Macroscopic Dynamics of Complex Metastable Systems: Theory, Algorithms, and Application to B-DNA *

Illia Horenko[1] Evelyn Dittmer[1] Filip Lankas[2]
John Maddocks[2] Phillip Metzner[1] Christof Schütte[1,†]

April 27, 2005

[1] *Institute of Mathematics II, Free University Berlin, Arnimallee 2–6, 14195 Berlin, Germany*

[2] *Mathematics Institute B, Ecole Polytechnique Federale de Lausanne CH-1015, Switzerland*

Abstract

This article is a survey of the present state of the transfer operator approach to the effective dynamics of metastable complex systems, and the variety of algorithms associated with it. We emphasize both conceptual foundations, and concrete application to the conformation dynamics of a biomolecular system. The algorithmic aspects are illustrated by means of several examples of various degrees of complexity, culminating in their application to a full-scale molecular dynamics simulation of a B-DNA oligomer.

keywords: Metastable states; Biomolecular conformations; Hidden Markov model; Maximum likelihood principle; Free energy; Stochastic differential equations; Molecular dynamics; B-DNA; Inter base pair parameters

1 Introduction

With the increasing availability of ever more powerful computational resources there is a current interest in performing long numerical simulations of large nonlinear dynamical systems, for example biomolecules, and in examining rather detailed properties of the results. For example there is a current effort to understand the sequence-dependent physical properties of B-form DNA via the construction and analysis of a self-consistent data base of thirty nine compatible simulations, each of a 15 base pair fragment or oligomer, with the oligomers constructed in such a way that each of the 136 possible independent tetramer

*Supported by the DFG research center "Mathematics for key technologies" (FZT 86) in Berlin, and the Swiss National Science Foundation.

† *Correspondence to:* Ch. Schütte; E-mail: schuette@math.fu-berlin.de or I. Horenko; E-mail: horenko@math.fu-berlin.de

sequences is present at least twice [5]. The time series generated in this particular project comprise more than half a terabyte of data. It is accordingly evident that there is an ever increasing need to analyze such time series efficiently, with mathematical algorithms that are practical for data sets of this order of magnitude. In particular many nonlinear dynamical systems, including biomolecules and specifically DNA, exhibit the phenomenon of *metastability*, i.e., the trajectory is localized in one sub-region of phase space for comparatively long time scales, before undergoing a rapid and rare transition to another region, where it then stays for a comparatively long residency time before eventually undergoing another rapid transition, and so on. Figure 1 illustrates this phenomenon via a plot of a single, scalar dependent variable, in this case a certain torsional angle in one of the DNA backbones between two particular base pairs in a simulation of a poly(AT) oligomer. The time series is of length 10^5, being a sampling of a 100 nanosecond simulation at every picosecond. The time series is evidently multi-well and exhibits metastability with essentially instantaneous sharp transitions between wells that are separated by 180^o or so, with rapid oscillations within each well during the long residency times.

The purpose of this article is to survey existing, and introduce new, methods for the identification of the metastable substates that are exhibited in a particular time series, and to estimate the transition probabilities between these sets. The primary messages of the article are the following. First, that the notion of the number of metastable states is a hierarchial concept–the appropriate number of metastable sets to be identified in a given time series depends upon the phenomena to be modelled. Second and nevertheless, the numbers of metastable states are not arbitrary; rather appropriate choices of the numbers of metastable sets can be identified via various clustering techniques. For example if the dimension of the time series is not too large a certain transfer operator can be explicitly computed and the appropriate possible numbers of metastable states can be associated with gaps in the spectrum of the operator via a Perron cluster analysis, while the metastable sets themselves can be identified from the associated eigenfunctions. Third, the usual methods for the computation of the transfer operator suffer from the *curse of dimensionality* which means that the methods are not practicable for large systems. However when the dimensionality of the time series is too high one may be able to instead project the time series onto a small dimension subspace via a technique such as Proper Orthogonal Decomposition (or POD), and then use Hidden Markov Models (HMM) or the new method of HMM Stochastic Differential Equations (HMMSDE) to identify metastable substates and to make good estimates of the associated transition probabilities.

