Modeling the Free Energy Landscape of Biomolecules via Dihedral Angle Principal Component Analysis of Molecular Dynamics Simulations

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich
Biochemie, Chemie und Pharmazie
der Goethe-Universität
in Frankfurt am Main

von
Alexandros Altis
aus Frankfurt am Main

Frankfurt am Main 2008
(D 30)

vom Fachbereich Biochemie, Chemie und Pharmazie der
Goethe-Universität Frankfurt am Main als Dissertation angenommen.

Dekan:	Prof. Dr. Dieter Steinhilber
1. Gutachter:	Prof. Dr. Gerhard Stock
2. Gutachter:	JProf. Dr. Karin Hauser
Datum der Disputation:	.....

Contents

Chapter 1

Introduction

Proteins can be regarded as the most important building blocks of our body. They function as mechanical tools, perform transport (e.g., hemoglobin) and communication, catalyze biochemical reactions, and are involved in many other essential processes of life. The native structure to which a protein folds by the process of protein folding determines its biological function. To answer the protein folding problem of how the amino acid sequence of a protein as synthesized by ribosomes dictates its structure, one has to understand the complex dynamics of protein folding. In the folding process the transition between metastable conformational states plays a crucial role. These are long-lived intermediates, which for proteins can have lifetimes up to microseconds before undergoing further transitions.

Experiments using nuclear magnetic resonance (NMR) spectroscopy or X-ray crystallography can provide structural information on the native state or sometimes metastable states [1]. But as a system quickly relaxes to a lower energy state, the dynamics of the process of folding is hard to assess by experiment. In addition, traditional experiments provide only average quantities such as mean structures, not distributions and variations. Molecular dynamics computer simulations are used to obtain a deeper understanding of the dynamics and mechanisms involved in protein folding [2].

Molecular dynamics simulations have become a popular and powerful approach to describe the structure, dynamics, and function of biomolecules in atomic detail. In the past few years, computer power has increased such that simulations of small peptides on the timescale of microseconds are feasible by now. With the help of worldwide distributed

computing projects as Folding@home [3] even folding simulations of small microsecond and submicrosecond folding proteins are possible [4]. Markov chain models constructed from molecular dynamics trajectories could prove promising for the modeling of the correct statistical conformational dynamics over much longer times than the molecular dynamics simulations used as input [5–7]. Unfortunately, it is neither trivial to define the discrete states for a Markov approach, nor is it clear whether the system under consideration obeys the Markov property.

As molecular dynamics simulations result in huge data sets which need to be analyzed, one needs methods which filter out the essential information. For example, biomolecular processes such as molecular recognition, folding, and aggregation can all be described in terms of the molecule's free energy [8–10]

$$\Delta G(r) = -k_{\mathrm{B}}T[\ln P(r) - \ln P_{\max}]. \tag{1.1}$$

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. The probability distribution along these "order parameters" may be obtained from experiment, from a theoretical model, or a computer simulation. The resulting free energy "landscape" has promoted much of the recent progress in understanding protein folding [8–12]. Being a very high-dimensional and intricate object with many free energy minima, finding good order parameters is essential for extracting useful low-dimensional models of conformational dynamics of peptides and proteins. For the decomposition of a system into a relevant (low-dimensional) part and an irrelevant part principal component analysis has become a crucial tool [13].

Principal component analysis (PCA), also called quasiharmonic analysis or essential dynamics method [14–17], is one of the most popular methods to systematically reduce the dimensionality of a complex system. The approach is based on the covariance matrix, which provides information on the two-point correlations of the system. The PCA represents a linear transformation that diagonalizes the covariance matrix and thus removes

the instantaneous linear correlations among the variables. Ordering the eigenvalues of the transformation decreasingly, it has been shown that a large part of the system's fluctuations can be described in terms of only a few principal components which may serve as reaction coordinates [14–20] for the free energy landscape.

Some PCA methods using internal (instead of Cartesian) coordinates [21–27] have been proposed in the literature. In biomolecules, in particular the consideration of dihedral angles appears appealing, because other internal coordinates such as bond lengths and bond angles usually do not undergo changes of large amplitudes. Due to the circularity of the angular variables it is nontrivial to apply methods such as PCA for the analysis of molecular dynamics simulations.

This work presents a contribution to the literature on methods in search of low-dimensional models that yield insight into the equilibrium and kinetic behavior of peptides and small proteins. A deep understanding of various methods for projecting the sampled configurations of molecular dynamics simulations to obtain a low-dimensional free energy landscape is acquired. Furthermore low-dimensional dynamic models for the conformational dynamics of biomolecules in reduced dimensionality are presented. As exemplary systems, mainly short alanine chains are studied. Due to their size they allow for performing long simulations. They are simple, yet nontrivial systems, as due to their flexibility they are rapidly interconverting conformers. Understanding these polypeptide chains in great detail is of considerable interest for getting insight in the process of protein folding. For example, K. Dill et al. conclude in their review [28] about the protein folding problem that "the once intractable Levinthal puzzle now seems to have a very simple answer: a protein can fold quickly and solve its large global optimization puzzle simply through piecewise solutions of smaller component puzzles".

The thesis is organized as follows: Chapter 2 provides the theoretical foundations of the dihedral angle principal component analysis (dPCA) for the analysis of the dynamics of the ϕ, ψ backbone dihedral angles. In an introduction to circular statistics we thoroughly discuss the implications of the proposed sin/cos-transformation of the dihedral angles which comes along with a doubling of variables from N angular variables to $2N$ Cartesian-like ones. It is shown that indeed this transformation can truthfully represent the original angle distribution without generating spurious results. Furthermore, we show

that the dPCA components can readily be characterized by the conformational changes of the peptide. For the trialanine system the equivalence between a Cartesian PCA and the dPCA is demonstrated. We then introduce a complex valued version of the dPCA which sheds some light on the doubling of variables occurring in the sin/cos dPCA. The developed concepts are demonstrated and applied to a 300 ns molecular dynamics simulation of the decaalanine peptide.

What follows is a detailed study of the similarities and differences of various PCA methods. The dPCA is evaluated in comparison to alternative projection approaches. In particular, it is shown that Cartesian PCA fails to reveal the true structure of the free energy landscape of small peptides, except for the conformationally trivial example trialanine. The smooth appearance of the landscape is an artifact of the mixing of internal and overall motion. This is demonstrated using a 100 ns and an 800 ns simulation of pentaalanine and heptaalanine, respectively. In addition, the dPCA is compared to a PCA which operates directly on the dihedral angles, thus avoiding a doubling of variables. Various drawbacks of such a method which doesn't properly take the circularity of the variables into account are discussed. The dPCA is also compared to a version using the correlation matrix instead of the covariance matrix. Finally, it is concluded that, for the cases studied, the dPCA provides the most detailed low-dimensional representation of the free energy landscape. The chapter ends with a correlation analysis for the dihedral angles of heptaalanine which is compared to results from the literature, and some remarks about nonlinear PCAs.

Based on the dPCA, Chapter 3 presents a systematic approach to construct a low-dimensional free energy landscape from a classical molecular dynamics simulation. Demonstrating that a representation of the free energy landscape in too less dimension can lead to serious artifacts and oversimplifications of this intricate surface, it is attempted to answer the question on how many dimensions or PCs need to be taken into account in order to appropriately describe a given biomolecular process. It is shown that this dimensionality can be determined from the distribution and the autocorrelation of the PCs. Employing an 800 ns simulation of heptaalanine using geometric and kinetic clustering techniques, it is shown that a five-dimensional dPCA energy landscape is appropriate for reproducing the correct number, energy, and location of the system's metastable states

and barriers. After presenting several ways to visualize the free energy landscape using transition networks and a disconnectivity graph, we close the chapter with conclusions.

Having constructed low-dimensional free energy landscapes, the remaining aim is to construct dynamic models in this reduced dimensionality. Chapter 4 is concerned with the construction of low-dimensional models for peptide and protein dynamics from the point of view of modern nonlinear dynamics. Using methods from nonlinear time series analysis a deterministic model of the dynamics is developed and applied to molecular dynamics simulations of short alanine polypeptide chains. The well-established concept of the complexity of a dynamical system is applied to folding trajectories. Very interestingly, while the dimension of the free energy landscape increases with system size, the Kaplan-Yorke dimension may decrease. This suggests that the molecular dynamics generates less and less chaotic orbits as the length of the peptide chains increases. Furthermore, we introduce a mixed deterministic stochastic model for the conformational dynamics in reduced dimensions which is based on the estimation of the drift and diffusion vector fields of a Langevin equation. This makes it possible to, e.g., study nonequilibrium dynamics as relaxation to the folded state of a protein.

Finally, in Chapter 5 we apply some of the developed techniques to a larger system, namely a variant of the villin headpiece subdomain (HP-35 NleNle). Using many hundreds of molecular dynamics trajectories as obtained from Folding@home, we analyze the resulting free energy landscape for this system. In a next step we attempt to find a good dynamic model using the Langevin ansatz as described in the last chapter. We finally estimate folding times for this system, and conclude with an outlook. Conclusions are drawn at the end of each chapter.

Chapter 2

Dihedral Angle Principal Component Analysis

Classical molecular dynamics (MD) simulations have become a popular and powerful method to describe the structure, dynamics, and function of biomolecules in microscopic detail [2]. As MD simulations produce a considerable amount of data (i.e., $3M$ coordinates of all M atoms for each time step), there has been an increasing interest to develop methods to extract the "essential" information from the trajectory. For example, one often wants to represent the molecule's free energy surface (the "energy landscape" [8–10]) as a function of a few important coordinates (the "reaction coordinates"), which describe the essential physics of a biomolecular process such as protein folding or molecular recognition. The reduction of the dimensionality from $3M$ atom coordinates to a few collective degrees of freedom is therefore an active field of theoretical research [5, 13–27, 29–38].

Recently, it has been suggested to employ internal (instead of Cartesian) coordinates in a PCA [21–27]. In biomolecules, in particular the consideration of dihedral angles appears appealing, because other internal coordinates such as bond lengths and bond angles usually do not undergo changes of large amplitudes. Studying the reversible folding and unfolding of pentaalanine in explicit water, Mu *et al.* [25] showed that a PCA using Cartesian coordinates did not yield the correct rugged free energy landscape due to an artifact of the mixing of internal and overall motion. As internal coordinates naturally provide a correct separation of internal and overall dynamics, they proposed a method, referred to as dPCA, which is based on the dihedral angles (ϕ_n, ψ_n) of the pep-

tide backbone. To avoid the 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 $\cos\varphi_n$. In a recent comment [39] to Ref. [25], the concern was raised that the dPCA method may lead to spurious results because of the inherent constraints ($\sin^2\varphi_n + \cos^2\varphi_n = 1$) of the formulation. While it is straightforward to show that the problem described in Ref. [39] was caused by numerical artifacts due to insufficient sampling [40], the discussion nevertheless demonstrates the need for a thorough general analysis of the dPCA.

In this chapter, we present a comprehensive account of various theoretical issues underlying the dPCA method. We start with a brief introduction to the basics of MD simulation and derive the basic concepts of PCA. In an introduction to the circular statistics of angle variables we discuss the transformation from an angle to the unit circle proposed in Ref. [25], and demonstrate that the transformation amounts to a one-to-one representation of the original angle distribution. Adopting the (ϕ, ψ) distribution of trialanine as a simple but nontrivial example, the properties of the dPCA are discussed in detail. In particular, it is shown that in this case the dPCA results are equivalent to the results of a Cartesian PCA, and that the dPCA eigenvectors may be characterized in terms of the corresponding conformational changes of the peptide. Furthermore, we introduce a complex-valued version of the dPCA, which provides new insights on the PCA of circular variables. Adopting a 300 ns MD simulation of the folding of decaalanine, we carry out a critical comparison of the various methods. The next two section are devoted to Cartesian PCA and possible PCAs that are applied directly to the angular variables, respectively. Here, adopting an 800 ns MD simulation of heptaalanine, we study the similarities as well as the differences between these methods. We show that the dPCA provides the most detailed representation of the free energy landscapes of the peptides under concern. After a thorough correlation analysis for the dihedral angles of heptaalanine, we conclude this chapter with some remarks about nonlinear PCA methods that have been recently proposed in the literature.

2.1 Introduction to molecular dynamics simulation

Molecular Dynamics (MD) Simulation is concerned with modeling molecular motion in atomic detail. MD simulations can provide detailed information on the fluctuations and conformational changes of proteins and nucleic acids. A *potential* or *force field* is assumed for the description of the interactions between the particles,

$$-\frac{\partial V(r)}{\partial r_i} = F_i, \tag{2.1}$$

where $V(r)$ typically has the form

$$V = V_{\text{bonds}} + V_{\text{angles}} + V_{\text{dihedrals}} + V_{\text{Coulomb}} + V_{\text{vdW}} \tag{2.2}$$

$$V_{\text{bonds}} = \sum_{\text{bonds}} \frac{1}{2} k^b_{ij} (r_{ij} - b^0_{ij})^2 \tag{2.3}$$

$$V_{\text{angles}} = \sum_{\text{angles}} \frac{1}{2} k^\theta_{ij} (\theta_{ijk} - \theta^0_{ijk})^2 \tag{2.4}$$

$$V_{\text{dihedrals}} = \sum_{\text{dihedrals}} \frac{1}{2} k^\phi_{ijkl} \cos(n_{ijkl}(\phi_{ijkl} - \phi^0_{ijkl})) \tag{2.5}$$

$$V_{\text{Coulomb}} = \sum_{\text{pairs}} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}} \tag{2.6}$$

$$V_{\text{vdW}} = \sum_{\text{pairs}} \frac{A_{ij}}{r^{12}_{ij}} - \frac{B_{ij}}{r^6_{ij}}. \tag{2.7}$$

The first three terms are the interactions for the covalent bonds, the bond angles, and the dihedral angles, respectively. The non-bonded interactions are described by the last to terms, the electrostatic Coulomb and the Van der Waals interactions. The parameters of the potential, e.g. bond length, force constants or atomic charges, determine the quality of the force field. They are obtained by fitting simulation data against detailed quantum chemical calculations and experimental measurements.

The second main assumption is that the atoms follow classical Newtonian dynamics.

MD algorithms then iteratively solve the equations of motion

$$\begin{aligned} F_i(t) &= m_i a_i(t) && (2.8)\\ v_i(t+\frac{\Delta t}{2}) &= v_i(t-\frac{\Delta t}{2}) + a_i(t)\Delta t && (2.9)\\ r_i(t+\Delta t) &= r_i(t) + v_i(t+\frac{\Delta t}{2})\Delta t, && (2.10) \end{aligned}$$

where velocities v_i and positions r_i of the atoms are commonly calculated by variants of the Verlet algorithm such as the *leap-frog* method presented here. The method derives its name from the fact that the computation of velocities and positions successively alternates at $\frac{1}{2}\Delta t$ time step intervals.

The most time consuming part in an MD simulation is the evaluation of the forces acting on every particle, with the major computational effort spent for the non-bonded interactions. To avoid the calculation of all $O(N^2)$ electrostatic interactions between pairs of atoms, one e.g. uses a cutoff radius, where one neglects interactions beyond the cutoff distance or variations of the particle-mesh Ewald's (PME) summation.

Experimental methods as X-ray crystallography and nuclear magnetic resonance (NMR) can provide atomic detailed structures which are long-lived and can hence be probed experimentally. But conformational states which make fast transitions between each other are still a challenge to experiments. The structural mechanism of transitions normally cannot be resolved. MD can provide insight into these processes.

Similar to experiments MD can simulate different ensembles. The microcanonical ensemble (NVE) is realized by simply integrating Newton's equation (2.8) in time. The sum of kinetic and potential energy is constant and there is no exchange of temperature or pressure with the sourrounding. To simulate e.g. the canonical ensemble (NVT) the system is coupled to a temperature bath or thermostat. At certain time steps all particle's velocities are scaled by a factor in order to guarantee constant temperature. Temperature in an MD simulation is obtained by equating the particle's total kinetic energy to $\frac{1}{2}N_f k_b T(t)$,

$$E_{\text{kin}}(t) = \sum_{i=1}^{N} \frac{1}{2} m_i v_i^2(t) = \frac{1}{2} N_f k_b T(t), \qquad (2.11)$$

with N_f being the number of degrees of freedom of the system.

2.2 Definition and derivation of principal components

Principal component analysis [13] (PCA), also called quasiharmonic analysis or essential dynamics method [14–17], is one of the most popular methods to systematically reduce the dimensionality of a complex system. The approach is based on the covariance matrix, which provides information on the two-point correlations of the system. The PCA represents a linear transformation that diagonalizes the covariance matrix and thus removes the instantaneous linear correlations among the variables. Ordering the eigenvalues of the transformation decreasingly, it has been shown that a large part of the system's fluctuations can be described in terms of only a few principal components, which may serve as reaction coordinates [14–20].

In this section we want to establish the basics of PCA and prove that the eigenvectors obtained by PCA point into directions of maximal variance in a data set (see also [13]). The main idea of PCA is to reduce the dimensionality of a given data set. This is achieved by finding a set of a few new variables which are linearly uncorrelated and describe most of the variation present in the originally very high dimensional data. The starting point is the covariance matrix $\mathbf{\Sigma} = \{\sigma_{ij}\}$ of the multidimensional random variable $\mathbf{q}$. For example $\mathbf{q}(t)$ can be a trajectory obtained from an MD experiment yielding realizations of the random variable. We are now looking for a vector $\mathbf{v}$ such that the projection of the original data

$$V(t) = \mathbf{v} \cdot \mathbf{q}(t) = \sum_i v_i q_i(t) \tag{2.12}$$

has maximum variance. Henceforward, we often omit to explicitly note the time t. As a normalization constraint we require $\mathbf{v}$ to have unit length, as we want to avoid an infinite variance of (2.12). For the variance of V we find

$$\begin{aligned} \mathrm{var}[V] &= \mathrm{var}\left[\sum_i v_i q_i\right] \\ &= \sum_i v_i^2 \mathrm{var}[q_i] + 2\sum_{i<j} v_i v_j \mathrm{cov}[q_i, q_j] \\ &= \sum_i v_i^2 \sigma_{ii} + 2\sum_{i<j} v_i v_j \sigma_{ij} \\ &= \mathbf{v} \cdot \mathbf{\Sigma}\mathbf{v}, \end{aligned} \tag{2.13}$$

where we used $\sigma_{ij} = \sigma_{ji}$ in the last equation. Hence, we want to maximize $\mathbf{v} \cdot \mathbf{\Sigma v}$ subject to $\mathbf{v} \cdot \mathbf{v} = 1$. This is done by using the method of Lagrange multipliers. Differentiating

$$\mathbf{v} \cdot \mathbf{\Sigma v} - \lambda(\mathbf{v} \cdot \mathbf{v} - 1) \tag{2.14}$$

with respect to $\mathbf{v}$ gives

$$\mathbf{\Sigma v} - \lambda \mathbf{v} = \mathbf{0}, \tag{2.15}$$

which shows that an optimal $\mathbf{v}$ must be an eigenvector of $\mathbf{\Sigma}$ with eigenvalue λ. From

$$\mathrm{var}[V] = \mathbf{v} \cdot \mathbf{\Sigma v} = \mathbf{v} \cdot \lambda \mathbf{v} = \lambda \mathbf{v} \cdot \mathbf{v} = \lambda \tag{2.16}$$

we learn that λ must be as large as possible as we aim at maximizing the variance. Hence, it follows that the optimal λ is the largest eigenvalue λ_1 of the covariance matrix $\mathbf{\Sigma}$, and we denote its corresponding eigenvector by $\mathbf{v^{(1)}}$. We have just shown that $\mathbf{v^{(1)}}$ points into the direction of maximum variance of our data set.

The projections

$$V_i = \mathbf{v}^{(i)} \cdot \mathbf{q} \tag{2.17}$$

are called *principal components* of $\mathbf{q}$, where $\mathbf{v}^{(i)}$ is the eigenvector of $\mathbf{\Sigma}$ which corresponds to the ith largest eigenvalue λ_i. In a similar way as above one can show that for all i

$$\mathrm{var}[V_i] = \lambda_i \tag{2.18}$$

holds, and that V_i has maximum variance subject to being instantaneously linearly uncorrelated with $V_1, \ldots, V_{i-1}$, i.e.,

$$\langle (V_i(t) - \langle V_i \rangle)\,(V_j(t) - \langle V_j \rangle) \rangle = 0, \quad j = 1, \ldots, i-1. \tag{2.19}$$

2.3 Circular statistics

Dihedral angles $\varphi \in [0°, 360°[$ represent circular (or directional) data [41]. Unlike to the case of regular data $x \in]-\infty, \infty[$, the definition of a metric is not straightforward, which makes it difficult to calculate distances or means. For example, the regular data $x_1 = 10$

and $x_2 = 350$ clearly give $\Delta x = |x_2 - x_1| = 340$ and $\langle x \rangle = (10 + 350)/2 = 180$. Visual inspection of the corresponding angles $\varphi_1 = 10°$ and $\varphi_2 = 350°$, on the other hand, readily shows that $\Delta\varphi = 20° \neq |\varphi_2 - \varphi_1|$ and $\langle \varphi \rangle = 0° \neq (\varphi_1 + \varphi_2)/2$. To recover the standard rules to calculate distances and the mean, we may assume that $\varphi \in [-180°, 180°[$. Then $\varphi_1 = 10°$ and $\varphi_2 = -10°$, and we obtain $\Delta\varphi = |\varphi_2 - \varphi_1| = 20°$ and $\langle \varphi \rangle = (\varphi_1 + \varphi_2)/2 = 0°$. This example manifests the general property that, if the range of angles covered by the data set is smaller than 180°, we may simply shift the origin of the angle coordinates to the middle of this range and perform standard statistics.

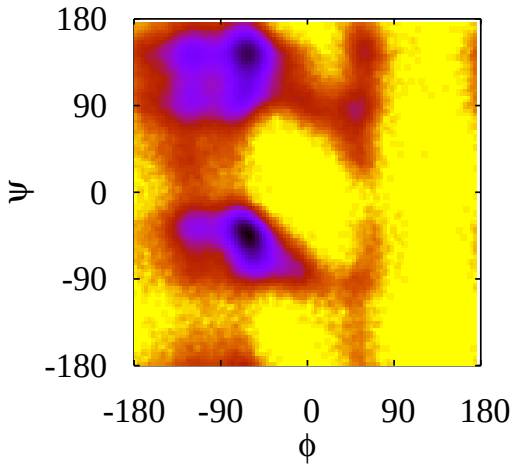

Figure 2.1: Typical Ramachandran plot for the backbone angles ϕ, ψ of a peptide backbone. The color code corresponds to the logarithmic population density.

The situation is more involved for "true" circular data whose range exceeds 180°. This is the case for folding biomolecules, since the ψ angle of the peptide backbone is typically distributed as $\psi_\alpha \approx -60° \pm 30°$ (for α_R helical conformations) and $\psi_\beta \approx 140° \pm 30°$ (for β extended conformations). If the values of the angles can be described by a normal distribution, one may employ the von Mise distribution [41], which represents the circular statistics' equivalent of the normal distribution for regular data. However, this method is not applicable to the description of conformational transitions, since the corresponding dihedral angle distributions typically can only be described by multi-peaked probability densities.

A general approach to circular statistics is obtained by representing the angle φ by its equivalent vector (x, y) on the unit circle. This amounts to the transformation

$$\varphi \mapsto \begin{cases} x = \cos\varphi \\ y = \sin\varphi \, . \end{cases} \tag{2.20}$$

Unlike to the periodic range of the angle coordinate φ, the vectors (x, y) are defined in a linear space, which means that we can define the usual Euclidean metric $\Delta^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2$ between any two vectors $(x_1, y_1)^T$ and $(x_2, y_2)^T$. The distance of two angles with an actually small distance, e.g. $\varphi_1 = 179°$ and $\varphi_2 = -179°$, is given by a small Δ in the (x, y)-space, since the corresponding vectors lie close on the unit circle. Hence, the problem of periodicity is circumvented. Furthermore, the vector representation of the angles allows us to unambiguously calculate mean values and other quantities. For example, to evaluate the mean of the angles φ_n, one simply calculates the sum of the corresponding vector components and then determines the mean angle by [41]

$$\tan\langle\varphi\rangle = \langle y\rangle / \langle x\rangle = \frac{\sum\limits_n \sin\varphi_n}{\sum\limits_n \cos\varphi_n}, \tag{2.21}$$

that is,

$$\langle\varphi\rangle = \begin{cases} \tan^{-1}\left(\dfrac{\sum\limits_n \sin\varphi_n}{\sum\limits_n \cos\varphi_n}\right), & \sum\limits_n \cos\varphi_n > 0 \\ \tan^{-1}\left(\dfrac{\sum\limits_n \sin\varphi_n}{\sum\limits_n \cos\varphi_n}\right) + 180°, & \sum\limits_n \cos\varphi_n < 0 \\ \dfrac{\pi}{2} \cdot \mathrm{sgn}\left(\sum\limits_{\mathrm{n}} \sin\varphi_{\mathrm{n}}\right), & \sum\limits_n \cos\varphi_n = 0. \end{cases} \tag{2.22}$$

Note that, even if the range of angles covered by the data set is smaller than 180° this definition of circular average can differ from the arithmetic average. For example, the arithmetic average of the 3 angles $0°, 0°, 90°$ is 30°, while the circular average equals to $\tan^{-1}\frac{1}{2} \approx 26.6°$.

Although the vector representation of angles in Eq. (2.20) appears straightforward and intuitively appealing, it has the peculiar property of doubling the variables: Given N angle coordinates φ_n, we obtain $2N$ Cartesian-like coordinates (x_n, y_n). In the example given

in Eq. (2.22), this does not lead to any problems, because in the end of the calculation we are able to calculate back from the averaged vector coordinates to the original angle coordinate, that is, the correctly averaged angle. Since Eq. (2.20) represents a nonlinear transformation, however, we will see that obtaining the peptide's angles in a direct way after a dPCA treatment of the data is not possible in general (see below). In this case, a subsequent analysis needs to be performed.

Having in mind to employ these coordinates for the description of peptide energy landscapes, the question arises of whether the resulting representation preserves the characteristics of the original energy landscapes. In particular, it is of interest if the number and structure of minima and transition states are preserved in the $2N$-dimensional (x_n, y_n) space. To answer these questions and to illustrate the properties of transformation (2.20), we consider a simple one-dimensional example described by the angular probability density (see Fig. 2.2A)

$$\rho(\varphi) = \frac{1}{2\pi}(1 - \cos 4\varphi) \tag{2.23}$$

with $\varphi \in [-180°, 180°[$. By construction, the density exhibits four maxima at $\varphi = \pm 45°, \pm 135°$. Employing transformation (2.20), we also want to express the density in terms of the transformed variables $x = \cos\varphi$ and $y = \sin\varphi$. Using that

$$\begin{aligned} \rho(\varphi) &= \frac{1}{2\pi}(1 - \cos 4\varphi) \\ &= \frac{1}{2\pi}(1 - \cos^2 2\varphi + \sin^2 2\varphi) \\ &= \frac{1}{2\pi} 2 \sin^2 2\varphi) \\ &= \frac{1}{\pi}(2\cos\varphi \sin\varphi))^2 \\ &= \frac{4}{\pi}\cos^2\varphi \sin^2\varphi, \end{aligned} \tag{2.24}$$

we obtain the corresponding probability density on a circle of unit radius

$$\rho(x, y) = \frac{4}{\pi} x^2 y^2 \delta(x^2 + y^2 - 1) \; . \tag{2.25}$$

The density plot of $\rho(x, y)$ displayed in Fig. 2.2B demonstrates that transformation (2.20) simply wraps the angular density $\rho(\varphi)$ around the circumference of the unit circle.

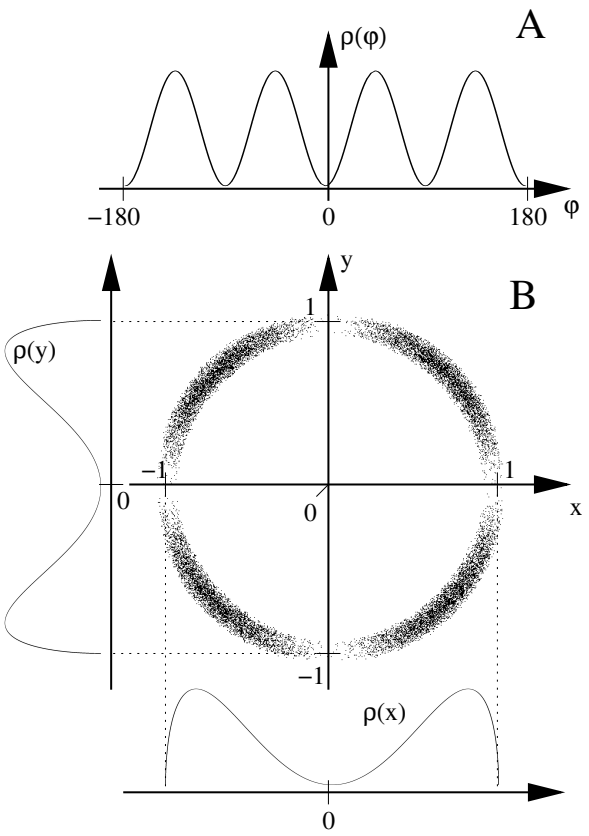

Figure 2.2: (A) Angular density $\rho(\varphi) = \frac{1}{2\pi}(1 - \cos 4\varphi)$. (B) Representation of $\rho(\varphi)$ through its probability density $\rho(x, y)$ on the unit circle (artificial width added for a better visualization). Also shown are the densities $\rho(x)$ and $\rho(y)$, which display the angular density along the single Cartesian-like variables x and y, respectively. Note that only $\rho(x, y)$ reproduces the correct number of extrema of $\rho(\varphi)$.

Hence, all features of $\rho(\varphi)$ are faithfully represented by $\rho(x, y)$, particularly the number and the structure of extrema. This is a consequence of the fact that transformation (2.20) is a bijection, which uniquely assigns each angle φ a corresponding vector (x, y) and vice versa.

We observe that this desirable feature is not obtained if we transform to only a single Cartesian-like variable, x or y. The corresponding densities

$$\rho(x) = \frac{8x^2\sqrt{1-x^2}}{\pi}, \tag{2.26}$$

$$\rho(y) = \frac{8y^2\sqrt{1-y^2}}{\pi} \tag{2.27}$$

are also shown in Fig. 2.2B and derived in the Appendix 6.1. As a consequence of the projection onto the x- or y-axis, each density exhibits only two instead of four maxima.

The above described properties of the one-dimensional example readily generalize to the N-dimensional case, $\varphi_n \mapsto (x_n, y_n)$. In direct generalization of the unit circle, the data points (x_n, y_n) are distributed on the surface of a $2N$-dimensional sphere with radius $\sqrt{N}$. This is because the distance of every data point $(x_1, y_1, \ldots, x_N, y_N)$ to the origin equals $(x_1^2 + y_1^2 + \cdots + x_N^2 + y_N^2)^{\frac{1}{2}} = (1 + \cdots + 1)^{\frac{1}{2}} = \sqrt{N}$. Since the transformation represents a bijection, there is a one-to-one correspondence between states in the N-dimensional angular space and in the $2N$-dimensional vector space. Again, the Euclidean metric of the $2N$-dimensional vector space guarantees that mean values and other quantities can be calculated easily.

We note that, alternatively to transformation (2.20), one may employ a complex representation $z_n = e^{\mathrm{i}\varphi_n}$ of the angles. As Euler's formula $e^{\mathrm{i}\varphi} = \cos\varphi + \mathrm{i}\sin\varphi$ provides a direct correspondence between the $2N$-dimensional real vectors $(x_1, y_1, \ldots, x_N, y_N)^T$ and the N-dimensional complex vectors $(z_1, \ldots, z_N)^T$, all considerations performed above can also be done using the complex representation. We will explore this idea in more detail in Sec. 2.7. Another straightforward way to use only N variables, is to use the angles φ_n directly. Therefore one may shift the origin of each angular variable in such a way that a minimal number of data points are at the periodic boundaries. We will also show the performance of such a method in Sec. 2.10.

2.4 Dihedral angle principal component analysis (dPCA)

Principal component analysis (PCA) is a well-established method to reduce the dimensionality of a high-dimensional data set [13]. In the case of molecular dynamics of M atoms, the basic idea is that the correlated internal motions are represented by the covariance matrix

$$\sigma_{ij} = \langle (q_i - \langle q_i \rangle)(q_j - \langle q_j \rangle) \rangle, \tag{2.28}$$

where $q_1, \ldots, q_{3M}$ are the mass-weighted Cartesian coordinates of the molecule and $\langle \ldots \rangle$ denotes the average over all sampled conformations [14–17]. By diagonalizing the covariance matrix we obtain $3M$ eigenvectors $\mathbf{v}^{(i)}$ and eigenvalues λ_i, which are rank-ordered descendingly, i.e., λ_1 represents the largest eigenvalue. The eigenvectors and eigenvalues of σ yield the modes of collective motion and their amplitudes, respectively. The principal components

$$V_i = \mathbf{v}^{(i)} \cdot \mathbf{q} \tag{2.29}$$

of the data $\mathbf{q} = (q_1, \ldots, q_{3M})^T$ can be used, for example, to represent the free energy surface of the system. Restricting ourselves to two dimensions, we obtain

$$\Delta G(V_1, V_2) = -k_\mathrm{B} T [\ln \rho(V_1, V_2) - \ln \rho_\mathrm{max}], \tag{2.30}$$

where ρ is an estimate of the probability density function obtained from a histogram of the data. ρ_max denotes the maximum of the density, which is subtracted to ensure that $\Delta G = 0$ for the lowest free energy minimum.

