JCTC Journal of Chemical Theory and Computation

Extracting Markov Models of Peptide Conformational Dynamics from Simulation Data

Verena Schultheis, Thomas Hirschberger, Heiko Carstens, and Paul Tavan*

*Lehrstuhl für Biomolekulare Optik, Ludwig-Maximilians-Universität,
Oettingenstrasse 67, 80538 München, Germany*

Received February 3, 2005

Abstract: A high-dimensional time series obtained by simulating a complex and stochastic dynamical system (like a peptide in solution) may code an underlying multiple-state Markov process. We present a computational approach to most plausibly identify and reconstruct this process from the simulated trajectory. Using a mixture of normal distributions we first construct a maximum likelihood estimate of the point density associated with this time series and thus obtain a density-oriented partition of the data space. This discretization allows us to estimate the transfer operator as a matrix of moderate dimension at sufficient statistics. A nonlinear dynamics involving that matrix and, alternatively, a deterministic coarse-graining procedure are employed to construct respective hierarchies of Markov models, from which the model most plausibly mapping the generating stochastic process is selected by consideration of certain observables. Within both procedures the data are classified in terms of prototypical points, the conformations, marking the various Markov states. As a typical example, the approach is applied to analyze the conformational dynamics of a tripeptide in solution. The corresponding high-dimensional time series has been obtained from an extended molecular dynamics simulation.

1. Introduction

The analysis of time series[1] is important in many areas of science. Depending on the data considered, different methods are applied.[1−4] For instance, in speech recognition[5] and other fields[6] hidden Markov models[7] found important applications. They describe a dynamical system by two parametric time-discrete processes: an underlying nonobservable Markov process[8] and an observation process, defined by a sequence of conditionally independent random variables depending at each time step only on the state of the Markov chain. In many of these applications, relatively low-dimensional data are analyzed. Frequently the treatment of higher dimensional data can be simplified by first reducing the dimension, for instance using a principal component analysis.[9] Generally, the analysis of high-dimensional data mapping complex dynamical systems requires special care and the application of methods, which by construction can cope with the peculiarities of the metrics in high-dimensional data spaces.

Here, we consider a class of extremely high-dimensional and complex dynamical systems, which exhibit a largely stochastic behavior and show Markovian transitions between coarse-grained states. A typical example for such systems is the thermal motion of proteins or peptides[10] in solution. Associated time series are generated by molecular dynamics (MD) simulations[11,12] of that motion.

MD simulations treat biological macromolecules and their solvent environments as classical many-body systems composed of atoms and account for the quantum mechanical forces acting on the nuclei and caused by the electrons through a parametrized molecular mechanics force field. In MD the coupled Newtonian equations of atomic motion are integrated numerically using time steps Δt of typically 1 fs. The result of such a simulation is a trajectory $\mathbf{x}_t = \mathbf{x}(t\cdot\Delta t)$, $t = 1, 2, ... T$, in a high-dimensional space $\mathbf{R}^D$ (e.g. the space $\mathbf{R}^{3N}$ of the Cartesian coordinates of all $N = 100-10\,000$ atoms of a protein) describing the time sequence of configurations $\mathbf{x}_t$ sampled by the macromolecule in solution upon thermal motion. Typical simulation times are nowadays in the range of a few tens of nanoseconds ($T \approx 10^7$).

* Corresponding author phone: +49-89-2180-9220; e-mail: tavan@physik.uni-muenchen.de.

10.1021/ct050020x CCC: $30.25
Published on Web 06/08/2005

Proteins are prototypes of complex dynamical systems in soft condensed matter. In addition to high-frequency thermal fluctuations of the atoms around their equilibrium positions that are also found in solids, they show large-scale low-frequency transitions between several metastable states, the so-called conformations.[13] This slow conformational dynamics is essential for protein function in biology. Various methods[14−21] have been suggested for the extraction of protein conformations from MD trajectories. Some make use of the fact that the conformations are marked by minima of the energy landscape,[15,19,22] some apply clustering procedures based on structural similarities,[16,21] and others[14] analyze the potential energy time series by the means of recurrence plot analysis.[23]

Grubmüller and Tavan[13] have demonstrated for a simplified protein model that its conformational dynamics can be described by a simple Markov model composed of only a few conformational states. Following this principle and considering only a few so-called essential degrees of freedom Dellnitz, Schütte, and others[18,20] chose a regular lattice for discretization of the thus reduced configuration space and determined the transfer matrix of the system by counting transitions between lattice cells. They identified the conformational states defining a coarse-grained Markov model by a rather complicated analysis of the eigenvectors and -values of the transfer matrix.

Following these general concepts we here propose an alternative approach toward the analysis of high-dimensional time series, which exhibit the characteristics of a Markov chain switching among a few states. In particular, the use of a density-oriented discretization of the data space[24−26] allows us to avoid the *curse of dimensionality* inherent to grid partitions. That curse expresses the common problem, that the number of parameters, which have to be statistically estimated from the data for the construction of a simplified model, grows exponentially with the dimension of the data space.

By modifying and expanding a self-organizing and biologically plausible neural network model originally suggested for the clustering of data sets[27] but without explicitly employing the language of neural networks, we construct from the time series a transfer matrix, whose dimension is kept relatively small due to the use of the density-oriented discretization.[25] As opposed to the Kohonen algorithm[28,29] used in ref 27 for discretization, our approach does not introduce distortions into the metrics of the data space.[24,30] The analysis of the transfer matrix is either performed by a nonlinear dynamics related to the neural network used previously for clustering[27] or by a deterministic coarse-graining procedure. Both methods generate hierarchies of Markov models at varying coarseness and provide the means to identify the particular hierarchy level which most plausibly maps the generating Markov process. We start with the explanation of the methods, and, to provide a relevant example, we subsequently analyze the MD trajectory of a small peptide in water.

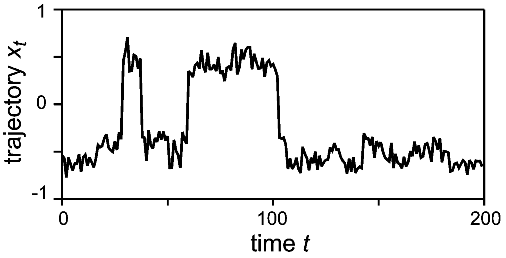

Figure 1. The first 200 steps of a time series of one-dimensional data created by a four-state Markov process. At the first glance, one can distinguish two ranges of frequent x_t values.

2. Method

For a simple graphical illustration of the employed concepts and methods, we first introduce a one-dimensional model time series, which, despite its simplicity and low-dimensionality, covers key ingredients of the problem. Figure 1 shows the first 200 steps of this time series $\mathscr{X} = \{x_t | t = 1, 2,..., T\}$, which covers $T = 10^6$ data points. The series has been generated from the Markov matrix

$$\mathbf{C}^{ex} = \begin{pmatrix} 0.8 & 0.17 & 0 & 0 \\ 0.2 & 0.8 & 0.03 & 0 \\ 0 & 0.03 & 0.8 & 0.2 \\ 0 & 0 & 0.17 & 0.8 \end{pmatrix} \tag{1}$$

by mapping the associated four-state Markov chain onto a one-dimensional dynamical system. The Markov chain generates *slow* transitions *among* the states i, $i = 1, ..., 4$. These transitions are differentiated by certain degrees of slowness: Very slow are the 2 ↔ 3 transitions, much faster but still slow are the transitions 1 ↔ 2 and 3 ↔ 4. A subsequent random process completes the mapping by creating *fast* one-dimensional jumps *within* the four coarse-grained states (jumps drawn from normal distributions $g(x|x_i,\sigma)$ of standard deviation $\sigma = 0.07$ and centered at $x_i \in \{\pm 0.4, \pm 0.6\}$, see Figure 2). The resulting one-dimensional time series shares the characteristics of fast fluctuations within and differently slow transitions among coarse-grained states with peptide and protein conformational dynamics.

Note that the Markov matrix (1) generating our model time series obeys the property of detailed balance,[8,31] which requires that there are nonzero numbers f_r with

$$C^{ex}_{r'r} f_r = C^{ex}_{rr'} f_r'$$

Up to a constant factor these numbers f_r are the components $p_{r,stat}$ of the stationary distribution $\mathbf{p}_{stat} = (0.17, 0.2, 0.2, 0.17)^T/0.74$, which is the right eigenvector of $\mathbf{C}^{ex}$ to the eigenvalue $\lambda_1 = 1$. In general, R-dimensional Markov matrices $\mathbf{C}$ generating a time discrete stochastic process

$$\mathbf{p}(t + \Delta t) = \mathbf{C}\mathbf{p}(t)$$

and obeying detailed balance have a set of nice mathematical properties:[31] (i) although they are usually nonsymmetric, their eigenvalues λ_r, $r = 1,..., R$, are all real with $1 \geq \lambda_r > 0$ ($\lambda_r \geq \lambda_{r'}$ for $r < r'$), (ii) for simply connected state spaces there is exactly one largest eigenvalue $\lambda_1 = 1$ marking the

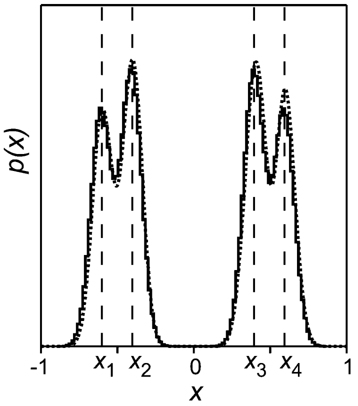

Figure 2. Histogram (solid line) and normal mixture density estimate (dotted line) of all $T = 10^6$ data points $x_t \in \mathcal{X}$(both estimates comprise the same number $R = 100$ of local components). The four density maxima marking the Markov states are clearly distinguished. Within the two pairs (1,2) and (3,4) of the generating Gaussians considerable overlaps and between the pairs a strict separation are observed.

stationary distribution $\mathbf{p}_{stat}$, and (iii) for every initial distribution $\mathbf{p}(0)$ the iteration of the process converges to $\mathbf{p}_{stat}$.