The theory developed in the article is illustrated with two examples. First there is an entirely tutorial and two-dimensional example involving the high friction or overdamped Brownian dynamics of a particle in a multi-well potential, in which all the conclusions are entirely explicit. Second, the theory is applied to the DNA simulation already mentioned above. That series is of length 10^5 and is of high dimension (for details see next section). In this context the metastability analysis plays an important role in identifying basins within which the base pair level, structural shape and stiffness parameters of DNA can be approximated. The DNA example lies within the class of problems that are too large for an explicit computation of the transfer operator for the full system. However we demonstrate that the essential dynamics of the system are encapsulated in a

sufficiently low dimensional system such that the metastable sets can be captured via a HMMSDE analyis. We apply the analysis to two descriptions of the system. In the first the coordinates are back-bone angles. In these coordinates the transitions are very rapid and the states could also be identified via a more standard HMM model. On the other hand the coordinates are periodic so that the usual assumed Gaussian distributions must be replaced by the analogous but periodic von Mises distribution. The structural parameters of DNA are more traditionally described using an inter base pair description, and our analysis reveals that in these coordinates the transitions between meta-stable sets are not sharp. Rather the base pair variables are slaved to the backbone parameters. Nevertheless the HMMSDE approach applied to the base pair parameters can identify metastable states and the transition rates between them, as well as the parameters of a stochastic model for the relaxation dynamics within each metastable well.

2 Molecular Dynamics of DNA segments

2.1 Dynamics and Statistics

In classical molecular dynamics atoms are described as mass points subject to forces that are generated by specified classical interaction potentials V. The dynamical behavior is described by a deterministic Hamiltonian system of the form

$$\dot{q} \;=\; M^{-1}\xi, \qquad \dot{\xi} \;=\; -\nabla_q V(q), \tag{1}$$

defined on the state space $\mathbf{X} = \mathbf{R}^{3N} \times \mathbf{R}^{3N}$ with M denoting the diagonal mass matrix. Eq. (1) models an energetically closed system, whose total energy, given by the Hamiltonian

$$H(q,\xi) \;=\; \frac{1}{2}\,\xi^T M^{-1}\xi \;+\; V(q), \tag{2}$$

is preserved under the dynamics.

It is well known that for every smooth function $\mathcal{F} : \mathbf{R} \to \mathbf{R}$ the probability measure $\mu(\mathrm{d}x) \propto \mathcal{F}(H)(x)\mathrm{d}x$ is invariant wrt. the Markov process X_t given by the solution of the Hamiltonian system (1). The most frequent choice is the canonical density or *canonical ensemble*

$$f(x) \;\propto\; \exp(-\beta H(x))$$

for some constant $\beta > 0$ that can be interpreted as inverse temperature. The associated measure $\mu(\mathrm{d}x) \propto f(x)\mathrm{d}x$ is called the *canonical measure*. The canonical ensemble is often used in modeling experiments on molecular systems that are performed under the conditions of constant volume and temperature $\mathcal{T} = \frac{1}{k_B\beta}$, where k_B is Boltzmann's constant. Obviously, a single solution of the Hamiltonian system (1) can never be ergodic wrt. the canonical measure, since it conserves the internal energy H, as defined in (2). One traditional aspect of molecular dynamics is the construction of (stochastic) dynamical systems that allow sampling of the canonical ensemble by means of long-term simulation. Several approaches have been discussed, most of them reducing to the construction of a Hamiltonian system in some slightly extended state space $\hat{\mathbf{X}}$, whose

projection onto the lower dimensional state space $\mathbf{X}$ of positions and momenta generates a sampling according to (2.1). One of the most prominent examples is the Nosé-Hoover thermostat [7].

2.2 100 ns timeseries of $GT(AT)_6C$ DNA

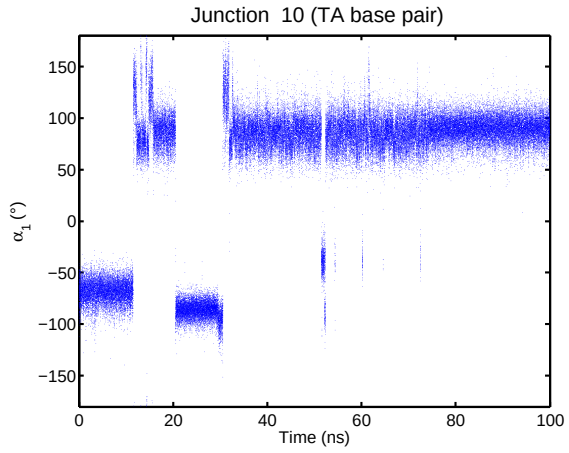

Figure 1: Timeseries of the 1st strand α-torsion angle for the junction 10. The dynamics exhibits sharp transitions between the metastable sets.