The basic idea of the dihedral angle principal component analysis (dPCA) proposed in Ref. [25] is to perform the PCA on sin- and cos-transformed dihedral angles

$$\begin{aligned} q_{2n-1} &= \cos \varphi_n, \\ q_{2n} &= \sin \varphi_n, \end{aligned} \tag{2.31}$$

where $n = 1, \ldots, N$ and N is the total number of peptide backbone and side-chain dihedral angles used in the analysis. Hence the covariance matrix (2.28) of the dPCA uses $2N$ variables q_n. The question then is whether the combination of the nonlinear transformation (2.31) and the subsequent PCA still gives a unique and faithful representation of

the initial angular data φ_n.

Let us first consider the above discussed example of a one-dimensional angular density $\rho(\varphi) = \frac{1}{2\pi}(1-\cos 4\varphi)$, which is mapped via transformation (2.31) on the two-dimensional density on the unit circle $\rho(x,y) = \frac{4x^2(1-x^2)}{\pi}\ \delta(x^2+y^2-1)$, where $x = q_1 = \cos\varphi$ and $y = q_2 = \sin\varphi$. Since in this case $\langle x\rangle = \langle y\rangle = \langle xy\rangle = 0$ and $\langle x^2\rangle = \langle y^2\rangle = \frac{1}{2}$, we find that the covariance matrix is diagonal with $\sigma_{11} = \sigma_{22} = \frac{1}{2}$. That is, we have degenerate eigenvalues $\lambda_{1/2} = \frac{1}{2}$ and may choose any two orthonormal vectors as eigenvectors. Choosing, e.g., the unit vectors $\mathbf{e}_x$ and $\mathbf{e}_y$, the PCA leaves the density $\rho(x,y)$ invariant, which —as discussed above— is a unique and faithful representation of the initial angular density $\rho(\varphi)$. In general, one does not obtain a diagonal covariance matrix for a one-dimensional angular density $\rho(\varphi)$ (e.g., for $\rho(\varphi) = \frac{1}{2\pi} + \frac{1}{9}\cos(\varphi) + \frac{1}{9}\sin(\varphi)$ we obtain $\sigma_{12} = -\frac{\pi^2}{81} \neq 0$). A sufficient condition for a diagonal covariance matrix for an N-dimensional angular density is that the latter factorizes in one-dimensional densities (i.e., $\rho(\varphi_1,\ldots,\varphi_N) = \rho(\varphi_1)\rho(\varphi_2)\cdots\rho(\varphi_N)$) and that $\langle\cos\varphi_n\rangle = 0$ or $\langle\sin\varphi_n\rangle = 0$ for all $n = 1,\ldots,N$. In these trivial cases, the dPCA method simply reduces to transformation (2.31).

2.5 A simple example - trialanine

The simplest nontrivial case of a dPCA occurs for a two-dimensional correlated angular density. As an example, we adopt trialanine whose conformation can be characterized by a single pair of (ϕ,ψ) backbone dihedral angles (see Fig. 2.3). Trialanine (Ala_3) in aqueous

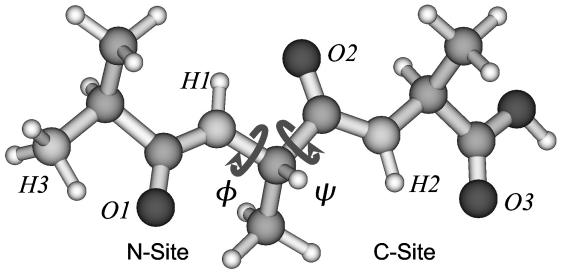

Figure 2.3: Molecular structure of trialanine.

solution is a model peptide which has been the subject of numerous experimental [42–45] and computational [46–48] studies. To generate the angular distribution of (ϕ,ψ) of trialanine, we performed a 100 ns MD simulation at 300 K. We used the GROMACS

program suite [49,50], the GROMOS96 force field 43a1 [51], the simple point charge (SPC) water model [52], and a particle-mesh Ewald [53] treatment of the electrostatics. Details of the simulation can be found in Ref. [47]. Figure 2.4A shows the (ϕ, ψ) distribution

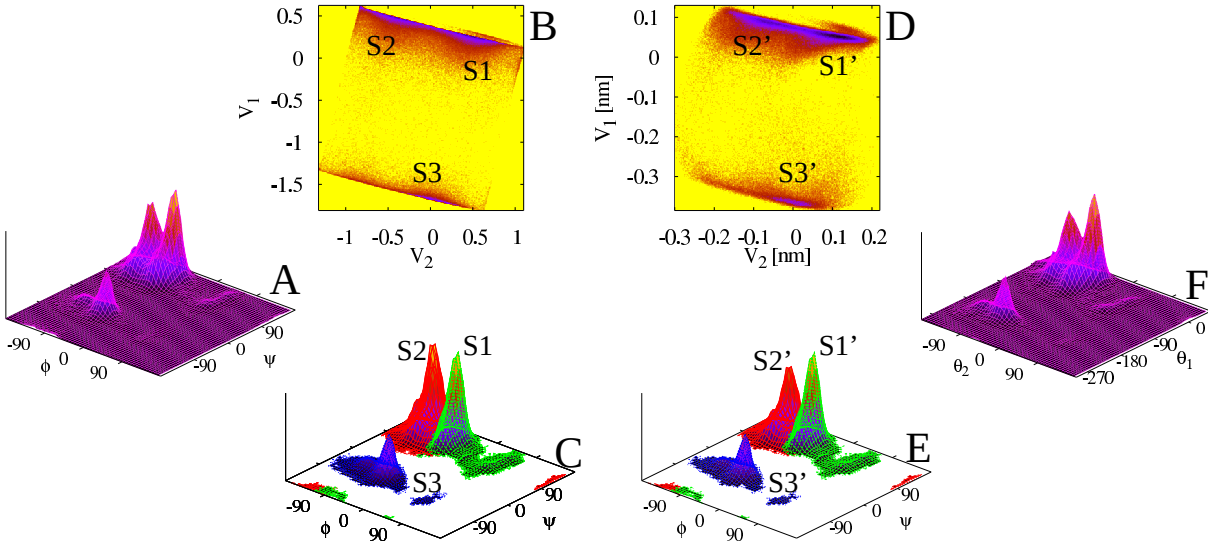

Figure 2.4: (A) Ramachandran (ϕ, ψ) probability distribution of Ala_3 in water as obtained from a 100 ns MD simulation. Performing a dPCA, the resulting free energy landscape along the first two principal components is shown in (B), the (ϕ, ψ) distributions pertaining to the labeled energy minima is shown in (C). Panels (D) and (E) show the corresponding results obtained for a Cartesian PCA. Panel (F) displays the (θ_1, θ_2) distribution obtained from the complex dPCA.

obtained from the simulation, which predicts that mainly three conformational states are populated: the right-handed helix conformation α_R (15 %), the extended conformation β (39 %), and the poly-L-proline II (P_{II}) helix-like conformation (42 %). Although recent experimental data [45] indicate that the simulation overestimates the populations of α_R and β, we nevertheless adopt the MD data as a simple yet nontrivial example to illustrate the performance of the dPCA method.

Performing the dPCA on the (ϕ, ψ) data, we consider the four variables $q_1 = \cos\phi$, $q_2 = \sin\phi$, $q_3 = \cos\psi$, and $q_4 = \sin\psi$. Diagonalization of the resulting covariance matrix yields four principal components $V_1, \ldots, V_4$, which contribute 51, 24, 15, and 10 % to the overall fluctuations of the system, respectively. To characterize the principal components, Fig. 2.5 shows their one-dimensional probability densities. Only the first two distributions are found to exhibit multiple peaks, while the other two are approximately unimodal. Hence

we may expect that the conformational states shown by the angular distribution of (ϕ, ψ) in Fig. 2.4A can be accounted for by the first two principal components.

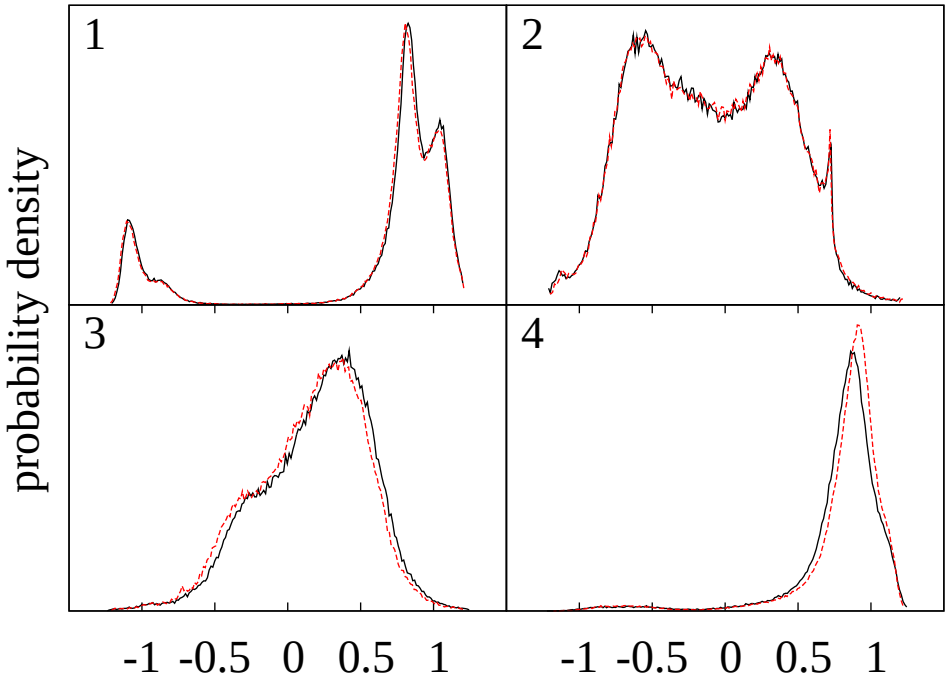

Figure 2.5: Probability densities of the four principal components obtained from the sin/cos (full lines) and the complex (dashed lines) dPCA of trialanine, respectively.

If we assume that V_1 and V_2 are independent (i.e., $\rho(V_1, V_2) = \rho(V_1)\rho(V_2)$), the three peaks found for $\rho(V_1)$ as well as for $\rho(V_2)$ give rise to $3 \times 3 = 9$ peaks of $\rho(V_1, V_2)$. To identify possible correlations, Fig. 2.4B shows the two-dimensional density along the first two principal components. For the sake of better visibility, we have chosen a logarithmic representation, thus showing the free energy landscape [Eq. (2.30)] of the system. The figure exhibits three (instead of nine) well-defined minima labeled S1, S2, and S3, revealing that the first two principal components are indeed strongly dependent. To identify the corresponding three conformational states, we have back-calculated the (ϕ, ψ) distributions of the minima from the trajectory [54]. As shown in Fig. 2.4C as well as by Table 2.1, the minima S1, S2, and S3 clearly correspond to $\mathrm{P_{II}}$, β, and α_{R}, respectively. A closer analysis reveals, that also fine details of the conformational distribution can be discriminated by the first two principal components. For example, the shoulder on the left side of the α_{R} state in Fig. 2.4A corresponds to the region around $V2 \approx -0.9$ of the S3

minimum. Moreover, the minor (3 %) population of the left-handed helix conformation α_{L} at $\phi \approx 60^{\circ}$ corresponds to the small orange region (outside of the square) of the S1 minimum.

It is instructive to compare the above results obtained by the dPCA to the outcome of a standard PCA using Cartesian coordinates. Restricting the analysis to the atoms CONH-$CHCH_3$-CONH around the central (ϕ, ψ) dihedral angles of trialanine, the first four principal components contribute 47, 28, 15, and 8 % to the overall fluctuations, respectively, and exhibit one-dimensional probability densities that closely resemble the ones obtained by the dPCA (data not shown). Figure 2.4D shows the resulting free energy surface along the first two principal components, which looks quite similar to the dPCA result. The three minima S1', S2', and S3' are identified in Fig. 2.4E as the conformational states $\mathrm{P_{II}}$, β, and α_{R}. Again, also the details of the conformational distribution such as the α_{L} state are resolved by the first two principal components.

In summary, it has been shown that both the Cartesian PCA and the dPCA reproduced the correct conformational distribution of the MD trajectory of trialanine. In both cases, the first two principal components were sufficient to resolve most details. Although only four coordinates were used, the dPCA was found to be equivalent to the Cartesian PCA using 33 coordinates.

	MD data		dPCA		Cartesian PCA	
state	P [%]	(ϕ, ψ) [deg]	P [%]	(ϕ, ψ) [deg]	P [%]	(ϕ, ψ) [deg]
$\mathrm{P_{II}}$	42	-67,132	45	-63,131	47	-64,132
β	39	-121,131	40	-121,131	38	-122,130
α_{R}	15	-75,-45	16	-74,-46	16	-75,-46

Table 2.1: Conformational states $\mathrm{P_{II}}$, β, and α_{R} of trialanine in water, characterized by their population probability P and the average dihedral angles (ϕ, ψ). The results from the dPCA and the Cartesian PCA are compared to reference data obtained directly from the MD simulation.

2.6 Interpretation of eigenvectors

In the simple example above, Fig. 2.4 demonstrates that the first two principal components V_1 and V_2 (or, equivalently, the first two eigenvectors $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$) are associated with

motions along the ψ and the ϕ dihedral angles, respectively. In the case of the Cartesian PCA, the structural changes of the molecule along the principal components are readily illustrated, even for high-dimensional systems. From

$$\begin{aligned} V_i &= \mathbf{v}^{(i)} \cdot \mathbf{q} \\ &= v_1^{(i)} q_1 + v_2^{(i)} q_2 + v_3^{(i)} q_3 + \ldots + v_{3M-2}^{(i)} q_{3M-2} + v_{3M-1}^{(i)} q_{3M-1} + v_{3M}^{(i)} q_{3M} \end{aligned}$$

we see that, e.g., the first three components $v_1^{(i)}$, $v_2^{(i)}$, and $v_3^{(i)}$ of the eigenvector $\mathbf{v}^{(i)}$ simply reflect the influence of the x, y, and z coordinates of the first atom on the ith principal component. Hence,

$$\Delta_1^{(i)} = (v_1^{(i)})^2 + (v_2^{(i)})^2 + (v_3^{(i)})^2 \tag{2.32}$$

is a suitable measure of this influence. The quantities $\Delta_2^{(i)}, \ldots, \Delta_M^{(i)}$ are defined analogously.

In the dPCA, the principal components are given by

$$\begin{aligned} V_k &= \mathbf{v}^{(k)} \cdot \mathbf{q} \\ &= v_1^{(k)} \cos\varphi_1 + v_2^{(k)} \sin\varphi_1 + \ldots + v_{2N-1}^{(k)} \cos\varphi_N + v_{2N}^{(k)} \sin\varphi_N. \end{aligned} \tag{2.33}$$

In direct analogy to Eq. (2.32), we may define

$$\Delta_1^{(k)} = (v_1^{(k)})^2 + (v_2^{(k)})^2 \tag{2.34}$$

as a measure of the influence of angle φ_1 on the principal component V_k (and similarly $\Delta_2^{(k)}, \ldots, \Delta_N^{(k)}$ for the other angles). The definition implies that $\sum_n \Delta_n^{(k)} = 1$, since the length of each eigenvector is one. Hence $\Delta_n^{(k)}$ can be considered as the percentage of the effect of the angle φ_n on the principal component V_k. Furthermore, Eq. (2.33) assures that only structural rearrangements along angles with nonzero $\Delta_n^{(k)}$ may change the value of V_k.

To demonstrate the usefulness of definition (2.34), we again invoke our example of trialanine with angles ϕ ($n = 1$) and ψ ($n = 2$), and consider the quantities $\Delta_n^{(k)}$ describing the effect of these angles on the four principal components ($k = 1, \ldots, 4$), see Fig. 2.6. We clearly see that the dihedral angle ϕ has almost no influence on V_1 ($\Delta_1^{(1)} \approx 0$), whereas ψ

has a very large one ($\Delta_2^{(1)} \approx 1$). As a consequence, the first principal component allows us to separate conformations with a different angle ψ, but does not separate conformations which differ in ϕ. Indeed, Fig. 2.4B reveals that V_1 accounts essentially for the $\alpha \leftrightarrow \beta/\mathrm{P_{II}}$ transition along ψ, but hardly separates conformations with different ϕ, such as β and $\mathrm{P_{II}}$. Considering the second principal component V_2, we obtain $\Delta_1^{(2)} \approx 1$ and $\Delta_2^{(2)} \approx 0$. This is again in agreement with Fig. 2.4B, which shows that the second principal component accounts essentially for transitions along ϕ. Recalling that V_1, V_2, V_3, and V_4, contribute 51, 24, 15, and 10 % to the overall fluctuations, respectively, the $\beta \leftrightarrow \mathrm{P_{II}}$ transitions described by the second principal component represent a much smaller conformational change than the $\alpha \leftrightarrow \beta/\mathrm{P_{II}}$ transitions described by V_1. Similarly, although the $\Delta_n^{(k)}$ of the third and forth principal component are quite similar to the previous ones, they only account for fluctuations within a conformational state and are therefore of minor importance in a conformational analysis.

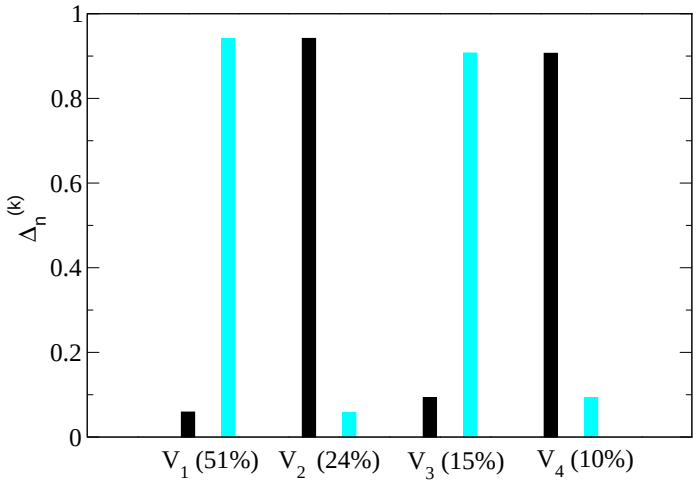

Figure 2.6: Influence of the dihedral angles ϕ (black bars) and ψ (gray bars) on the principal component V_k ($k = 1, \ldots, 4$) of the cos/sin dPCA of trialanine. Shown are the quantities $\Delta_1^{(k)}$ (for ϕ) and $\Delta_2^{(k)}$ (for ψ) defined in Eq. (2.34), representing the percentage of the effect of the two dihedral angels on V_k. Also shown are the contributions (in %) of each principal component to the overall fluctuations of the system.

2.7 Complex dPCA

Alternatively to the sin/cos transformation in Eq. (2.31) which maps N angles on $2N$ real numbers, one may also transform from the angles φ_n to the complex numbers

$$z_n = e^{\mathrm{i}\varphi_n} \quad (n = 1, \ldots, N), \tag{2.35}$$

which give an N-dimensional complex vector $\mathbf{z} = (z_1, z_2, \ldots, z_N)^T$. In what follows, we develop a dPCA based on this complex data ("complex dPCA"), and discuss its relation to the real-valued dPCA ("sin/cos dPCA") considered above.

The covariance matrix pertaining to the complex variables z_n is defined as

$$C_{mn} = \langle (z_m - \langle z_m \rangle)(z_n^* - \langle z_n^* \rangle) \rangle \tag{2.36}$$

with $m, n = 1, \ldots, N$, and z^* being the complex conjugate of z. Being an in principle observable quantity, C is a Hermitian matrix with N real-valued eigenvalues μ_n and N complex eigenvectors $\mathbf{w}^{(n)}$

$$C\mathbf{w}^{(n)} = \mu_n \mathbf{w}^{(n)}, \tag{2.37}$$

where the eigenvectors are unique up to a phase θ_0. We define the complex principal components to be

$$W_n = {\mathbf{w}^{(n)}}^T \mathbf{z} = r_n e^{\mathrm{i}(\theta_n + \theta_0)}, \tag{2.38}$$

where we use vector-vector multiplication instead of a Hermitian inner product (see Appendix for details). Two nice features of the complex dPCA are readily evident. First, the complex representation of N angular variables directly results in N eigenvalues and eigenvectors, that is, there is no doubling of variables as in the sin/cos dPCA. Second, the representation of the complex principal components by their weights r_n and angles θ_n in Eq. (2.38) may facilitate their direct interpretation in terms of simple physical variables.

From Euler's formula $e^{\mathrm{i}\varphi} = \cos\varphi + \mathrm{i}\sin\varphi$, one would expect an evident correspondence between the sin/cos and the complex dPCA. That is, there should be a relation between the N complex eigenvectors $\mathbf{w}^{(n)}$ and the $2N$ real eigenvectors $\mathbf{v}^{(k)}$. Furthermore, the N real eigenvalues μ_n of the complex dPCA should be related to the $2N$ real eigenvalues λ_k of the sin/cos dPCA. However, this general correspondence turned out to be less obvious than

expected (see Appendix 6.2), and we were only able to find an analytical relation in some limiting cases. In these cases, one indeed may construct suitably normalized eigenvectors $\mathbf{w}^{(n)}$ such that the real and imaginary parts of the resulting principal components W_n of the complex dPCA are equal to the $2N$ principal components V_k of the sin/cos dPCA. In other words, for every $n \in \{1, \ldots, N\}$ there are two indices $k_n, k'_n \in \{1, \ldots, 2N\}$ such that

$$\text{Re } W_n = V_{k_n}, \quad \text{Im } W_n = V_{k'_n}, \tag{2.39}$$

and the union of the indices k_n, k'_n gives the complete set $\{1, \ldots, 2N\}$. Moreover, the eigenvalues μ_n of the complex dPCA are given by the sum of the two corresponding eigenvalues λ_{k_n} and $\lambda_{k'_n}$ of the sin/cos dPCA

$$\mu_n = \lambda_{k_n} + \lambda_{k'_n}. \tag{2.40}$$

Apart from the limiting cases of completely uncorrelated and completely correlated variables, we could not establish general conditions under which Eqs. (2.39) and (2.40) hold. Empirically, Eq. (2.40) was always satisfied, while Eq. (2.39) was found to hold in many (but not all) cases under consideration, see Figs. 2.5 and 2.9 below. We note that even in numerical studies it may be cumbersome to establish the correspondences, since the accuracy of (2.39) and (2.40) depends on the number of data points one uses to calculate the covariance matrices in both methods, i.e., on the overall sampling of the MD trajectory.

To demonstrate the performance of the complex dPCA, we first apply it to the above discussed example of trialanine. Comparing the $2N = 4$ eigenvalues of the sin/cos dPCA $\lambda_1, \ldots, \lambda_4$ to the two eigenvalues μ_1 and μ_2 of the complex dPCA, we obtain

$$\begin{aligned} \mu_1 &= 0.630 = 0.489 + 0.141 = \lambda_1 + \lambda_3, \\ \mu_2 &= 0.338 = 0.237 + 0.101 = \lambda_2 + \lambda_4, \end{aligned}$$

that is, equation (2.40) is fulfilled. Choosing suitable normalization constants θ_0 for the

complex eigenvectors, we furthermore find the correspondence

$$\begin{aligned} \mathrm{Re}\, W_1 &\approx V_1, \quad \mathrm{Re}\, W_2 \approx V_2, \\ \mathrm{Im}\, W_1 &\approx V_3, \quad \mathrm{Im}\, W_2 \approx V_4. \end{aligned}$$

As shown by the probability densities of the principal components in Fig. 2.5, both formulations lead to virtually identical principal components.

Finally, it is interesting to study if the representation of the complex principal components by their weights r_n and angles θ_n in Eq. (2.38) facilitates their interpretation. In the case of our trialanine data, it turns out that the weights are approximately constant, i.e., $r_1 \approx r_2 \approx 1$. Hence, the probability distribution of the two angles (θ_1, θ_2) contains all the conformational fluctuations of the data. Indeed, Fig. 2.4 reveals that $\rho(\theta_1, \theta_2)$ is almost identical to the original (ϕ, ψ) density from the MD simulation. In this simple case, the complex dPCA obviously has managed to completely identify the underlying structure of the data.

2.8 Energy landscape of decaalanine

We finally wish to present an example which demonstrates the potential of the dPCA method to represent the true multidimensional energy landscape of a folding biomolecule. Following earlier work on the folding of alanine peptides [25,36,45], we choose decaalanine (Ala_{10}) in aqueous solution. Employing similar conditions as in the case of trialanine described above (GROMOS96 force field 43a1 [51], SPC water model [52], and particle-mesh Ewald [53] treatment of the electrostatics), we ran a 300 ns trajectory of Ala_{10} at 300 K and saved every 0.4 ps the coordinates for analysis.

Let us first consider the free energy landscape ΔG [Eq. (2.30)] obtained from a PCA using all Cartesian coordinates of the system. The calculations of $\Delta G(V_1, V_2)$ and $\Delta G(V_3, V_4)$ presented in Fig. 2.7A and B show that the resulting energy landscape is rather unstructured and essentially single-peaked, indicating a single folded state and a random ensemble of unfolded conformational states. However, as will be discussed in detail in the next section, this smooth appearance of the energy landscape in the Cartesian PCA merely represents an artifact of the mixing of internal and overall motion.

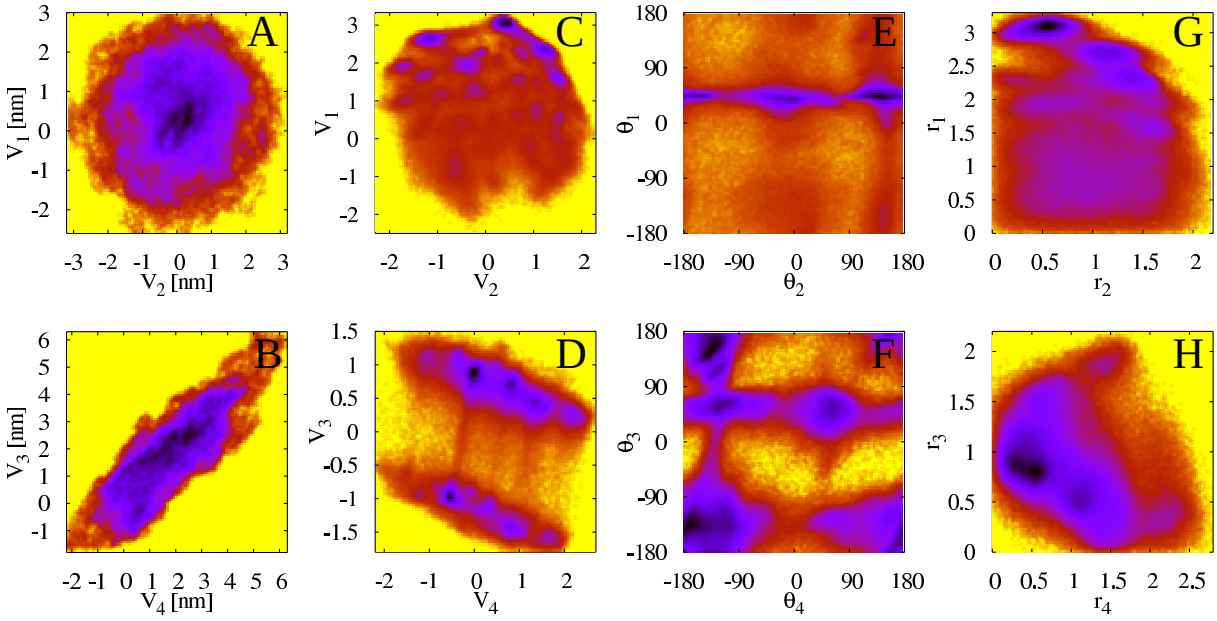

Figure 2.7: Free energy landscapes of Ala_{10} in water as obtained from a 300 ns MD simulation. The first column, (A) and (B), shows the results along the first four principal components obtained from a Cartesian PCA, the second column, (C) and (D), the corresponding landscapes calculated from the sin/cos dPCA. Panels (E), (F), (G), and (H) display the landscapes along the angles (θ_1, θ_2) and (θ_3, θ_4) and the weights (r_1, r_2) and (r_3, r_4) of the complex dPCA, respectively.

This becomes clear when a sin/cos dPCA of the $N = 18$ inner backbone dihedral angles $\{\varphi_n\} = \{\psi_1, \phi_2, \psi_2, \ldots, \phi_9, \psi_9, \phi_{10}\}$ is performed. The resulting dPCA free energy surfaces $\Delta G(V_1, V_2)$ and $\Delta G(V_3, V_4)$ shown in Fig. 2.7C and D exhibit numerous well-separated minima, which correspond to specific conformational structures. By back-calculating from the dPCA free energy minima to the underlying backbone dihedral angles of all residues [54], we are able to discriminate and characterize 15 such states [55]. The most populated ones are the all α_{R}-helical conformation (8 %), a state (15 %) with the inner seven residues in α_{R} (and the remaining residues in $\beta/\mathrm{P_{II}}$), and two states (each 8 %) with six inner residues in α_{R}. Well-defined conformational states are also found in the unfolded part of the free energy landscape, revealing that the unfolded state of decaalanine is rather structured than random.

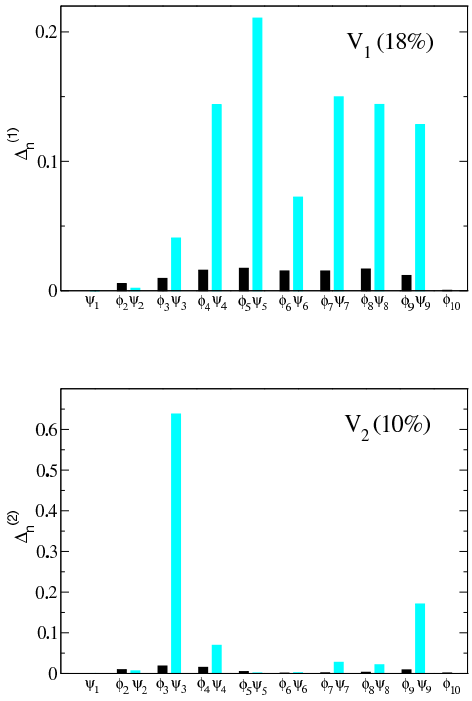

Figure 2.8: Influence of the 18 inner backbone dihedral angles $\{\varphi_n\} = \{\psi_1, \phi_2, \psi_2, \ldots, \phi_9, \psi_9, \phi_{10}\}$ on the first two principal component V_1 and V_2 of the cos/sin dPCA of $\mathrm{Ala_{10}}$. Shown are the quantities $\Delta_n^{(1)}$ (for V_1) and $\Delta_n^{(2)}$ (for V_2) defined in Eq. (2.34), representing the percentage of the effect of the dihedral angles on V_k. The black and gray bars correspond to the ϕ and ψ angles, respectively. Also shown are the contributions (in %) of each principal component to the overall fluctuations of the system.

To obtain an interpretation of the kth principal component in terms of the dihedral angles φ_n, Fig. 2.8 shows the quantities $\Delta_n^{(k)}$ defined in Eq. (2.34) which describe the effect of these angles on the first two principal components. The first principal component V_1 is clearly dominated by motion along the ψ angles (gray bars), while fluctuations of the ϕ angles (black bars) hardly contribute. Hence, going along V_1 we will find conformations which mainly differ in ψ angles. Considering the second principal component V_2, we find a dominant $\Delta_n^{(2)}$ for the angle ψ_3 (and a smaller value for ψ_9), revealing that V_2 separates mainly conformation that differ in ψ_3. Similarly, the $\Delta_n^{(k)}$ obtained for next few principal components are dominated by the contribution of a single ψ angle. For example, we find that $\Delta_n^{(3)}$, $\Delta_n^{(4)}$, $\Delta_n^{(5)}$, and $\Delta_n^{(6)}$ depend mostly on the angles ψ_2, ψ_9, ψ_4 (and ψ_8), and ψ_5, respectively (data not shown). Together with the percentage of the fluctuations (18, 10, 8, 7, 6, and 5 % for $V_1, \ldots, V_6$) the quantities $\Delta_n^{(k)}$ therefore give a quick and valuable interpretation of the conformational changes along the principal components V_k.

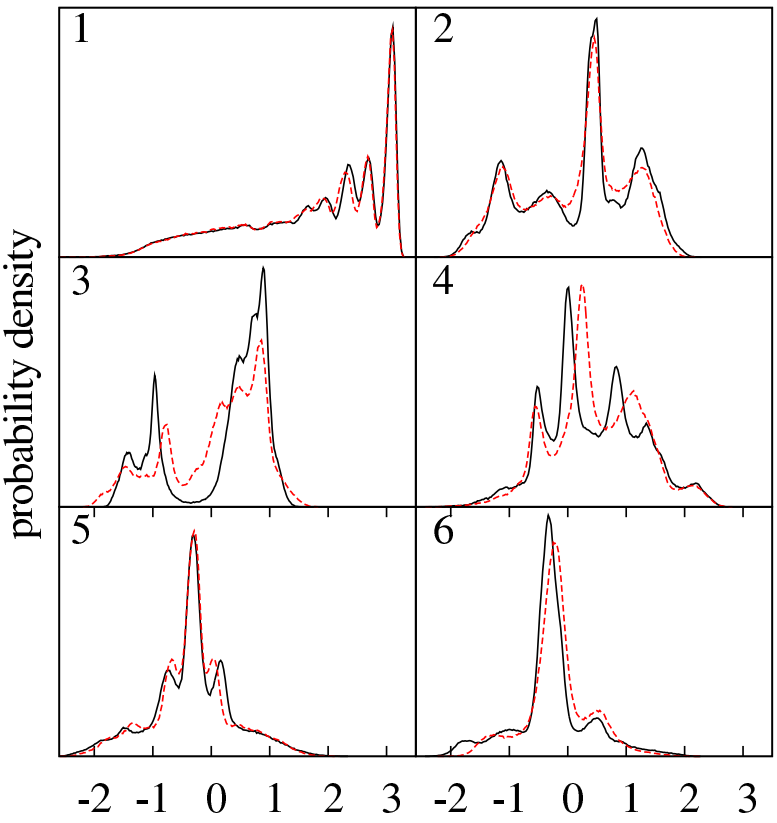

Figure 2.9: Probability densities of the first six principal components obtained from the sin/cos (full lines) and the complex (dashed lines) dPCA of Ala_{10}, respectively.