For physical systems in thermal equilibrium the property of detailed balance frequently applies and then derives from the principle of microscopic reversibility. By applying the arguments in chapter 5.3.6b of ref 8, the following sufficient condition may be formulated for the equilibrium conformational fluctuations of a peptide sampled by an MD simulation: If the resulting trajectory $\mathcal{X}$ provides a statistically sufficient sampling of the accessible configuration space and if the observed transitions among arbitrarily defined coarse-grained states are statistically independent of the previous history of the process, i.e., are Markovian, then the associated conformational dynamics obeys detailed balance (like our simple model process does by construction).

For this simple example the particular task of time series analysis treated in this paper can now be stated as follows: Identify and reconstruct the generating Markov model (1) from the observed time series $\mathcal{X}$ as well as possible!

2.1. Partitioning the Data Space. For an ergodic system the distribution of the configurations $\mathbf{x}_t$ sampled by the trajectory $\mathcal{X}$ in the limit $t \rightarrow \infty$ defines the so-called invariant density $p_{inv}(\mathbf{x})$.[32] A parametric model for $p_{\mathrm{inv}}(\mathbf{x})$ can be estimated from a sufficiently extended sample trajectory $\mathcal{X}$ by using a mixture

$$\hat{p}(\mathbf{x}|\mathcal{W}, \sigma) = \frac{1}{R}\sum_{r=1}^{R} g(\mathbf{x}|\mathbf{w}_r, \sigma) \qquad (2)$$

of R univariate normal distributions $g(\mathbf{x}|\mathbf{w}_r, \sigma)$ of identical widths σ and statistical weights $1/R$ centered at points $\mathbf{w}_r \in \mathbf{R}^D$. With the exception of the number R, the model parameters, i.e., the *codebook* $\mathcal{W} \equiv \{\mathbf{w}_r | r = 1,...., R\}$ and the common width σ, are adapted to the data set $\mathcal{X}$ according to the *maximum likelihood* criterion[33] by a safely converging deterministic annealing algorithm.[24-26,34] The extraordinary robustness of this quite simple algorithm critically depends on the choice of identical widths σ for the normal distributions, although an extension toward more complicated multivariate mixture models is available.[25,26] The algorithm guarantees that the univariate normal distributions associated with the resulting optimal parameters $\mathcal{W}^{ML}$ and σ^{ML} represent roughly the same number of data points each. This property of the optimal density estimate (2) is called *load balance*[24-26] and induces a first guideline for the choice of the remaining model parameter R through the following considerations.

The components of the mixture model (2) are R class-conditional probability densities and indicate how the data belonging to class r are distributed. By Bayes' theorem every point $\mathbf{x} \in \mathbf{R}^D$ is assigned to the class r with the probability[25]

$$\hat{P}(r|\mathbf{x}, \mathcal{W}, \sigma) = \frac{(1/R)\; g(\mathbf{x}|\mathbf{w}_r, \sigma)}{\hat{p}(\mathbf{x}|\mathcal{W}, \sigma)} \qquad (3)$$

Due to the normalization

$$\sum_{r=1}^{R} \hat{P}(r|\mathbf{x}, \mathcal{W}, \sigma) = 1 \qquad (4)$$

the probabilities (3) define a fuzzy partition of the data space when considered as functions of $\mathbf{x}$. In the limit $\sigma \rightarrow 0$ this partition becomes a crisp Voronoi tessellation[35] of the data space. Because of the load balance, each of the partition volumes covers approximately the same number T/R of data points, independently of the dimension D of the data. For a given data set of size T, the choice of the codebook size R determines T/R and thus defines the statistical quality, at which each $\mathbf{w}_r \in \mathcal{W}$ is estimated from the data $\mathbf{x}_t \in \mathcal{X}$.[26] Therefore, this type of density-oriented data space discretization can avoid the curse of dimensionality mentioned in the Introduction. For our one-dimensional example, Figure 2 compares a grid discretization (histogram) with the mixture model (2) and demonstrates the quality of the mixture estimate. Note that, because of load balance, the distribution $\tilde{p}(w)$ of codebook vectors closely models the distribution $p_{inv}(x)$ of the data ($\tilde{p} \approx p_{inv}$).[24]

2.2. Transfer Operator. The transfer operator describing the observed dynamical system is estimated using the partition described above. To simplify the notation, we extract from the trajectory $\mathcal{X}$ the set Y of all $T - 1$ pairs $y_t \equiv (\mathbf{x}_t, \mathbf{x}_{t+1})$ and define the correlation product

$$\langle f(\mathbf{x}_{t+1}) g(\mathbf{x}_t)\rangle_Y \equiv \frac{1}{T-1}\sum_{y_t \in Y} f(\mathbf{x}_{t+1}) g(\mathbf{x}_t) \qquad (5)$$

where f and g are functions of $\mathbf{x}_t$. The transfer matrix $\mathbf{C}$ defined by the partition (3) then is[36]

$$C_{rr'} = \frac{\langle \hat{P}(r|\mathbf{x}_{t+1}, \mathcal{W}, \sigma)\hat{P}(r'|\mathbf{x}_t, \mathcal{W}, \sigma)\rangle_Y}{\langle \hat{P}(r'|\mathbf{x}_t, \mathcal{W}, \sigma)\rangle_Y} \qquad (6)$$

Clearly, $\mathbf{C}$ depends on the parameters $\{\mathcal{W}, \sigma\}$ as well as on the choice of the codebook size R. There are R^2 matrix-elements $C_{rr'}$, which have to be statistically estimated from the $T - 1$ data points $y_t \in Y$ by evaluation of the correlation products in eq 6. To ensure sufficient statistics one should therefore demand that $R^2/T \ll 1$: This requirement thus represents a second guideline for the choice of R. Note that large values of σ, though helping to improve the statistics,

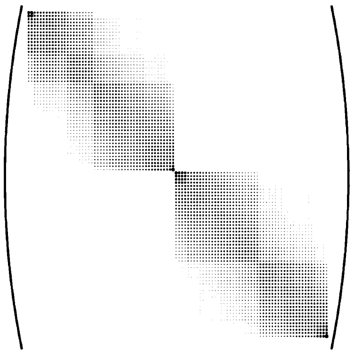

Figure 3. Transfer matrix (6) for the sample trajectory $\mathscr{X}$. The radii of the circles code the sizes of the matrix elements. On a coarse level, two diagonal blocks with non-zero elements are seen. The elements outside these blocks are substantially smaller and, therefore, are invisible here. Each of the coarse diagonal blocks decomposes into two diagonal sub-blocks, partly linked by off-diagonal blocks.

decrease the information content of **C** by smoothening. This aspect of the dependence of **C** on the fuzziness parameter σ is further discussed in section 2.7.

Because the partition functions $\hat{P}(r|\mathbf{x}, \mathscr{W}, \sigma)$ are centered around the points $\mathbf{w}_r \in \mathbf{R}^D$ and assume values close to 1 for points **x** near $\mathbf{w}_r$, the matrix (6) codes the spatial correlations between consecutive points. The elements of **C** are non-negative, and its columns are normalized to 1 ($\forall r'$: $\sum_{r=1}^{R} C_{rr'} = 1$). Therefore, **C** is an R-state Markov matrix. As it is generated from a trajectory, the associated state space is simply connected. Correspondingly, **C** has only one right eigenvector $\mathbf{p}_{stat}$ to the eigenvalue 1 marking the stationary state. As one can easily show from the definition (6) of the transfer matrix **C** and using the normalization (4) of the partition functions the stationary distribution is given by the loads of the partition volumes, i.e., $p_{r,stat} \approx \langle \hat{P}(r|\mathbf{x}_t, \mathscr{W}, \sigma)\rangle_Y$ up to corrections smaller than $1/T$. The property of load balance characteristic for our partition then implies that $\mathbf{p}_{stat}$ approximately represents a uniform distribution, that is $p_{r,stat} \approx 1/R$.

Figure 3 shows the transfer operator (6) for the time series of Figure 1. This matrix, like all other Markov matrices discussed further below, obeys detailed balance to a very good approximation: the statistical errors $|C_{rr'}p_{r',stat} - C_{r'r}p_{r,stat}|/\max\{C_{r'r}p_{r,stat}\}$ are all smaller than 1%. Because $\mathbf{p}_{stat}$ is nearly uniform, it is nearly symmetric. Apart from the eigenvalue 1, the matrix has three sizable eigenvalues (0.969, 0.442, 0.421), whereas all the remaining 96 eigenvalues are smaller than 0.002. According to refs 20 and 18 such a distribution of eigenvalues indicates the existence of two long-lived or four somewhat shorter lived metastable states. This dynamical structure of the sample trajectory is also visible in the hierarchical block structure of the depicted matrix, which clearly reveals the underlying Markov process (1). The visibility of that Markov process results from ordering the codebook elements w_r according to size ($w_r < w_{r'} \Rightarrow r < r'$), which is only feasible in one dimension.