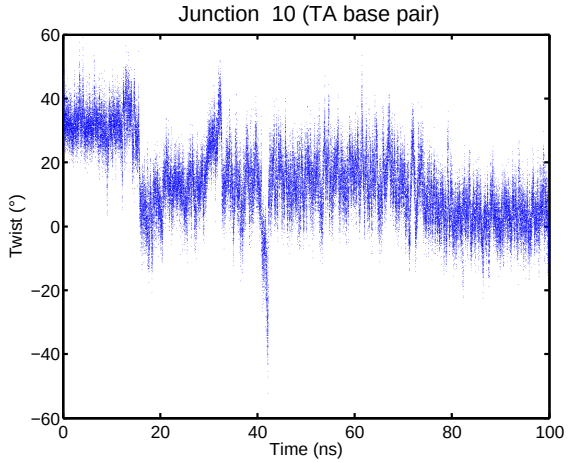

Figure 2: Time series of the inter base pair basepair coordinate for the junction 10. The dynamics exhibits slow relaxational transitions (in a ns region) between the metastable sets.

Our primary objective is to analyse time series arising from biomolecular simulations. In particular our largest data set was generated via a Molecular Dynamics (MD) simulation of a 15 base pair fragment, or oligomer, using the Amber package [9]. The detailed protocol was that of the ABC project as described in detail in [5]. That simulation provided a 100 ns time series of the oligomer with the sequence $GT(AT)_6C$ with explicit water and counter-

ions. The MD delivers a time series of the cartesian coordinates of all atoms (about 23.000 atoms, including solvent). The MD trajectory was sampled every picosecond to obtain a series of length 10^5. The variables in the time series that we work with are the physically motivated projections onto two different sets of coarse-grained internal coordinates: either the torsion angle series of the backbone data [36] or the inter base pair step parameters [20]. In either case the dimension of the time series is 84 arising from the six degrees of freedom at each of the fourteen junctions between fifteen base pairs. At each sampling time the coarse grain variables were extracted from the full set of Cartesian coordinates following standard conventions. Dependent upon the projection chosen, two basic temporal patterns of dynamics were found: abrupt, almost instantaneous change of the backbone torsion angles (see Fig. 1), but slow relaxations of the inter base pair parameters with a relaxtion time on the order of 1-2 ns (cf. Fig. 2).

Both the backbone angle and base pair parameter descriptions of the DNA oligomer take standard, sequence-independent values on the idealized B-form Watson-Crick double helix. One of the motivations for the development of the time-series analysis developed here, is to extract and understand deviations from these standard values both as a function of sequence and as a function of time.

3 Metastability

The evolution of a single microscopic system is assumed to be given by a *homogeneous Markov process* $X_t = \{X_t\}_{t\in\mathbf{T}}$ in either continuous or discrete time. We write $X_0 \sim \mu$, if the Markov process X_t is initially distributed according to the probability measure μ. The motion of X_t is given in terms of the *stochastic transition function*

$$p(t, x, A) \;=\; \mathbf{P}[X_{t+s} \in A \,|X_s = x], \tag{3}$$

for every $t, s \in \mathbf{T}$, $x \in \mathbf{X}$ and $A \subset \mathbf{X}$ that satisfies the well-known Chapman-Kolmogoroff equation $p(t+s, x, A) = \int_{\mathbf{X}} p(t, x, \mathrm{d}z)\, p(s, z, A)$ [15].

We say that the Markov process X_t admits an *invariant probability measure* μ, or μ is invariant wrt. X_t, if $\int_{\mathbf{X}} p(t, x, A)\mu(\mathrm{d}x) \;=\; \mu(A)$. In the following we always assume that the invariant measure of the process under investigation exists and is unique. A Markov process is called *reversible* wrt. an invariant probability measure μ if $\int_A p(t, x, B)\mu(\mathrm{d}x) \;=\; \int_B p(t, x, A)\mu(\mathrm{d}x)$ for every $t \in \mathbf{T}$ and $A, B \subset \mathbf{X}$.

3.1 Transition probabilities and Transfer Operators

Metastability of some subset of the state space is characterized by the property that the dynamical system is likely to remain within the subset for a long period of time, until it exits and a transition to some other region of the state space occurs. There are in fact several related but different definitions of metastability in literature (see, e.g., [8, 11, 40, 41]); we will focus on the so-called ensemble concept introduced in (4), for a comparison with, e.g., the exit time concept, see [39].