It is interesting to compare the above results to the outcome of a complex dPCA of the Ala_{10} trajectory. To check the similarity of the complex and the sin/cos dPCA in this case, Fig. 2.9 compares the distributions of the sin/cos principal components V_k to the distributions of the corresponding principal components Re W_n and Im W_n, obtained from the complex dPCA using suitably normalized eigenvectors. Although we find good overall agreement, the correspondence (2.39) is not perfect in all cases (see Appendix). Finally, we wish to investigate whether the polar representation (2.38) of the complex principal components facilitates the interpretation of the energy landscape of Ala_{10}. To this end, Fig. 2.7E-H shows the free energy surfaces (E) $\Delta G(\theta_1, \theta_2)$, (F) $\Delta G(\theta_3, \theta_4)$, (G) $\Delta G(r_1, r_2)$, and (H) $\Delta G(r_3, r_4)$. Similarly as found for Ala_3, the energy landscape is only little structured along the weights r_n (mainly along r_1), thus leaving the main information on the conformational states to the angles θ_n (mainly θ_2, θ_3, and θ_4). A closer analysis reveals, e.g., that θ_2 separates conformational states with different dihedral angle ψ_3, while θ_3 separates conformations with different dihedral angle ψ_2. Unlike to the simpler case of trialanine, where the (θ_1, θ_2) representation of the complex dPCA was found to directly reproduce the original (ϕ, ψ) distribution, however, the polar principal components of Ala_{10} appear to be equivalent to the results of the standard sin/cos dPCA. Roughly speaking, in both formulations we need about the same number of principal components to identify the same number of conformational states.

2.9 Cartesian PCA

In section 2.5 Cartesian PCA was found to be equivalent to dPCA for the trialanine system. Going to longer peptide chains which adopt much more conformational states, it has been demonstrated by Mu et al. [25] that a PCA on the Cartesian coordinates fails to reveal the true structure of the free energy landscape in the case of pentaalanine. The smooth appearance in the Cartesian PCA represents an artifact of the mixing of internal and overall motion. In this section we discuss the several problems of Cartesian PCA for very flexible peptides.

In order to study dynamic structural changes of a peptide by a Cartesian PCA one has to remove rotational and translational motion from an MD trajectory. This is usually done by least-squares superpositioning. The full trajectory is fitted to a single reference

structure. After the trivial removal of translational motion by subtracting the center of mass from all configurations of the MD run, the overall rotation can be removed at each time t by minimizing the function

$$\Delta(t) = \sum_{n=1}^{\#\text{atoms}} m_n \|R(t) r_n(t) - c_n\|^2 \tag{2.41}$$

with respect to the rotation matrix $R(t) \in \mathbb{R}^{3\times 3}$, where $r_n(t)$ is the position of the nth atom at time t with its mass being m_n, and the c_n's are the one chosen reference structure for all the MD trajectory. In Fig. 2.10A and B we see the free energy landscape $\Delta G(V_1, V_2)$ for Ala_5 where we fitted the trajectory to the starting configuration of the MD simulation, $c_n = r_n(0)$ for all n, which is a mostly unfolded β structure. Depending on the reference structure chosen we obtain slightly different landscapes, but anyways the various conformations of the peptide cannot be resolved. In contrast to this, the dPCA landscape shows many multiple peaks (Fig. 2.10C and D) which correspond to different conformational states of the peptide. See [25] for a thorough discussion. Note that the dPCA landscape is unique in the sense that its shape does not depend on a reference structure as it is explicitly constructed from internal coordinates, the dihedral angles.

The failure of Cartesian PCA to provide the true free energy landscape seems to be ubiquitous for small very flexible peptides, with the exception of the conformationally trivial trialanine. We will see more examples of landscapes obtained by a Cartesian PCA on larger systems in Chapter 5 of this thesis. But let us now get to the root of the problem. While fitting of an MD trajectory is straightforward in the case of small fluctuations around a mean structure, it is not possible to define an appropriate reference structure of a molecule undergoing large amplitude motion. The fit will alter the coordinates and, for example, artificial correlations between atoms may be introduced.

Another problem is the least-squares superposition itself. The least-squares treatment implicitly requires that atoms are uncorrelated and that each atom has the same variance, which is not given of course. Theobald et al. [56, 57] propose a maximum likelihood method to overcome these drawbacks. They show that their maximum likelihood superposition provides markedly more accurate structural correlations than those extracted from least-squares superpositions. Their method is implemented in the THESEUS package [58].

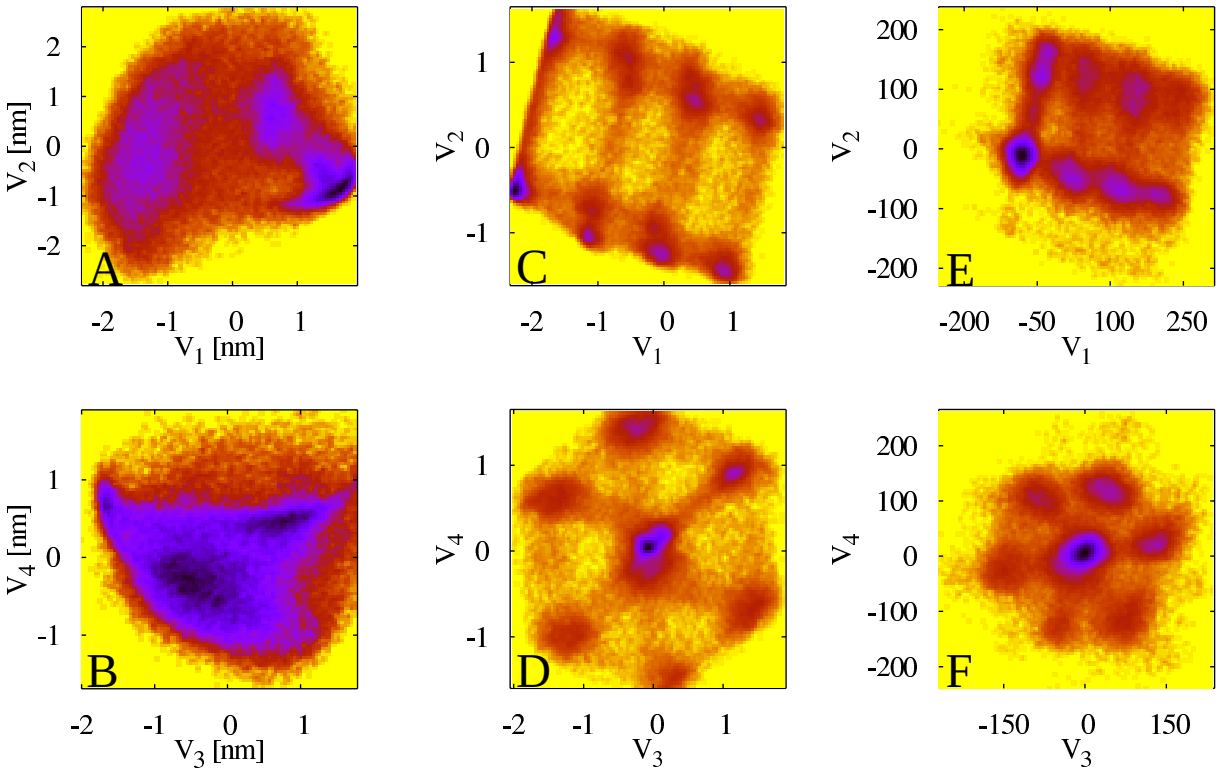

Figure 2.10: Free energy landscapes of Ala_5 in water as obtained from a 100 ns MD simulation. The first column, (A) and (B), shows the results along the first four principal components obtained from a Cartesian PCA, the second column, (C) and (D), the corresponding landscapes calculated from the sin/cos dPCA. Panels (E) and (F) display the landscape along the principal components of a PCA directly on the dihedral angels without prior sin/cos transformation.

However the problem of the absence of an appropriate reference structure remains. Nevertheless we tried out a maximum likelihood fit in order to see if we obtain a more detailed free energy landscape than with a least-squares fit. Comparing Figs. 2.11 (A) and (B),

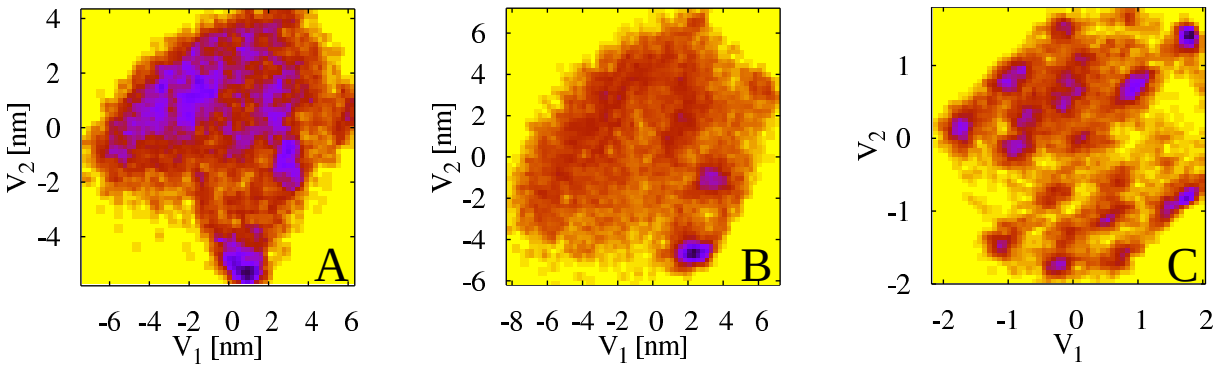

Figure 2.11: Comparison of different superposition methods for Cartesian PCA, and dPCA free energy landscape for Ala_7 as obtained from a 200 ns MD simulation. (A) least-squares fit, (B) THESEUS maximum likelihood fit, and (C) dPCA (no fit necessary).

it seems that a more accurate fit does not provide a more detailed picture of the free energy landscape when Cartesian coordinates are used. As already seen for Ala_5 above, also the dPCA free energy landscape $\Delta G(V_1, V_2)$ of Ala_7 as shown in 2.11 (C), provides a much more detailed picture. Again the distinct free energy minima correspond to various conformational states of the system. The two superposition methods seem to fail to correctly describe the correlations between the atoms in the case of peptides undergoing large amplitude motion, and hence result in artificial free energy landscapes.

It is interesting to take a look at the covariance matrices of the Cartesian atoms for Ala_7 as visualized in Fig. 2.12 for the two different superposition methods. Even though they seem qualitatively similar, covariances are stronger among the first five as well as among the last five atoms for the maximum likelihood fit in panel (B). Also certain covariances are more pronounced in the least squares fit (A), but also vice versa. Nevertheless one can say that these differences are not significant for the construction of the free energy landscapes $\Delta G(V_1, V_2)$ (Fig. 2.11A and B) which are constructed along the eigenvectors of the respective covariance matrix. The two landscapes resemble each other and both provide an artificial picture of the true free energy landscape of Ala_7.

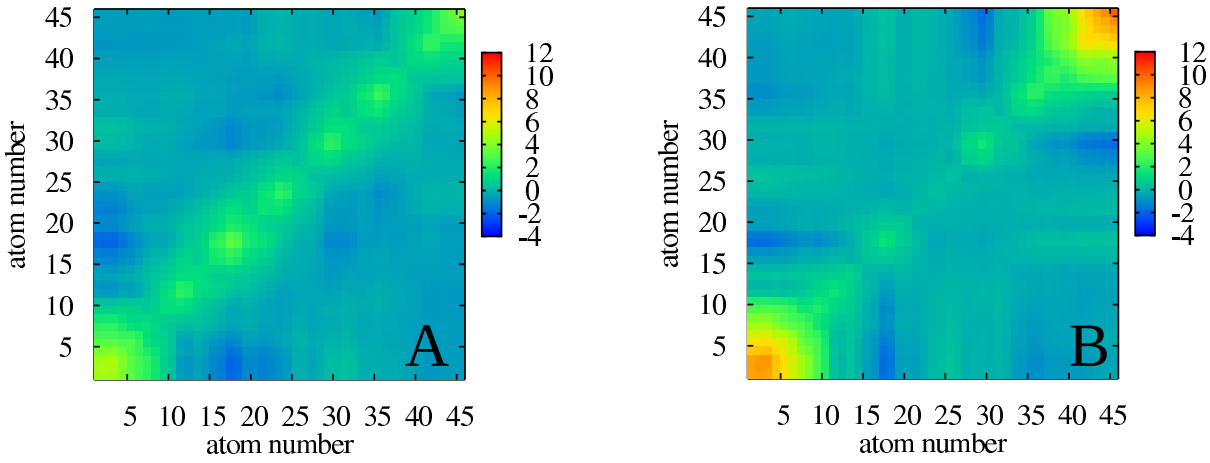

Figure 2.12: Covariance matrices for Ala_7 (200ns data) (A) least-squares fit, (B) THESEUS maximum likelihood fit.

2.10 Direct angular PCA

One may pose the question whether a PCA performed directly on the dihedral angles can result in a correct free energy landscape as already seen for the dPCA treatment. To answer to this question we will discuss several questions at issue. One point is that treating the dihedral angles with a data range larger than 180° like Cartesian variables, as detailed in Sec. 2.3, the average angles and hence also covariances are not correctly calculated. A possible consequence might be that the eigenvectors of the covariance matrix are not optimally chosen to obtain equally good results as from the dPCA.

In order to minimize the error which is due to the fact that circularity of the angles is not taken into account, we shift each angular variable in such a way that a minimal number of data points are at the periodic boundaries. Algorithmically, for all angles separately, one finds the angular value φ_0 with minimum density, and shifts all values above φ_0 by $-360°$. By doing so the interval of the circular data becomes $[\varphi_0 - 360°, \varphi_0]$, and not anymore the somewhat arbitrarily chosen $[-180°, 180°]$. This preprocessing of the data is visualized in Fig. 2.13. Note that this shifting (by 360°!) does does not change the circular mean of the data, nor does it change the cosine or sine values of φ. Hence the dPCA on the shifted angles is exactly the same as before shifting, but for a PCA directly on the angles we have now minimized the errors coming from a states which split up at the periodic boundaries.

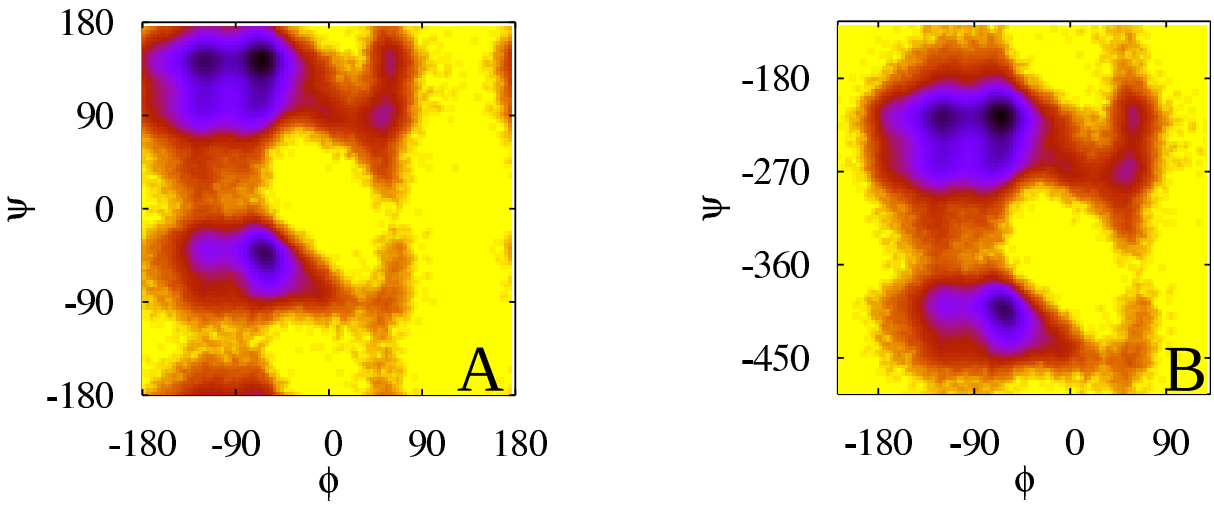

Figure 2.13: (A) Exemplary Ramachandran plot in the interval $[-180°, 180°] \times [-180°, 180°]$. (B) Same Ramachandran plot periodically shifted in order to minimize data points at the periodic boundaries. Here, angular values of minimum density are $\phi_0 = 135°$ and $\psi_0 = -125°$.

Employing a PCA directly on the shifted dihedral angles, without a prior sin/cos transformation as in the dPCA, in Fig. 2.10E and F, we find a qualitatively similar picture of the free energy landscape, nevertheless with a seemingly lower resolution - it seems as if the peaks are kind of smeared out and not as well separated as in the dPCA representation. The same phenomenon can be seen for the landscapes of the heptaalanine system Ala_7 in Fig. 2.14A and B. Even though there is quite some resemblance between the two landscapes, which we will discuss next, the dPCA landscape provides the more detailed picture. Later in Sec. 3.3 and the following sections in Chapter 3, we will provide a full analysis of the dPCA free energy landscape of Ala_7.

Recall that after centering the data in the direct angular PCA we calculated the covariance matrix as we would do for Cartesian coordinates,

$$\sigma_{ij} = \langle (\varphi_i - \langle \varphi_i \rangle)(\varphi_j - \langle \varphi_j \rangle) \rangle, \tag{2.42}$$

where $\langle \cdot \rangle$ denotes the arithmetic average over the shifted angles. We have already argued multiple times why this is not the correct average. One may wonder if we get qualitatively better landscapes using the circular average instead. From Table 2.2A we see that the circular averages for ψ-angles deviate $\approx$ 40-50 degrees from the arithmetic averages.

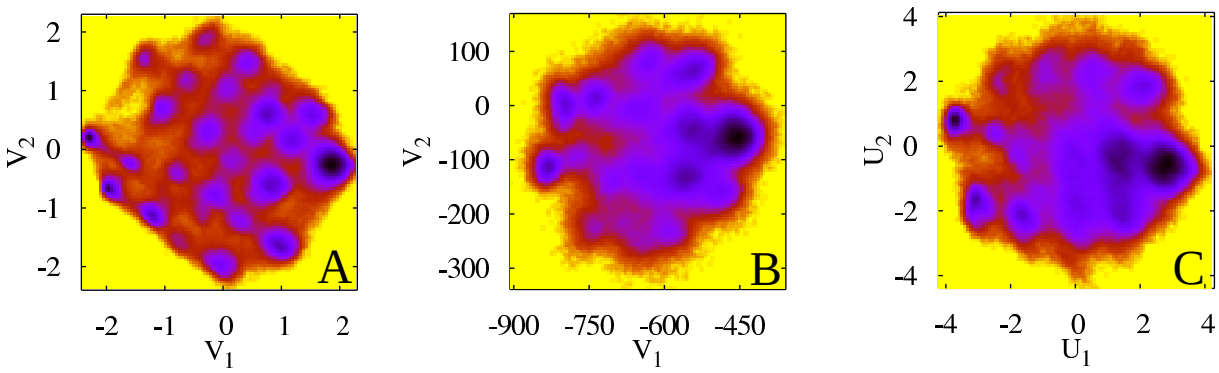

Figure 2.14: Free energy landscapes of Ala_7 in water as obtained from an 800 ns MD simulation. (A) shows the results along the first two principal components obtained from a dPCA, the second panel, (B) the corresponding landscapes calculated from a direct PCA without prior sin/cos transformation. Panel (C) displays the landscape along the principal components of a dPCA using the correlation instead of the covariance matrix.

Nevertheless we found that replacing the average by the circular one in (2.42) does not significantly change the free energy landscape (data not shown). Also the covariance may be calculated in a circular fashion. We will compare the circular with the standard correlation in the following section 2.11.

But the problem of a PCA directly on circular variables must rather be the use of eigenvectors, which are Cartesian by nature, as reaction coordinates for the free energy landscape for circular data. Next, in an example we detail this problem. We also reason why the landscapes obtained from a dPCA and a direct angular PCA have such a high degree of resemblance with each other.

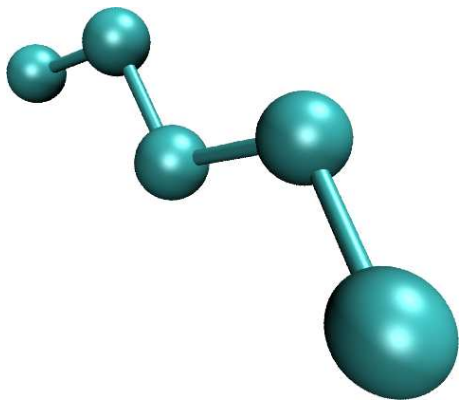

Figure 2.15: Structure of Carbon chain with 5 Carbon atoms.

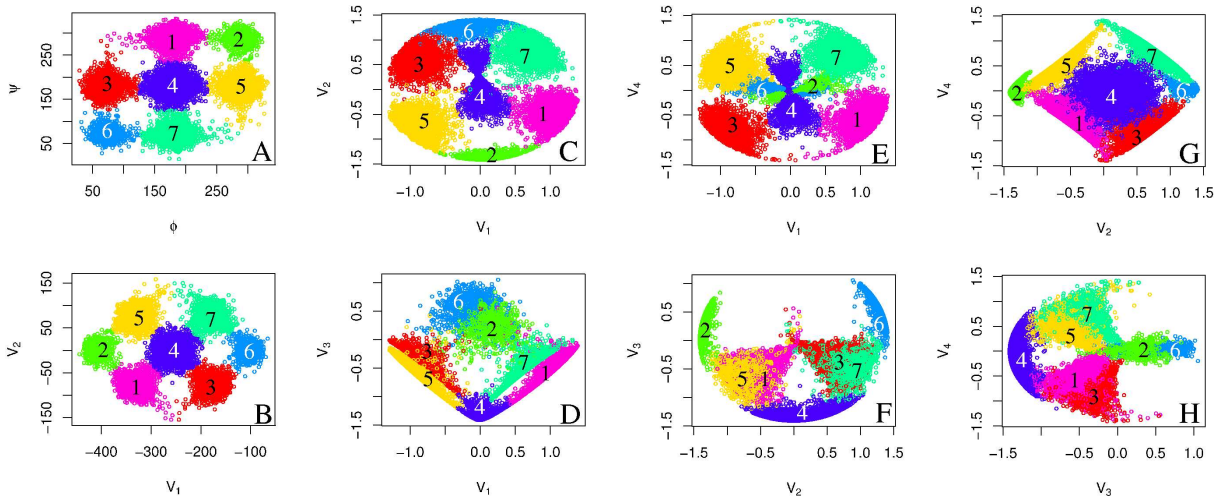

Figure 2.16: Results from a Carbon chain simulation. (A) Angular distribution of the two dihedral angles. Panel (B) shows the data after PCA transformation. Panels (C)-(H) present the data along all possible combinations of eigenvectors of the sin/cos dPCA.

To obtain a toy model we performed an MD simulation of a Carbon chain with 5 Carbon atoms as seen in Fig. 2.15. In Fig. 2.16A we see the resulting distribution for the two dihedral angles of our model. From the distribution we clearly see the symmetric cosine potential resulting in seven symmetric conformational states. In panel B we see the same data after a direct PCA on A, which simply is a rotation by $-135°$ in the plane. Panels C-H show the data after a sin/cos transformation and subsequent projections onto all possible 2D-planes of dPCA eigenvectors. In this way one can get an idea of how the data is distributed on the 4-dimensional sphere with radius $\sqrt{2}$ (see Sec. 2.3). We now want to learn how the PCAs perform on the angles, which will help understanding the advantages of dPCA over a PCA directly on the angles. Let us now examine these plots in more detail. In B the periodic boundaries are no longer at the x- and y-axes as in A, but they are at the diagonals, that is, e.g. states 1 and 7 and states 3 and 5 are neighbors, respectively. The problem of such an illustration of the data is that in general we do not know where the periodic boundaries are as the eigenvectors are a combination of the original angular variables. Hence, the circular geometry is mixed with the Cartesian nature of the eigenvectors, e.g. after the PCA transform it is not clear anymore that state 1 and 7 are geometric neighbors. In contrast to that, in the dPCA treatment in C, state 4 is twisted to simultaneously flip around states 3, 6, and 7. In such a way it is ensured that

e.g. state 1 can be represented as a direct neighbor of state 7, and that the geometric proximity of states 3 and 5 can also be truthfully kept. Note that similar phenomena can be observed from other perspectives on the data in panels D-H, but in addition due to the fact that these representations are only projections, non-neighboring states can overlap as e.g. states 1 and 3 in panel H. This cannot be regarded as a drawback of the dPCA as this is a general problem when visualizing a high dimensional data set on a 2D projected plane. Because of the one-to-one mapping to the $2N$-dimensional sin/cos space geometrically close states stay together even if they are originally separated by a periodic boundary. In general, for N angles, this is not possible in an N-dimensional Cartesian space only, this is a reason why the the dPCA needs up to $2N$ Cartesian variables.

As can be seen from Figs. 2.16C-H, the shape of the 7 states in the dPCA can be represented as in the original distribution A, as seen e.g. for states 1, 3, 5, 7 in panel C, states 2 and 6 in panel D, and state 4 in panel G. But they can also be kind of squeezed or twisted as e.g. states 2, 4, 6 in C. That is because the states are wrapped on the surface of a sphere, and as a sphere can locally be regarded as a plane, depending on from which direction one looks at it, the states appear either similar to the original distribution or in a way squeezed. Now it is important to recall that dPCA is looking for directions with largest variance for the first modes. Thus, the V_1/V_2 plane will most likely be such that a maximum number of conformational states will be represented in their original unfolded shape, and not squeezed, because squeezing them would decrease their variance. If the data is localized such that the curvature of the sin/cos sphere can be neglected we obtain very similar results as for a PCA directly on the angles. In other cases the mixing of Cartesian eigenvectors and circular geometry of the angles can result in seemingly less detailed landscapes than the ones obtained by dPCA, as seen from a comparison in Fig. 2.14. The fact that a low-dimensional dPCA landscape provides a truthfull representation of the true free energy landscape with the correct number, energy, and location of the metastable conformational states and barriers is thoroughly studied in the next Chapter.

2.11 Correlation analysis

In this section we want to analyze the correlations between the dihedral angles of Ala_7. We would like to compare our result to the one obtained by J. E. Fitzgerald et al. in [59], where

they report a strong anticorrelated motion of the ϕ angle of the ith residue (ϕ_i) and the ψ angle of the residue $i-1$ (ψ_{i-1}). Only a slight correlation was found between the motions of the two backbone dihedral angles of the same residue. They used a 200 ns simulation of Ala_7 with implicit solvent, N2 nonbonded interactions, and the GS-AMBER94 force field. We used our 800 ns GROMOS simulation (see Appendix 6.4). Similar to [59] we calculate the correlation coefficients between angles φ_i and φ_j as follows:

$$c_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}, \tag{2.43}$$

where

$$\sigma_{ij} = \langle(\varphi_i - \langle\varphi_i\rangle)(\varphi_j - \langle\varphi_j\rangle)\rangle \tag{2.44}$$

is the covariance between angles φ_i and φ_j as used in the PCA which is performed directly on the dihedral angels. Note that by doing so angles are treated like Cartesian variables. In order to minimize the error which is due to the fact that circularity of the angles is not taken into account in this definition of correlation, unlike Fitzgerald et al., we shift the angular variables in such a way that a minimal number of data points are at the periodic boundaries as detailed in Sec. 2.10.

Despite this shifting of angles, the arithmetic mean of the angles φ_i and also the correlations are not properly calculated if circularity of the angular variables is not explicitly taken care of. In Table 2.2A we already see that the circular averages for ψ-angles deviate $\approx$ 40-50 degrees from the arithmetic averages. In contrast to that, the averages for the ϕ-angles deviate only less than 5° from each other. We now compare the standard correlations to circular correlations which are calculated as above in Eq. (2.43), however, $(\varphi_i - \langle\varphi_i\rangle)$ and $(\varphi_j - \langle\varphi_j\rangle)$ in Eq. (2.44) are replaced by $\sin(\varphi_i - \langle\varphi_i\rangle)$ and $\sin(\varphi_j - \langle\varphi_j\rangle)$, where now $\langle\cdot\rangle$ denotes the circular mean of a variable as defined in Eq. (2.22). This is the measure for correlation of circular variables as proposed by Jammalamadaka in [60].

Our results are presented in Table 2.2B. According to a circular correlation coefficient of -0.01, ψ_2 and ψ_4 can be considered as uncorrelated, whereas we obtain an artificially increased correlation of 0.12 with the standard correlation measure. For ϕ_2 and ψ_2 this is similar. Examining the circular correlations, we see that in contrast to Fitzgerald et al. [59], in both cases we observe the strongest correlations between the pairs ϕ_i and ψ_i,

(A)	ϕ_2	ψ_2	ϕ_3	ψ_3	ϕ_4	ψ_4	ϕ_5	ψ_5	ϕ_6	ψ_6
standard	-81.9	92.4	-75.6	71.5	-75.1	62.3	-77.8	65.3	-81.4	82.2
circular	-83.5	131.0	-78.5	121.5	-79.2	109.8	-81.7	108.2	-86.2	115.8

(B)	ϕ_2	ψ_2	ϕ_3	ψ_3	ϕ_4	ψ_4	ϕ_5	ψ_5	ϕ_6	ψ_6
ϕ_2	**1**	**-0.19**	0.04	-0.08	0.02	-0.08	0.03	-0.09	0.03	-0.06
ψ_2	0.02	**1**	0.04	0.05	-0.06	**0.12**	-0.02	0.10	-0.05	0.08
ϕ_3	0.05	0.06	**1**	**-0.26**	0.04	**-0.13**	0.02	**-0.12**	0.01	**-0.11**
ψ_3	-0.07	-0.02	**-0.11**	**1**	0.03	0.20	-0.06	**0.22**	-0.01	**0.16**
ϕ_4	0.04	-0.01	0.06	0.05	**1**	**-0.29**	0.01	**-0.12**	0.02	**-0.11**
ψ_4	-0.08	-0.01	**-0.18**	0.05	**-0.22**	**1**	0.07	**0.25**	-0.02	**0.22**
ϕ_5	0.05	0.00	0.04	-0.02	0.02	0.06	**1**	**-0.23**	0.04	-0.07
ψ_5	-0.09	-0.01	**-0.12**	0.06	**-0.17**	**0.14**	**-0.20**	**1**	0.10	**0.19**
ϕ_6	0.03	0.01	0.02	-0.02	0.01	-0.04	0.00	0.06	**1**	**-0.20**
ψ_6	-0.06	0.00	-0.10	0.04	**-0.11**	**0.12**	**-0.12**	**0.15**	**-0.13**	**1**

(C)	ϕ_2	ψ_2	ϕ_3	ψ_3	ϕ_4	ψ_4	ϕ_5	ψ_5	ϕ_6	ψ_6
standard	0.07	0.08	0.09	0.12	0.08	0.15	0.06	0.16	0.05	0.13
circular	0.05	0.02	0.08	0.04	0.08	0.10	0.06	0.11	0.04	0.09

Table 2.2: Comparison of the standard and the circular mean and correlation for the dihedral angles of Ala_7. (A) Arithmetic average (given in $[-180°, 180°]$) calculated from the centered/shifted dihedral data and the circular average. (B) Circular correlation are listed below the diagonal of the table. The standard correlation above the diagonal is calculated for the shifted data. Entries $|c_{ij}| > 0.1$ are denoted bold for clarity. (C) Averaged absolute correlation coefficients from (B) for each angle, i.e., $\frac{1}{9}\sum_{j\neq i}|c_{ij}|$.

whereas ψ_{i-1} and ϕ_i tend to have a only small correlations ($|c_{ij}| \lesssim 0.1$) with each other. It is also notable that relatively high correlation coefficients are found between ϕ_3 and ψ_4 ($c_{ij} = -0.18$), and also for ϕ_4 and ψ_5 ($c_{ij} = -0.17$), whereas Fitzgerald et al. found no correlations between backbone dihedral angles that are separated by more than one torsion angle. In 2.2C we listed the averaged absolute correlation coefficients. From there one can see that the standard correlation measure tends to overestimate the correlations, especially for ψ-angles. If this value is close to zero, as is the case for ψ_2, it means that the respective angle is almost uncorrelated to any of the other angles, and thus could be omitted in a PCA analysis. For larger systems this quantity might help in reducing the number of variables before a PCA.

It is also interesting to see the performance of a PCA that uses the correlation matrix rather than the covariance matrix. Recall that in the standard dPCA using the covariance matrix $\mathbf{\Sigma}$ of the cos/sin transformed angles $\mathbf{q}$ we obtain their principal components by

$$V_i(t) = \mathbf{v}^{(i)} \cdot \mathbf{q}(t), \tag{2.45}$$

where $\mathbf{v}^{(i)}$ are the eigenvectors of $\mathbf{\Sigma}$.