2.3. Analysis of the Transfer Operator. Since there is no natural ordering of the codebook vectors $\mathbf{w}_r$ in higher-dimensional cases, the analysis of transfer matrices requires other means than simple visual inspection. For this purpose

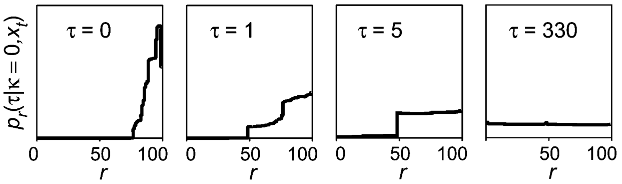

Figure 4. Linear ($\kappa = 0$) dynamics eq 8 elicited by the randomly chosen point $x_t = 0.66$ for **C** from Figure 3. Left to right: The initial distribution ($\tau = 0$) associated with x_t spreads rapidly filling predominantly the right quarter of the state space ($\tau = 1$); within the next five time steps a second metastable state appears covering predominantly the right half of the state space; at $\tau = 330$ the nearly uniform stationary distribution is reached.

we define the time-dependent probability vector[37]

$$\mathbf{p}(\tau|\mathbf{x}_t) \equiv \begin{pmatrix} \hat{P}(1,\tau|\mathbf{x}_t, \mathscr{W}, \sigma) \\ \vdots \\ \hat{P}(R,\tau|\mathbf{x}_t, \mathscr{W}, \sigma) \end{pmatrix} \tag{7}$$

whose initial components $\hat{P}(r, 0|\mathbf{x}_t, \mathscr{W}, \sigma) = \hat{P}(r|\mathbf{x}_t, \mathscr{W}, \sigma)$ are given by the posterior probabilities (3) of a given point $\mathbf{x}_t$. Furthermore we consider the evolution of the components $p_r(\tau)$ of $\mathbf{p}(\tau|\mathbf{x}_t)$ described by the following family of nonlinear differential equations

$$\frac{d}{d\tau} p_r = (\mathbf{L}\mathbf{p})_r + \kappa p_r(p_r - \mathbf{p}^2) \tag{8}$$

where the family parameter $\kappa \geq 0$ scales the nonlinear term. The matrix **L** derives from the transfer operator **C** and from the associated sampling time step $\Delta\tilde{t}$ according to

$$\mathbf{L} = \frac{1}{\Delta\tilde{t}} \ln \mathbf{C} \tag{9}$$

Note that the nonlinear dynamics (8) conserves the normalization $\sum_{r=1}^{R} p_r(\tau|\mathbf{x}_t) = 1$ of the probabilities. Since the time evolution of $\mathbf{p}(\tau|\mathbf{x}_t)$ depends on κ, we extend the notation to $\mathbf{p}(\tau|\kappa, \mathbf{x}_t)$. To calculate that evolution numerically, a discretization of (8) is used as described in Appendix A.

For an understanding of the dynamics (8), we look at the linear and the nonlinear terms of eq 8 separately. The purely linear dynamics (i.e. $\kappa = 0$) describes a Markov process of probability redistribution. Independent of the initial condition $\mathbf{x}_t$, the distribution $\mathbf{p}(\tau|\kappa = 0, \mathbf{x}_t)$ is temporarily caught in some metastable intermediate states but eventually converges toward the single stationary right eigenvector $\mathbf{p}_{stat}$ of **C**. This process is illustrated in Figure 4, which also demonstrates that $\mathbf{p}_{stat}$ is nearly uniform as claimed above.

As we explain in Appendix B, the purely nonlinear dynamics has $2^R - 1$ stationary points, each given by distributions $\mathbf{p}^{\mathscr{M}}$, which are uniform on a nonempty subset $\mathscr{M} \subset \{1,...., R\}$ and vanish elsewhere. However, only R of these distributions, the δ-distributions $p_r = \delta_{rs}$, are stable attractors of the nonlinear dynamics. The attractor δ_{rs} selected by the dynamics is defined by the largest component $p_s(0|\kappa,\mathbf{x}_t)$ of the initial distribution. Thus, the nonlinearity generates a winner-takes-all dynamics of *Darwinian selection*[27] and may be considered as the inverse of the diffusion operator.

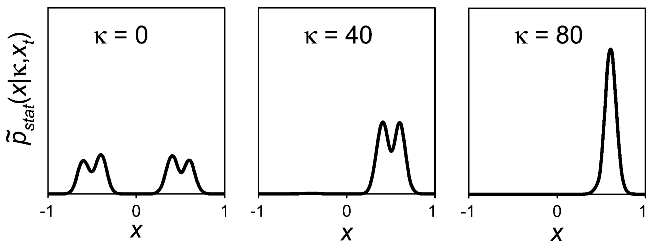

Figure 5. Stationary virtual densities associated with the starting point $x_t = 0.66$ at various strengths κ of the nonlinearity. At $\kappa = 0$ the attractor is the mixture model (2) of the invariant density depicted in Figure 2, because the generating Markov model (1) obeys detailed balance. At increasing κ the virtual density first ($\kappa = 40$) becomes confined to the two overlapping Gaussian components $i = 3$, 4 and eventually ($\kappa = 80$) to component $i = 4$ of the generating model shown in Figure 2.

By combining these mutually counteracting processes as given by eq 8 one obtains a dynamics capable of stabilizing and focusing metastable intermediates of the linear ($\kappa = 0$) relaxation process. It exhibits N_κ attractors $\mathbf{p}^n_{stat}(\kappa)$, $n = 1,...., N_\kappa$, where N_κ increases with κ ($1 \leq N_\kappa \leq R$). The specific attractor $\mathbf{p}^n_{stat}(\kappa)$ selected by the dynamics depends on the initial condition $\mathbf{x}_t$ and, therefore, classifies these initial conditions by $n \equiv n(\mathbf{x}_t|\kappa)$. Figure 1 in the Supporting Information illustrates how larger strengths κ of the nonlinearity stabilize increasingly short-lived metastable intermediates, prevent their diffusive spreading, and correspondingly identify metastable states at a decreasing level of coarse-graining.

2.4. Virtual Density. The distributions $\mathbf{p}(\tau|\kappa, \mathbf{x}_t) \in \mathbf{R}^R$ can be mapped onto *virtual*[38] probability densities

$$\tilde{p}(\mathbf{x}|\tau,\kappa,\mathbf{x}_t) \equiv \sum_{r=1}^{R} p_r(\tau|\kappa,\mathbf{x}_t) g(\mathbf{x}|\mathbf{w}_r,\sigma) \tag{10}$$

in the data space. For a given parameter set $\mathcal{Z} = \{\mathcal{W}, \sigma, \mathbf{C}\}$ the virtual density $\tilde{p}(\mathbf{x}|\tau,\kappa,\mathbf{x}_t)$ depends on the time τ, the nonlinearity parameter κ, and the initial condition $\mathbf{x}_t$. By eq 10, the dynamics (8) of the distributions $p_r(\tau|\kappa,\mathbf{x}_t)$ is mapped onto an equivalent temporal evolution of the virtual densities. At convergence one obtains the stationary virtual density $\tilde{p}^n_{stat}(\mathbf{x}|\kappa)$, which is associated with the initial data point $\mathbf{x}_t$ by the dynamics [$n \equiv n(\mathbf{x}_t|\kappa)$]. Particularly in the linear case ($\kappa = 0$) and for a transfer matrix (6) obeying detailed balance,[31] the virtual density $\tilde{p}(\mathbf{x}|\tau,\kappa,\mathbf{x}_t)$ converges for each $\mathbf{x}_t$ toward the mixture model (2) of the invariant density (cf. Figure 2 in the Supporting Information).

For increasing values of the nonlinearity κ, Figure 5 depicts the mapping (10) of the stationary distributions $\mathbf{p}^n_{stat}(\kappa)$ (see Figure 1 of the Supporting Information) onto the corresponding stationary virtual densities $\tilde{p}^n_{stat}(x|\kappa)$. At growing nonlinearity κ these densities $\tilde{p}^n_{stat}(x|\kappa)$ are confined to increasingly narrow and short-lived substructures of our model (2) for the invariant density. Depending on the strength κ of the nonlinearity, differently coarse-grained classes $\tilde{p}^n_{stat}(x|\kappa)$ are associated with the initial condition x_t. Thus, the stationary virtual densities turn out to represent density models for the metastable states.