The objective is an identification of a *decomposition of the state space into metastable subsets* and the corresponding "flipping dynamics" between these

sub-states. In general, a *decomposition* $\mathbf{D} = \{D_1, \ldots, D_m\}$ of the state space $\mathbf{X}$ is a collection of subsets $D_k \subset \mathbf{X}$ with the properties: (1) positivity $\mu(D_k) > 0$ for every k, (2) disjointness up to null sets, and (3) the covering property $\cup_{k=1}^{m} \overline{D_k} = \mathbf{X}$. In particular the appropriate number m of metastable subsets must be identified. Within a transfer operator approach this can be achieved via spectral analysis (see *key idea* on page 6).

We define the *transition probability* $p(t, B, C)$ from $B \subset \mathbf{X}$ to $C \subset \mathbf{X}$ within the time span t as the conditional probability

$$p(t, B, C) \;=\; \mathbf{P}_\mu[X_t \in C \;|X_0 \in B] \;=\; \frac{\mathbf{P}_\mu[X_t \in C \text{ and } X_0 \in B]}{\mathbf{P}_\mu[X_0 \in B]}, \qquad (4)$$

where $\mathbf{P}_\mu$ indicates that initially $X_0 \sim \mu$. Then (4) may be rewritten as

$$p(t, B, C) \;=\; \frac{1}{\mu(B)} \int_B p(t, x, C)\, \mu(\mathrm{d}x). \qquad (5)$$

In other words, the transition probability quantifies the dynamical fluctuations within the stationary ensemble μ. Concomitant with our ensemble dynamics approach to metastability, we call a subset $B \subset \mathbf{X}$ *metastable* on the time scale $\tau > 0$ if

$$p(\tau, B, B^c) \approx 0, \quad \text{or equivalently,} \quad p(\tau, B, B) \;\approx\; 1,$$

where $B^c = \mathbf{X} \setminus B$ denotes the complement of B.

Transfer Operator. We define the *semigroup of propagators* or forward transfer operators $P^t : L^r(\mu) \to L^r(\mu)$ with $t \in \mathbf{T}$ and $1 \le r < \infty$ as follows:

$$\int_A P^t v(y)\; \mu(\mathrm{d}y) \;=\; \int_{\mathbf{X}} v(x) p(t, x, A) \mu(\mathrm{d}x) \qquad (6)$$

for $A \subset \mathbf{X}$. As a consequence of the invariance of μ, the characteristic function $\mathbf{1}_\mathbf{X}$ of the entire state space is an invariant density of P^t, i.e., $P^t \mathbf{1}_\mathbf{X} = \mathbf{1}_\mathbf{X}$. Furthermore, P^t is a Markov operator, i.e., P^t conserves both norm $\|P^t v\|_1 = \|v\|_1$ and positivity $P^t v \ge 0$ if $v \ge 0$, which is a simple consequence of the definition. Due to (6), the semigroup of propagators mathematically models the evolution of sub–ensembles in time.

The *key idea of the transfer operator approach* wrt. the identification of metastable decompositions can be described as follows:

> Metastable subsets can be detected via eigenvalues of the propagator P close to its maximal eigenvalue $\lambda = 1$; moreover they can be identified by exploiting the corresponding eigenfunctions. In doing so, the number of metastable subsets is equal to the number of eigenvalues close to 1, including $\lambda = 1$ and counting multiplicity.

This strategy was first proposed by Dellnitz and Junge [12] for discrete dynamical systems with weak random perturbations, and has been successfully applied to molecular dynamics in different contexts [37, 38, 39]. Its justification is given below. The key idea requires the following two *conditions on the propagator* P:

- (C1) The essential spectral radius of P is less than one, i.e., $r_{\mathrm{ess}}(P) < 1$.

- (C2) The eigenvalue $\lambda = 1$ of P is simple and dominant, i.e., $\eta \in \sigma(P)$ with $|\eta| = 1$ implies $\eta = 1$.

In this article, two types of Markov process will be considered: (1) high-friction Langevin processes, and (2) (Nose-Hoover) constant temperature molecular dynamics. For both cases the dynamics is reversible and the transfer operator is self-adjoint. For type (1) examples, conditions (C1) and (C2) are known to be satisfied under rather weak condition on the potential [39]. For type (2) examples, it is unknown whether or not the conditions are satisfied; however, it normally is assumed in molecular dynamics that they are valid for realistically complex systems in solution.