In practice it is also common to use the correlation matrix of $\mathbf{q}$, and (2.45) becomes

$$U_i(t) = \mathbf{u}^{(i)} \cdot \mathbf{q}_{norm}(t), \tag{2.46}$$

where $\mathbf{u}^{(i)}$ are the eigenvectors of the correlation matrix, and $\mathbf{q}_{norm}$ is the standardized version of $\mathbf{q}$, with $\mathbf{q}_{norm}$ having the nth element $q_n/\sigma_{nn}^{1/2}$ (see also [13]). Note that the covariance matrix for $\mathbf{q}_{norm}$ is the correlation matrix of $\mathbf{q}$, that is, (2.46) can be regarded as a covariance PCA for the standardized data $\mathbf{q}_{norm}$.

Applying the correlation PCA to the sin/cos transformed variables for Ala_7, Fig. 2.14C shows the resulting free energy landscape $\Delta G(V_1, V_2)$. Interestingly, but without having a deeper meaning, the landscape resembles the one obtained from the direct angular PCA in Fig. 2.14B. In any case, the detailed structural appearance of the free energy landscape is a particular property of the dPCA using the covariance matrix only. To conclude this section we note that using the correlation may be advantageous for a PCA analysis in some cases [13, 57], we think that for our purpose the dPCA using the covariance matrix of the cos/sin transformed data shows the best results.

2.12 Nonlinear principal component analysis

Finally we want to mention that recently there have been efforts to study the free energy landscapes of peptides by a nonlinear principal component analysis (NLPCA) [35]. This can be advantageous if the conformational states are nonlinearly distributed in the given data set.

The basic idea of this method is that hierarchically arranged neural networks are designed and these networks are trained to build a set of adequate nonlinear mapping functions that map an input vector to its counterpart in the principal component space.

Without going into detail, in Fig. 2.17 we present a quick comparison of the free energy landscapes for hexaalanine Ala_6 as obtained by NLPCA and the dPCA, respectively. For a detailed analysis of the landscapes see [35]. It is interesting that we observe a strong correlation between the first first modes of the NLPCA and the dPCA as seen from Fig. 2.17(C). Beyond the scope of this thesis, it can be an interesting topic to further analyze nonlinear PCA methods, and to demonstrate possible advantages of such methods for the construction and interpretation of free energy landscapes of biomolecules.

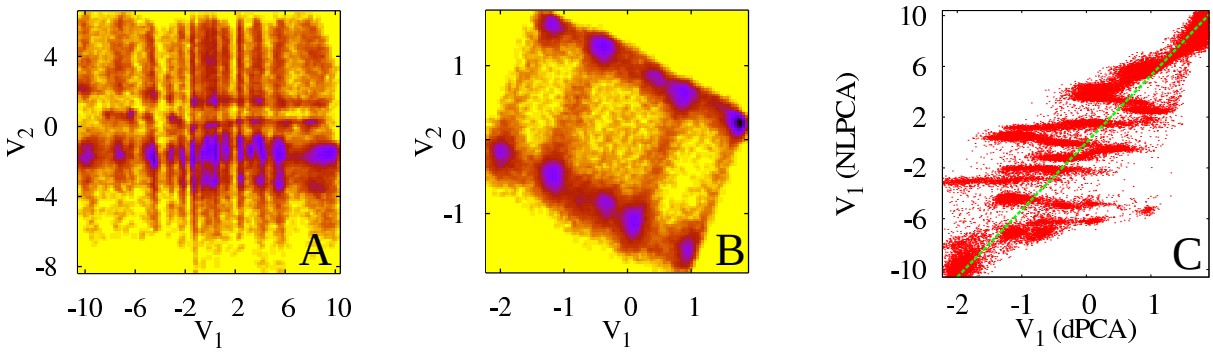

Figure 2.17: Free energy landscapes for hexaalanine Ala_6 as obtained by (A) NLPCA and (B) dPCA. Panel (C) compares the first eigenmode V_1 of NLPCA with V_1 of the dPCA.

2.13 Conclusions

We have studied the theoretical foundations of the dPCA in order to clarify the validity and the applicability of the approach. In particular, we have shown that dPCA amounts to a one-to-one representation of the original angle distribution and that its principal components can be characterized by the corresponding conformational changes of the peptide. Furthermore, we have investigated a complex version of the dPCA which sheds some light on the mysterious doubling of variables occurring in the sin/cos dPCA. One learns that N angular variables actually can be represented by N complex variables, which then naturally lead to N eigenvalues and eigenvectors. Despite its similarity to the sin/cos dPCA, the complex dPCA might be advantageous because the representation of the complex principal components by their weights and angles may facilitate their direct interpretation in terms of simple physical variables. Furthermore we have thoroughly studied the similarities and differences of Cartesian PCA, PCAs performed directly on

the angular variables, and the dPCA.

To demonstrate the potential of the dPCA, we have applied it to construct the energy landscape of Ala_{10} from a 300 ns MD simulation. The resulting free energy surface exhibits numerous well-separated minima corresponding to specific conformational states, revealing that the unfolded state of decaalanine is rather structured than random. The smooth appearance of the energy landscape obtained from a PCA using Cartesian coordinates was found to be caused by an artifact of the mixing of internal and overall motion. Hence the correct separation of internal and overall motion is essential for the construction and interpretation of the energy landscape of a biomolecule undergoing large structural rearrangements. Internal coordinates such as dihedral angles fulfill this requirement in a natural way. Performing and analyzing an 800 ns MD simulation of Ala_7 we could show that the dPCA provided the most detailed low-dimensional representation of the free energy landscape. A correlation study for the dihedral angles of Ala_7 using a circular correlation measure could show that, in contrast to a study performed by Fitzgerald et al. in [59], that the correlated motion of the ϕ angle of the ith residue (ϕ_i) and the ψ angle of the residue $i-1$ (ψ_{i-1}), is much weaker. Furthermore, we found the strongest correlations for neighboring torsion angles of the same residue.

Recently, several nonlinear approaches have been proposed [33–36] which may account for nonlinear correlations not detected by a standard PCA. For example, it has been discussed in Ref. [34] that completely correlated motion such as two atoms oscillating in parallel direction but with a 90° phase shift is not monitored by a linear PCA, since $\langle \sin(\omega t) \sin(\omega t + \pi/2) \rangle = 0$. This geometrical artifact caused by the relative orientation of the atomic fluctuations was found to lead to a considerable (≈ 40 %) underestimation of the correlation of protein motion [34]. Because of the use of dihedral angles and the inherent nonlinear transformation, the dPCA represents a nonlinear PCA with respect to Cartesian atomic coordinates and is therefore able to identify this type of fluctuations.

Furthermore, various methods have been suggested which allow for a identification of metastable conformational states [5, 20, 30–32]. By calculating the transition matrix that connects these states, one may then model the conformational dynamics of the system via a master-equation description. While the dPCA also allows us to calculate metastable conformational states and their transition matrix [25], it moreover provides a way to

represent the free energy landscape as well as all observables of the system in terms of well-defined collective coordinates [61]. This way the dPCA free energy surface can be used to perform (equilibrium or nonequilibrium) Langevin simulations of the molecular dynamics [62, 63] as well as a simulation using a nonlinear dynamic model [36]. As all quantities of interest can be converged to the desired accuracy by including more principal components, the approach avoids problems associated with the use of empirical order parameters (such as the number of native contacts) or low-dimensional reaction coordinates (such as the radius of gyration), which may lead to artifacts and an oversimplification of the free energy landscape [64].

Chapter 3

Free Energy Landscape

In this chapter we present a systematic approach to construct a low-dimensional free energy landscape from a classical molecular dynamics (MD) simulation. The approach is based on the in Chapter 2 discussed dihedral angle principal component analysis (dPCA), which avoids artifacts due to the mixing of internal and overall motion in Cartesian coordinates and circumvents problems associated with the circularity of angular variables. 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 distribution and autocorrelation of the principal components. Performing an 800 ns MD simulation of the folding of heptaalanine in explicit water and using geometric and kinetic clustering techniques, it is shown that a five-dimensional dPCA energy landscape is a suitable and accurate representation of the full-dimensional landscape. In a second step, the dPCA energy landscape can be employed (e.g., in a Langevin simulation) to facilitate a detailed investigation of biomolecular dynamics in low dimensions. Finally, several ways to visualize the multidimensional energy landscape are discussed.

3.1 Introduction

As we have seen in Chapter 2, assuming a time scale separation of the slow motion along the first few PCs and the fast motion along the remaining PCs, the first few PCs may

serve as reaction coordinates to represent the free energy landscape of a biomolecular system. Since the eigenvectors of the covariance matrix form a complete basis, it is clear that the representation of the conformational space in terms of PCs becomes exact when sufficiently many PCs are taken into account. In practice, on the other hand, one- and two-dimensional representations are commonplace, which may lead to serious artifacts and oversimplifications of the free energy landscape of small folding peptides [64]. This rises the important question on how many dimensions or PCs need to be taken into account in order to appropriately describe a given biomolecular process. An energy landscape may be characterized in terms of its minima which represent the metastable conformational state of the systems, and its barriers which connect these states. Hence a suitable reduced representation of the energy landscape should (at least) reproduce the correct number, energy and location of the metastable states and barriers. Unfortunately, these crucial quantities often get lost when the energy landscape is projected on a low-dimensional subspace.

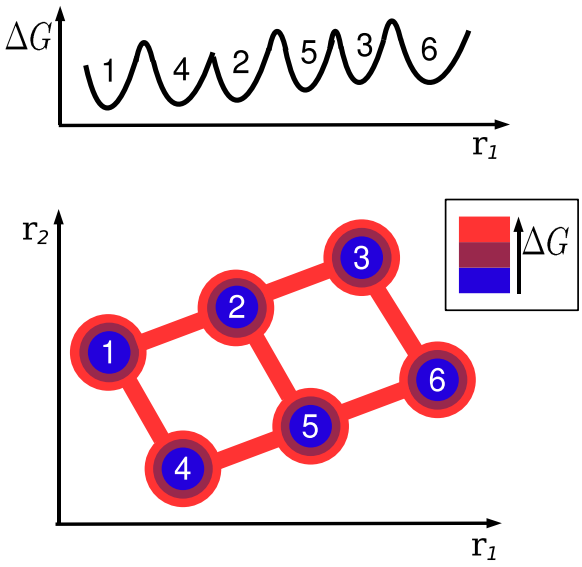

Figure 3.1: Schematic one- and two-dimensional representations of a model free energy landscape. Although the reduced dimensionality representation reproduces the correct number of minima and their energies, the connectivity of these states and their barriers are obscured in a single dimension.

To illustrate the problem, Fig. 3.1 shows schematic one- and two-dimensional rep-

resentations of a model free energy landscape. The two-dimensional energy landscape $\Delta G(r_1, r_2)$ exhibits $n = 6$ minima of energy ΔG_i corresponding to metastable conformational states of the system. The minima are connected by barriers of height ΔG_{ij}. The projection of the two-dimensional surface on its first coordinate is given by

$$\Delta G(r_1) \propto -k_{\mathrm{B}} T \ln \int \mathrm{d}r_2 P(r_1, r_2). \tag{3.1}$$

The one-dimensional representation is found to reproduce the correct number of minima and their energies. The former is clearly a consequence of the fact that all minima are located at different values of r_1. In general, however, we may obtain less minima in lower dimensions because several minima may overlap along the reduced coordinate r_1. More importantly, though, Fig. 3.1 reveals that the true nature of the barriers may be obscured in reduced dimensionality. As a typical example, consider minima **2** and **4**. In two dimensions, there exist two pathways of minimal energy between these two states, $\mathbf{2} \to \mathbf{1} \to \mathbf{4}$ and $\mathbf{2} \to \mathbf{5} \to \mathbf{4}$. Projecting on a single dimension, however, this connectivity gets lost. Now states **2** and **4** are direct neighbors connected by a single barrier and states **1** and **2** are only connected via state **4**. The energies ΔG_{24} and ΔG_{12} of these spurious barriers and the corresponding transition rates k_{24} and k_{12} may be smaller or larger than in full dimensionality as detailed next.

We adopt a simple example to show that the barrier heights in reduced dimensionality may be smaller or larger than in full dimensionality. The idea is given in Fig. 3.2 which shows two-dimensional population maps $P(i, j)$ $(i, j = 1, 2, 3)$ and their one-dimensional projections $P(i) = \sum_j P(i, j)$. The corresponding free energies are again calculated via $\Delta G \propto \ln P$. In case (a), there are two states at (1,1) and (3,3) with populations 4/12 and 6/12, respectively, which are separated by a barrier at (2,2) with a population 2/12. Projecting on one dimension, the energies of states and the barrier are retained. In panel (b), the minimum-energy path has a barrier at $(2, 2)$. Projecting on the horizontal axis, the barrier between right and left states becomes higher. Finally, in panel (c) we have constructed an example in which the barrier in one dimension becomes smaller compared to the true barrier in full dimensionality.

Further circumstances under which a PCA-based free energy landscape may appear simpler as it actually is have been thoroughly discussed in Chapter 2.

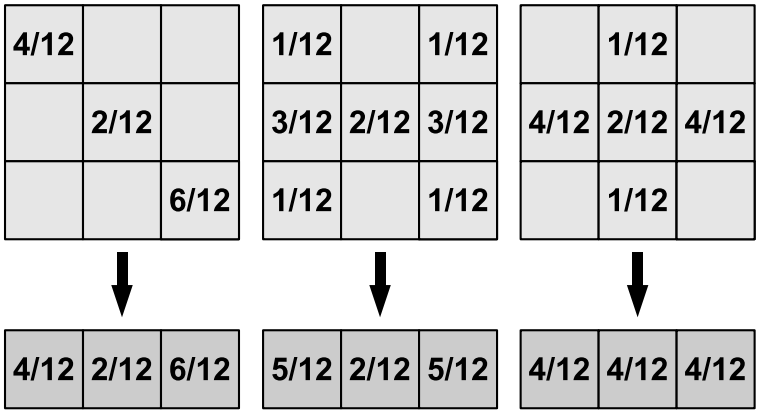

Figure 3.2: Two-dimensional population maps $P(i,j)$ $(i,j=1,2,3)$ (upper panels) and their one-dimensional projections $P(i)=\sum_j P(i,j)$ (lower panels), where $P(i,j)=0$ for empty fields. Projecting on the horizontal axis, the barrier between right and left states (a) remains, (b) becomes higher, and (c) becomes smaller.

In this chapter, we employ the dPCA to systematically construct a low-dimensional free energy landscape from a classical molecular dynamics (MD) simulation. Being based on the backbone dihedral angles, the dPCA naturally distinguishes between the kinetically well-separated main conformational states of the peptide, such as the α_R helical and the β extended conformations. 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 [62,63,65,66] or a nonlinear dynamic model [33–36] as also detailed in Chapters 4 and 5 of this thesis. Adopting an 800 ns MD simulation of the folding of heptaalanine (Ala_7) in explicit water, we show that a five-dimensional dPCA energy landscape is a suitable and accurate representation of the full-dimensional landscape of Ala_7. In particular, geometric and kinetic clusterings yield approximately the same metastable states and barriers as for the full-dimensional surface. Finally, we present several ways to visualize the multidimensional energy landscape.

3.2 Clustering

To visualize the multidimensional energy landscape and identify its metastable states, we have employed k-means clustering [67]. The k-means algorithm aims at finding a partition $C=(C_1,\ldots,C_k)$ of a given data set into k subsets that minimizes the sum of squares of

distances between the objects and their corresponding cluster centroids

$$\sigma^2 = \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2, \tag{3.2}$$

where x_j are the objects contained in cluster C_i with centroid μ_i. The algorithm is initialized with k random centers. For every object, all distances to the k centroids are determined, and the object is assigned to the centroid with minimum distance. When all objects have been assigned to a group, k new centroids are calculated as the average over all objects in their corresponding groups. These steps are repeated until the objects no longer switch clusters. As this method is sensitive to the initial conditions, one can become trapped in a local minimum of Eq. (3.2). A simple solution to this problem is to run the algorithm several times and to choose the best solution, i.e., with minimal value of σ^2. In the calculations shown below, typically 200 runs were performed.

As the number of clusters must be known beforehand in k-means, we need to establish a criterion to determine this number. For example, we may request that a suitable clustering should give a large fraction (say, larger that 90 %) of "good" clusters. To define such a good cluster, it is useful to introduce the circular variance [41] which provides an appropriate measure of the spread of angular variables. It is defined as

$$\mathrm{var}(\varphi) = 1 - R/L, \tag{3.3}$$

$$R^2 = \left(\sum_{i=1}^{L} \cos\varphi(l) \right)^2 + \left(\sum_{i=1}^{L} \sin\varphi(l) \right)^2,$$

i.e., R is the resultant length of the vector sum of the vectors $(\cos\varphi(l), \sin\varphi(l))$, where $\varphi(l)$ are realizations of angle φ. Note that $\mathrm{var}(\varphi) \in [0, 1]$. A cluster is "good," if the average circular variance of all its N dihedral angles is below a certain limit

$$\frac{1}{N} \sum_{i=1}^{N} \mathrm{var}(\varphi_i) < \sigma^2_{\max}, \tag{3.4}$$

where the maximal circular variance $\sigma^2_{\max}$ is chosen such that angular fluctuations *within* a conformational state are significantly smaller than $\sigma^2_{\max}$, while transitions *between* different conformational states result in a circular variance much larger than $\sigma^2_{\max}$. For example,

in order to clearly separate α_R helical and the β extended conformations, a value of $\sigma^2_{\max} \approx 0.2$ is suitable.

By dividing the conformational space of the molecule into k clusters, discrete states are defined for which we calculate the $k \times k$ transition matrix $\mathbf{T}(\tau)$ of the process. 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 [68,69]. Hence its diagonal elements $T_{ii}(\tau)$ are a measure for the metastability of state i. To estimate $\mathbf{T}(\tau)$ from a MD simulation, we represent the conformational state of the system at time t by vector $\mathbf{c}(t)$, where $c_i(t) = 1$ if the system is in state i and $c_i(t) = 0$ if not. Then the transition probability $T_{ij}(\tau)$ is given by [7]

$$T_{ij}(\tau) = \frac{\langle c_j(\tau) c_i(0) \rangle}{\langle c_i \rangle}, \tag{3.5}$$

where $\langle \ldots \rangle$ denotes an equilibrium average of the MD trajectory. If the process under consideration can be described by a Markov chain [68], a master equation using transition matrix $\mathbf{T}(\tau)$ provides the complete information of the time evolution of the system (see Refs. [5–7, 20, 30, 32, 70, 71] for recent applications of this approach to biomolecular processes). We note that the estimates of T_{ij} satisfy detailed balance, that is, time-reversed information gave the same transition probabilities. Throughout this article we use $\tau = 1$ ps and omit the τ-dependence of the transition matrix for notational convenience.

The eigenvalues μ_k ($0 \leq \mu_k \leq 1$) of the transition matrix can be used to construct a *kinetic* clustering of the process, that is, a clustering that defines its states through their metastability rather than through geometric similarity [6, 69, 71]. In systems governed by hierarchical dynamics [72], one expects a separation of time scales which allows us to define metastable clusters which exhibit fast intracluster motion and slow intercluster motion. Eigenvalues close to unity, the so-called Perron eigenvalues, correspond to such metastable clusters, while small eigenvalues indicate the existence of kinetically unstable clusters. Systems showing hierarchical dynamics typically exhibit a clear gap between Perron and small eigenvalues.

A popular means to illustrate the energy landscape of biomolecules are disconnectivity graphs [10, 73]. To construct a free energy disconnectivity graph [74], one needs to calculate the free energies ΔG_i of the k clusters as well as the free energy barriers ΔG_{ij} along the minimum-energy path connecting states i and j. Using Eq. (1.1), the ΔG_i are

readily obtained from the population probabilities P_i of the corresponding conformational states. The barriers ΔG_{ij} can be estimated from transition state theory, which gives for the transition from state i to state j the rate

$$k_{ij} = k_0 e^{-\Delta G_{ij}/k_B T}. \tag{3.6}$$

Following Ref. [74], we estimate the transition state prefactor as $k_0 = k_B T/h$, which results in $k_0 \approx 1/(0.16\text{ ps})$ at $T = 300$ K. Furthermore, we estimate the transition rates from the transition matrix through $k_{ij} = [T(\tau)]_{ij}/\tau$. This gives for the barrier heights

$$\Delta G_{ij} = -k_B T \ln\left(\frac{T_{ij}}{k_0 \tau}\right). \tag{3.7}$$

Owing to the numerous approximations involved, this expression is not meant to provide an accurate description of free energy barriers, but solely serves as a qualitative estimate for the disconnectivity graph. The disconnectivity graph shown below was generated using the program of M. Miller [75].

3.3 Dimensionality of the free energy landscape

In what follows, we employ the above described methods to construct and analyze the free energy landscape of heptaalanine (Ala_7), which is obtained from an 800 ns MD simulation in aqueous solution at 300 K. We restrict the analysis to the backbone dihedral angles $\phi_2, \psi_2, \ldots, \phi_6, \psi_6$ of the inner residues (Fig. 3.3), since the dihedral angles of both end-groups were found to be virtually uncorrelated to the rest of the system.

Generally speaking, the goal of any reduced-dimensionality representation is to appropriately describe a given problem by using a minimum number of dimensions. As explained above, we consider ten (sin- and cos-transformed) dihedral angles $\phi_2, \psi_2, \ldots, \phi_6, \psi_6$ in the dPCA of Ala_7, thus resulting in a 20-dimensional vector space. For the dPCA representation of the free energy landscape, this amounts to the question of how many PCs are needed in order to (at least) reproduce the correct number, energy, and location of the metastable states and barriers. To address this question, Fig. 3.4 presents two-dimensional dPCA representations of the free energy landscape of Ala_7, including (A)

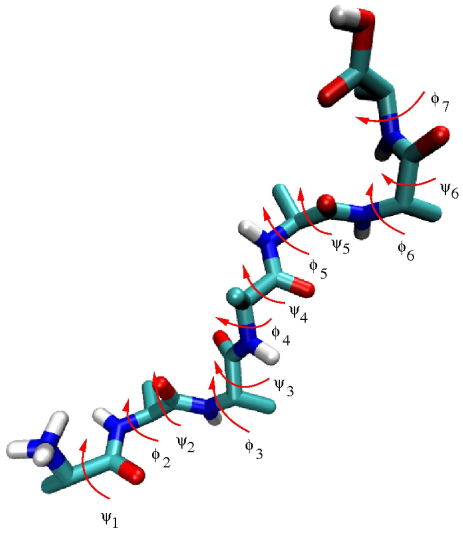

Figure 3.3: Structure and dihedral angles labeling of Ala_7.

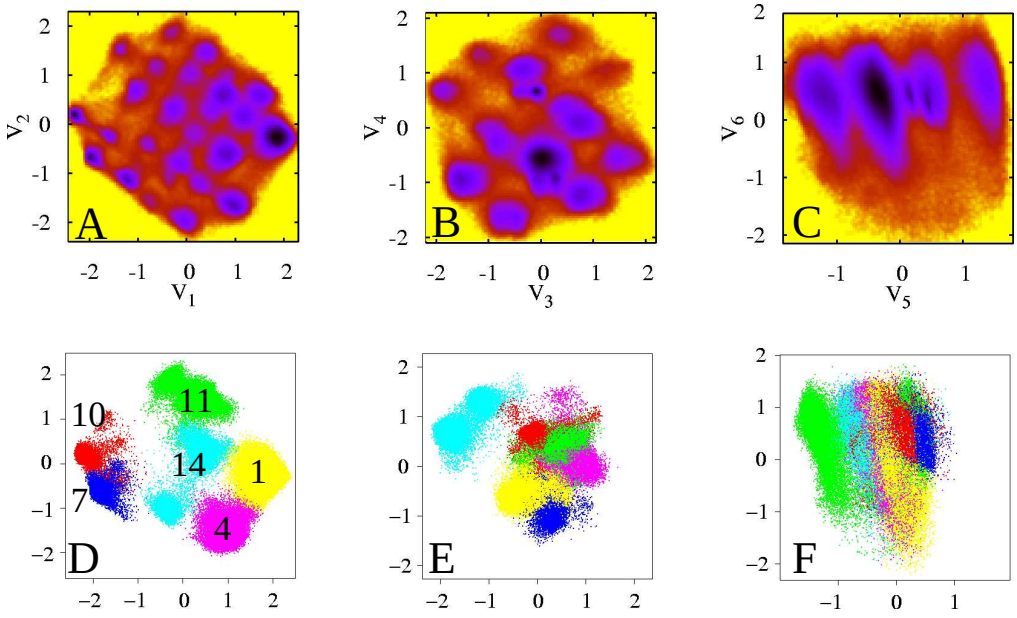

Figure 3.4: Two-dimensional representations of the free energy landscape of Ala_7 as obtained by dPCA: (A) $\Delta G(V_1, V_2)$, (B) $\Delta G(V_3, V_4)$, and (C) $\Delta G(V_5, V_6)$. The color coding in panels (D)-(F) illustrates some prominent conformational states which are described in Table 3.1, visualized on the upper landscape.

$\Delta G(V_1, V_2)$, (B) $\Delta G(V_3, V_4)$, and (C) $\Delta G(V_5, V_6)$. While the free energy exhibits several minima corresponding to distinct metastable conformational states along the first five PCs, there is only a single minimum found along V_6, reflecting intrastate fluctuations.

As a further indication of the number of "essential" PCs, we may consider the percentage of overall fluctuations covered by the first n PCs (i.e., the sum of the first n eigenvalues of the PCA). Interestingly, Fig. 3.5(A) reveals three kinds of PCs: The first

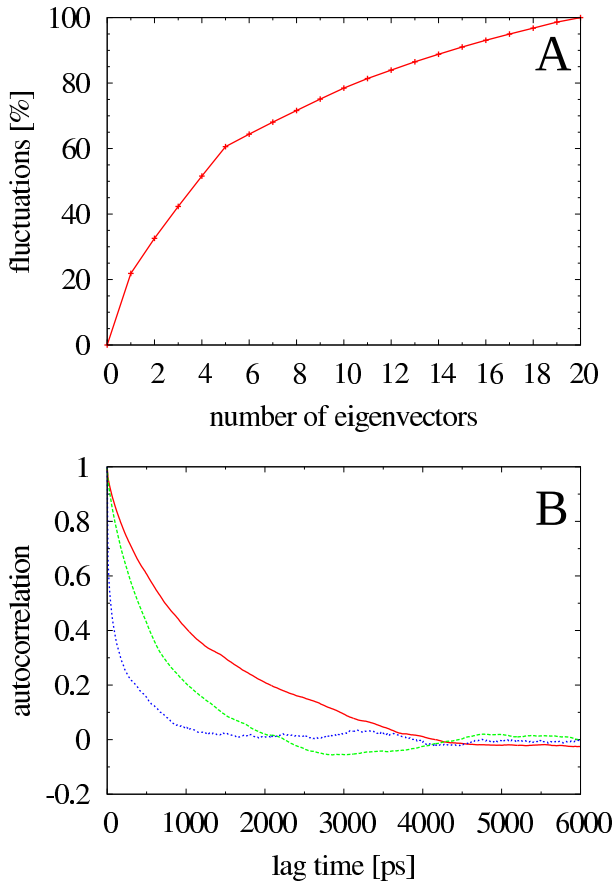

Figure 3.5: The principal components of Ala_7 as obtained by the dPCA, characterized by (A) their cumulative fluctuations and (B) their normalized fluctuation autocorrelation functions. The latter is shown for the principal components V_1 (full line), V_2 (dashed line), and V_6 (dotted line). The size of the statistical error is similar to the line width of the plots.

one covers 22 % of all fluctuations, each of the next four contribute about 10 %, while the remaining PCs contribute less than 4 % each. A similar behavior is found for the time scales of the fluctuations, revealed by the normalized fluctuation autocorrelation function $(\langle V_n(t)V_n\rangle - \langle V_n\rangle^2)/(\langle V_n^2\rangle - \langle V_n\rangle^2)$ shown in Fig. 3.5(B). Judged by their initial

time evolution, the first five PCs decay on a time scale of 1 ns, whereas the decay time of the higher PCs is clearly shorter.

From the above results we expect that a five-dimensional dPCA representation of the free energy surface of Ala_7 suffices to correctly describe its main features. This is because higher PCs with unimodal probability distribution account for fluctuations rather than for conformational transitions. Appendix 6.3 shows that this is strictly true for Gaussian-distributed degrees of freedom. For other apparently unimodal distributions, where insufficient statistics might obscure smaller substructures, the situation is less clear-cut and introduces a certain ambiguity. Considering $\Delta G(V_5, V_6)$ in Fig. 3.4, for example, we observe a weak correlation between the two principal components V_5 and V_6, although the probability distribution along V_6 is unimodal. This residual correlation of essential PCs ($V_1 - V_5$) with (apparently) non-essential PCs ($V_6 - V_{20}$) therefore may also somewhat change the definition of metastable states as well as their barriers. To investigate this effect, in the following we employ various clustering techniques to study the metastable states obtained from a five- and the full-dimensional energy landscape of Ala_7.

3.4 Geometric and kinetic clustering

To characterize the metastable states of the reduced free energy landscape of Ala_7 shown in Fig. 3.4, we employ the k-means algorithm [67] as a well-established simple and fast geometric clustering method. As the number of clusters must be known beforehand in k-means, we first need to decide how many clusters should be considered in the analysis. From a visual inspection of Fig. 3.4A it is already clear that we should include at least $\approx$ 20 clusters to distinguish all states shown by the $\Delta G(V_1, V_2)$ surface. However, since the two-dimensional representations in Fig. 3.4 do not reveal possible correlations between each other, we cannot tell if the $\approx$ 20 states in $\Delta G(V_1, V_2)$ split up further in $\Delta G(V_3, V_4)$ or not. To test if a clustering in k states is suitable, we request that such a clustering should give a large fraction (say, larger that 90 %) of good clusters. As explained in Sec. 3.2, we call a cluster "good" when the average circular variance of all dihedral angles is less than a certain threshold, thus discriminating fluctuations *within* a conformational state from transitions *between* different conformational states. Figure 3.6(A) shows the resulting percentage of good clusters as a function of k, the number of clusters used in

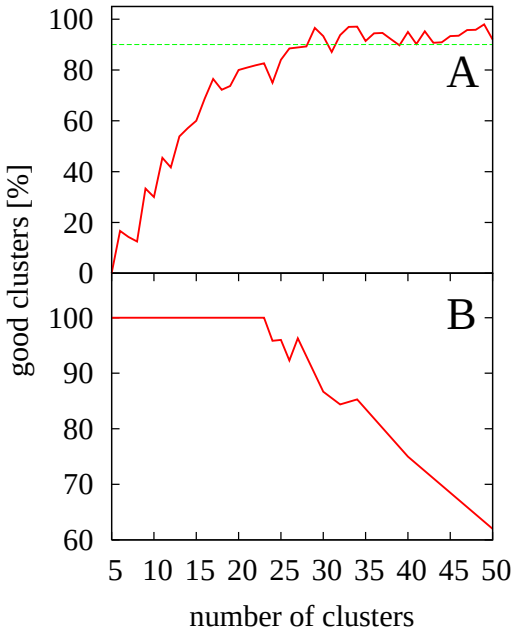

Figure 3.6: Geometric and kinetic clustering of the free energy landscape of Ala_7. (A) Percentage of "geometrically good" clusters as a function of the number of clusters considered in k-means. (B) Percentage of metastable states (i.e., "kinetically good" clusters) as a function of the number of clusters.

the algorithm. The fluctuations of the curve reflect the fact that k-means is a stochastic algorithm and that therefore the best out of 200 k-means runs is shown for each value of k. (A smooth curve is obtained by averaging over the n best runs.) The percentage of good clusters exhibits a steep increase for small k and saturates for $k \gtrsim 20$. In accordance with the visual inspection of Fig. 3.4, this suggest that twenty represents a lower limit for the number of clusters.

It is interesting to compare the above findings to the results of a *kinetic* clustering of the process, that is, a clustering that defines its states through their metastability rather than through geometric similarity [6, 69, 71]. To this end, we have calculated the number of Perron eigenvalues of the transition matrix, which reflects the number of metastable states of the partitioning (see Sect. 3.2). Plotting the fraction of metastable clusters P_{Perron} (i.e., the number of Perron eigenvalues divided by k) as a function of k, Figs. 3.6(B) and 3.7 reveal that for $k \leq 23$ all clusters are metastable. For larger k, we observe an approximately linear decrease of $P_{\text{Perron}}(k)$. From Fig. 3.7 we can see up to 31 metastable clusters if we use $k = 100$ for the clustering. Note that the linear decrease

indicates that number of metastable clusters hardly increases anymore with the number of clusters used. That is, an additional cluster does not increase the number of kinetically stable clusters anymore. In fact, metastable clusters are split up in two ore more unstable clusters.