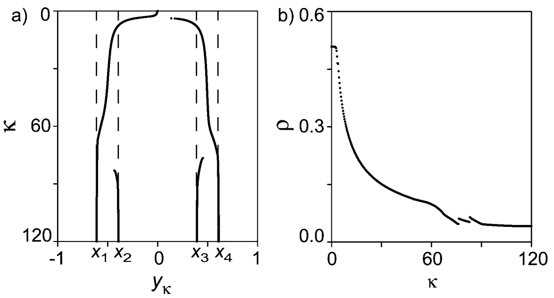

Figure 6. (a) Prototypes y_κ for $\kappa \in [0,120]$ and all data $x_t \in \mathcal{X}$. For comparison also the centers x_i for the four Gaussians associated with the states of the original Markov chain (1) are indicated by dashed lines. (b) Dependency of the average spread ρ of the stationary virtual densities on κ. The initial value $\rho \approx 0.5$ decreases almost monotonously with growing κ. At values of k near bifurcations in (a) $\rho(\kappa)$ is steeper. The small discontinuities of $\rho(\kappa)$ in the vicinity of bifurcations are due to numerics.

2.5. Moments of the Virtual Density. The first moments

$$\mathbf{y}^n_\kappa = \int \mathbf{x}\ \tilde{p}^n_{stat}(\mathbf{x}|\kappa)dx \tag{11}$$

of the stationary virtual densities (10) are obtained by integrating over the local normal distributions as[39]

$$\mathbf{y}^n_\kappa = \sum_{r=1}^{R} \mathbf{w}_r\, p^n_{r,stat}(\kappa) \tag{12}$$

Because at each κ the label n classifies the $\mathbf{x}_t$, the stationary solutions of (8) thus associate to each data point $\mathbf{x}_t \in \mathbf{R}^D$ a prototypical point $\mathbf{y}^n_\kappa \in \mathbf{R}^D$.

For our sample trajectory $\mathcal{X}$, Figure 6a depicts all prototypes y^n_κ associated with the $x_t \in \mathcal{X}$ as a function of the nonlinearity parameter $\kappa \in [0, 120]$. The figure shows that they remain invariant over wide ranges of κ while exhibiting bifurcations at certain critical values κ_ℓ. The prototypes y^n_κ mark metastable states, characterized by fast transitions within, and slow transitions among the states of the original two-stage Markovian dynamics x_t. According to Figure 6, the boundary between *slow* and *fast* shifts toward shorter time scales with increasing κ because more and more short-lived metastable states are identified. At large κ the nonlinear dynamics eventually identifies the four prototypical points x_i characterizing the states of the original Markov model (1). Thus, the depicted bifurcation pattern reflects the hierarchical block structure of the transfer matrix (cf. Figure 3) analyzed by the nonlinear dynamics (8) at varying κ.

Higher moments of the stationary virtual densities can be calculated analogously. For a given initial condition $\mathbf{x}_t$ the variance is given by

$$\mathcal{V}[\mathbf{x}_t] = \sum_{r=1}^{R} p^{n(\mathbf{x}_t)}_{r,stat}(\kappa)[\mathbf{y}^{n(\mathbf{x}_t)}_\kappa - \mathbf{w}_r]^2 + \sigma^2 \tag{13}$$

and is—apart from the constant variance σ^2 of the Gaussians g in (10)—the sum of the squared distances between the prototypes $\mathbf{y}_\kappa$ and the codebook vectors $\mathbf{w}_r$ weighted by the probabilities (7). The value of $\rho(\mathbf{x}_t) \equiv \sqrt{\mathcal{V}[\mathbf{x}_t] - \sigma^2}$ mea-

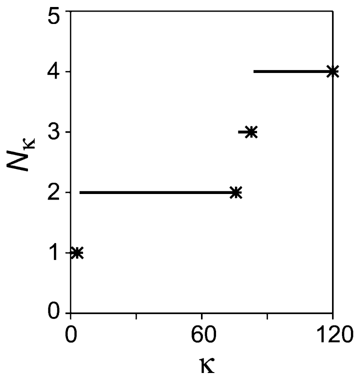

Figure 7. Number N_κ of prototypes of the trajectory from Figure 1. The asterisks mark the values of κ_ℓ, at which bifurcations occur in Figure 6a, and the value $\kappa_{max} = 120$.

sures the spread of the virtual density in data space. The dependency of the data average spread $\rho \equiv \langle \rho(\mathbf{x}_t) \rangle_Y$ on κ is plotted in Figure 6b and clearly indicates the contraction of the $\tilde{p}_{stat}$ with increasing κ.

2.6. Hierarchical Classification. For higher-dimensional dynamics one cannot visualize bifurcation patterns. Therefore other means are required to obtain insight into the hierarchy of classes identified by the nonlinear dynamics at increasing κ. A generally applicable procedure is to determine all prototypes (12) for all vectors $\mathbf{x}_t \in \mathcal{X}$, which results for each value of κ in a prototype set $\mathcal{P}_\kappa = \{\mathbf{y}_\kappa^n | n = 1,..., N_\kappa\}$ (cf. Figure 6a). The number N_κ of different stationary solutions can then be plotted as a function of κ and gives a first insight into the coarse-grained structure of the dynamics.

Figure 7 shows such a plot for our sample dynamics. The number N_κ grows monotonically with κ and remains constant within certain intervals $[\kappa_\ell, \kappa_{\ell+1}]$ ($\ell = 1, 2,...$ and $\kappa_1 < \kappa_2 < ...$). These intervals differ strongly in widths. Two large intervals belong to the values $N_\kappa = 2$ and $N_\kappa = 4$. They indicate that the system has two and four differently coarse-grained states with strongly different lifetimes. Thus, the corresponding two or four classes may be good choices for the intended construction of coarse-grained models, and we know in this case, of course, that they are. There is also a very small interval marking a three-state model, which finds no correspondence in the generating process given by eq 1. This three-state model is due to statistical fluctuations affecting the elements of the 100-dimensional transfer matrix and, therefore, the classification of the data points by the nonlinear dynamics. However, the small range of κ-values, within which the three-state model is predicted, indicates that it is not an intrinsic feature of the monitored time series. Similar structures are expected to be found in such plots whenever a reasonably clear-cut separation of time scales happens to exist in the dynamics represented by the transfer matrix (6). Also here large intervals with constant N_κ will point to plausible models.

It now remains to be seen at which values of κ these models should be determined. For this purpose we use the observation (cf. Figure 6a) that the prototypes $\mathbf{y}_\kappa^n$ do not vary much as κ approaches a critical bifurcation value κ_ℓ from below. Therefore, we reduce the continuous family $\{\mathcal{P}_\kappa | \kappa \in \mathbf{R}_0^+\}$ of prototype sets $\mathcal{P}_\kappa$ to a minimal discrete family $\{\mathcal{P}_\ell | \ell = 1,..., \ell_{max}\}$ of ℓ_{max} prototype sets $\mathcal{P}_\ell$ by selecting the prototype sets $\mathcal{P}_\ell \equiv \mathcal{P}_{\kappa_\ell}$, which are located just before the jumps of N_κ to higher values. In our simple example these values κ_ℓ are marked by asterisks in Figure 7.

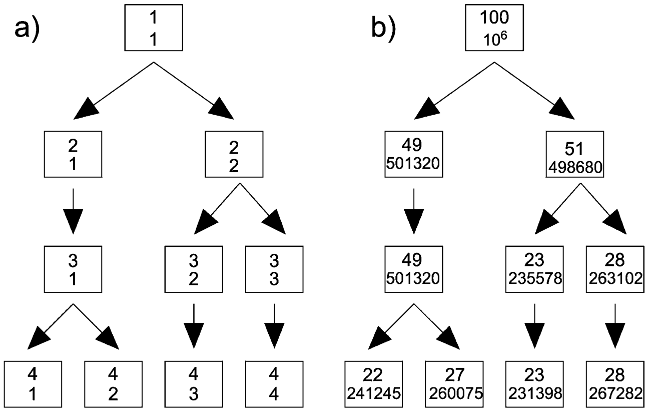

Figure 8. (a) The hierarchy of prototypes as a tree. The nodes (boxes) characterize the prototypes $\mathbf{y}_\ell^n$ by the upper index ℓ and lower index n. The edges (opposite to the arrow) denote the association of a prototype $\mathbf{y}_{\ell+1}^n$ to a prototype $\mathbf{y}_\ell^{n'}$. (b) The classification of the data set $\mathcal{X}$. Here, the upper index denotes the number of codebook vectors, $\mathbf{w}_r$ associated with the respective prototype, and the lower index the corresponding number of data $\mathbf{x}_t$.

In the next step we arrange the thus determined discrete family of prototype sets into a hierarchy by associating with a higher-level prototype $\mathbf{y}_\ell^{n'}$ each lower-level prototype $\mathbf{y}_{\ell+1}^n$, whose probability vector $\mathbf{p}(\tau|\kappa, \mathbf{y}_{\ell+1}^n)$ converges at $\kappa = \kappa_\ell$ under the dynamics (8) to $\mathbf{p}^{stat}(\kappa_\ell, \mathbf{y}_\ell^{n'})$. For our standard example the resulting hierarchy is drawn as a directed tree in Figure 8a.

Analogously we can classify the codebook $\mathcal{W}$ and the data set $\mathcal{X}$ on the hierarchy level ℓ by calculating for each of their elements the first moment (12) of the stationary virtual density at $\kappa = \kappa_\ell$. The result of this classification for the sample data set is shown in Figure 8b. As a result of the load balance of the mixture model (2) mentioned in 2.1, the percentage of codebook vectors $\mathbf{w}_r$ associated with a prototype $\mathbf{y}_\ell^n$ reproduces approximately the respective percentage of data points $\mathbf{x}_t$.

