.
(25) Garcia, A. E. *Phys. Rev. Lett.* **1992**, *68*, 2696–2699.
(26) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. *Proteins* **1993**, *17*, 412–425.
(27) Kitao, A.; G ō, N. *Curr. Opin. Struct. Biol.* **1999**, *9*, 164–169.
(28) de Groot, B. L.; Daura, X.; Mark, A. E.; Grubmüller, H. *J. Mol. Biol.* **2001**, *309*, 299–313.
(29) Mu, Y.; Nguyen, P. H.; Stock, G. *Proteins* **2005**, *58*, 45.
(30) An alternative approach to circumvent problems associated with the mixing of internal and overall motion is the isotropic reorientational eigenmode dynamics method of Prompers and Bruschweiler,[85] which aimes to calculate NMR data.
(31) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. *J. Chem. Phys.* **2007**, *126*, 244111.
(32) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
(33) Altis, A.; Otten, M.; Nguyen, P. H.; Hegger, R.; Stock, G. *J. Chem. Phys.* **2008**, *128*, 245102.
(34) Nguyen, P. H.; Stock, G.; Mittag, E.; Hu, C.-K.; Li, M. S. *Proteins* **2005**, *61*, 795–808.
(35) Maisuradze, G. G.; Leitner, D. M. *Proteins* **2007**, *67*, 569–578.
(36) Nguyen, P. H.; Li, M. S.; Stock, G.; Straub, J. E.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 111–116.
(37) Hegger, R.; Altis, A.; Nguyen, P. H.; Stock, G. *Phys. Rev. Lett.* **2007**, *98*, 028102.
(38) Hartigan, J. A.; Wong, M. A. *Appl. Stat.* **1979**, *28*, 100–108.
(39) Van Kampen, N. G. *Stochastic Processes in Physics and Chemistry*; Elsevier: Amsterdam, 1997.
(40) Noe, F.; Krachtus, D.; Smith, J. C.; Fischer, S. *J. Chem. Theory Comput.* **2006**, *2*, 840–857.

(41) Noe, F.; Horenko, I.; Schütte, C.; Smith, J. C. *J. Chem. Phys.* **2007**, *126*, 155102.
(42) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101.
(43) Lange, O. F.; Grubmüller, H. *J. Chem. Phys.* **2006**, *124*, 214903.
(44) Yang, S.; Onuchic, J. N.; Levine, H. *J. Chem. Phys.* **2006**, *125*, 054910.
(45) Horenko, I.; Hartmann, C.; Schütte, C.; Noe, F. *Phys. Rev. E* **2007**, *76*, 016706.
(46) Hegger, R.; Stock, G. *J. Chem. Phys.* **2009**, *130*, 034106.
(47) Das, P.; Moll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 9885–9890.
(48) Nguyen, P. H. *Proteins* **2006**, *65*, 898.
(49) Lange, O. F.; Grubmüller, H. *Proteins* **2006**, *62*, 1053–1061.
(50) The pseudo-rotational phase angle is defined as $P = P_0 + 180°$ (if $v_2 < 0$), or $P = P_0 - 180°$ (if $P_0 < 0$), or $P = P_0$ (otherwise), where $P_0 = \arctan(((v_4 - v_0) - (v_3 - v_1))/(2v_2(\sin 36° + \sin 72°)))$ and $v_i$ ($i = 1-4$) represents the dihedral angles of the sugar ring.
(51) Saenger, W. *Principles of nucleic acid structure*; Springer-Verlag: New York, 1988.
(52) Reijmers, T. H.; Wehrens, R.; Buydens, L. M. C. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 61–71.
(53) Duarte, C. M.; Pyle, A. M. *J. Mol. Biol.* **1998**, *284* (5), 1465–1478.
(54) Biochemical Nomenclature (JCBN), I.-I. J. C. Eur. J. Biochem. 1983, *131* (1), 5−7.
(55) Eaton, J. W. *GNU Octave Manual*; Network Theory Limited: NY, 2002.
(56) Duchardt, E.; Schwalbe, H. *J. Biomol. NMR* **2005**, *32*, 295–308.
(57) Ferner, J.; Villa, A.; Durchardt, E.; Widjajakusuma, E.; Stock, G.; Schwalbe, H. *Nucleic Acids Res.* **2008**, *38*, 1928–1940.
(58) Miller, J.; Kollman, P. *J. Mol. Biol.* **1997**, *270*, 436–450.
(59) Williams, J.; Hall, K. *J. Mol. Biol.* **2000**, *297*, 1045–1061.
(60) Zacharias, M. *Curr. Opin. Struct. Biol.* **2000**, *10*, 311–317.
(61) Villa, A.; Widjajakusuma, E.; Stock, G. *J. Phys. Chem. B* **2008**, *112*, 134–142.
(62) Uhlenbeck, O. C. *Nature* **1990**, *346*, 613–614.
(63) Jagath, J. R.; Matassova, N. B.; De Leeuw, E.; Warnecke, J. M.; Lentzen, G.; Rodnina, M. V.; Luirink, J.; Wintermeyer, W. *RNA-A* **2001**, *7*, 293–301.
(64) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141.
(65) Okabe, T.; Kawata, M.; Okamoto, Y.; Mikami, M. *Chem. Phys. Lett.* **2001**, *335*, 435–439.
(66) Frenkel, D.; Smit, B. *Understanding Molecular Simulations*; Academic: San Diego, 2002.
(67) Cheatham III, T. E.; Cieplak, P.; Kollman, P. A. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845.
(68) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
(69) Lawrence, D. C.; Stover, C. C.; Noznitsky, J.; Wu, Z.; Summers, M. F. *J. Mol. Biol.* **2003**, *326*, 529–542.
(70) Schiemann, O.; Piton, N.; Mu, Y.; Stock, G.; Engels, J.; Prisner, T. F. *J. Am. Chem. Soc.* **2004**, *126*, 5722.
(71) Piton, N.; Schiemann, O.; Mu, Y.; Stock, G.; Prisner, T. F.; Engels, J. *Nucleic Acids Res.* **2007**, *35*, 3128–3143.
(72) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. *J. Chem. Phys.* **1983**, *79*, 926.
(73) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21* (12), 1049–1074.
(74) Widjajakusuma, E. C.; Villa, A.; Stock, G. In preparation.
(75) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
(76) Darden, T.; York, D.; Petersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.
(77) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
(78) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph* **1990**, *14*, 33–38.
(79) Jolliffe, I. T. *Principal Component Analysis*; Springer: New York, 2002.
(80) Fisher, N. I. *Statistical Analysis of Circular Data*; Cambridge University Press: Cambridge, 1996.
(81) Frickenhaus, S.; Kannan, S.; Zacharias, M. *J. Comput. Chem.* **2009**, *30*, 479–492.
(82) R: A language and environment for statistical computing. *R Development Core Team*; R Foundation for Statistical Computing: Vienna, Austria, 2005.
(83) Riccardi, L.; Nguyen, P. H.; Stock, G. To be published.
(84) Crick, F. H. *J. Mol. Biol.* **1966**, *19* (2), 548–555.
(85) Prompers, J. J.; Brüschweiler, R. *Proteins* **2002**, *46*, 177.

JP9076036