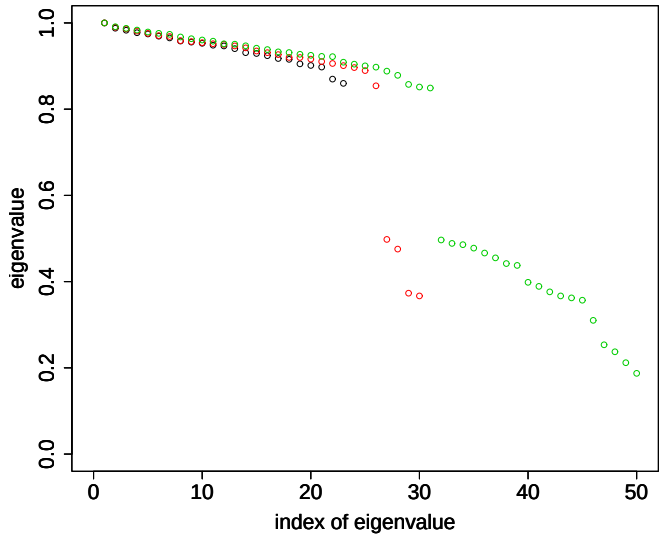

Figure 3.7: Eigenvalues of transition matrices calculated for the states obtained from k-means clustering of the five-dimensional free energy landscape of Ala_7. The black spectrum was calculated for $k = 23$ clusters, whereas we used $k = 30$ and $k = 50$ for the red and the green eigenvalue spectrum, respectively.

In what follows, we adopt $k = 23$ in order to obtain kinetically metastable states that are at the same time geometrically well separated. To characterize these states, Table 3.1 comprises their population probability P_i, their metastability T_{ii}, and a rough description of the conformational state (α, β) of the five inner amino acids. Here α denotes the right-handed helix conformation and β accounts for both extended β and poly-L-proline II (P_{II}) helix-like conformations, since most biomolecular force fields discriminate these states only weakly [47]. α/β means that the circular variance of the corresponding ψ-angle exceeded the threshold $\sigma^2 = 0.2$, i.e., the corresponding amino acid adopts both α and β conformations. All clusters are found to have a high metastability, ranging from 88-97 %. The largest cluster, and hence the global free energy minimum, is the all-β/P_{II} conformation with a population of 23.2%, followed by mostly extended conformations with one amino acid in α. The all-α state has a population of 3.5%. The occurrence of

cluster	aa 2	aa 3	aa 4	aa 5	aa 6	P_i (%)	T_{ii} (%)	P_i^{d} (%)	T_{ii}^{d} (%)
1	β	β	β	β	β	23.2	97	23.3	95
2	β	β	α	β	β	7.6	96	7.5	92
3	β	β	β	α	β	7.5	96	7.4	93
4	β	α	β	β	β	7.0	95	7.1	92
5	α	α/β	β	β	β	6.4	96	1.4/5.4	89/92
6	β	β	β	β	α	5.3	95	5.2	91
7	β	α	α	α	α	3.9	97	3.8	95
8	β	α	α	α	β	3.8	95	3.8	92
9	β	β	α	α	β	3.5	94	3.4	89
10	α	α	α	α	α	3.5	97	3.3	96
11	α	β	α	β	α/β	3.5	96	0.8/2.7	88/90
12	β	α	α	β	β	3.3	94	3.3	90
13	β	α	β	α	β	3.3	94	3.1	90
14	α	α/β	β	α	β	2.5	95	0.8/1.8	88/90
15	β	β	α	α	α	2.4	95	2.3	91
16	β	β	α	β	α	2.2	93	2.0	88
17	α	β	α	α	α/β	2.1	94	0.9/1.2	89/84
18	α/β	α	β	β	α	1.9	91	0.3/1.8	83/87
19	β	β	β	α	α	1.8	91	1.7	88
20	α	α	α	α/β	β	1.7	92	1.3/0.6	87/81
21	α/β	α	α	β	α	1.5	93	0.3/1.1	79/87
22	α	β	β	α/β	α	1.4	88	0.4/0.9	85/84
23	α/β	α	β	α	α	1.0	88	0.3/0.7	87/85

Table 3.1: Conformational states of Ala_7 as obtained from a k-means clustering on the five-dimensional dPCA space. The states are characterized by the structure of their five inner amino acids (α for helical conformations and β for extended or poly-L-proline II conformations), their population probability P_i and their metastability T_{ii}. The k-means results for P_i and T_{ii} in reduced space are compared to the results P_i^{d} and T_{ii}^{d} of a direct clustering on the full-dimensional free energy landscape of Ala_7. Statistical errors are $\pm 0.2\%$ for populations and $\pm 1\%$ for metastabilities.

several clusters with mixed α/β states demonstrates the limits of the k-means algorithm in obtaining a physically meaningful clustering. On one hand, one needs a larger cluster number k to resolve the conformations combined in such a state. On the other hand, by increasing k, the metastabilities of the resulting clusters decrease, indicating that also conformations with well-defined structure split up.

We are now in a position to assess the quality of the five-dimensional (5D) landscape with respect to the true full-dimensional free energy landscape of Ala_7. To this end,

we employ a simple "direct" clustering of the full-dimensional dihedral angle space by considering the two conformational states α and β/P_{II} for each individual residue i. For simplicity, we choose the definitions $-180^\circ \leq \psi_i < 25^\circ$ for α and $25^\circ \leq \psi_i < 180^\circ$ for β/P_{II}. This results in a total number of $2^5 = 32$ possible conformational states for the whole peptide. (Note that, due to this exponential scaling, direct clustering is only feasible for small systems, while the linearly scaling k-means algorithm can be employed to truly many-dimensional systems such as proteins.) The population probabilities and metastabilities of these 32 states were calculated from the trajectory and are listed in the two last columns of Table 3.1.

Regarding the direct clustering calculations as full-dimensional reference results, we find that k-means clustering on the reduced dPCA energy landscape nicely reproduces the population probabilities of all conformational states. This is also true for the mixed α/β states in k-means, whose subpopulations can be determined by visual inspection (data not shown). It is important to note that the latter analysis is not possible, if less than five principal components are used. Furthermore, we find a good agreement for the metastabilities of both methods, although the metastability of the k-means clusters are typically a few percent higher. The latter is most likely due to the simple definition of states used in the direct clustering. As a consequence, the "direct" barriers are consistently ≈ 10 % lower as the k-means barriers (data not shown). The latter findings, however, represent mostly the shortcomings of the two simple clustering schemes, which could be improved by invoking advanced kinetic clustering techniques as recently suggested in Refs. [6] and [71].

Taking together the cluster analysis presented in Table 3.1, the distribution of the PCs displayed in Fig. 3.4, and their fluctuations and time scales shown in Fig. 3.5, it has been demonstrated that the 5D dPCA energy landscape is a suitable and accurate representation of the full-dimensional landscape of Ala_7. That is, by using only five dimensions, we correctly account for all populations and metastabilities of the conformational states as well as for all slow motions of the system. In a second step, this reduced-dimensionality representation may be employed to schematically illustrate the main features (states, barriers, connectivities, energy basins, etc.) of the biomolecular system, see Sec. 3.6. Furthermore, the free energy surface can be used to perform (equilibrium or nonequilib-

rium) simulations of the molecular dynamics using the Langevin approach [62, 63, 65] or a nonlinear dynamic model [36].

3.5 Markovian modeling

In order to describe the conformational dynamics of the system, one might wonder whether an explicit simulation in a five-dimensional coordinate space is even necessary or if it is sufficient to resort to a much simpler master equation modeling using the above described conformational states and their transition matrix $\mathbf{T}(\tau)$. The latter is correct if the dynamics is Markovian, that is, if the Chapman-Kolmogorov property

$$\mathbf{P}(n\tau) = \mathbf{P}(0)\mathbf{T}(n\tau) = \mathbf{P}(0)\mathbf{T}^n(\tau) \tag{3.8}$$

holds, where $\mathbf{P}(t) = (P_1(t), \ldots, P_{23}(t))$ comprises the time-dependent population probabilities of the conformational states. To check this condition, we used the discrete state space as obtained by the k-means clustering with 23 states as listed in Table 3.1.

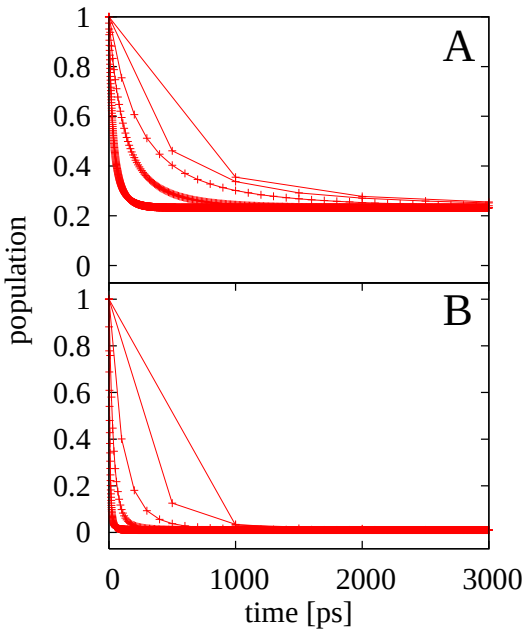

Figure 3.8: Master equation results for the decay of (A) the most stable state **1** and (B) the least stable state **23**, assuming lag times $n\tau = 1$, 10, 100, 500, and 1000 ps (from left to right). The size of the statistical error is similar to the line width of the plots.

Choosing the most stable state **1** and the least stable state **23** as representative examples, Fig. 3.8 shows the master equation results for the decay of these two states assuming that (A) $P_1(0) = 1$ and (B) $P_{23}(0) = 1$, respectively. In contrast to condition (3.8), this decay depends significantly on the chosen lag time $n\tau$ = 1, 10, 100, 500, and 1000 ps. Only for lag times longer than several hundreds of picoseconds the Chapman-Kolmogorov property is found to hold at least approximately. We note that memory times of 100-3000 ps are also expected from the decay of the PC autocorrelation functions shown in Fig. 3.5(B).

To study if lag times $\gtrsim$ 100 ps are suited to describe the conformational dynamics of Ala_7, we estimate the life times of the conformational states directly from the MD simulation. Assuming an exponential waiting time distribution, we estimate mean life times of $\approx$ 20 ps for states **22** and **23**, while the most stable states **1**, **7**, and **10** live for about 70 ps. That is, for our choice of discrete state space a Markov model of the conformational dynamics of Ala_7 is hardly appropriate, since the minimal lag time to assure Markovian dynamics considerably exceeds the life times of the conformational states. Although relatively long correlation times seem ubiquitous in biomolecular simulations, this finding is of course not general but depends on the specific choice of discrete state space as well as on the molecular system under consideration. For example, Chodera *et al.* found a suitable time scale separation for the alanine dipeptide and the α-helical Fs peptide, whereas the trpzip2 hairpin defied a Markovian treatment [6].

3.6 Visualization of the free energy landscape

A part of the reason that most authors focus on one- and two-dimensional energy landscapes lies in the problem of visualizing $\Delta G(q)$ in higher dimensions. Adopting the above established five-dimensional free energy landscape of Ala_7, in the following we discuss several options to do so. As shown in Fig. 3.4, a straightforward way is to consider two-dimensional cuts of the full-dimensional energy landscape. By color coding various conformational states of interest (panels D-G), it is seen that k-means clustering nicely reveals the correlation of the free energy minima in the respective representations. As an example, consider state **11** which is clearly separated from the other states in the (V_1, V_2) representation, while it overlaps with states **4** and **10** in (V_3, V_4), and partly overlaps with

other states in (V_5, V_6). By including all necessary five dimension for the description of the free energy landscape, we take all these correlations into account, e.g., when we evaluate the distances between the clusters during a k-means run.

For illustrative purposes, nevertheless, one often wants to restrict the representation of the full-dimensional energy landscape to two dimensions (2D). A simple way to do so is to plot the energy landscape $\Delta G(V_1, V_2)$ along the first two components. Calculating the geometric centers of all clusters and connecting all clusters that make transitions to each other with transition probability $\mathbf{T}_{ij} > 0.1\%$, Fig. 3.9(A) shows that the arrows mostly connect neighboring states. That is, kinetically well separated clusters are also geometrically distinct in the first two PCs of the dPCA.

As the distances of the cluster centers in (V_1, V_2) subspace do not reflect the true distances in full-dimensional space, one may ask for a representation that yields the best possible approximation of these distances in 2D. Here we use "best possible" in the sense that we aim at finding the plane, on which the distances obtained through the projection of the cluster centers deviates minimally from the original distances. This is obtained by a PCA on the cluster centers (sometimes referred to as principal coordinate analysis [18,76]) and subsequent projection on its first two eigenvectors. (We note that, in general, the latter are different from the first two eigenvectors of the whole data set [13, 76].) Figure 3.9 reveals that the distances in the resulting 2D representation (panel B) may differ from the distances in the (V_1, V_2) subspace (panel A).

For clarity, furthermore, Fig. 3.9(B) only displays arrows between clusters i and j, if their transition probability $\mathbf{T}_{ij} > 1.5\%$. While most transitions again occur between geometrically close clusters, there are also geometrically close clusters which only show very infrequent transitions, e.g., clusters **2** and **14** or **11** and **14**. From Table 3.1 we learn that those states are actually quite distinct as they differ in the conformations of several amino acids. Their geometrical similarity therefore represents an artifact of the projection on only two dimensions in the principal coordinate plot. Nevertheless, the transitions are correctly represented, as they were calculated from the clusters in five dimensions. In particular, this visualization clearly separates the all-α state **10** from the all-extended conformations (states **1** and its neighbors).

As a popular alternative, one may construct a free energy disconnectivity graph [10,

73, 74] of the conformational states of Ala_7 (see Sec. 3.2). As shown in Fig. 3.9(C), the disconnectivity graph directly displays the connectivity and the barriers between all states. Dividing up the energy landscape in six "basins", this representation readily reveals the hierarchy of the states. Moreover, we find many similarities with the 2D principal coordinate representation in panel (B). For example, the large geometric and kinetic separation of clusters **10** and **20** from all other states in the principal coordinate plot shows up as the highest barrier separation in the disconnectivity graph. Furthermore, the directed arrows point from states **2**, **3**, and **6** to state **1** but not vice versa. This corresponds to the fact that these states share the same basin and that the free energy

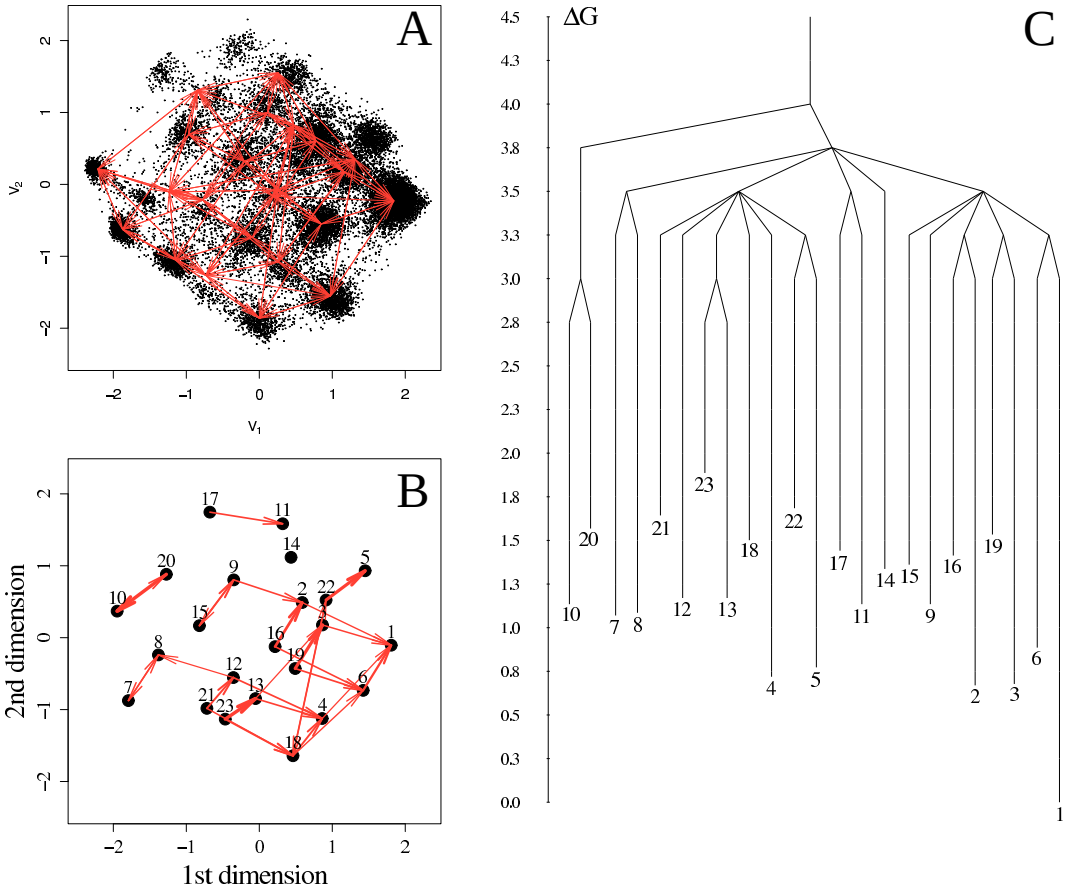

Figure 3.9: Visualization of the free energy landscape of Ala_7. Shown are (A) a two-dimensional cut $\Delta G(V_1, V_2)$ along the first two components including transitions with transition probability $\mathbf{T}_{ij} > 0.1\%$ between cluster centers, (B) a two-dimensional principal coordinate representation where only transitions with probability $\mathbf{T}_{ij} > 1.5\%$ are indicated (using a line width that is proportional to $\mathbf{T}_{ij}$), and (C) a disconnectivity graph of the system.

barrier to state **1** is almost one kcal/mol lower than from state **1** to the other states. As another example, we find that state **14** has its own basin in the disconnectivity graph, which is reflected by the absence of an edge with any other state in the principal coordinate plot.

3.7 Conclusions

We have outlined a systematic approach to construct a low-dimensional free energy landscape from a classical MD simulation. For this purpose, we have employed the dPCA. The dimensionality of the free energy landscape (i.e., the minimal number of PCs along which the energy is considered) results from the condition that the energy landscape reproduces the correct number, energy, and location of the system's metastable states and barriers. Restricting the analysis to a one- and two-dimensional energy surface may completely obscure the true connectivity of the conformational states (Fig. 3.1) and result in spurious barriers that can be smaller or larger than in full dimensionality (Fig. 3.2).

We have studied several criteria to determine the minimal number of PCs or number of "essential" components. As a simple rule, it is clear that all PCs with multi-peaked distributions need to be taken into account (Fig. 3.4). This is because the various peaks correspond to distinct metastable conformational states, while unimodal distributions intrastate fluctuations. The number of essential components is also reflected by their overall fluctuations and the distribution of time scales as shown by their autocorrelation functions (Fig. 3.5). Employing these criteria, it has been found that a five-dimensional dPCA energy landscape is a suitable and accurate representation of the full-dimensional landscape of Ala_7. In particular, we have performed various clusterings on the 5D landscape (Fig. 3.6) and obtained approximately the same metastable states and barriers as for a clustering of the full-dimensional surface (Table 3.1).

The resulting free energy landscape may be employed for interpretative purposes to schematically illustrate the main conformational states, barriers, and reaction pathways of a biomolecular system. With this end in mind, we have studied several approaches to visualize energy landscapes. Considering various two-dimensional cuts, we have shown that a color coding of k-means clusters nicely reveals the correlation of the free energy minima in the various representations (Fig. 3.4). To restrict the visualization of the energy

landscape to two dimensions, we have considered a principal coordinate analysis [i.e., a PCA on the cluster centers in (V_1, V_2)] which yields the best possible 2D approximation of the distances in full-dimensional space (Fig. 3.9). Connecting all clusters that make transitions to each other, this representation also facilitates a simple scheme of the transition network of the system. We have found that mostly neighboring states are connected, i.e., kinetically well separated clusters are also geometrically distinct in the first two principal components of the dPCA. The transition network of the principal coordinate analysis yields in many aspects similar information as a free energy disconnectivity graph, which directly displays the connectivity and the barriers between all states and also reveals the energy basins of the system.

The ultimate goal of this work is to construct a model of the dynamics in reduced dimensionality [33–36, 62, 63, 65]. If the process under consideration can be described by a Markov chain of metastable states, this effort is obsolete since a suitable clustering combined with a simple master equation provides the complete information of the time evolution of the system. In many biomolecular systems, however, the underlying assumption of a time scale separation between fast intrastate and slow interstate transitions may break down. As seen in the following chapters, in these cases, the dPCA energy landscape combined with, e.g., a Langevin simulation may facilitate a detailed investigation of biomolecular dynamics in low dimensions.

Chapter 4

Dynamics Simulations

In this chapter we will be concerned with the modeling of the dynamics of molecular dynamics (MD) simulations using methods from nonlinear time series analysis. We start with elaborating the necessary concepts of dynamical systems and time series analysis. Conducting a proof of principle, demonstrating that it is possible to first decompose the dynamics from an MD simulation in a relevant and an irrelevant part and then describe simpler models in reduced dimensionality, we aim at answering the question: How "complex" is the dynamics of peptide folding? Therefore we make use of the well-established concept of the complexity of a dynamic system in the theory of nonlinear dynamics. It is often associated with the fact that the "effective dimension" of the system [77], that is, the dimension of the subspace a trajectory $\vec{x}(t) \in \mathbb{R}^n$ will occupy in the course of its time evolution $\dot{\vec{x}}(t) = \vec{f}(\vec{x}(t))$, can be much smaller than n, the dimension the problem is formulated in. This dimensionality reduction is caused by nonlinear couplings which give rise to cooperative or synchronization effects and consequently reduce the effective number of degrees of freedom. In the case of MD simulations hard constraints such as covalent bonds and softer constraints such as intramolecular hydrogen bonds restrict the motion of the atoms, thus reducing the dimensionality.

The decomposition into "system" and "bath" variables is a crucial step before modeling the dynamics because it should ensure a time scale separation of these variables. The system variables should contain all slow large-amplitude motions of the molecule and hence represent conformational transitions while the bath variables only account for high-frequency oscillations which trigger the transitions. In this chapter we will apply

a deterministic model to describe the dynamics of peptide folding for various alanine chains. The significance of the concept becomes apparent in the case of a dissipative chaotic system, whose effective dimension typically is a noninteger number. Apart from its conceptional value, the effective dimension of a dynamic system is of practical interest since it may be calculated from measured or simulated data, e.g., by estimating the correlation dimension [78] or the Lyapunov exponents from which the Kaplan-Yorke dimension [79] might be obtained. While the dimension of the free energy landscape of the alanine peptides increases with system size, a Lyapunov analysis shows that the effective dimension of the dynamic system is rather small and even decreases with chain length. The observed reduction of phase space is a nonlinear cooperative effect that is caused by intramolecular hydrogen bonds that stabilize the secondary structure of the peptides.

In section 4.3 we will introduce another approach to describe the dynamics of peptide folding by a mixed deterministic and stochastic model. The method is based on the local estimation of the drift and diffusion Langevin vector fields.

4.1 Dynamical systems and time series analysis

In this section we want to provide the basic concepts of dynamical systems and time series analysis which we need for the interpretation and modeling of MD simulations. For more in-depth discussions of this broad subject see e.g. the books [80–83].

A continuous-time *dynamical system* is given by a set of ordinary differential equations

$$\dot{\vec{x}}(t) = \vec{f}(\vec{x}(t), t), \tag{4.1}$$

together with an initial condition $\vec{x}(0) = x_0 \in \mathbb{R}^d$.

As the data from MD trajectories is output at certain time-steps only (e.g. every 1 ps), we obtain a time series which is discrete in time. Henceforward we will restrict ourselves to discrete-time dynamical systems. For discrete-time dynamical system the time evolution is determined by a map

$$\vec{x}_{n+1} = \vec{f}(\vec{x}_n). \tag{4.2}$$

A dynamical system is called nonlinear if $\vec{f}$ is a nonlinear function. The space $\mathbb{R}^d$ is

referred to as the *phase space*. The sequence $x_0, x_1, x_2, \ldots$ obtained by iteration of (4.2) is called *orbit* or *trajectory*.

We distinguish between *conservative* or *Hamiltonian* dynamical systems and *dissipative* dynamical systems. A conservative system is volume preserving in the sense that the volume of an arbitrary volume element of phase space is preserved when it is evolved in time. This is equivalent to

$$|\det D\vec{f}(\vec{x})| = 1 \tag{4.3}$$

for all $\vec{x}$, where $D\vec{f}$ is the Jacobian matrix of partial derivatives of $\vec{f}$. If in some region

$$|\det D\vec{f}(\vec{x})| \neq 1, \tag{4.4}$$

then the system is called *dissipative*, and a small phase space volume either shrinks or expands. Consider the one-dimensional case where $|f'(x)| < 1$. Then, for a point y in a small neighborhood of x it holds

$$f(y) \approx f(x) + f'(x)(y - x), \tag{4.5}$$

which implies

$$\begin{aligned} |f(y) - f(x)| &\approx |f'(x)||y - x| && (4.6)\\ &< |y - x|, && (4.7) \end{aligned}$$

and hence the distance of x and y decreases after one iteration of the map f. It is typical for a dissipative system that many trajectories (depending on the initial condition x_0) are attracted by one or several certain subsets of phase space, that is, the trajectories come arbitrarily close and never leave a so called *attractor* for large enough times.

An attractor can simply be a stable fixed point of $\vec{f}$, for example, where the vicinity of the fixed point contracts in all directions. But often attractors reveal a much more complicated geometrical structure. They might even be *fractals*, a set showing self-similarity on arbitrary length scales, having noninteger dimension. An attractor is called *chaotic* if $\vec{f}$ displays exponentially sensitive dependence on initial conditions, that is, the distance between to nearby points on the attractor grows exponentially fast with time when the

dynamical system is evolved. Chaotic attractors are often fractals.

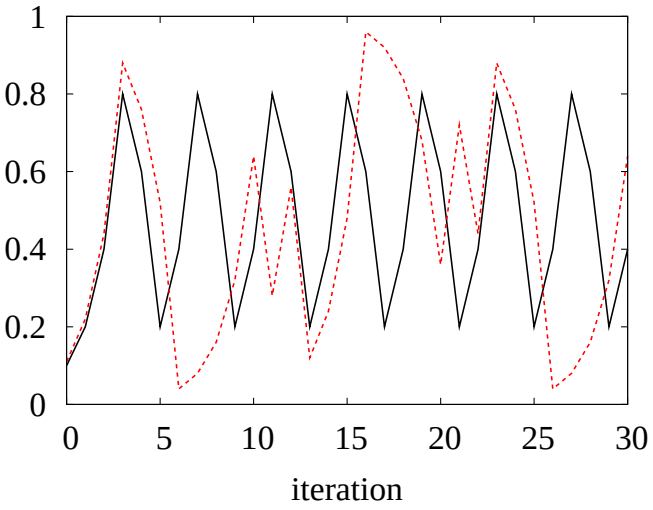

Figure 4.1: Orbits of the $2x$ modulo 1 map for initial conditions $x_0 = 0.10$ (full line) and $x_0 = 0.11$ (dashed line).

As an example for the sensitive dependence on initial conditions in one dimension we consider the $2x$ modulo 1 map [83]

$$x_{n+1} = 2x_n \text{ modulo } 1. \tag{4.8}$$

Fig. 4.1 shows two trajectories of the map which initially deviate by 10^{-2}. After only six iterations the difference between the orbits is almost 0.4 which is forty (!) times as much. Hence also small error in the initial conditions grows at a large rate rendering the exact long-term prediction impossible for computer simulations (since numbers are only stored up to a certain accuracy). For a chaotic attractor, this rate of divergence of nearby trajectories can be measured by the *Lyapunov exponents*. They are defined as

$$h_i = \lim_{N\to\infty} \frac{1}{N} \ln |\lambda_i(Df^N(\vec{x}_0))|, \quad i = 1, \ldots, d, \tag{4.9}$$

where the λ_i's are the eigenvalues of the matrix

$$Df^N(\vec{x}_0) = D\vec{f}(\vec{x}_{N-1}) \cdot D\vec{f}(\vec{x}_{N-2}) \cdot \ldots \cdot D\vec{f}(\vec{x}_0). \tag{4.10}$$

The existence of (at least one) positive Lyapunov exponents implies exponentially sensitive

dependence on initial conditions and thus chaotic behavior of the dynamical system. They correspond to expanding directions in phase space, whereas negative exponents correspond to contracting directions.

Using (4.9) and (4.10), we derive the Lyapunov exponent of the $2x$ modulo 1 map as follows,

$$\begin{aligned} h &= \lim_{N\to\infty} \frac{1}{N} \ln |f'(x_{N-1}) \cdot \ldots \cdot f'(x_0)| && (4.11)\\ &= \lim_{N\to\infty} \frac{1}{N} \ln(2^N) && (4.12)\\ &= \ln 2. && (4.13) \end{aligned}$$

As a rule, the Lyapunov exponent h is the average of the separation rate of an initial difference

$$|x_N - y_N| \approx \exp(hn)|x_0 - y_0|. \tag{4.14}$$

Thus, in the above example with $h = \ln 2$ we can expect an error $2^N\varepsilon$ after N iterations, if ε was the initial error.

To determine the complexity of an attractor, which is often a fractal if the attractor is chaotic, various dimension measures can be defined. Besides the topological dimension, the information dimension, and the box-counting dimension, we want to point out the *Kaplan-Yorke* or *Lyapunov dimension* [79]. It is defined as follows:

$$d_{\mathrm{KY}} = k + \frac{1}{|h_{k+1}|} \sum_{i=1}^{k} h_i, \tag{4.15}$$

where k is the number of Lyapunov exponents such that (if they are ordered decreasingly) the sum of the first k exponents is still positive or zero, whereas the sum of the first $k+1$ exponents is already negative. Loosely speaking, the definition of the leading term k in d_{KY} assures that the phase-space expanding ($h_i > 0$) directions just counterbalance the phase-space contracting ($h_i < 0$) directions, thus warranting an overall invariant phase-space volume, and thus an invariant set which is the attractor.

In experiments one cannot always or one does not want to measure all the components of the phase space vector $\vec{x}(t)$. Usually only one (or a few) component of a function of

$\vec{x}(t)$ is available,

$$g(t) = G(\vec{x}(t)) \in \mathbb{R}. \tag{4.16}$$

The aim of *delay* or *phase space reconstruction* is to convert these observations into state vectors to obtain phase space information on the geometry of the attractor. To allow for the reconstruction of the deterministic system from the projection given by (4.16), one might use *delay coordinates*. Therefore, the m-dimensional *embedding vector*

$$\vec{y}(t) = (g(t), g(t - \Delta t), g(t - 2\Delta t), \ldots, g(t - (m-1)\Delta t)) \tag{4.17}$$

is formed, where Δt is called the *lag* or *delay time* [84]. Provided that the embedding dimension m is large enough the attractor formed by $\vec{y}(t)$ has a qualitatively similar structure as the unknown attractor formed by the original trajectory $\vec{x}(t)$. This can be motivated by noting that $\vec{y}(t)$ actually can be seen as a function of $\vec{x}(t)$,

$$\vec{y}(t) = \vec{H}(\vec{x}(t)), \tag{4.18}$$

as $g(t - n\Delta t) = G(\vec{x}(t - n\Delta t))$, and $\vec{x}(t - n\Delta t)$ can be regarded as a function of $\vec{x}(t)$ by integrating Eq. (4.1) backwards in time by an amount $n\Delta t$. Under very general conditions H is well-defined and provides a one-to-one image between the two trajectories. For example, Lyapunov exponents do not change under this coordinate transformation, and hence it is feasible to calculate them from the embedded dynamics. If the dimension m of the embedding is too small, the mapping of $\vec{x}(t)$ to $\vec{y}(t)$ can produce self-intersections of $\vec{y}(t)$. This would violate the uniqueness of the orbit of a dynamical system. F. Takens could show in [84] that if m is larger than twice the box counting dimension, this is sufficient to avoid such effects. Note that this result is irrespective of the chosen lag time Δt for the embedding, but strictly valid only for perfectly noise-free data. In practice, when data is contaminated with noise, it is rather difficult to obtain good estimates of the lag time. Choosing a too large lag time, successive elements of the embedding vector will be almost independent (see Fig. 4.2A), and will give almost no further information than the single dimension. On the other hand a too small lag time will result in a strong correlation and similarity between successive elements as seen in Fig. 4.2B. Hence, unless m is very large, the deterministic structures of the dynamical system may become hard

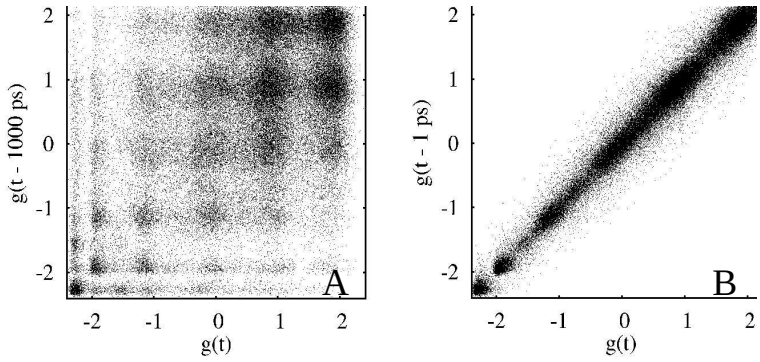

Figure 4.2: Exemplary delay embedding for the first dPCA mode of Ala_7. In (A) the delay time is 1 ns and in (B) $\Delta t = 1$ ps.

to distinguish. Here, visual inspection can help finding a reasonable lag time for the embedding.