2.7. Extracting a Markov Model at a Hierarchy Level. Having set up a hierarchy of classifiers, which, at each hierarchy level ℓ, associate the codebook vectors $\mathbf{w}_r$ and data points $\mathbf{x}_t$ to one of N_ℓ prototypes $\mathbf{y}_\ell^n$, $n = 1,..., N_\ell$, it remains to be clarified as to how one should calculate correspondingly coarse-grained N_ℓ-state Markov matrices $\mathbf{C}^\ell = \{C_{nn'}^\ell\}$. As discussed above, such matrices can represent plausible coarse-grained descriptions of the observed dynamics, if the associated number N_ℓ of prototypes has been found to be stable over a wide range of the nonlinearity parameter κ.

There are two different choices for the computation of the $\mathbf{C}^\ell$. One can (i) reduce the original R-state Markov matrix (6) by using the classification of the codebook vectors $\mathbf{w}_r$ or (ii) directly set up the coarse-grained matrices by employing the classification of all data $\mathbf{x}_t$. To make notation simpler we consider a single selected hierarchy level ℓ and discuss choice (i) first.

Let $I_n = \{r|\mathbf{y}_{\ell}(\mathbf{w}_r) = \mathbf{y}^n_{\ell}\}$ be the set of all indices r of codebook vectors $\mathbf{w}_r$, which are classified to a given prototype $\mathbf{y}^n_{\ell}$. Summing the associated partition volumes $\hat{P}(r|\mathbf{x}_t,\mathcal{W}, \sigma)$ according to

$$\tilde{P}_n(\mathbf{x}_t) \equiv \sum_{r\in I_n} \hat{P}(r|\mathbf{x}_t,\mathcal{W}, \sigma) \tag{14}$$

we obtain the partition function of the prototype $\mathbf{y}^n_{\ell}$, which measures the *posterior* probability that the point $\mathbf{x}_t$ belongs to $\mathbf{y}^n_{\ell}$. Like the original *R*-state Markov model defined by eq 6 also the correspondingly reduced Markov matrix $C^{\ell}_{nn'}$ should fulfill the analogous relation

$$C^{\ell}_{nn'} = \frac{\langle \tilde{P}_n(\mathbf{x}_{t+1})\, \tilde{P}_{n'}(\mathbf{x}_t)\rangle_Y}{\langle \tilde{P}_{n'}(\mathbf{x}_t)\rangle_Y} \tag{15}$$

with $n, n' \in \{1,..., N_{\ell}\}$. Inserting the *posterior* probabilities (14) into (15) and taking into account the definition (6) as well as the fact that the index sets I_n and $I_{n'}$ are disjoint for all classes $n \neq n'$, we obtain a reduced Markov matrix $C^{\ell}_{nn'}$ from the original matrix $C_{rr'}$ by

$$C^{\ell}_{nn'} = \frac{\sum_{r'\in I_{n'}} [\langle \hat{P}(r'|\mathbf{x}_t,\mathrm{W},\sigma)\rangle_Y \sum_{r\in I_n} C_{rr'}]}{\sum_{r'\in I_{n'}} \langle \hat{P}(r'|\mathbf{x}_t,\mathcal{W}, \sigma)\rangle_Y} \tag{16}$$

Note that this coarse graining procedure of Markov matrices preserves detailed balance, i.e., if detailed balance holds for $C_{rr'}$, it also holds for $C^{\ell}_{nn'}$ as can be seen by a few lines of algebra. The stationary distribution at level ℓ follows by $p^{\ell}_{n,stat} = \sum_{r\in I_n} p_{r,stat}$ from $\mathbf{p}_{stat}$ associated with $\mathbf{C}$. For our synthetic sample time series, in particular, one can additionally show that in the limit of infinite sampling the detailed balance of the generating Markov matrix (1) induces detailed balance also into the discretized transfer operator $\mathbf{C}$ given by eq 6. In this case one therefore expects that detailed balance holds at all levels of coarse graining up to statistical errors.

Following choice (ii) we can alternatively count all initial data pairs $(\mathbf{x}_{t+1},\mathbf{x}_t)$ which the nonlinear dynamics maps to the prototype pairs $(\mathbf{y}^n_{\ell},\mathbf{y}^{n'}_{\ell})$ and all initial data points $\mathbf{x}_t$ mapped to $\mathbf{y}^{n'}_{\ell}$. Calling the respective numbers $T_{nn'}$ and $T_{n'}$, with $\sum_n T_{nn'} = T_{n'}$, the reduced transfer matrix is

$$\tilde{C}^{\ell}_{nn'} = \frac{T_{nn'}}{T_{n'}} \tag{17}$$

For overlapping coarse-grained classes both choices will overestimate the transition probabilities due to unavoidable Bayesian decision errors. For explanation consider our standard example, in which the classes associated with the Gaussians $i = 1$, 2 and $i = 3$, 4 of the generating dynamics exhibit considerable overlaps (cf. Figure 2). Even an optimal Bayesian classifier[26] will, e.g., erroneously associate data $x_t > -0.5$ that have been drawn from the normal distribution 1 to class 2. As a result, fast transitions within class 1 are erroneously counted as 1 → 2 transitions, and the corresponding off-diagonal element $\tilde{C}^4_{21}$ of a four-state Markov model is overestimated at the expense of the diagonal element $\tilde{C}^4_{11}$. The size of this Bayesian decision error can be estimated by comparing the four-state Markov matrix

$$\tilde{\mathbf{C}}^4 = \begin{pmatrix} 0.72 & 0.24 & 0.00 & 0.00 \\ 0.28 & 0.73 & 0.03 & 0.00 \\ 0.00 & 0.03 & 0.72 & 0.28 \\ 0.00 & 0.00 & 0.25 & 0.72 \end{pmatrix} \tag{18}$$

which has been calculated by eq 17 at the highly plausible level $\ell = 4$ of the hierarchy in Figure 8, with the generating Markov model (1). In fact, a Bayesian classification of the data $x_t \in \mathcal{X}$ (using the knowledge on the four class-conditional distributions from which the data have been drawn) numerically reproduces the four-state Markov matrix (18). As a result of the Bayesian decision error the estimated lifetimes

$$\tau_n = \frac{\Delta \tilde{t}}{1 - \tilde{C}^{\ell}_{nn}} \tag{19}$$

of the various coarse-grained states n are lower bounds to the true lifetimes of the generating dynamics.

For a related reason, small additional errors of this type will be introduced, if a Markov model on a given hierarchy level is estimated by the efficient reduction algorithm (16) instead by the computationally more demanding counting algorithm (17). The additional errors arising in the description of transitions among overlapping states are now due to the fuzziness σ of the partition (3) used both for the original discretization (6) of the transfer operator and for the coarse-grained partition functions (14). Correspondingly, they can be reduced by decreasing σ beyond the value σ^{ML} determined by the maximum likelihood estimate (cf. section 2.1). For our standard example and $\sigma = \sigma^{ML}$, they can be estimated by comparing the four-state matrix

$$\mathbf{C}^4 = \begin{pmatrix} 0.69 & 0.26 & 0.00 & 0.00 \\ 0.31 & 0.71 & 0.03 & 0.00 \\ 0.00 & 0.03 & 0.71 & 0.32 \\ 0.00 & 0.00 & 0.26 & 0.68 \end{pmatrix} \tag{20}$$

extracted by eq 16 with the optimal estimate (18) and the underlying Markov model (1). For instance, due to the Bayesian decision error the lifetime of state 1 is underestimated in (18) by about 30%, to which the fuzziness affecting (20) adds another 5%.

Fortunately, overlapping coarse-grained states are unlikely in high-dimensional data spaces, particularly in the ones one may use for the characterization of peptide conformational dynamics. Therefore, the unavoidable Bayesian decision errors are expected to be small. For the same reason the use of a fuzzy partitioning should not introduce large errors here, because overlapping partition volumes will mainly occur in the mapping of statistically predominant states and will then be combined by eq 14 into the associated coarse partition volumes. Because they then belong to the same state, they cannot affect the critical statistics of interstate transitions.

As a result, both algorithms should be equally applicable here and Bayesian decision errors will hardly deteriorate the results.

2.8. Alternative Construction of a Hierarchy. To check our results, we now will explain an alternative, deterministic, and very fast algorithm for constructing a hierarchy of coarse-grained Markov models from the original *R*-dimensional transfer matrix (6) and for identifying most plausible levels within that hierarchy. Here, the basic idea is to consecutively unite those Markov states that are mutually connected by the fastest transitions.

The alternative procedure is solely applicable to dynamical processes obeying *detailed balance* (see the introductory remarks to section 2), because this principle allows us to uniquely assign a *time scale* to the mutual transitions at the various levels of the hierarchy. The *R*-state Markov model (6) obeys detailed balance, if

$$C_{rr'}\langle \hat{P}(r'|\mathbf{x}_t, \mathcal{W}, \sigma)\rangle_Y = C_{r'r}\langle \hat{P}(r|\mathbf{x}_t, \mathcal{W}, \sigma)\rangle_Y \tag{21}$$

meaning that the probability flow between any two states r and r' is equal in the stationary distribution. Dividing eq 21 by the components of the stationary distribution we immediately see that the matrix

$$D_{rr'} \equiv \frac{C_{rr'}}{\langle \hat{P}(r|\mathbf{x}_t, \mathcal{W}, \sigma)\rangle_Y} \tag{22}$$

is symmetric. Therefore, its off-diagonal elements measure flow rates of the mutual transitions $r \leftrightarrow r'$, and we denote the maximal rate by D_{max}.

If we collect the index pair $\{r, r'\}$ belonging to D_{max} into an index set I_{R-1} and define one-member index sets I_n, $n = 1,..., R - 2$, to contain the indices r'' of the remaining states, we obtain the $R - 1$ index sets I_n, required for a first coarse-graining of partition volumes (14) and Markov matrices (16). The resulting $(R - 1)$-state Markov matrix $\hat{\mathbf{C}}^{R-1}$ can be considered as the level $\ell = R - 1$ of a model hierarchy, whose basis is formed by $\mathbf{C} \equiv \hat{\mathbf{C}}^R$. At this level, the coarse-grained partition volumes (14) provide a Bayesian classifier for the data $\mathcal{X}$ in terms of $R - 1$ Markov states.

The above process of combining the fastest mixing states into new and coarser states can be iterated until the level $\ell = 2$ just below the top of the hierarchy is reached. In this recursive coarse-graining scheme, the set $\hat{\mathcal{P}}^{\ell-1}$ of prototypes $\mathbf{y}_n^{\ell-1}$ is obtained for $n = 1,..., \ell - 2$ by

$$\hat{\mathbf{y}}_n^{\ell-1} = \mathbf{y}_{r''}^{\ell}, \quad \text{if } I_n = \{r''\} \tag{23}$$

and for $n = \ell - 1$ by

$$\hat{\mathbf{y}}_n^{\ell-1} = \frac{A_r^{\ell}\hat{\mathbf{y}}_r^{\ell} + A_{r'}^{\ell}\hat{\mathbf{y}}_{r'}^{\ell}}{A_r^{\ell} + A_{r'}^{\ell}}, \quad \text{if } I_n = \{r, r'\} \tag{24}$$

from $\hat{\mathcal{P}}^{\ell}$, where the coefficient A_r^{ℓ} denotes the number of codebook vectors $\mathbf{w}_r$ previously united into the prototype $\mathbf{y}_r^{\ell}$. Note here the initial conditions $\hat{\mathcal{P}}^R = \mathcal{W}$ and $\hat{\mathbf{y}}_r^R = \mathbf{w}_r$.

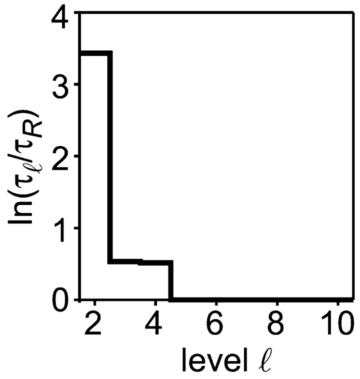

Figure 9. Time scales τ_ℓ for the last nine steps of the recursive coarse-graining procedure applied to our standard example. As time unit we have chosen the fastest time scale τ_R, which is given by the smallest eigenvalue λ_{min}^R of **C**. For explanation see the text.

At each level ℓ the fastest relaxation time scale can be characterized by considering the quantity

$$\tau_\ell \equiv \frac{1}{1 - \lambda_{min}^{\ell}} \tag{25}$$

where λ_{min}^{ℓ} is the smallest eigenvalue of $\hat{\mathbf{C}}^{\ell}$. Due to the consecutive removal of the most rapidly mixing states during our recursive coarse-graining, τ_ℓ is expected to increase in the sequence $\ell = R,..., 2$. Therefore the question, whether a given level of the resulting hierarchy furnishes a plausible coarse-grained model for the observed dynamics, can be decided by considering the ℓ-dependence of the fastest relaxation time scale τ_ℓ remaining at level ℓ. If $\tau_\ell \gg \tau_{\ell+1}$, then the model at level ℓ is considered to be plausible, because a large jump toward slower time scales indicates the presence of slowly mixing, i.e., metastable states at ℓ and of rapidly mixing states at the preceding level $\ell + 1$.

Figure 9 shows such a plot for our standard example using a logarithmic time scale. Large jumps of $\ln(\tau_\ell/\tau_R)$ occur when ℓ approaches the levels four and two from above. Thus the plot clearly reveals the hierarchical four- and two-scale structure of our example. Although the model hierarchy obtained by recursive coarse-graining generally differs from that generated by the nonlinear dynamics, the two procedures predict essentially identical models at the relevant levels $\ell = 2, 4$. Here, particularly the matrices $\hat{\mathbf{C}}^{\ell}$ calculated by the recursion are identical to the $\mathbf{C}^{\ell}$ obtained by version (i) of the dynamics-based procedure. Thus, the aim of validating the latter procedure has been reached.

2.9. Merits and Deficiencies of the Various Schemes. Up to this point we have introduced three algorithmic schemes by which one can derive a hierarchy of coarse-grained Markov models $\mathbf{C}^{\ell}$ from the transfer operator **C** defined by eq 6.

Scheme 1, which represents version (ii) of the dynamics-based procedure, relies at each hierarchy level ℓ on a crisp partitioning of the data $\mathbf{x}_t \in \mathcal{X}$ into N_ℓ coarse-grained classes n by the nonlinear dynamics (8) and a subsequent counting (17) of transitions among classes. Remarkably, in this scheme the fuzziness of the partition employed for the evaluation of the transfer operator **C** does not introduce errors into the

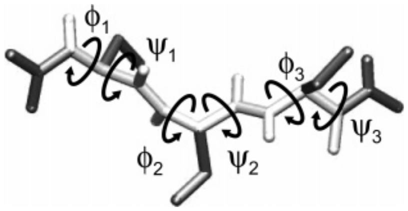

Figure 10. Tripeptide consisting of the backbone (light gray) and the serine side chains (dark gray) and definition of the dihedral angles (ϕ_i,ψ_i), $i = 1, 2, 3$. Most of the hydrogen atoms and all surrounding water molecules are omitted for clarity of representation.

computation of the coarse grained models $\mathbf{C}^{\ell}$. Experience has even shown that the nonlinear dynamics for classification of the data becomes more stable, if one increases the fuzziness of the partition in the computation of $\mathbf{C}$. In scheme 1, solely the limited statistics and Bayesian decision errors, which are unavoidable in the case of overlapping coarse-grained states, are sources of errors. The scheme is computationally expensive, because all T data points have to be classified by iteration of eq 8 at the $\ell = 1,..., \ell_{max}$ stages of the hierarchy.

Scheme 2, which represents version (i) of the dynamics-based procedure, classifies solely the R codebook vectors $\mathbf{w}_r$ by the nonlinear dynamics (8) and builds the hierarchy of ℓ_{max} Markov models $\mathbf{C}^{\ell}$ by a corresponding coarse-graining (16) of $\mathbf{C}$. Because the original partition used for the computation of $\mathbf{C}$ is preserved, small errors may be induced by its fuzziness. Therefore, one should reduce the fuzziness of the partition for the computation of $\mathbf{C}$ in this case below the value σ^{ML} obtained from codebook optimization. The computational effort is smaller by a factor R/T than in scheme 1.

Scheme 3 directly constructs the $\mathbf{C}^{\ell}$ from $\mathbf{C}$ by a deterministic and iterative unification of partition volumes and, thus, avoids costly iterations of the nonlinear dynamics (8). It is the computationally fastest procedure, shares the fuzziness errors with scheme 2, but is applicable solely to transfer operators exhibiting detailed balance to a good approximation. Because detailed balance requires extended trajectories this requirement limits the applicability of scheme 3.

In contrast, the other two schemes can also cope with a less extensive sampling and will render reasonable Markov models for a trajectory simulating equilibrium fluctuations (or for a set of trajectories starting from a given nonequilibrium state) as long as the data exhibit Markovian transitions among the various coarse-grained states. Therefore, they are also capable of modeling nonequilibrium relaxation processes. All three schemes provide the means to distinguish relevant levels of the hierarchy from spurious ones, and in all cases the computational effort is very small as compared to the cost of generating an extended MD-trajectory for a peptide in solution.

3. Sample Application

In the above explanation of algorithms we have considered a simple one-dimensional time series for purposes of illustration. To give an impression of more realistic applications, we will now consider a six-dimensional time series obtained from a 50 ns MD simulation of a serine tripeptide in water at room temperature and ambient pressure. Details of the simulation are given in Appendix C.

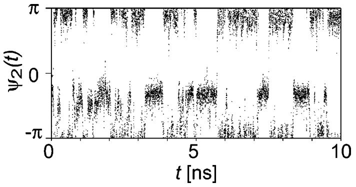

Figure 11. Time evolution of the angle ψ_2 (cf. Figure 10) during the first 10 ns of the MD simulation. Note that ψ_2 is a circular variable.

Figure 10 shows a configuration of the tripeptide molecule. At physiological temperatures its backbone (light gray) exhibits only six large-scale torsional degrees of freedom around chemical bonds, which are described by the dihedral angles ϕ_i and ψ_i. Thus, the temporal fluctuations of these six angles can be employed to determine the conformational dynamics of the backbone sampled by the MD trajectory.