4.2 How complex is peptide folding?

In this section, we now wish to apply the concept of dimensionality to the interpretation of classical MD simulations [85]. While MD simulations describe biomolecular processes such as folding and molecular recognition in atomic detail (i.e., $3N$-6 coordinates for an N-atomic system), it is clear that the many geometrical constraints of the molecule (e.g., covalent and hydrogen bonds) result in a considerable reduction of the effective number of degrees of freedom. As detailed in the previous chapters, in practice, the structural dynamics of biomolecules is often described in terms of the molecule's free energy landscape, which is represented as a function of empirically introduced reaction coordinates. As already thoroughly studied in the course of this thesis, alternatively, one may employ a principal component analysis of the trajectory. While these coordinates in some sense represent the essential dynamics of the system [16], in general it is not clear how to determine the effective dimension of a biomolecular MD simulation, since there always is some ambiguity in the choice of the reaction coordinates. As a first attempt to assess the complexity of a biomolecular system, in this work we (i) perform MD simulations of various peptide systems and extract time series that account for their structural dynamics, (ii) construct a deterministic model of the dynamics using methods from nonlinear time series analysis, and (iii) perform a Lyapunov analysis to calculate

their effective dimension.

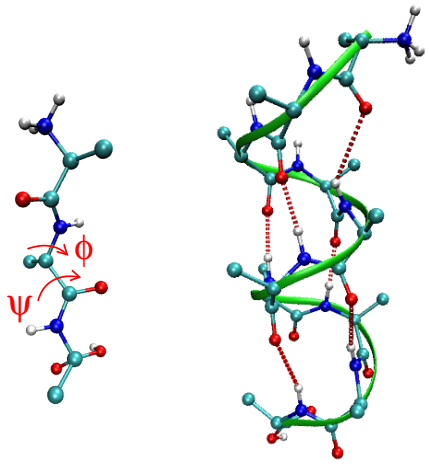

Figure 4.3: MD snapshots of (left) an extended conformation of Ala_3 showing the central backbone dihedral angles ϕ and ψ, and (right) the α_R helix conformation of Ala_{10} indicating the stabilizing $n-(n+4)$ hydrogen bonds.

As molecular systems we have chosen the alanine peptides Ala_n with $n = 3, 5, 7$ and 10 in aqueous solution (see Fig. 4.3), for which 100 ns MD simulations at 300 K were performed using the GROMACS program suite [50], the GROMOS96 force field 43a1 [51], and the SPC water model [52] (for details see appendix 6.4). Unlike to proteins, these systems are too small to adopt a stable native structure, but exhibit reversible folding and unfolding of their secondary structure. Since this large amplitude motion results in a strong mixing of internal and global motion (while only the internal motion is of interest), we choose internal coordinates to describe the peptide structure, i.e., their (ϕ_k, ψ_k) backbone dihedral angles ($k = 2, \ldots, n-1$), see Fig. 4.3. To circumvent problems associated with the fact that angles are circular variables we employ the dPCA procedure as detailed in Chap. 2, i.e. the angles are mapped onto a Cartesian-like space via $x_{4k} = \cos\phi_k$, $x_{4k-1} = \sin\phi_k$, $x_{4k-2} = \cos\psi_k$, and $x_{4k-3} = \sin\psi_k$, resulting in $4(n-2)$ variables [25,37]. To remove linear correlations, a PC analysis of the MD trajectory $\vec{x}(t)$ is performed, yielding the PCA eigenvectors $\vec{u}_i$ and the corresponding PCs $v_i(t) = \vec{x}(t) \cdot \vec{u}_i$, which serve as a time series for the subsequent analysis.

As a first example, Fig. 4.4 shows the time series $v_i(t)$, the distributions $P(v_i)$, and the autocorrelation functions $C_i(t)$ obtained for the first two PCs of the Ala_3 system. Both

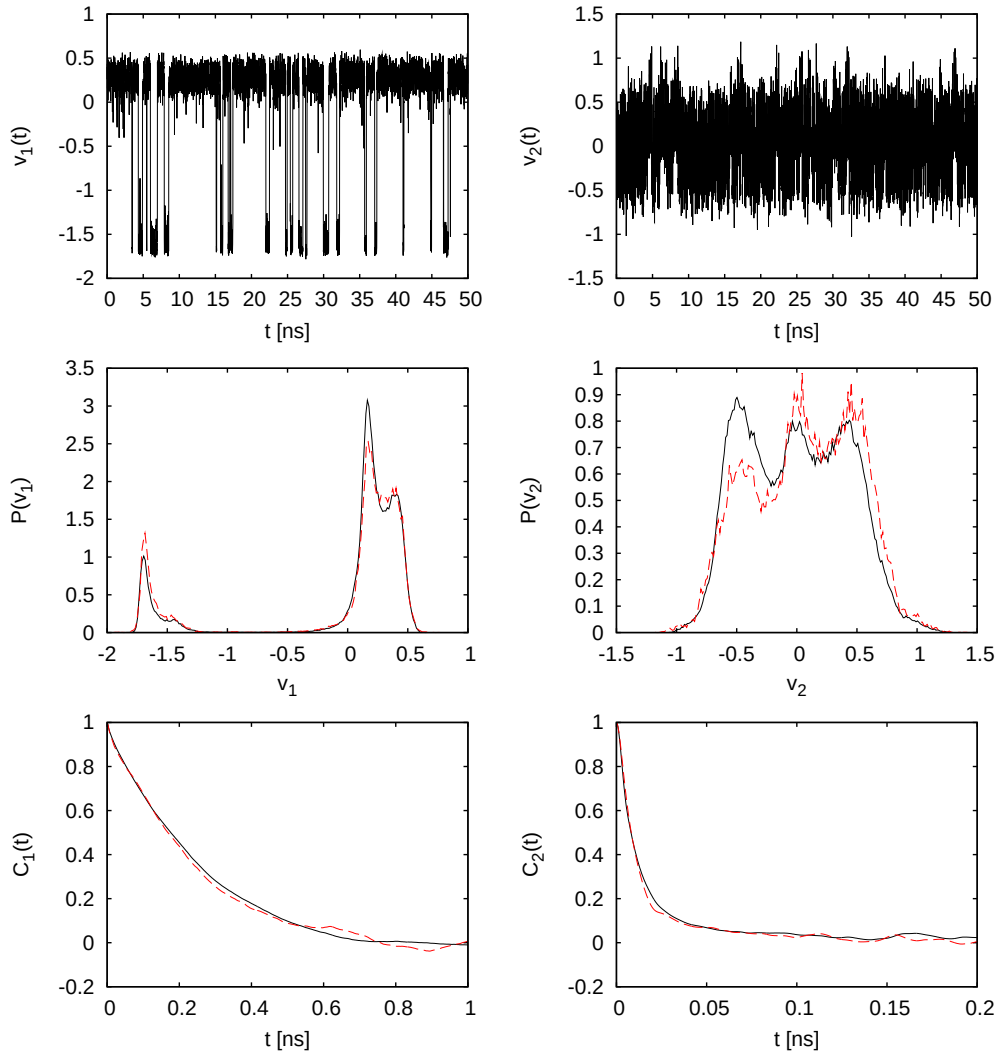

Figure 4.4: Time series $v_i(t)$, distributions $P(v_i)$, and autocorrelation functions $C_i(t)$ obtained for the first two principal components of the Ala_3 system. The solid black lines represent the results of the MD simulation, the dashed red lines correspond to results from the nonlinear model of the dynamics.

distributions exhibit multiple peaks which correspond to different conformational states of the peptide. For the first component, the peak at $v_1 \approx -1.7$ reflects the right-handed helix conformation α_{R}, while the peak at $v_1 \approx 0.2$ reflects extended conformations of the peptide (see Fig. 4.3). Invoking the second PC, the latter can be decomposed in the poly-L-proline II ($\mathrm{P_{II}}$) conformation and the fully extended (β) conformation [46]. The

transitions between these states occur on a 200 ps ($\alpha_{\mathrm{R}} \leftrightarrow \beta$) and 20 ps ($\mathrm{P_{II}} \leftrightarrow \beta$) time scale, respectively. While the three conformational states of $\mathrm{Ala_3}$ can be described using only two PCs, the situation is more involved for the longer peptides. For the $\mathrm{Ala_{10}}$ system, for example, Fig. 4.5 shows that the distributions of the first two PCs are characterized by a prominent double peak corresponding to an α_{R}-type folded state (see Fig. 4.3), and a large range of extended and intermediate states corresponding to unfolded structures of the peptide. To discriminate these states, in total eight PCs are required. An analysis of the time evolution of the first PC reveals collective conformational transitions, accounting for the reversible folding and unfolding of the secondary structure of the peptide.

Performing a PC analysis of a MD trajectory, only the distribution of the first, say d_{EL}, PCs exhibit multiple peaks, while the remaining distributions $P(v_i)$ with $i > d_{\mathrm{EL}}$ are single-peaked and approach a Gaussian shape with increasing i [16]. That is, the distributions $P(v_i)$ with $i > d_{\mathrm{EL}}$ describe the fluctuations of the peptide within a specific conformational state, while the distributions with $i \leq d_{\mathrm{EL}}$ define these conformation states. Hence d_{EL} can be considered as the dimension of the free energy landscape $\Delta G(\{v_i\}) \propto -\ln P(\{v_i\})$, because ΔG shows nontrivial structure only along the first d_{EL} PCs. As listed in Table 4.1, d_{EL} increases with system size, i.e., from 2 for $\mathrm{Ala_3}$ to 8 for $\mathrm{Ala_{10}}$.

It should be emphasized, however, that the energy landscape dimension d_{EL} is conceptionally different from the effective dimension of the dynamics in phase space. In principle, the latter can be obtained directly from a Lyapunov analysis of the MD trajectory [86]. In practice, though, the ubiquitous noise on the data prevents an accurate calculation of the Lyapunov exponents, which account for the sensitivity of the trajectory with respect to infinitesimal small deviations of its initial conditions. To overcome this problem, we employ the methods of nonlinear time series analysis [80] and construct a deterministic model of the dynamics which reproduces the main features of the MD data, but at a much better signal to noise ratio.

With this end in mind, we assume that the dynamics of the system that produces the the time series $\vec{v}(t)$ can be expressed by the Langevin equation

$$\dot{\vec{v}}(t) = \vec{f}(\vec{v}(t)) + \vec{\eta}(t). \tag{4.19}$$

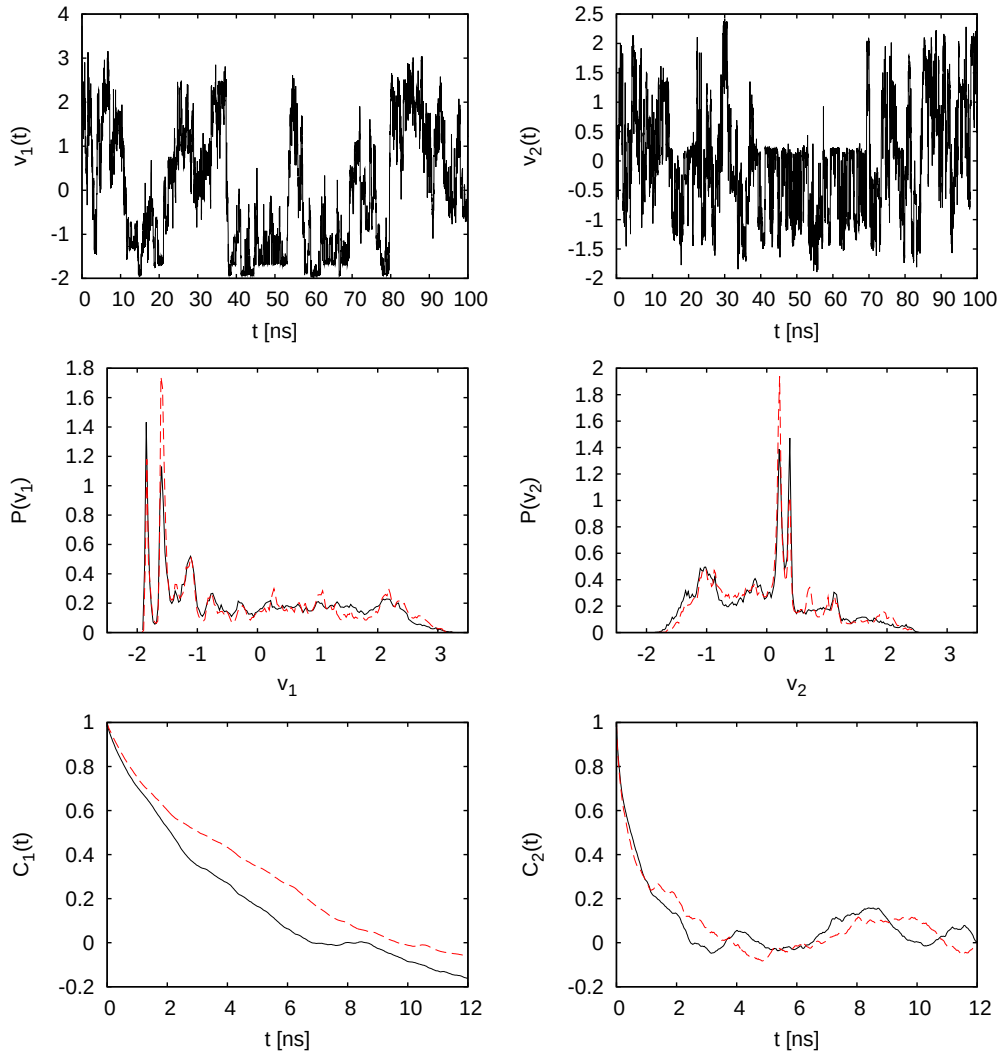

Figure 4.5: Same as in Fig. 4.4, but for the Ala_{10} system.

Here $\vec{f}$ describes the deterministic part of the dynamics, while $\vec{\eta}$ denotes a stochastic driving term which represents the fluctuations of all degrees of freedom we want to ignore, including, e. g., high-frequency bond oscillations, the motion of the solvent, and the realization of the external heat bath. In order to obtain a simple model for the deterministic part $\vec{f}$ of the dynamics, the following steps are taken. First, we restrict the analysis to the first d_{EL} PCs, thus disregarding all components accounting for simple Gaussian fluctuations. Since only the deterministic part is subject of the dimensionality

reduction (the noise term $\vec{\eta}(t)$ by definition explores all directions of phase space), we also neglect the stochastic driving in Eq. (4.19). This is realized by applying a simple noise reduction scheme, i.e., the Savitzky-Golay or least-squares filter [87] to the resulting trajectory $\vec{v}(t) \in \mathbb{R}^{d_{EL}}$. The filter is applied to each dimension separately. To understand the Savitzky-Golay filter consider a single data point, e.g. $v_1(t_j)$, which we want to replace by some kind of local average (in time) of surrounding data points. Taking a window of n data points earlier than t_j and n data points later than it, we obtain a window of length $2n + 1$. The idea of the noise reduction filter is to approximate the data within the window by a polynomial of typically quadratic or quartic order. This is realized by least-squares fitting a polynomial to the $2n + 1$ data points in the window, and then replacing the point $v_1(t_j)$ by the value of the polynomial at point t_j. In Fig. 4.6 we see

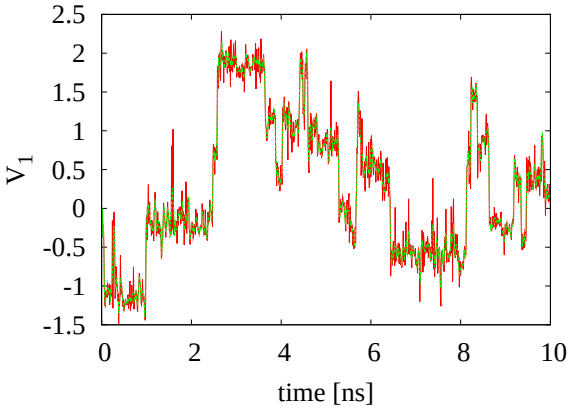

Figure 4.6: Time series of the MD simulation along V_1 of a dPCA for Ala_7 (full line), together with its noise reduced time series after application of the Savitzky-Golay filter (dashed line).

10 ns of MD simulation of Ala_7 along the first principal component of a dPCA together with the filtered data, using a window of 100 ps length and polynomials of quartic order for the Savitzky-Golay noise reduction scheme. The noisy data is clearly smoothed while the conformational transitions are still correctly reproduced. The third and final step is to construct a state space in which the trajectory $\vec{v}(t)$ shows a deterministic behavior. Since $\vec{v}(t)$ represents a projection of the original phase space (that explicitly includes the positions and momenta of all atoms of the system), in general we can not expect that the dimension d_{EL} of the trajectory is sufficient for this purpose. To account for a possibly

higher-dimensional phase space, we use an extension of the Takens embedding [84] and embed the first (and most important) PC until this component is reconstructed sufficiently well. Adding the remaining $d_{\mathrm{EL}} - 1$ components and using a time delay Δt, the resulting embedding vector at time step t_j reads

$$\begin{aligned}\vec{v}_j \equiv (\quad & v_1(t_j) \quad , v_1(t_j - \Delta t), \ldots, v_1(t_j - (m-1)\Delta t), \\ & v_2(t_j) \quad , v_3(t_j), \ldots, v_{d_{\mathrm{EL}}}(t_j))^T , \end{aligned} \tag{4.20}$$

where m denotes the embedding dimension of the first PC, resulting in a dimension $d_{\mathrm{RS}} = d_{\mathrm{EL}} + m - 1$ for the reconstructed state space. For all systems considered, $m = 9$ and Δt from 0.2 ps (Ala_3) to 4 ps (Ala_{10}) were used.

Knowing the state space, we are now in a position to fit a deterministic nonlinear model to the data. Following Farmer and Sidorowich [88], we employ a locally linear model defined by the map

$$\vec{v}_{j+1} = \mathbf{A}_j \vec{v}_j + \vec{b}_j, \tag{4.21}$$

where $\mathbf{A}_j$ is a $d_{\mathrm{RS}} \times d_{\mathrm{RS}}$ matrix. Locally linear means that, given the vector $\vec{v}_j$ at time t_j, the subsequent vector $\vec{v}_{j+1}$ at time t_{j+1} is obtained in linear approximation from Eq. (4.21). The model parameters $\mathbf{A}_j$ and $\vec{b}_j$ are obtained by a least squares fit which only uses the spatial neighbors of $\vec{v}_j$ [80, 89]. As a consequence, the model parameters need to be calculated for every time step of the model trajectory. To validate the model, we again consider the distributions and autocorrelation functions of the first two PC of Ala_3 (Fig. 4.4) and Ala_{10} (Fig. 4.5) and compare the modeled data to the results obtained from the MD simulations. Reproducing the time scales of the dynamics as well as the conformational distribution in almost all details, the model accounts nicely for the essential features of the MD data.

Let us now turn to the Lyapunov exponents λ_i, $(i = 1, \ldots, d_{\mathrm{RS}})$ of the peptide dynamics, which are calculated through the Jacobian matrix $\mathbf{A}_j$ of the map (4.21). For all systems considered, we found two positive exponents λ_1 and λ_2, which quantify the chaoticity of the dynamics in phase space. We first consider the Kolmogorov-Sinai entropy h_{KS}, which is given by the sum of the positive Lyapunov exponents [80]. It's reciprocal value $\tau_{\mathrm{KS}} = 1/h_{\mathrm{KS}}$ is an estimate for the time span the evolution of the trajectory can be

	Ala_3	Ala_5	Ala_7	Ala_{10}
d_{EL}	2	3	6	8
d_{KY}	5.0	4.7	4.9	3.3
n_{HB}	-	0.03	0.6	2.4
τ_{KS} [ps]	3.8	3.7	5.9	8.0

Table 4.1: Comparison of d_{EL}, the dimension of the energy landscape, and d_{KY}, the effective dimension of the dynamics, as obtained for various alanine peptides. Also shown are n_{HB}, the average number of α_R-type $i-(i+4)$-intramolecular hydrogen bonds, and τ_{KS}, the reciprocal value of the Kolmogorov-Sinai entropy.

forecasted. As shown in Table 4.1, this picosecond time scale increases with system size, thus indicating that the structural dynamics of the larger peptides is less chaotic than the dynamics exhibited by the smaller systems.

To estimate the effective dimension d_{KY} from the Lyapunov exponents, we employ the Kaplan-Yorke conjecture [79] as given by Eq. (4.15). Table 4.1 lists the resulting values of the effective dimension d_{KY} obtained for Ala_3 through Ala_{10}. Ranging from ≈ 3 to 5, the dimensions appear to be quite small, considering that it accounts for the motion of thousands of atoms. Most intriguing, though, is the fact that the effective dimension *decreases* with system size, from 5 for Ala_3 to 3.3 for Ala_{10}. This is in striking contrast to the behavior of the energy landscape dimension d_{EL} which –as expected– increases with chain length.

To explain this finding, detailed analyses of the all-atom MD trajectories were performed, which revealed that the effect is caused by intramolecular interactions that stabilize the secondary structure of the peptide. Most importantly, this is achieved by intramolecular hydrogen bonds connecting the ith and $(i+4)$th residues of the amino acid chain, thus stabilizing the α_R helix structure (see Fig. 4.3). As shown in Table 4.1 as well as in Fig. 4.7, the average number of these hydrogen bonds increases significantly, once the number of possible α_R-type bonds reaches three for Ala_7. Remarkably, the formation of stabilizing hydrogen bonds seems to significantly reduce the effective dimension, although these bonds are not stable but formed and broken on a nanosecond time scale.

It is interesting to note that this decrease of the effective dimension is not observed for the energy landscape dimension d_{EL}. Apparently, this is because the latter quantity is defined in the linear framework of PC analysis theory, whereas the effective dimension

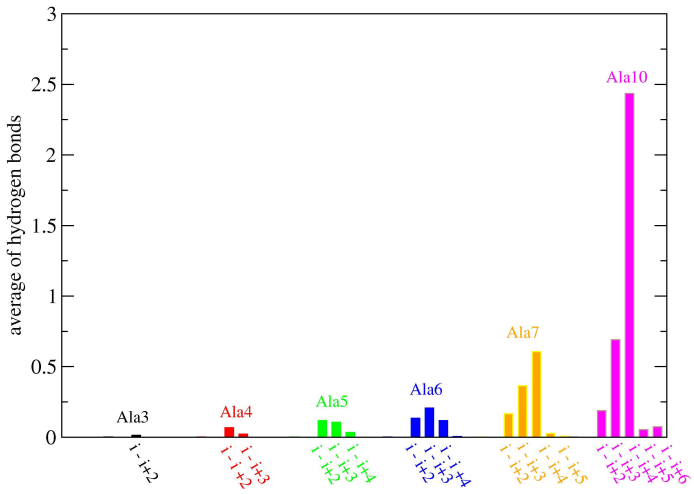

Figure 4.7: Number of hydrogen bonds connecting the ith and $(i+j)$th residues of the alanine chains averaged over all snapshots of the respective trajectories.

d_{KY} is obtained from a nonlinear description of the dynamics. In a similar vain, other nonlinear methods for the analysis of biomolecular dynamics have been proposed that are sensitive to nonlinear correlations and therefore may reduce the dimensionality of the problem [33, 34]. The effect of nonlinear dimensionality reduction is supposedly even more important for the folding of larger peptides and proteins, which exploit a variety of stabilizing interactions and exhibit significant cooperativity [90].

4.3 Multidimensional Langevin modeling

In the deterministic approach described in the last section we have eliminated the influence of the bath variables by using a noise reduction scheme. By doing so, we obtained a deterministic model which allowed for calculating e.g. Lyapunov exponents and hence the estimation of the effective dimension. The more general approach is to first rewrite the general multidimensional Langevin equation (4.19) as

$$\dot{\vec{v}}(t) = \vec{f}(\vec{v}(t)) + \mathbf{D}(\vec{v}(t))\vec{\epsilon}(t), \tag{4.22}$$

where we replaced the stochastic driving $\vec{\eta}(t)$ by the diffusion operator $\mathbb{D}(\vec{v}(t))$ which contains all spatial and temporal dependencies of the driving and a Gaussian-distributed white noise process $\vec{\epsilon}(t)$ which has variance σ. The goal of the Langevin approach is to estimate drift and diffusion from the MD data. R. Hegger et al. showed [66] that under very weak assumptions one can locally obtain the vector fields $\vec{f}$ and $\mathbf{D}$ of the discretized version of (4.22)

$$\Delta\vec{v}_n = \vec{v}_{n+1} - \vec{v}_n = \vec{f}(\vec{v}_n) + \mathbf{D}(\vec{v}_n)\vec{\epsilon}_n, \tag{4.23}$$

by the local average and covariance matrix of the position difference $\Delta\vec{v}_n$. That is

$$\vec{f}(\vec{v}_n) = \langle\Delta\vec{v}_n\rangle \tag{4.24}$$

$$\sigma^2\mathbf{D}(\vec{v}_n)\mathbf{D}^T(\vec{v}_n) = \langle\Delta\vec{v}_n\Delta\vec{v}_n^T\rangle - \langle\Delta\vec{v}_n\rangle\langle\Delta\vec{v}_n^T\rangle, \tag{4.25}$$

where the average $\langle\cdot\rangle$ is taken over spatial neighbors of the point $\vec{v}_n$. This method has been implemented and tested by R. Hegger et al. and showed promising results for the modeling of the dynamics on the free energy landscapes for Ala_3 and Ala_7. The distributions and the autocorrelation functions of all the principal components (serving as reaction coordinates) as well as the lifetimes of metastable states have been correctly reproduced by the Langevin model. The approach can be used for obtaining a continuous trajectory from many short replica exchange MD simulations as it uses only pairs of adjacent trajectory points for the estimation of the drift and diffusion. Also nonequilibrium simulations can be easily conducted. One can e.g. restart several trajectories from the same nonequilibrium point and study relaxation times. We will use it in the following chapter when we model the dynamics of a variant of the villin headpiece subdomain.

4.4 Conclusions

After having presented the basic concepts of dynamical systems and nonlinear time series analysis, we presented a deterministic model for the dynamics of short alanine chains. This allowed for calculating the "effective dimension" of the systems. A Lyapunov analysis revealed that, while the dimensionality of the free energy landscape increases with system size, the effective dimension of the dynamic system remains rather small and even

decreases with chain length. This effect was shown to be caused by intramolecular hydrogen bonds causing a nonlinear cooperative effect. We also presented a mixed deterministic and stochastic computational approach to describe the conformational dynamics in reduced dimensionality. This method was based on the local estimation of the drift and diffusion vector fields of a general Langevin equation for the dynamics. While the work presented here is only a first step towards a nonlinear analysis of MD data, it may open ways to address the larger problem of describing folding processes. For example, we wish to study if the folding of various structural motifs such as α_{R}-helices and β-sheets results in distinguishable properties of the corresponding dynamical model. Another next step is to go beyond the locally linear ansatz and construct analytical models of the dynamics. Such analytical models would contain a set of parameters which presumably depend on, e.g., experimental conditions, amino-acid sequence, and folding motifs. The study of this parameter dependence could then shed some light on the still elusive mechanism of folding.

Chapter 5

Applications to larger systems - an outlook

So far we have developed statistical and dynamical methods for the construction, interpretation, and modeling of the free energy landscape of relatively small peptides. Analyzing these relatively well-understood systems put us in a position to extensively test our methods. For example, in Sec. 3.4 of Chap. 3 we assessed the quality of the free energy landscape of Ala_7 in reduced dimension with respect to the full-dimensional landscape. Serving as a reference, a direct clustering of the full-dimensional dihedral angle space was only feasible or reasonable because of the small system size.

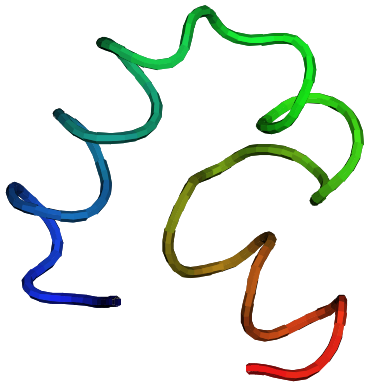

Figure 5.1: Experimental crystal structure of the 35 residue variant of the villin headpiece subdomain (HP-35 NleNle). The backbone is colored from red at the N terminus to blue at the C terminus.

In this chapter we want to analyze molecular dynamics simulations generated by the world-wide distributed computing project Folding@home [3]. Every user of a personal computer can download a client software that utilizes unused computer power to perform the simulations in the background or via a screen saver. In this way Folding@home became the world's most powerful distributed computing cluster according to Guinness World Records.

In April 2008 molecular dynamics trajectories of a villin variant, as described in [4], became available for download [91]. The variant of the villin headpiece subdomain (HP-35 NleNle) is the fastest-folding protein yet discovered, folding on a time scale of 1 μs. Its native state is shown in Fig. 5.1. Using Folding@home, despite the large system-size of about 10,000 atoms, hundreds of all-atom, explicit solvent MD simulations of the 35 residue subdomain could be performed, each on a time scale comparable to experimental folding time, resulting in a total simulation time of almost half a millisecond.

In the following we will perform Cartesian PCA and dihedral angle PCA on the villin trajectories. Applying our methods to a much larger system than before, we point out the differences and similarities. Thereafter we construct a multidimensional Langevin model for the dynamics of the system from which we can estimate folding times that we compare to the folding times obtained by Ensign et al. [4].

5.1 Free energy landscapes for the villin system

The many hundreds of trajectories of the villin project are organized in 2 projects, one starting from 9 different unfolded conformations (PROJ3036) and one starting from the experimental structure (PROJ3037). The trajectories starting from structure k, are found in RUNk. Each RUN contains up to 100 continuous trajectories with maximum length of 2 μs each.

We start our analysis using the first 5 trajectories (CLONE0-CLONE4) of the unfolded conformation 0 (RUN0, see Fig. 5.2) of PROJ3036. Therefore we simply concatenate the 5 trajectories resulting in approximately 9 μs of simulation time. First we perform a dPCA on the 66 backbone dihedral angles $\{\phi_2, \psi_2, \ldots, \phi_{34}, \psi_{34}\}$. The resulting free energy landscape $\Delta G(V_1, V_2)$ is presented in Fig. 5.3A. We clearly distinguish several free energy minima on the 2D projected landscape. We note that until the 10th PCs we

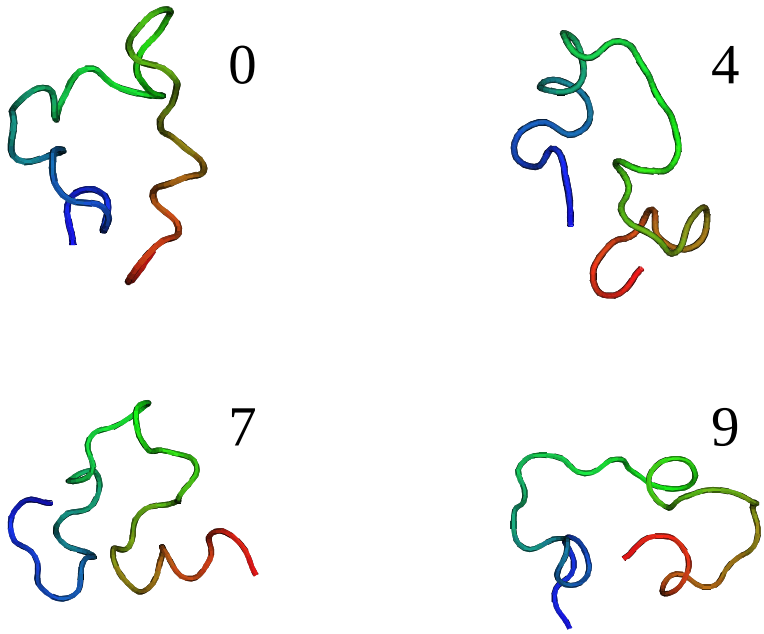

Figure 5.2: Starting structures for RUNs 0,4,7,9 the 35 residue variant of the villin headpiece subdomain (HP-35 NleNle).

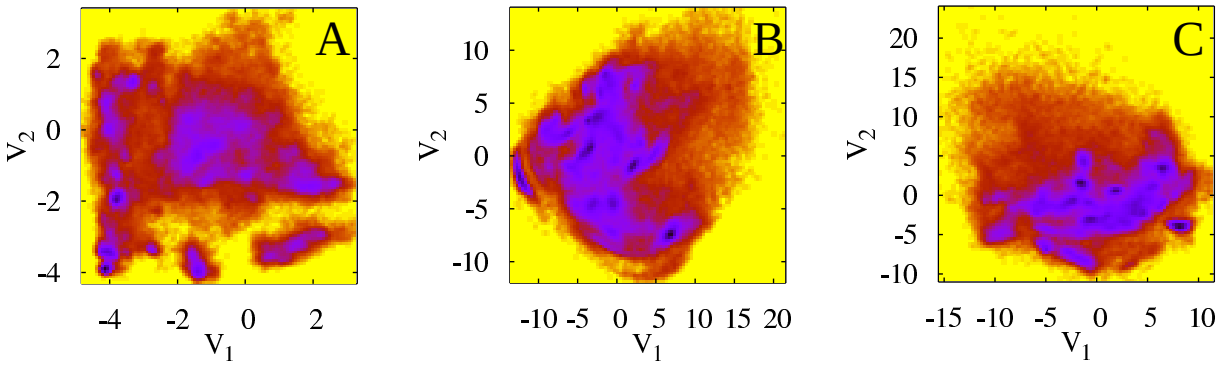

Figure 5.3: Free energy landscapes of the villin headpiece subdomain as obtained from the first 5 trajectories of RUN0. (A) shows the results along the first two principal components obtained from a dPCA. (B) and (C) display the landscape obtained by a Cartesian PCA using the starting and the native conformation as reference structure for the least-squares fit, respectively.

obtain clearly multipeaked distributions, i.e. structural information in the free energy surface. Interestingly, a Cartesian PCA on the C_α atoms (Fig.5.3B and C) reveals several minima on the landscape as well. This finding clearly differs from the case for short alanine chains were the free energy landscapes obtained by Cartesian PCA appeared to be smooth and unstructured. This was shown to be caused by a mixing of internal and

overall motion. Here, for the villin system also the reference structure seems not to make a qualitative difference, as the landscape obtained by fitting the trajectory to the unfolded state has a comparable amount of structure when fitting to the native state. Noting that the trajectory of CLONE1 was the only one of this set that folded from unfolded structure 0 ($V_1 \approx -1.3, V_2 \approx 0.6$ in (A)) to a native-like conformation ($V_1 \approx -3.6, V_2 \approx -3.6$ in (A)) we move on to analyzing the full data set of the many hundreds of trajectories all at once.

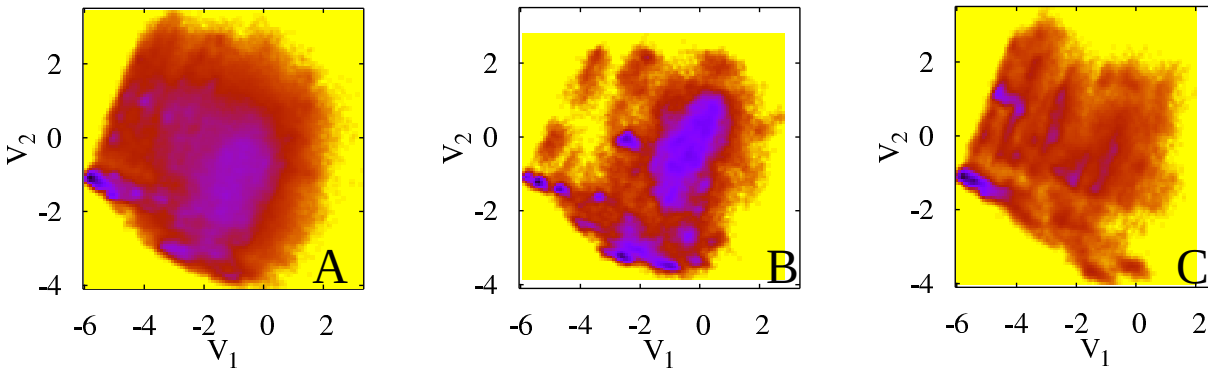

Figure 5.4: Free energy landscapes of the villin headpiece subdomain as obtained from (A) all trajectories of PROJ3036 ($\approx 400\mu$s), (B) 5 trajectories (CLONE0-CLONE4, $\approx 9\mu$s) starting from the unfolded conformation 0, and (C) 10 trajectories (CLONE0-CLONE9, $\approx 15\mu$s) starting from conformation 4. All landscapes are visualized along the same first two principal components as obtained from a dPCA on all the trajectories of PROJ3036.

We performed a dPCA on all the trajectories from PROJ3036 which consists of almost 400 μs of simulation. The free energy landscape $\Delta G(V_1, V_2)$ in Fig. 5.4A exhibits one clear minimum corresponding to native-like structures, while the rest of the landscape seems to be quite structureless at a first glance. In the full data set the population of native-like states at $V_1 \approx -6$ is very high compared to a single unfolded conformation. This is one reason why it is hard to distinguish structure in the unfolded part of the landscape. Excluding the native state from the landscape the remaining part still looks quite smeared out (data not shown). Does this suggest that the unfolded part of the free energy landscape for the villin system is unstructured or even random? Projecting only a small number of trajectories onto this landscape in 5.4B we see that actually the landscape is quite structured. As already mentioned above only one of the 5 latter trajectories samples the native state, hence the native state is not as much populated

as when taking the whole set of trajectories into consideration. Thus, the landscape is not dominated by the native structure and free energy minima can be distinguished in the unfolded region. Also the first 10 trajectories starting from structure 4 (CLONE0-CLONE9 of RUN4) reveal even more peaks on the landscape as can be seen in 5.4C.

We can conclude two things. The first observation is somewhat trivial. The dominant native structure renders it almost impossible to distinguish structure in the unfolded region of the free energy landscape. The second result is that taking more and more trajectories into account which were simulated starting from different structures, the 2D representation of the free energy landscape $\Delta G(V_1, V_2)$ kind of fills up with energy minima which lie geometrically close in this representation, resulting in a landscape which looks smeared out. Along other modes V_k the free energy landscape doesn't look more structured either (data not shown). It can well be that even though in the full dimensional sin/cos space (2*66 angles = 132 dimensions) the free energy minima can be clearly distinguished from each other, on every 2D projection of the landscape the minima come together giving this smeared out picture. This effect did not occur in the case of the shorter alanine peptide chains. But there, there were not as many conformational states as there seem to be for the villin system. Even if every amino acid is only treated as a two-state system being either in the α- or the β/P_{II}-region, we already have $2^{33} \approx 10^{10}$ theoretically possible conformations. To compare, we distinguished 32 conformational states for the heptaalanine system. It still remains a challenge for future work to classify the conformational diversity of the villin system.

5.2 Langevin dynamics for the villin system

We now wish to model the dynamics for the villin system using the multidimensional Langevin model as described in section 4.3 in Chap. 4. To obtain a model we restrict ourselves to a subset of trajectories as it would be computationally too costly to estimate the drift and diffusion vector fields from the whole data set. In order to sample well the phase space we choose one trajectory for every starting structure, that is, CLONE0 of all 10 RUNs of PROJ3036 and CLONE0 from PROJ3037, and concatenate them as seen in Fig. 5.5. Note that, as the Langevin model does not require a continuous trajectory, this is a feasible approach. When concatenating trajectories we mark the last point of

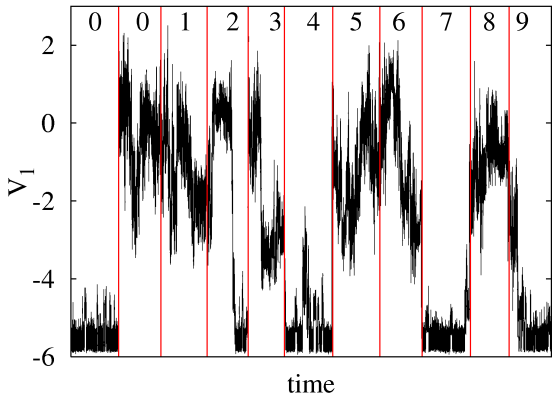

Figure 5.5: Time evolution of 11 concatenated trajectories along V_1 as obtained by a dPCA of all 11 trajectories. Each trajectory corresponds to a simulation time of $\lesssim 2$ μs. The first trajectory of RUN0 (PROJ3037) is followed by the first trajectories of RUNk, $k = 0, \ldots, 9$ (PROJ3036), respectively. The numbering given in the upper row of the figure is according to the starting structure k. Native-like structures correspond to a value of $V_1 \approx -5.5$.

each trajectory in order not to use it for the estimation of the drift and diffusion, as these points undergo false transitions originating only from the concatenation. We can observe that trajectories which reach native-like structures tend to stay there as far as one can tell from the maximum continuous simulation time of 2 μs. This is the case for the trajectories from RUNs 0 (PROJ3037),2,4,7,9. This is in agreement with the fact that the trajectories of PROJ3037 show a stable behavior staying close to the native structure.

Using the 11 trajectories we now wish to find the parameters for our dynamic model. Therefore we need to determine the embedding vector with its dimension m and an appropriate delay time Δt, as well as the number of spatial neighbors k for the estimation of the drift and diffusion fields. We tried out various embeddings in order to find a suitable model. For example, we used an 8-dimensional embedding vector where we embed the first dPCA component 5 times with lag time $\Delta t = 50$ ps and then add the next 3 components, i.e.

$$\vec{v}_j \equiv (v_1(t_j), \ldots, v_1(t_j - 4\Delta t), v_2(t_j), v_3(t_j), v_4(t_j)) \,. \tag{5.1}$$

The evolution of the resulting Langevin dynamics using $k = 5$ neighbors for the local estimations of the drift and diffusion fields can be seen in 5.6A. This model cannot be appropriate as it frequently leaves the native state and makes transitions to the unfolded

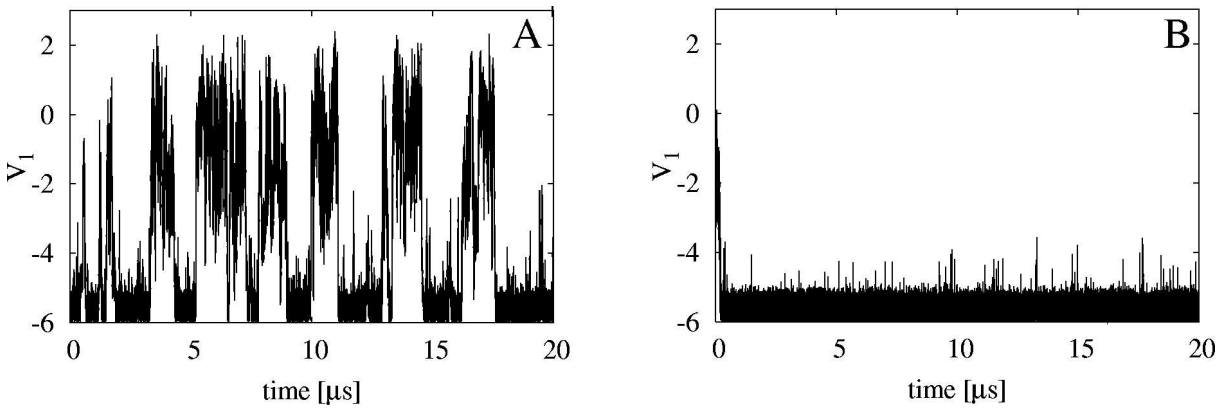

Figure 5.6: Time evolution of the first principal component using a Langevin simulation of the villin headpiece. (A) was generated using a 5D model, while for (B) the dynamics was modeled in 14 dimensions. Native-like structures correspond to a value of $V_1 \approx -5.5$.

part of the free energy landscape. A truthful model should tend to stay in the native state once it is reached. Using the first 10 modes of the dPCA, which are all multipeaked modes, we embed the first 5 modes in 2 dimensions, respectively, with a much larger lag time of 500 ps. We then add the modes 6-10 to the embedding vector. From 5.6B we see that the resulting Langevin model with $k = 50$ after folding to the native state stays there for all the simulation time of 20 μs. Thus, the necessary condition for a good model is fulfilled.

Now we want to apply this model to estimate folding times from a reduced data set. Therefore we used the first ten trajectories of RUNs 4, 7, and 9, respectively. For these three different starting we calculated the dPCA free energy landscape, and ran 1000 Langevin simulations for each starting structures with the above derived model. We stopped a Langevin run when it reached a native-like state which we determined by the free energy minimum on the respective landscape which corresponds to the native structure. In such a way we obtained the distribution of folding times as presented in Fig. 5.7. The mean of the folding times are 450 ns, 100 ns, and 1.1 μs for structures 4, 7, and 9, respectively.

Let us compare our results to the time scales found by Ensign et al. [4] which were calculated by analyzing all trajectories (instead of only 10) from the respective RUN. They estimate folding times of structures 4 and 7 to be 746 ns and 417 ns, respectively. For structure 9 the timescale was estimated to be of the order of 5 μs. Note, that our

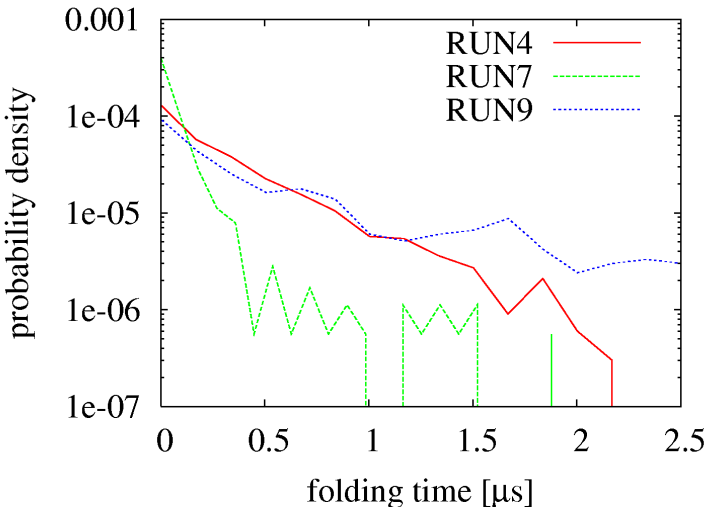

Figure 5.7: Distribution of relaxation times to native-like structures from unfolded structures 4 (450 ns), 7 (100 ns), and 9 (1.1 μs). The mean of the distribution is given in brackets. For every starting structure the distribution was calculated from 1000 Langevin simulations which were stopped when reaching a native-like state.

folding times qualitatively reproduce the reference results. Nevertheless, in comparison, our approach underestimates the folding time for structure 4 by about a factor of 2, and for structures 7 and 9 by a factor of 4-5. This can have various explanations. One reason is that we stopped the Langevin RUNs if the folded state was not reached after 5 μs. Although relatively few trajectories did not fold within this time, by doing so, the real lifetimes are a little underestimated by the mean given above. This effect is larger for RUN9 where the folding time is longer than for the other RUNs. Now recall that we used as a necessary condition for a good model (embedding dimension, lag time, number of neighbors) that the trajectories stay close to the native structure once they reach it. Our experience in running Langevin simulations is that when using a too low-dimensional embedding the lifetimes of the states can be considerably underestimated [66]. So it may well be that the 14-dimensional embedding vector we chose for modeling the folding process is still of too small dimension. Another possible explanations can be that we used only 10 trajectories for each RUN as input for our Langevin modeling, whereas the folding times estimated from the MD by Ensign et al. used all (up to 100) simulations.

5.3 Outlook

In this chapter we have detailed the first application of the methods developed in this thesis to such a large system with hundreds of microseconds of MD simulation available from the Folding@home project. In contrast to the case of smaller peptides that we analyzed so far, besides the free energy landscape as obtained by dPCA the landscape as obtained by Cartesian PCA seems to be structured as well. In order to give more quantitative results more detailed analyses in that respect are needed. As the 2D representations of the free energy landscapes seem to fill up the more simulations one takes into account, this could indicate the existence of a very large number of conformational states in the case of the villin headpiece. One would need to cluster the landscape in a way such that a manageable number of clusters is obtained, yet providing enough detailed information on the processes one is interested to study. We here only made the distinction between the unfolded structures 0-9, and native-like states in order to study folding times. Therefore a Langevin simulations of the folding process showed very promising first results. This is only a first step towards modeling the villin dynamics by means of nonlinear time series analysis.

In future works, one should verify whether one Langevin model is appropriate for the description of the simulations from all different RUNs. Therefore one could derive a model for each RUN separately that well-reproduces the folding times as estimated from the MD trajectories, and then see whether the models are similar. More than 10 trajectories should be taken into account, or at least it should be shown that the model one obtains by resticting oneself to a subset of trajectories is appropriate. An additional careful k-means analysis together with a transition a matrix analysis using ideas as presented in Chap. 3 will help to gain deeper insight in the structure of the free energy landscape and the dynamics on it.

Chapter 6

Appendix

6.1 Transformation of probability densities

In this section we derive the probability densities for the random variables $\cos\varphi$ and $\sin\varphi$, given the density for the angular variable

$$\rho(\varphi) = \frac{1}{2\pi}(1 - \cos 4\varphi), \quad \varphi \in [-180°, 180°]. \tag{6.1}$$

We first need to consider the density (6.1) in the interval $[-180°, 0°]$ in order to ensure invertibility of the cosine function, obtaining

$$\rho(\varphi) = \frac{1}{\pi}(1 - \cos 4\varphi), \quad \varphi \in [-180°, 0°] \tag{6.2}$$

by rescaling the original ρ by a factor of 2 in order to fulfill the condition $\int_{0°}^{180°} \rho(\varphi)d\varphi = 1$ of a probability density. Note that for obtaining the correct probability density for $x = \cos\varphi$ in the interval $[-1, 1]$ of the cosine, we need to add up the contribution of (6.1) in $[-180°, 0°]$ and $[0°, 180°]$, which is consistent with our approach to consider the doubled density in only one of the intervals (see Fig. 2.2B).

Now, in $[-180°, 0°]$, we have

$$\varphi = -\arccos x, \quad x \in [-1, 1]. \tag{6.3}$$

Hence we find by substitution

$$\int_{0^\circ}^{180^\circ} \rho(\varphi)d\varphi = \int_{-1}^{1} \rho(-\arccos x)\frac{1}{\sqrt{1-x^2}}dx. \tag{6.4}$$

Finally, we obtain for the probability density $h(x)$ of $x = \cos\varphi$

$$\begin{aligned}
h(x) &= \rho(-\arccos x)\frac{1}{\sqrt{1-x^2}} \\
&= \frac{1}{\pi}\left(1-\cos(4\arccos x)\right)\frac{1}{\sqrt{1-x^2}} \\
&= \frac{1}{\pi\sqrt{1-x^2}}\left(1-\cos^2(2\arccos x)+\sin^2(2\arccos x)\right) \\
&= \frac{1}{\pi\sqrt{1-x^2}}2\sin^2(2\arccos x) \\
&= \frac{2}{\pi\sqrt{1-x^2}}\left(2\cos(\arccos x)\sin(\arccos x)\right)^2 \\
&= \frac{8}{\pi\sqrt{1-x^2}}\sin^2(\arccos x)x^2 \\
&= \frac{8}{\pi\sqrt{1-x^2}}\left(1-\cos^2(\arccos x)\right)x^2 \\
&= \frac{8(1-x^2)x^2}{\pi\sqrt{1-x^2}} \\
&= \frac{8x^2\sqrt{1-x^2}}{\pi}
\end{aligned} \tag{6.5}$$

Analogously, one derives the probability density for $\sin\varphi$.

Alternatively, we could have derived Eq. (6.5) from the two-dimensional density (2.25) $\rho(x,y) = \frac{4}{\pi}x^2y^2\delta(x^2+y^2-1)$ by integrating over y from -1 to 1. We originally had a problem with this approach, which we make clear with the following example. When integrating the uniform density $\rho(\varphi) = \frac{1}{2\pi}$ over the unit circle in 2D, that is,

$$\int_{-1}^{1}\int_{-1}^{1}\frac{1}{2\pi}\delta(x^2+y^2-1)dydx, \tag{6.6}$$

using the substitution $z := x^2+y^2-1$, we obtain 0.5 instead of 1. This is due to the use of this Delta function to describe the unit circle. Thus, when using it, one has to rescale the two-dimensional density by a factor of 2 to obtain the correct result.

6.2 Complex dPCA vs. dPCA

The purpose of this section is to discuss the relations of the principal components (2.39) and the eigenvalues (2.40) between the sin/cos and the complex dPCA, respectively. To this end, we first establish a correspondence between the covariance matrices of the two formulations. Using Euler's formula, we express the matrix elements of the covariance matrix (2.36) as

$$\begin{aligned} C_{mn} &= \langle (e^{\mathrm{i}\varphi_m} - \langle e^{\mathrm{i}\varphi_m}\rangle)(e^{-\mathrm{i}\varphi_n} - \langle e^{-\mathrm{i}\varphi_m}\rangle)\rangle \\ &= \mathrm{cov}(\cos\varphi_m, \cos\varphi_n) + \mathrm{cov}(\sin\varphi_m, \sin\varphi_n) \\ &\quad -\mathrm{i}\,\mathrm{cov}(\cos\varphi_m, \sin\varphi_n) + \mathrm{i}\,\mathrm{cov}(\sin\varphi_m, \cos\varphi_n), \end{aligned} \tag{6.7}$$

where $\mathrm{cov}(a,b) = \langle ab\rangle - \langle a\rangle\langle b\rangle$. Without loss of generality (since the generalization is straightforward), we restrict ourselves in the following to the case of two angles ($N = 2$). Using Eq. (6.7) and the definition (2.28) of σ together with (2.31), it is easy to see that one can transform the sin/cos covariance matrix σ into the complex covariance matrix C according to

$$T\sigma T^\dagger = C, \tag{6.8}$$

where

$$T = \begin{pmatrix} 1 & -\mathrm{i} & 0 & 0 \\ 0 & 0 & 1 & -\mathrm{i} \end{pmatrix}. \tag{6.9}$$

Let us next derive Eqs. (2.39) and (2.40) for the limiting case of two uncorrelated angle variables. The resulting covariance matrix of the sin/cos dPCA exhibits a block-diagonal structure with 2×2 blocks A and B. Assuming that $(x_1, x_2)^T$ is an eigenvector of A with eigenvalue λ_1, then, due to orthogonality, $(-x_2, x_1)^T$ is an eigenvector of A, too. Let its eigenvalue be λ_2. Analogously, let $(x_3, x_4)^T$ and $(-x_4, x_3)^T$ be the eigenvectors of B with eigenvalues λ_3 and λ_4. It follows that

$$\begin{aligned} \mathbf{v}^{(1)} &= (x_1, x_2, 0, 0)^T, \quad \mathbf{v}^{(2)} = (-x_2, x_1, 0, 0)^T, \\ \mathbf{v}^{(3)} &= (0, 0, x_3, x_4)^T, \quad \mathbf{v}^{(4)} = (0, 0, -x_4, x_3)^T \end{aligned} \tag{6.10}$$

are eigenvectors of σ with eigenvalues $\lambda_1, \ldots, \lambda_4$. Using Eq. (6.8), it is now straightforward

to verify that the eigenvectors $\mathbf{w}^{(n)}$ of the complex dPCA can be defined as follows

$$\begin{aligned} C\mathbf{w}^{(1)} &:= C\left(x_1 - \mathrm{i}x_2, 0\right)^T = (\lambda_1 + \lambda_2)\mathbf{w}^{(1)} =: \mu_1\mathbf{w}^{(1)}, \\ C\mathbf{w}^{(2)} &:= C\left(0, x_3 - \mathrm{i}x_4\right)^T = (\lambda_3 + \lambda_4)\mathbf{w}^{(2)} =: \mu_2\mathbf{w}^{(2)}, \end{aligned} \tag{6.11}$$

which reveals the simple relation (2.40) between the eigenvalues λ_k of the sin/cos dPCA and the eigenvalues μ_n of the complex dPCA. By comparing the principal components $W_n = \mathbf{w}^{(n)T}\mathbf{z}$ $(n = 1, 2)$ and $V_k = \mathbf{v}^{(k)} \cdot \mathbf{q}$ $(k = 1, \ldots, 4)$, we finally obtain the equality (2.39) of the principal components of the two formulations

$$\begin{aligned} \mathrm{Re}\, W_1 &= V_1, \quad \mathrm{Im}\, W_1 = V_2, \\ \mathrm{Re}\, W_2 &= V_3, \quad \mathrm{Im}\, W_2 = V_4. \end{aligned} \tag{6.12}$$

We note that the above definition of the principal components W_n is not equivalent to the projection $\mathbf{w}^{(n)} \cdot \mathbf{z}$ given by a Hermitian inner product. However, the appealingly simple relation (2.39) between the principal components of the two dPCA methods only holds when the W_n are defined that way.

While a 2×2 block-diagonal structure of the sin/cos covariance matrix σ represents a sufficient condition, it is certainly not a necessary requirement to yield relations (2.39) and (2.40). In the case of trialanine, where the latter equations were satisfied to high accuracy (see Fig. 2.5), the covariance matrix σ was indeed approximately block-diagonal. On the other hand, our second example Ala_{10} also satisfied the equalities quite well (see Fig. 2.9), although σ revealed only little block-diagonal structure. Finally, we found cases where the correspondence holds for covariance matrices that are not block-diagonal at all. For example, it can be shown that two completely correlated angle variables (say, φ_1 and $\varphi_2 = \varphi_1 +$ const.) result in dPCA covariance matrices that satisfy Eqs. (2.39) and (2.40).

6.3 Integrating out Gaussian-distributed degrees of freedom

We wish to reduce a high-dimensional energy surface to a lower dimensional one by integrating out coordinates which only exhibit a single minimum and therefore do not

describe conformational transitions. The question arises if the barriers of the landscape are reproduced correctly when the lower dimensional surface is considered. As an illustrative example, we consider the two-dimensional model

$$E(x,y) = V(x) + \frac{1}{2}\omega(x)y^2 + c(x)y, \tag{6.13}$$

consisting of a general potential $V(x)$ coupled via $c(x)y$ to a harmonic potential $\frac{1}{2}\omega(x)y^2$, where $V(x)$, $c(x)$, and $\omega(x)$ are general functions of coordinate x. This corresponds to the case that the probability distribution along coordinate y is a Gaussian. Since

$$\frac{\partial E}{\partial y} = \omega(x)y + c(x) = 0 \quad \rightarrow \quad y_e = -c(x)/\omega(x), \tag{6.14}$$

the one-dimensional function

$$E_e(x) = E(x, y = y_e) = V(x) - c^2(x)/2\omega(x) \tag{6.15}$$

connects all extrema of the two-dimensional surface. The reduced free energy landscape ($N = const.$)

$$\begin{aligned} G(x) &= -kT \ln N \int_{-\infty}^{\infty} dy\; e^{-\beta E(x,y)} \\ &= V(x) - kT \ln N \int_{-\infty}^{\infty} dy\; e^{-\beta[\frac{1}{2}\omega(x)y^2 + c(x)y]} \\ &= V(x) - c^2(x)/2\omega(x) + const. \end{aligned} \tag{6.16}$$

is apart from a constant equivalent to $E_e(x)$ and therefore reproduces correctly all barriers and other extremal points of the free energy landscape.

6.4 Molecular dynamics simulation details

All MD simulations of the polyalanine chains were generated using the GROMACS program suite [92]. What all simulations have in common is that the respective peptide was solvated in a box of simple point charge (SPC) water [52], keeping a minimum distance of 10 Å between the solute and each face of the box. 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 [93] with a relative geometric tolerance of 0.0001. We employed a particle-mesh Ewald treatment for the long-range electrostatics with a real-space cutoff of 1.2 nm, a grid of 0.12 nm, spline interpolation of order four, and direct sum tolerance of 10^{-5}. The Lennard-Jones interactions were cut off at 1.2 nm without using shift or switch functions. The nonbonded interaction pair-list was updated every 5 fs. The solute and solvent were separately weakly coupled to external temperature baths at 300 K. [94] The temperature coupling constant was 0.1 ps. The total system was weakly coupled to an external pressure bath at 1 atm using a coupling constant of 0.5 ps.

Ala_3: For the trialanine simulation as introduced in Sec. 2.5 we used the GROMOS96 force field 43A1 [95] to perform a 100 ns MD simulation. The final system contained 2914 atoms within a cubic box of dimension 25 Å. The coordinates were saved every 0.5 ps for analysis. For the analysis of the dihedral angles, throughout the thesis we only used the two dihedral angles ϕ_2, ψ_2. The data can be found in /data /MD_ANA/ALA3_Aleko /phipsi.dat. As we needed to observe the fast interstate dynamics between the α, β, and P_{II} configurations for the nonlinear modeling in Chap. 4, we also ran a simulation were we saved the data every 0.2 ps. This simulation is saved in /data /MD_ANA/ALA3_Aleko_0.2ps.

Ala_5: The details for the 100 ns pentalanine simulation used in Sec. 2.9 are given in Mu et al. [25].

Ala_7: The GROMOS force field 45A3 [95] was used in the simulations of Ala_7 in the zwitterionic state. The final system contained 3775 atoms within a cubic box of dimension 37 Å. Starting with an extended configuration of heptaalanine, the system was minimized using the conjugate gradient method, followed by followed by 50 ps of MD simulation at 300 K and constant pressure at 1 atm.

We ran two simulations for the heptaalanine system. The first one has length 600 ns, and the second one is $\approx$ 200 ns long (191.2 ns to be exact). The data were saved every 0.1 ps, but the timestep used in this thesis is 1 ps. In Sec. 2.9 we used the 200 ns simulation for the comparison between the landscapes as obtained by dPCA and the Cartesian PCAs, respectively. As the THESEUS fit required a too high amount of memory, we used a larger timestep of 20 ps, thus only around 10,000 data points for the analysis as presented in Fig. 2.11. Henceforward, from Sec. 2.9 we used the concatenation between the two

trajectories (the 600 ns one is followed by the 200 ns trajectory), thus obtaining an 800 ns simulation. The reason for concatenating these was that the 600 ns very rarely sampled the all-α configuration, whereas the 200 ns one did well-sample that region. One should be aware of the discontinuity or false transition after 600 ns when modeling the data or, more importantly, when calculating autocorrelation functions. We calculated the autocorrelation functions for the two parts of the 800 ns simulation separately, and then averaged the function values. The dihedral angles $\{\phi_2, \psi_2, \ldots, \phi_6, \psi_6\}$ for the concatenated trajectory can be found in /data /MD_ANA/ALA7 /ala7_phipsi_1ps.dat.

Ala_{10}: For the decaalanine simulation as introduced in Sec. 2.8 we used the GROMOS96 force field 43A1 [95] to perform a $\approx$ 300 ns (more exact, 309.5 ns) MD simulation. The final system contained 9073 atoms within an octahedral box of dimension given by the vector $(46, 47, 40)$ Å.

The coordinates were saved every 0.2 ps for analysis and can be found in /data /MD_ANA/ALA10_dihedral_angle_more.dat. We used a timestep of 0.4 ps for our analyses.

Villin headpiece subdomain: The details for the Folding@home simulations of the villin headpiece subdomain HP-35 NleNle are given in Chap. 5, Ref. [4], and references therein.

6.5 Source code in R

In this section we provide implementations of the most important PCA and clustering methods we presented in this thesis. The code is written in the R program package [96] using the circular statistics library [97]. This is exemplary code for heptaalanine which can easily be adjusted for other peptides. The input file contains the 10 angles $\{\phi_2, \psi_2, \ldots, \phi_6, \psi_6\}$.

Method 1: Source code for performing dPCA.

```
rm(list=ls())
mem.limits(2000000000)
```

```
a<-read.table("/data/aleko/CL_PAPER/ala7_phi2-psi6_10ps.dat")
a<-a/180*pi

nangles<-length(a[1,])
npoints<-length(a[,1])

#cos/sin transformation of angles
y<-matrix(nrow=npoints,ncol=2*nangles);
for (i in seq(1,2*nangles,2)) {
  y[,i]<-cos(a[,(i+1)/2]);
  y[,i+1]<-sin(a[,(i+1)/2]);
}

s<-svd(cov(y)) #diagonalize covariance matrix

V<-y %*% s$u   #projection on eigenvectors

#write out dPCA modes
write.table(round(V,5),"ala7_phi2-psi6_10ps.dpca",row.names=F,col.names=F)
```

■

Method 2: Source code for performing a PCA directly on the dihedral angles which are shifted in order to minimize the points at the periodic boundaries as described in section 2.10.

```
rm(list=ls())
mem.limits(2000000000)

a<-read.table("/data/aleko/CL_PAPER/ala7_phi2-psi6_10ps.dat")

nangles<-length(a[1,])
npoints<-length(a[,1])
```

```
#shift angles such that minimum density is on the periodic boundaries
for (i in 1:nangles) {
  hista<-hist(a[,i],breaks=50,plot=F)

  #position of minimum density
  minpos<-hista$mids[hista$counts==min(hista$counts)][1]

  a[,i][a[,i]>minpos]<-a[,i][a[,i]>minpos]-360
}

s<-svd(cov(a)) #diagonalize covariance matrix

V<-as.matrix(a) %*% s$u   #projection on eigenvectors

#write out PCA modes
write.table(round(V,5),"~/ala7_phi2-psi6_10ps.apca",row.n=F,col.n=F)
```

■

Method 3: This is an implementation of the clustering method using the circular variance to determine the number of clusters as proposed in section 3.4. The output is a table similar to Table 3.1.

```
rm(list=ls())
mem.limits(2000000000)
library(circular)

x<-read.table("/data/aleko/CL_PAPER/ala7_phi2-psi6_10ps.dpca")
a<-read.table("/data/aleko/CL_PAPER/ala7_phi2-psi6_10ps.dat")
acirc<-as.circular(a,units="degrees")

nangles<-length(a[1,])
```

```
npoints<-length(a[,1])

ndim<-5          #dimensions used for clustering
nseeds<-2        #number of independent k-means runs
cutoff<-0.2      #threshold for circular variance
clfrac<-0.9      #fraction of good clusters

ncluster<-20     #number of clusters to start with
maxcluster<-30   #upper limit for cluster number
cstep<-1         #step to increase cluster number
tstep<-1         #lag for transition matrix

trackcl<-matrix(nrow=ceiling((maxcluster-ncluster)/cstep)+1,ncol=2)
count<-1
ok<-FALSE

#perform clustering
while(ok==FALSE) {
  print(ncluster)
  goodcl<-0

  vmatrix<-matrix(nrow=ncluster,ncol=nangles)

  cl<-kmeans(x[,1:ndim],ncluster,nstart=nseeds,iter.max=40)

  for (k in 1:ncluster)
    for (l in 1:nangles)
      vmatrix[k,l]<-var.circular(subset(acirc,cl$cluster==k)[,l])

  for (i in 1:ncluster)
    if (mean(vmatrix[i,])<cutoff) goodcl=goodcl+1
```

```
  trackcl[count,1]<-ncluster
  trackcl[count,2]<-goodcl/ncluster
  count<-count+1

  if (goodcl/ncluster>clfrac || ncluster>=maxcluster)
    ok<-TRUE
    else ncluster<-ncluster+cstep
}

round(trackcl,2) #show number of clusters and fraction of good clusters

#calculate transition matrix
tcount <- array(0,c(ncluster,ncluster))
tmatrix <- array(0,c(ncluster,ncluster))
for (n in 1:(length(cl$cluster)-tstep)) {
  i<-cl$cluster[n]
  j<-cl$cluster[n+tstep]
  tcount[i,j] <- tcount[i,j]+1
}

for (n in 1:ncluster)
   tmatrix[n,]<-tcount[n,]/sum(tcount[n,])

#calculate circular averages
amatrix<-matrix(nrow=ncluster,ncol=nangles)
for (k in 1:ncluster)
  for (l in 1:nangles)
    amatrix[k,l]<-mean.circular(subset(acirc,cl$cluster==k)[,l])

#Calculate table with sequence, population and metastability of clusters
```

```
#"1": alpha, "2": beta/PII, "3": circular variance of psi angle too large
seqmatrix<-matrix(nrow=ncluster,ncol=7)
for (i in 1:ncluster) {
  seqmatrix[i,6]=round(cl$size[i]/npoints*100,1)
  seqmatrix[i,7]=round(tmatrix[i,i]*100,0)
  for (j in 1:5) {
    if (vmatrix[i,2*j]<cutoff) {
      if (amatrix[i,2*j]<25) seqmatrix[i,j]=1
      else seqmatrix[i,j]=2
    }
    else seqmatrix[i,j]=NA
  }
}

#show table ordered by population of clusters
seqmatrix<-seqmatrix[sort(cl$size,index.return=T,decreasing=T)$ix,]
seqmatrix
```

■

Bibliography

[1] Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., Bourne, P. The Protein Data Bank. Nucleic Acids Research 28(1):235–242, JAN 1 2000.