Correspondingly we have generated a time series $\mathcal{X}$ from the MD trajectory, which consists of $T = 50\,001$ six-dimensional vectors $\mathbf{x}_t = [\phi_1(t),\psi_1(t),...,\psi_3(t),\psi_3(t)] \in (-\pi,\pi]^6$ and represents the backbone configurations at sampling intervals $\Delta\tilde{t} = 1$ps. Note that the torsion angles are circular variables and have to be treated accordingly.[21,40]

While the ϕ-angles fluctuate around $\approx -\pi/2$ (data not shown), the ψ-angles show a more interesting behavior. As an example Figure 11 shows the angle $\psi_2(t)$ during the first 10 ns of the simulation. Two ranges of values for $\psi_2(t)$ can be distinguished. One is given by the interval $I_\alpha = [-5\pi/6, \pi/6]$ and the other by its complement I_β. The angles ψ_1 and ψ_3 exhibit a similar bimodal behavior (data not shown) as is typical for polypeptides or proteins. Following the usual nomenclature[41] we classify local backbone configurations as α-helical, if $\psi_i \in I_\alpha$, and otherwise as extended or β-strandlike. Because each ψ-angle is either in the α- or in the β-range, we a priori expect the tripeptide to populate $2^3 = 8$ different conformations.

For time series analysis we first modeled the data distribution by a 25-component mixture density $\hat{p}(\mathbf{x}_t|\mathcal{W}^{ML},\sigma^{ML})$ as described in section 2.1. Here, the value $R = 25$ was chosen, because the quotient $R^2/T \approx 1.25\%$ appeared to be small enough as to enable a reasonably accurate statistics in the estimation of the transfer operator $\mathbf{C}$ by (6). $\mathbf{C}$ turned out to have eight large eigenvalues in the range [1.0,0.78]. The remaining eigenvalues were all smaller than 0.48. As discussed in section 2.2 such a structure of the eigenvalue spectrum points toward an eight-state model in agreement with our above expectation.

The plausibility of an eight-state Markov model was subsequently confirmed by classifying the data $\mathbf{x}_t$ through the nonlinear dynamics (8) at varying κ, because $N_\kappa = 8$ prototypes $\mathbf{y}_\kappa^n$ were found to be stable attractors of that dynamics over a wide range of κ-values. A classification of the three ψ_i values of these prototypes in terms of the α- and β-ranges introduced above then revealed that the $\mathbf{y}_\kappa^n$ are

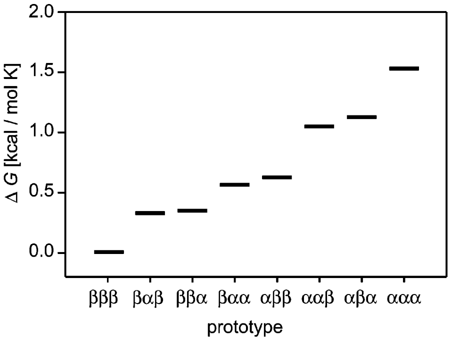

Figure 12. Free energy differences $\Delta G_n = G_n - G_1$ for an eight-state model with states n ordered according to increasing free energy $G_n = -k_BT\ln(P_n)$ and labeled by the simple (α/β)-classification of the three ψ-angles occurring in the prototypes $\mathbf{y}_\kappa^n$.

characterized by the eight possible triples ααα, βαα, ..., which can be formed from the symbols α and β. Also the recursive coarse-graining of **C** explained in section 2.8 and a time scale analysis analogous to that in Figure 9 (see Figure 3 of the Supporting Information) indicated an eight-state model. Furthermore, the two thus determined eight-state Markov matrices turned out to be identical, i.e., $\mathbf{C}^8 = \hat{\mathbf{C}}^8$ (data not shown).

The nonlinear dynamics classification of the data $\mathbf{x}_t \in \mathcal{X}$ at the eight-state level of the model hierarchy yielded the statistical weights P_n of the states $n = 1, ..., 8$. According to the arguments in ref 13 they determine the free energies $G_n = -k_BT \ln(P_n)$ of these states, where k_B is the Boltzmann constant and T is the temperature. The resulting relative energies of the eight conformational states are depicted in Figure 12. Interestingly the fully extended conformation $\beta\beta\beta$ is seen to be energetically most favorable and, therefore, is most frequently encountered in the trajectory. Furthermore, a $\beta \rightarrow$ α transition is seen to be energetically most favorable at ψ_2 and least favorable at ψ_1.

However, the dynamical connectivity of the eight states, which is visualized in Figure 13 by a plot of the Markov matrix $\mathbf{C}^8$, does not simply reflect the energetic state ordering. For instance, transitions $\beta\beta\beta \rightarrow \beta\beta$α are seen to be more likely than $\beta\beta\beta \rightarrow \beta$α$\beta$ although the latter target state has a slightly lower free energy than the former. Furthermore, the various conformations are mainly connected by single $\beta \rightarrow$ α transitions at individual angles ψ_i, whereas correlated transitions at pairs of these angles are quite rare.

By looking at further details of the connectivity displayed in Figure 13 and of the energetics shown in Figure 12, by analyzing the structures of the prototypes $\mathbf{y}_\kappa^n$ through molecular graphics, etc. one could now derive a lively picture and physical understanding concerning the conformational dynamics of serine tripeptide in water. However, these issues are beyond the scope of this paper. In the present context the given example solely serves to illustrate the kind of insights into complex dynamical systems, which now can be gained by applying the methods of time series analysis outlined above.

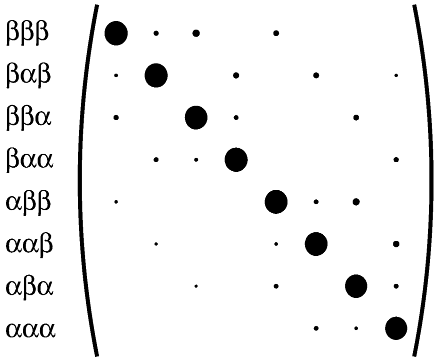

Figure 13. Graphical representation of the Markov matrix $\mathbf{C}^8$ extracted from the trajectory. The diameters of the dots code the sizes of the matrix elements $C^8_{nn'}$. (Matrix elements $C^8_{nn'}$ < 0.01 not drawn.)

4. Summary and Conclusion

For the analysis of high-dimensional time series in terms of coarse-grained Markov models we first have applied a density-oriented discretization of the data space. The properties of this partition ensure a balanced statistics for the estimation of all elements of the correspondingly discretized transfer operator **C**. The nonlinear dynamics eq 8 involving **C** was shown to classify the elements $\mathbf{x}_t$ of the time series in terms of prototypical points $\mathbf{y}_\ell^n$ marking the states $n = 1,..., N_\ell$ of coarse-grained Markov models $\mathbf{C}^\ell$. By varying the strength κ of the nonlinearity a hierarchy of such models is obtained, in which the number of states monotonically increases with κ in the range $N_\ell = 1,..., R$. Here, the case $N_\ell = 1$ is the trivial stationary model, and $N_\ell = R$ recovers the original discretization. Two different algorithms have been introduced to construct coarse-grained transfer operators $\mathbf{C}^\ell$ at the intermediate levels ℓ of the hierarchy. Here, the more time-consuming but accurate approach applies a classification of the data $\mathcal{X}$, whereas the other variant rests on a classification of prototypical points. The correctness of the latter procedure has been demonstrated by comparison with a deterministic and stepwise coarse-graining of the original R-state transfer operator $\mathbf{C} \equiv \mathbf{C}^R$.

For all these approaches observables were introduced, which allow for identifying the most plausible level within the thus constructed hierarchies of models. Their validity has been checked using a synthetic one-dimensional time series, which, apart from its low-dimensionality, exhibits all the characteristics of the relevant model class.

As an example for a more realistic application we have analyzed a six-dimensional time series obtained from a MD simulation of a tripeptide in aqueous solution. In this case the most plausible Markov model could be a priori guessed by physical knowledge on the conformational dynamics of such systems, and our approach actually recovered this guess by analysis of the simulation data. Although in practical applications questions concerning e.g. the number of partition volumes, by which the data space is discretized, the size of the time step, at which a dynamics is sampled, or the validity of the Markovian assumption for the coarse-grained time series have to be additionally addressed, the presented results demonstrate that our approach toward the identification of the most plausible coarse-grained Markov model compatible with the observations is actually viable. We would like to stress that the approach is applicable also to extremely high-dimensional data sets, which could result, e.g., from simulations of protein folding.

Appendix A: Discrete Dynamics

To solve the differential eq 8 numerically for given $\mathbf{x}_t$ and κ, we use the following algorithm, where $\tau = 1, 2, ...$ denotes discrete time steps of widths $\Delta\tilde{t}$.