[2] van Gunsteren, W. F., Bakowies, D., Baron, R., Chandrasekhar, I., Christen, M., Daura, X., Gee, P., Geerke, D. P., Glättli, A., Hünenberger, P. H., Kastenholz, M. A., Oostenbrink, C., Schenk, M., Trzesniak, D., van der Vegt, N. F. A., Yu, H. B. Biomolecular modelling: goals, problems, perspectives. Angew. Chem. Int. Ed. 45:4064–4092, 2007.

[3] Shirts, M., Pande, V. S. Computing - Screen savers of the world unite! Science 290:1903–1904, 2000.

[4] Ensign, D. L., Kasson, P. M., Pande, V. S. Heterogeneity even at the speed limit of folding: large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. J. Mol. Biol. 374(3):806–816, 2007.

[5] Chodera, J. D., Swope, W. C., Pitera, J. W., Dill, K. A. Obtaining long-time protein folding dynamics from short-time molecular dynamics simulations. Multiscale Modeling & Simulation 5:1214–1226, 2006.

[6] Chodera, J. D., Singhal, N., Pande, V. S., Dill, K. A., Swope, W. C. Automatic discovery of metastable states for the construction of markov models of macromolecular dynamics. J. Chem. Phys. 126:155101, 2007.

[7] Swope, W., Pitera, J., Suits, F. Describing protein folding kinetics by molecular dynamics simulations. 1. Theory. Journal of Physical Chemistry B 108(21):6571–6581, 2004.

[8] Onuchic, J. N., Schulten, Z. L., Wolynes, P. G. Theory of protein folding: The energy landscape perspective. Annu. Rev. Phys. Chem. 48:545–600, 1997.

[9] Dill, K. A., Chan, H. S. From levinthal to pathways to funnels: The "new view" of protein folding kinetics. Nat. Struct. Bio. 4:10–19, 1997.

[10] Wales, D. J. Energy Landscapes. Cambridge: Cambridge University Press. 2003.

[11] Ball, K. D., Berry, R. S., Kunz, R. E., Li, F.-Y., Proykova, A., Wales, D. J. From topographies to dynamics on multidimensional potential energy surfaces. Science 271:963–965, 1996.

[12] Gruebele, M. Protein folding: The free energy surface. Curr. Opin. Struct. Biol. 12:161–168, 2002.

[13] Jolliffe, I. T. Principal Component Analysis. New York: Springer. 2002.

[14] Ichiye, T., Karplus, M. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. Proteins 11:205–217, 1991.

[15] Garcia, A. E. Large-amplitude nonlinear motions in proteins. Phys. Rev. Lett. 68:2696–2699, 1992.

[16] Amadei, A., Linssen, A. B. M., Berendsen, H. J. C. Essential dynamics of proteins. Proteins 17:412–425, 1993.

[17] Hayward, S., Kitao, A., Hirata, F., Go, N. Effect of solvent on collective motions in globular proteins. J. Mol. Biol. 234:1207–1217, 1993.

[18] Becker, O. M. Geometric versus topological clustering: An insight into conformation mapping. Proteins 27:213–226, 1997.

[19] Lange, O. F., Grubmüller, H. Can principal components yield a dimension reduced description of protein dynamics on long time scales. J. Phys. Chem. B 110:22842–22852, 2006.

[20] Noe, F., Krachtus, D., Smith, J. C., Fischer, S. Transition networks for comprehensive characterization of complex conformational change in proteins. J. Chem. Theory Comput. 2:840–857, 2006.

[21] Abseher, R., Nilges, M. J. Mol. Biol. 279:911, 1998.

[22] van Aalten, D. M. D., de Groot, B. L., Finday, J. B. C., Berendsen, H. J. C., Amadei, A. A comparison of techniques for calculating protein essential dynamics. J. Comput. Chem. 18:169–181, 1997.

[23] Elmaci, N., Berry, R. S. Principal coordinate analysis on a protein model. J. Chem. Phys. 110:10606–10622, 1999.

[24] Reijmers, T. H., Wehrens, R., Buydens, L. M. C. Circular effects in representations of an RNA nucleotides data set in relation with principal component analysis. Chemom. Intell. Lab. Syst. 56:61–71, 2001.

[25] Mu, Y., Nguyen, P. H., Stock, G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. Proteins 58:45, 2005.

[26] Sims, G. E., Choi, I.-G., Kim, S.-H. Protein conformational space in higher order $\phi - \psi$ maps. Proc. Natl. Acad. Sci. USA 102:618–621, 2005.

[27] Wang, J., Brüschweiler, R. 2D entropy of discrete molecular ensembles. J. Chem. Theory Comput. 2:18–24, 2006.

[28] Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, J. D., Voelz, V. A. The protein folding problem: when will it be solved? Current Opinion in Structural Biology 17(3):342–346, JUN 2007.

[29] Alakent, B., Doruker, P., Camurdan, M. C. Application of time series analysis on molecular dynamics simulations of proteins: A study of different conformational spaces by principal component analysis. J. Chem. Phys. 121:4756–4769, 2004.

[30] Schultheis, V., Hirschberger, T., Carstens, H., Tavan, P. Extracting Markov models of peptide conformational dynamics from simulation data. J. Chem. Theory Comput. 1:515–526, 2005.

[31] Ma, A., Dinner, A. R. Automatic method for identifying reaction coordinates in complex systems. J. Phys. Chem. B 109:6769–6779, 2005.

[32] Meerbach, E., Dittmer, E., Horenko, I., Schütte, C. Multiscale modelling in molecular dynamics: Biomolecular conformations as metastable states. Lect. Notes Phys. 703:475, 2006.

[33] Das, P., Moll, M., Stamati, H., Kavraki, L. E., Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. Proc. Natl. Acad. Sci. USA 103:9885–9890, 2006.

[34] Lange, O. F., Grubmüller, H. Generalized correlation for biomolecular dynamics. Proteins 62:1053–1061, 2006.

[35] Nguyen, P. H. Complexity of free energy landscapes peptides revealed by nonlinear principal component analysis. Proteins 65:898, 2006.

[36] Hegger, R., Altis, A., Nguyen, P. H., Stock, G. How complex is the dynamics of peptide folding? Phys. Rev. Lett. 98:028102, 2007.

[37] Altis, A., Nguyen, P. H., Hegger, R., Stock, G. Dihedral angle principal component analysis of molecular dynamics simulations. J. Chem. Phys. 126:244111, 2007.

[38] Altis, A., Otten, M., Nguyen, P. H., Hegger, R., Stock, G. Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. J. Chem. Phys. 128:245102, 2008.

[39] Hinsen, K. Comment on "Energy Landscape of a Small Peptide Revealed by Dihedral Angle Principal Component Analysis". Proteins 64:795–797, 2006.

[40] Mu, Y., Nguyen, P. H., Stock, G. Reply to the Comment on "Energy Landscape of a Small Peptide Revealed by Dihedral Angle Principal Component Analysis". Proteins 64:798–799, 2006.

[41] Fisher, N. I. Statistical Analysis of Circular Data. Cambridge: Cambridge University Press. 1996.

[42] Woutersen, S., Hamm, P. Structure determination of trialanine in water using polaization sensitive two-dimensional vibrational spectroscopy. J. Phys. Chem. B 104:11316, 2000.

[43] Woutersen, S., Pfister, R., Hamm, P., Mu, Y., Kosov, D., Stock, G. Peptide conformational heterogeneity revealed from nonlinear vibrational spectroscopy and molecular dynamics simulations. J. Chem. Phys. 117:6833, 2002.

[44] Schweitzer-Stenner, R., Eker, F., Huang, Q., Griebenow, K. Dihedral angles of trialanine in D_2O determined by combining FTIR and polarized visible Raman spectroscopy. J. Am. Chem. Soc. 123:9628, 2001.

[45] Graf, J., Nguyen, P. H., Stock, G., Schwalbe, H. Structure and dynamics of the homologues series of alanine peptides: A joint molecular-dynamics/NMR study. J. Am. Chem. Soc. 129:1179–1189, 2007.

[46] Mu, Y., Stock, G. Conformational dynamics of trialanine in water: A molecular dynamics study. J. Phys. Chem. B 106:5294, 2002.

[47] Mu, Y., Kosov, D. S., Stock, G. Conformational dynamics of trialanine in water II: Comparison of AMBER, CHARMM, GROMOS, and OPLS force fields to NMR and infrared experiments. J. Phys. Chem. B 107:5064, 2003.

[48] Gnanakaran, S., Garcia, A. E. Validation of an all-atom protein force field: from dipeptides to larger peptides. J. Phys. Chem. B 107:12555–12557, 2003.

[49] Berendsen, H. J. C., van der Spoel, D., van Drunen, R. Gromacs: A message-passing parallel molecular dynamics implementation. Comp. Phys. Comm. 91:43, 1995.

[50] van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., Berendsen, H. J. C. Gromacs; fast, flexible and free. J. Comput. Chem. 26:1701–1718, 2005.

[51] van Gunsteren, W. F., Billeter, S. R., Eising, A. A., Hünenberger, P. H., Krüger, P., Mark, A. E., Scott, W. R. P., Tironi, I. G. Biomolecular Simulation: The GROMOS96 Manual and User Guide. Zürich: Vdf Hochschulverlag AG an der ETH Zürich. 1996.

[52] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Hermans, J. Interaction models for water in relation to protein hydration. In: Intermolecular Forces. Pullman, B. ed. . D. Reidel Publishing Company Dordrecht 1981 331–342.

[53] Darden, T., York, D., Petersen, L. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. J. Chem. Phys. 98:10089, 1993.

[54] A direct back-calculation of the dihedral angles is not possible. But since the time indices of the original trajectory and the principal components are identical, we can use these indices to identify corresponding dihedral angles.

[55] Details of the identification of the metastable conformational states and their transition matrix are given in Ref. [25].

[56] Theobald, D. L., Wuttke, D. S. Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem. Proc. Natl. Acad. Sci. USA 103(49):18521–18527, 2006.

[57] Theobald, D. L., Wuttke, D. S. Accurate structural correlations from maximum likelihood superpositions. PLoS Comput. Biol. 4(2):0001–0008, 2008.

[58] Theobald, D. L., Wuttke, D. S. THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. Bioinformatics 22(17):2171–2172, 2006.

[59] Fitzgerald, J. E., Jha, A. K., Sosnick, T. R., Freed, K. F. Polypeptide motions are dominated by peptide group oscillations resulting from dihedral angle correlations between nearest neighbors. Biochemistry 46(3):669–682, JAN 23 2007.

[60] Jammalamadaka, S. R., SenGupta, A. Topics in Circular Statistics. Singapore: World Scientific Press. 2001.

[61] As the complete analysis is performed in the space of dihedral angle principal components, there is no need to invoke the Jacobian transformation between these coordinates and the atomic Cartesian coordinates. [98].

[62] Lange, O. F., Grubmüller, H. Collective Langevin dynamics of conformational motions in proteins. J. Chem. Phys. 124:214903, 2006.

[63] Yang, S., Onuchic, J. N., Levine, H. Effective stochastic dynamics on a protein folding energy landscape. J. Chem. Phys. 125:054910, 2006.

[64] Krivov, S. V., Karplus, M. Hidden complexity of free energy surfaces for peptide (protein) folding. Proc. Natl. Acad. Sci. USA 101:14766–14770, 2004.

[65] Horenko, I., Hartmann, C., Schütte, C., Noe, F. Data-based parameter estimation of generalized multidimensional Langevin processes. Phys. Rev. E 76:016706, 2007.

[66] Hegger, R., Stock, G. Multidimensional Langevin modeling of biomolecular dynamics. J. Chem. Phys., *accepted for publication.*

[67] Hartigan, J. A., Wong, M. A. A k-means clustering algorithm. Applied Statistics 28:100–108, 1979.

[68] Van Kampen, N. G. Stochastic Processes in Physics and Chemistry. Amsterdam: Elsevier. 1997.

[69] Schütte, C., Fischer, A., Huisinga, W., Deuflhard, P. A direct approach to conformational dynamics based on hybrid Monte Carlo. J. Comp. Phys 151:146–168, 1999.

[70] de Groot, B. L., Daura, X., Mark, A. E., Grubmüller, H. Essential dynamics of reversible peptide folding: Memory-free conformational dynamics governed by internal hydrogen bonds. J. Mol. Biol. 309:299–313, 2001.

[71] Noe, F., Horenko, I., Schütte, C., Smith, J. C. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. J. Chem. Phys. 126:155102, 2007.

[72] Levy, Y., Jortner, J., Berry, R. S. Phys. Chem. Chem. Phys. 4:5052, 2002.

[73] Becker, O. M., Karplus, M. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. J. Chem. Phys. 106 (4):1495–1517, 1997.

[74] Krivov, S. V., Karplus, M. Free energy disconnectivity graphs: Application to peptide models. J. Chem. Phys. 117:10894–10903, 200.

[75] The program can be downloaded from www-wales.ch.cam.ac.uk.

[76] Gower, J. C. Multivariate analysis and multidimensional geometry. The Statistician 17:13–28, 1967.

[77] Farmer, J. D., Ott, E., Yorke, J. A. The dimension of chaotic attractors. Physica D 7:153–180, 1983.

[78] Grassberger, P., Procaccia, I. On the characterization of strange attractors. Phys. Rev. Lett. 50:346–349, 1983.

[79] Frederickson, P., Kaplan, J. L., Yorke, E. D., A., Y. J. The Lyapunov dimension of strange attractors. J. Diff. Eqns. 49:185, 1983.

[80] Kantz, H., Schreiber, T. Nonlinear Time Series Analysis. Cambridge, UK: Cambridge Univ. Press. 1997.

[81] Ott, E. Chaos in Dynamical Systems. Cambridge, UK: Cambridge Univ.. 1993.

[82] Beck, C., Schlögl, F. Thermodynamics of chaotic systems. Cambridge, UK: Cambridge Univ. Press. 1993.

[83] Alligood, K., Sauer, T. D., Yorke, J. A. Chaos: An Introduction to Dynamical Systems. Heidelberg: Springer. 1996.

[84] Takens, F. Detecting strange attractors in turbulence. In: Dynamical Systems and Turbulence (Warwick 1980) (Lecture Notes in Mathematics). Vol. 898. Rand, D. A., Young, L.-S. eds. Vol. 898. . Springer-Verlag Berlin, Heidelberg 1980 366–381.

[85] Frenkel, D., Smit, B. Understanding Molecular Simulations. San Diego: Academic. 2002.

[86] Villani, V. Complexity of polypeptide dynamics: chaos, brownian motion and elasticity in aqueous solution. J. Mol. Struct. 621:127–139, 2002.

[87] Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. Numerical Recipies, Second Edition. Cambridge: Cambridge University Press. 1992.

[88] Farmer, J. D., Sidorowich, J. J. Predicting chaotic time series. Phys. Rev. Lett. 59:845, 1987.

[89] Hegger, R., Kantz, H., Schreiber, T. Practical implementation of nonlinear time series methods: The TISEAN package. Chaos 9(2):413, 1999.

[90] Dill, K. A. Dominant forces in protein folding. Biochem. 29:7133–7155, 1990.

[91] The client software can be downloaded from http://wiki.simtk.org/foldvillin/.

[92] van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., Berendsen, H. J. C. Gromacs; fast, flexible and free. J. Comput. Chem. 26:1701–1718, 2005.

[93] Ryckaert, J. P., Ciccotti, G., Berendsen, H. J. C. Numerical-integration of cartesian equations of motions of a system with constraints-molecular dynamics of n-alkanes. J. Comput. Phys. 23:327–341, 1977.

[94] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Dinola, A., Haak, J. R. Molecular dynamics with coupling to an external bath. J. Chem. Phys. 81:3684, 1984.

[95] Schuler, L. D., Daura, X., van Gunsteren, W. F. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. J. Comput. Chem. 22:1205–1218, 2001.

[96] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, 2005.

[97] Lund, U., Agostinelli, C. circular: Circular statistics 2006. R package version 0.3-6.

[98] He, S., Scheraga, H. A. Macromolecular conformational dynamics in torsional angle space. J. Chem. Phys. 108:271–286, 1998.

Acknowledgments

At first I would like to express my gratitude to Prof. Dr. Gerhard Stock for his constant support and excellent supervision. The many fruitful scientific discussions with him became an integral part of my PhD.

I would like to thank Dr. Rainer Hegger for his mentoring and close collaboration. His expert knowledge in the field of nonlinear time series analysis and programming as well as his job as our group's system administrator have been invaluable.

I am very grateful to Dr. Phuong Nguyen for a lot of assistance especially in the beginning of my PhD and for the production of several molecular dynamics trajectories that I analyzed.

Many thanks to Dr. Roman Gorbunov for being an always helpful colleague with many creative ideas to discuss.

My thanks also go to Miriam Kreth from the mathematical institute of numerical analysis in Frankfurt for several discussions, her opinions, and encouragement.

I also thank all (former) group members of AK Stock, AK Dreuw, and AK Wachtveitl, especially Sang-Min Park, Dr. Elisabeth Widjajakusuma, Dr. Hiroshi Fujisaki, Dr. Jessica Biedinger, Moritz Otten, Dr. Stefan Knippenberg, Heike Staudt, Thomas Köhler, Miriam Kohagen, Dr. Radhan Ramadass, Laura Riccardi, Abhinav Jain, Maja Kobus, and Dr. Alessandra Villa for creating a very nice working atmosphere.

My deep gratitude goes to my family for their love and caring support making it possible to get this kind of education. I thank my lovely girl-friend Elena Schmidt and my friends outside the institute for the work-life balance they provided me.

I thank all the people that I forgot to mention which were involved in the successful completion of this thesis.

Finally I would like to thank trialanine Ala_3 for being a simple, yet nontrivial system

from which I learned a lot about the methods that we developed, applied, and tested.

This work has been supported by the Frankfurt Center for Scientific Computing, the Fonds der Chemischen Industrie, and the Deutsche Forschungsgemeinschaft.

Zusammenfassung

Das Ziel der vorliegenden Arbeit ist es, einen Beitrag zur Entwicklung von Methoden zur Modellierung von freien Energieflächen von Biomolekülen zu leisten. Ausgehend von Molekulardynamik-Simulationen geht es insbesondere darum, niedrig-dimensionale Modelle für die Beschreibung von Konformationen und der Kinetik von Peptiden und kleinen Proteinen zu erhalten.

Molekulardynamik-Simulationen haben sich als gängige und leistungsstarke Methode zur Modellierung der Struktur, Dynamik und Funktion von Biomolekülen auf atomistischer Ebene etabliert. In den letzten Jahren hat sich die Rechenleistung von Computern so weit entwickelt, dass Simulationen von kleinen Peptiden auf einer Zeitskala von Mikrosekunden heutzutage kein Problem mehr darstellen. Mit Hilfe von Projekten wie Folding@home, welche die benötigte Rechenleistung weltweit auf möglichst viele Rechner verteilen, ist es mittlerweile sogar möglich, die Faltung von kleinen Proteinen im Mikrosekunden und Sub-Mikrosekundenbereich zu simulieren.

Molekulardynamik-Simulationen erzeugen allerdings riesige Datenmengen (3M Koordinaten bei M Atomen für jeden Zeitschritt), die analysiert werden müssen. Es ist daher von großer Bedeutung, Methoden zur Verfügung zu haben, um diese Daten zu handhaben und die entscheidenden Informationen herauszufiltern. Beispielsweise ist man daran interessiert, die freie Energiefläche eines Moleküls als Funktion von einigen wenigen, aber wichtigen, Koordinaten auszudrücken. Diese Reaktionskoordinaten sollen die wesentliche Physik hinter den betrachteten biomolekularen Prozessen beschreiben können. Beliebte Wahlen hierfür sind die Zahl nativer Kontakte, der Gyrationsradius und die mittlere quadratische Abweichung des Moleküls von seiner nativen Struktur. In letzter Zeit hat die resultierende freie Energiefläche das Verständnis von Proteinfaltung sehr vorangetrieben. Ursprünglich ist diese Fläche jedoch ein sehr hoch-dimensionales und kompliziertes Ob-

jekt mit einer Vielzahl von freien Energie-Minima. Daher ist es unerlässlich, gute Reaktionskoordinaten zu finden, um brauchbare niedrig-dimensionale Modelle für die freie Energiefläche und der sich auf ihr abspielenden konformationellen Dynamik zu erhalten. Um ein System in einen wichtigen (niedrig-dimensionalen) und einen belanglosen Teil zu zerlegen, hat sich als Methode die Hauptkomponentenanalyse (principal component analysis, PCA) als äußerst hilfreich bewährt. Ein Vorteil der Verwendung von PCA-Moden als Reaktionskoordinaten gegenüber der oben genannten Möglichkeiten ist, dass es prinzipiell möglich ist, durch einfache Hinzunahme von mehr Moden alle interessierenden Größen mit der erwünschten Genauigkeit anzunähern.

Als sehr beliebte Methode, um die Dimensionalität eines komplexen Systems zu reduzieren, wird die PCA häufig auf kartesische Koordinaten angewendet. Es wurde gezeigt, dass ein Großteil der Fluktuationen des Systems durch einige wenige Hauptkomponenten beschrieben werden kann. Diese Hauptkomponenten können direkt mit Konformationsänderungen des betrachteten Moleküls in Zusammenhang gebracht werden und somit als Reaktionskoordinaten für die freie Energiefläche dienen. Das Problem bei der Verwendung von kartesischen Koordinaten ist, dass es eine große Herausforderung sein kann, die interessante interne Bewegung, welche Konformationsänderungen entspricht, von der globalen Gesamtbewegung zu trennen. Mu et al. [25] zeigten, dass aufgrund dieser Schwierigkeit eine PCA auf kartesischen Koordinaten nicht die korrekte freie Energiefläche für das Peptid Pentaalanin liefert. Um Probleme dieser Art zu vermeiden, wurden in der Literatur einige Hauptkomponentenanalysen vorgeschlagen, die mit internen Koordinaten arbeiten. Für Moleküle ist die Verwendung von Torsionswinkeln naheliegend, da andere interne Koordinaten wie Bindungslängen oder Bindungswinkel sich normalerweise bei Faltungsprozessen nicht so stark verändern. Aufgrund der Periodizität von Winkeln ist es jedoch nicht trivial, eine PCA auf solche Koordinaten anzuwenden. Beispielsweise können Mittelwerte von Winkeln nicht ohne Weiteres wie bei kartesischen Koordinaten als arithmetisches Mittel gebildet werden, was sich ebenfalls auf die Berechnung von Korrelationen auswirkt.

In dieser Arbeit verwenden wir als Beispiele, um unsere Methoden zu entwickeln und zu testen, hauptsächlich Molekulardynamik-Simulationen von kurzen Poly-Alanin-Ketten. Aufgrund ihrer Größe ist es uns möglich gewesen, hinreichend lange Simulationen als Aus-

gangspunkt zu erhalten. Diese Systeme sind aufgrund der Anzahl ihrer Konformationen überschaubar, jedoch nicht trivial, denn sie besitzen wegen ihrer Beweglichkeit eine sehr schnelle Konformationsdynamik. Diese Bausteine von größeren Systemen genau zu verstehen ist von erheblicher Bedeutung, um Erkenntnisse über den Prozess der Proteinfaltung zu gewinnen. Aber auch größere Systeme wie das 36 Aminosäuren lange Kopfstück des Villin-Proteins werden betrachtet. Hunderte von Molekulardynamik-Trajektorien wurden hierzu durch das Projekt Folding@home bereitgestellt.

Nach einer einführenden Einleitung erarbeiten wir uns im zweiten Kapitel dieser Arbeit ein tiefes Verständnis verschiedener PCA-Methoden, um von Molekulardynamik-Simulationen erzeugte Konformationen in niedrig-dimensionale Räume zu projizieren. Der Schwerpunkt liegt hierbei auf der genauen theoretischen Beschreibung der Dihedral Angle Principal Component Analysis (dPCA). Die dPCA verwendet als interne Koordinaten den Sinus und den Kosinus der phi/psi-Winkel des Peptid- bzw. Protein-Rückgrats. Die Auswirkungen dieser nichtlinearen Transformation, welche mit einer Verdopplung von N phi/psi-Winkelkoordinaten auf 2N kartesische Koordinaten einhergeht, wird sorgfältig behandelt. Hierfür benutzen wir Konzepte aus der zirkulären Statistik. Wir zeigen, dass diese Transformation die Winkelverteilungen originalgetreu abbildet ohne beispielsweise künstliche freie Energieminima zu erzeugen. Ausserdem zeigen wir, dass die dPCA-Moden, ähnlich wie im kartesischen Fall, in direkten Zusammenhang mit Konformationsänderungen gebracht werden können. Eine alternative Version der dPCA im komplexen Zahlenraum liefert weitere Erkenntnisse über die Zusammenhänge der 2N Variablen der sin/cos-dPCA. Wie wir ausführen, kann man damit N Winkelkoordinaten durch N komplexe Variablen beschreiben, was von Vorteil für die physikalische Interpretation der PCA-Moden sein kann. Dies wird am Beispiel einer 300 ns langen Molekulardynamik-Simulation von Decaalanin erläutert. Es folgt ein Vergleich der dPCA mit kartesischen PCA-Varianten und es wird gezeigt, dass eine kartesische PCA, außer für das konformationell triviale Trialanin, für alle betrachteten Poly-Alanin-Ketten die falsche freie Energiefläche liefert.

Es mag die Frage aufkommen, ob in der Praxis eine Verdopplung der Variablen, wie sie durch die Sinus/Kosinus-Transformation in der dPCA zustande kommt, überhaupt notwendig ist oder ob man direkt auf den Winkeln arbeiten kann ohne die Periodizität

explizit zu behandeln. Wir zeigen daher im Vergleich zu solch einer direkten Methode, dass für die von uns studierten Fälle die dPCA die detailliertesten freien Energieflächen liefert. Kapitel 2 schließt mit einer Korrelationsanalyse der Torsionwinkel von Heptaalanin, welche in Zusammenhang mit Ergebnissen aus der Literatur gebracht wird, und einigen Bemerkungen zu nichtlinearen PCA-Methoden ab.

Aufbauend auf den vorangegangenen Resultaten erarbeiten wir in Kapitel 3 eine systematische Vorgehensweise, um freie Energieflächen mit Hilfe der dPCA zu erhalten und zu charakterisieren. Einleitend zeigen wir, welche Probleme mit zu stark vereinfachten, d.h. zu niedrig-dimensionalen, Darstellungen der freien Energiefläche einher gehen können. Es wird versucht, die notwendige Anzahl der dPCA-Moden zu bestimmen, um einen gegebenen biomolekularen Prozess mit Hilfe der resultierenden freien Energiefläche korrekt beschreiben zu können. Dazu fordern wir, dass zumindest die Anzahl, die Lage und die Energie der metastabilen Zustände sowie die Energiebarrieren richtig wiedergegeben werden. Diese notwendige Dimensionalität kann durch die Verteilungs- und Autokorrelationsfunktionen der dPCA-Moden bestimmt werden. Anhand der Molekulardynamik einer 800 ns langen Trajektorie von Heptaalanin zeigen wir, dass eine 5-dimensionale dPCA-Energiefläche eine angemessen exakte Beschreibung der genauen hoch-dimensionalen freien Energiefläche darstellt. Zur Charakterisierung dieser Flächen verwenden wir geometrische und kinetische Clustering-Verfahren. Wir stellen dabei fest, dass, zumindest mit unserer Charakterisierung der Zustände, eine Markov'sche Modellierung der Dynamik nicht in Frage kommt. Dies führt uns, nach Untersuchung verschiedener Visualisierungen der freien Energiefläche, zu Kapitel 4.

Das letztendliche Ziel dieser Arbeit ist es, niedrig-dimensionale Modelle für die Dynamik auf der freien Energiefläche auszuarbeiten. Wir verwenden hierzu moderne Konzepte der nichtlinearen Dynamik und Methoden der nichtlinearen Zeitreihenanalyse. Für die Poly-Alanin-Ketten modellieren wir die Dynamik zunächst mit einem deterministischen, lokal linearen Modell. Diese Auffassung der Faltungsprozesse als dynamisches System im mathematischen Sinne ermöglicht eine Betrachtung der Komplexität ihrer Dynamik. Besipielsweise errechnen wir die effektive Dimension (Kaplan-Yorke Dimension), die wir mit der Dimension der freien Energieflächen vergleichen. Interessanterweise nimmt die effektive Dimension bei ansteigender Systemgröße (Länge der Polypeptid-Kette) tenden-

ziell ab, wenngleich die Dimension der freien Energieflächen zunimmt. Dies deutet auf eine niedrigere Komplexität der Trajektorien für größere Systeme hin, welche durch die ansteigende Anzahl von Wasserstoffbrücken erklärt wird. Zum Schluss des Kapitels führen wir ein Modell für die Dynamik ein, welches sowohl eine deterministische als auch eine stochastische Komponente hat. Es basiert auf der Schätzung der Drift- und Diffusionsvektorfelder einer allgemeinen multidimensionalen Langevin-Gleichung.

Im abschließenden 5. Kapitel wenden wir einige der bisher entwickelten Methoden auf Trajektorien des Kopfstücks des Villin-Proteins an. Wir betrachten insbesondere freie Energieflächen für dieses System und weisen auf Unterschiede zu den Poly-Alanin-Ketten hin, die unter anderem aus der Größe dieses Systems resultieren. Mit dem Langevin-Ansatz unternehmen wir erste erfolgversprechende Versuche, die Dynamik niedrig-dimensional zu modellieren, und schätzen Faltungszeiten ab. Mit einem kurzen Ausblick beschließen wir dieses Kapitel und damit auch diese Arbeit.

Lebenslauf

Alexandros Altis

Persönliche Angaben:

Geburtsdatum und -ort	18.08.1981 in Frankfurt am Main
Familienstand	ledig
Staatsangehörigkeit	deutsch

Schulausbildung:

- 08.1987 - 07.1991: Diesterweg Grundschule in Frankfurt am Main
- 08.1991 - 07.2000: Wöhler-Gymnasium in Frankfurt am Main

Hochschulausbildung:

- WS 00/01 - WS 04/05: Studium der Mathematik mit Nebenfach Informatik
- 01.2005: Erhalt Diplomzeugnis, Betreuer: Prof. Johann Baumeister, Bewertung: sehr gut
- 02.2005 - laufend: Naturwissenschaftliche Promotion am Institut für Theoretische und Physikalische Chemie der Goethe-Universität Frankfurt am Main, Betreuer: Prof. Gerhard Stock

Publikationen:

- Hegger, R., Altis A., Nguyen, P.H., Stock, G. How complex is the dynamics of peptide folding? Phys. Rev. Lett. 98:028102, 2007.
- Altis A., Nguyen, P.H., Hegger, R., Stock, G. Dihedral angle principal component analysis of molecular dynamics simulations. J. Chem. Phys. 126:244111, 2007.
- Altis A., Otten, M., Nguyen, P.H., Hegger, R., Stock, G. Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. J. Chem. Phys. 128:245102, 2008.

Posterpräsentationen:

- September 2005, Symposium für Theoretische Chemie, Innsbruck
- März 2006, Tagung Deutsche Physikalische Gesellschaft e.V., Frankfurt
- April 2006, Workshop Computer Simulation and Theory of Macromolecules, Hünfeld
- September 2006, Workshop Methods of Molecular Simulation, IWR Heidelberg
- April 2008, Workshop Computer Simulation and Theory of Macromolecules, Hünfeld

Vorträge:

- März 2006, Institutsseminar Theoretische und Physikalische Chemie Frankfurt, Hirschegg
- September 2006, Workshop Methods of Molecular Simulation, IWR Heidelberg
- März 2007, Institutsseminar Theoretische und Physikalische Chemie Frankfurt, Hirschegg
- April 2007, Workshop Computer Simulation and Theory of Macromolecules, Hünfeld
- Mai 2008, Institutsseminar Theoretische und Physikalische Chemie Frankfurt, Universität Frankfurt
- Mai 2008, Mini-Symposium Molecular Dynamics Simulation, FIAS Frankfurt
- 2005-2008 mehrere Vorträge im Gruppenseminar AK Stock, Frankfurt