- Calculate the probability vector (7).
- While $|\mathbf{p}(\tau + 1) - \mathbf{p}(\tau)| > \epsilon$, $0 \leq \epsilon \ll 1$:

$$\mathbf{p}^{(1)} = \mathbf{C}\mathbf{p}(\tau) \tag{26}$$

$$p_r^{(2)} = p_r^{(1)} + \tilde{\kappa} p_r^{(1)}[p_r^{(1)} - (\mathbf{p}^{(1)})^2] \tag{27}$$

$$p_r^{(3)} = \begin{cases} 0, & \text{if } p_r^{(2)} < 0 \\ 1, & \text{if } p_r^{(2)} > 1 \\ p_r^{(2)}, & \text{otherwise} \end{cases} \tag{28}$$

$$\mathbf{p}(\tau + 1) = \frac{\mathbf{p}^{(3)}}{\sum_{r'} p_{r'}^{(3)}} \tag{29}$$

Equations 26 and 27 discretize the differential eq 8 under the approximation

$$\ln[(\mathbf{C} - 1) + 1] = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}(\mathbf{C}\text{-}1)^n}{n} \approx (\mathbf{C} - 1) \tag{30}$$

and with $\tilde{\kappa} = \Delta\tilde{t}\kappa$. The cutoff (28) and the renormalization (29) serve to stabilize the algorithm numerically.

Appendix B: Attractors of the Nonlinear Dynamics

Here, we consider the nonlinear part

$$\frac{d}{d\tau} p_r = p_r(p_r - \mathbf{p}^2) \tag{31}$$

of (8) for the probability distribution (7). The uniform distribution

$$p_r = \begin{cases} \frac{1}{|\mathscr{M}|}, & r \in \mathscr{M} \\ 0, & \text{otherwise} \end{cases} \tag{32}$$

over a nonempty subset $\mathscr{M} \subset \{1,..., R\}$ of indices $r \in \mathscr{M}$ is stationary under (31), as can be seen by inserting (32) into (31). For each $M \equiv |\mathscr{M}|$ there are

$$\binom{R}{M}$$

possibilities to choose an index set $\mathscr{M}$. Therefore, there are a total of

$$\sum_{M=1}^{R} \binom{R}{M} = 2^R - 1$$

stationary solutions of (31). By applying a small deviation to one component p_s, $s \in \mathscr{M}$, one can easily show that only the R δ-distributions ($M = 1$) are stable attractors of (31).

Appendix C: Simulation Method

As a simulation system we have chosen a periodic rhombic dodecahedron (inner radius $R_I = 17$ Å) filled with 930 water molecules and one serine tripeptide molecule with acetylated N- and amidated C-termini. A Berendsen thermostat and barostat[42] were used to control the temperature at 300 K and the pressure at 1 atm. The molecular mechanics of the system was described by means of the all-atom force field CHARMM22.[43] Toroidal boundary conditions were applied to the computation of the electrostatics. As described in detail in refs 44 and 45 they comprise a moving-boundary reaction field description for electrostatic interactions beyond a distance of about R_I and fast hierarchical multipole expansions combined with a multiple time step integrator[46] at smaller distances. The basic integration time step was $\Delta t =$ 1 fs. By periodically saving the peptide configuration the sampling time step was set to $\Delta\tilde{t} = 1$ ps.

Acknowledgment. This work was supported by the Deutsche Forschungsgemeinschaft (SFB 533/C1).

Supporting Information Available: Stationary distributions $\mathbf{p}_{\text{stat}}$ (Figure 1), temporal evolution of the virtual density (Figure 2), and relaxation time scales (Figure 3). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Hamilton, J. D. *Time series analysis;* Princeton University Press: Princeton, 1994.

(2) Bloomfield, P. *Fourier analysis of time series;* Wiley: New York, 2000.

(3) Percival, D. B.; Walden, A. T. *Wavelet methods for time series analysis;* Cambridge University Press: Cambridge, 2000.

(4) Bradley, E. Analysis of time series. In *Intelligent data analysis*; Berthold, M., Hand, D. J., Eds.; Springer: Berlin, 2003, 199−227.

(5) Rabiner, L. R. *Proc. IEEE* **1989**, *77*, 257−286.

(6) Coast, D. A.; Stern, R. M.; Cano, G. G.; Briller, S. A. *IEEE Trans. Biomed. Eng.* **1990**, *37*, 826−836.

(7) Ephraim, Y.; Merhav, N. *IEEE Trans. Inform. Theory* **2002**, *48*, 1518−1569.

(8) Gardiner, C. W. *Handbook of stochastic methods;* Springer: Berlin, 1990.

(9) Taylor, P. Statistical methods. In *Intelligent data analysis*; Berthold, M., Hand, D. J., Eds.; Springer: Berlin, 2003; pp 69−129.

(10) Branden, C.; Tooze, J. *Introduction to protein structure*; Garland: New York, 1999.

(11) Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids;* Clarendon Press: Oxford, 1987.

(12) van Gunsteren, W. F.; Berendsen, H. J. C. *Angew. Chem., Int. Ed. Engl.* **1990**, *29*, 992−1023.

(13) Grubmüller, H.; Tavan, P. *J. Chem. Phys.* **1994**, *101*, 5047−5057.

(14) Giuliani, A.; Manetti, C. *Phys. Rev. E* **1996**, *53*, 6336–6340.

(15) Becker, O. M.; Karplus, M. *J. Chem. Phys.* **1997**, *106*, 1495–1517.

(16) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Angew. Chem., Int. Ed.* **1999**, *38*, 236–240.

(17) Huisinga, W.; Best, C.; Roitzsch, R.; Schütte, C.; Cordes, F. *J. Comput. Chem.* **1999**, *20*, 1760–1774.

(18) Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. *Linear Algebra Appl.* **2000**, *315*, 39–59.

(19) Hamprecht, F. A.; Peter, C.; Daura, X.; Thiel, W.; van Gunsteren, W. F. *J. Chem. Phys.* **2001**, *114*, 2079–2089.

(20) Schütte, C.; Huisinga, W.; Deuflhard, P. Transfer operator approach to conformational dynamics in biomolecular systems. In *Ergodic theory, analysis, and efficient simulation of dynamical systems*; Fiedler, B., Ed.; Springer: Berlin, 2001; pp 191–224.

(21) Carstens, H.; Renner, C.; Milbradt, A.; Moroder, L.; Tavan, P. *Biochemistry* **2005**, *44*, 4829–4840.

(22) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598–1603.

(23) Eckmann, J. P.; Kamphorst, S. O.; Ruelle, D. *Europhys. Lett.* **1987**, *4*, 973–977.

(24) Dersch, D. R.; Tavan, P. Load balanced vector quantization. In *Proc. ICANN'94, Sorrento*; Moreno, M., Morasso, P. G., Eds.; Springer: London, 1994.

(25) Kloppenburg, M.; Tavan, P. *Phys. Rev. E* **1997**, *55*, 2089–2092.

(26) Albrecht, S.; Busch, J.; Kloppenburg, M.; Metze, F.; Tavan, P. *Neural Networks* **2000**, *13*, 1075–1093.

(27) Hillermeier, C.; Kunstmann, N.; Rabus, B.; Tavan, P. *Biol. Cybern.* **1994**, *72*, 103–117.

(28) Kohonen, T. *Biol. Cybern.* **1982**, *43*, 59–69.

(29) Kohonen, T. *Biol. Cybern.* **1982**, *44*, 135–140.

(30) Dersch, D. R.; Tavan, P. *IEEE Trans. Neural Networks* **1995**, *6*, 230–236.

(31) Haken, H. *Synergetics;* Springer: Berlin, 1983.

(32) Dellnitz, M.; Hohmann, A.; Junge, O.; Rumpf, M. *CHAOS: Interdiscip. J. Nonlinear Sci.* **1997**, *7*, 221–228.

(33) Duda, R. O.; Hart, P. E. *Pattern classification and scene analysis;* Wiley: New York, 1973.

(34) The optimization of $\tilde{p}(\mathbf{x}|\mathscr{W};\sigma)$ can be interpreted as a self-organizing learning process of forward connections within a two-layer generalized radial basis functions (GRBF) network.[26,47]

(35) Voronoi, G. F. *J. Reine Angew. Math.* **1908**, *134*, 198–287.

(36) The matrix $\mathbf{C}$ can be interpreted as a self-organizing matrix of lateral connections in the central layer of a recurrent three-layer GRBF network (cf. ref 27).

(37) In the following the parameter τ denotes the time of a probability dynamics. It should not be confused with the time steps t of the analyzed time series.

(38) In the theory of neural networks the centers $\mathbf{w}_r$ of the Gaussians in (2) are also called the *virtual positions* of the associated mapping neurons.[48]

(39) The $\mathbf{y}_\kappa^n$ can be interpreted as the output of the neural network addressed in ref 36.

(40) Jammalamadaka, S. R.; Gupta, A. S. *Topics in circular statistics;* World Scientific: Singapore, 2001.

(41) Hu, H.; Elstner, M.; Hermans, J. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 451–463.

(42) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(43) MacKerell, A. D. et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(44) Mathias, G.; Tavan, P. *J. Chem. Phys.* **2004**, *120*, 4393–4403.

(45) Mathias, G.; Egwolf, B.; Nonella, M.; Tavan, P. *J. Chem. Phys.* **2003**, *118*, 10847–10860.

(46) Eichinger, M.; Grubmüller, H.; Heller, H.; Tavan, P. *J. Comput. Chem.* **1997**, *18*, 1729–1749.

(47) Bishop, C. *Neural networks for pattern recognition*; Clarendon Press: Oxford, 1997.

(48) Tavan, P.; Grubmüller, H.; Kühnel, H. *Biol. Cybern.* **1990**, *64*, 95–105.

CT050020X