We define the *metastability of a decomposition* $\mathbf{D}$ as the sum of the metastabilities of its subsets. That is, suppose that the time scale τ of interest is fixed. Then, for each arbitrary decomposition $\mathcal{D}_m = \{A_1, \ldots, A_m\}$ of the state space $\mathbf{X}$ into m sets we define its metastability measure by

$$\mathrm{meta}(\mathbf{D}_m) = \sum_{j=1}^{m} p(\tau, A_j, A_j)/m.$$

For given m the optimal metastable decomposition into m sets can then be defined as that decomposition into m sets which maximizes the functional meta.

The next result [28] justifies the above key idea:

Theorem 3.1 *Let $P^\tau : L^2(\mu) \to L^2(\mu)$ denote a reversible propagator satisfying (C1) and (C2). Then P^τ is self–adjoint with spectrum of the form*

$$\sigma(P^\tau) \subset [a, b] \cup \{\lambda_m\} \cup \ldots \cup \{\lambda_2\} \cup \{1\}$$

with $-1 < a \leq b < \lambda_m \leq \ldots \leq \lambda_1 = 1$ and λ_i isolated, eigenvalues that are counted according to their finite multiplicities. Denote by $v_m, \ldots, v_1$ the corresponding eigenfunctions, normalized to $\|v_k\|_2 = 1$. Let Q be the orthogonal projection of $L^2(\mu)$ onto $\mathrm{span}\{\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_m}\}$. Then the metastability of an arbitrary decomposition $\mathbf{D}_m = \{A_1, \ldots, A_m\}$ of the state space $\mathbf{X}$ can be bounded from above by

$$p(\tau, A_1, A_1) + \ldots + p(\tau, A_m, A_m) \leq 1 + \lambda_2 + \ldots + \lambda_m,$$

while it is bounded from below according to

$$1 + \kappa_2\lambda_2 + \ldots + \kappa_m\lambda_m + c \leq p(\tau, A_1, A_1) + \ldots + p(\tau, A_m, A_m),$$

where $\kappa_j = \|Qv_j\|^2_{L^2(\mu)}$ and $c = a\ ((1 - \kappa_2) + \ldots + (1 - \kappa_n))$.

Theorem 3.1 highlights the strong relation between a decomposition of the state space into metastable subsets and a *Perron cluster* of dominant eigenvalues close to 1. It states that the metastability of an arbitrary decomposition $\mathbf{D}_m$ cannot be larger than $1+\lambda_2+\ldots+\lambda_m$, while it is at least $1+\kappa_2\lambda_2+\ldots+\kappa_m\lambda_m+c$, which is close to the upper bound whenever the dominant eigenfunctions $v_2, \ldots, v_m$ are almost constant on the metastable subsets $A_1, \ldots, A_m$ implying $\kappa_j \approx 1$ and $c \approx 0$. The term c can be interpreted as a correction that is small whenever $a \approx 0$ or $\kappa_j \approx 1$. It is demonstrated in [28] that the lower and upper bounds are sharp and asymptotically exact.

3.2 Metastability Analysis is Hierarchical

The last theorem and the illustrations and examples below contain one main message about metastability analysis: *it has to be hierarchical*. Whenever we approximate the optimal metastable decomposition $\mathbf{D}_2$ of state space into, say, two sets, we should always be aware that there could be a decomposition $\mathbf{D}_3$ into three sets for which meta($\mathbf{D}_3$) is almost as large as meta($\mathbf{D}_2$). For example, one or both of the two subsets in $\mathbf{D}_2$ could decompose into two or several metastable subsets from which exit is comparably difficult for the system under investigation.

However, whenever there is a gap in the spectrum of the transfer operator after m dominant eigenvalues, then the results of, e.g., [27, 8] tell us that any decomposition into more than m sets will be associated with a significantly larger drop in metastability as measured by the function meta. In the context of applications to molecular dynamics, however, one should always be aware that particular aspects of interest may make it desirable to explore the hierarchy of metastable decompositions up to a certain depth that is not necessarily selected only on the values of the functional meta.

3.3 Discretization and PCCA

Let $\chi = \{\chi_1, \ldots, \chi_n\} \subset L^2(\mu)$ denote a set of *non–negative* functions that are a partition of unity, i.e., $\sum_{k=1}^n \chi_k = \mathbf{1}_\mathbf{X}$. The *Galerkin projection* $\Pi_n : L^2(\mu) \to \mathcal{S}_n$ onto the associated finite dimensional ansatz space $\mathcal{S}_n = \text{span}\{\chi_1, \ldots, \chi_n\}$ is defined by

$$\Pi_n v \;=\; \sum_{k=1}^{n} \frac{\langle v, \chi_k \rangle_\mu}{\langle \chi_k, \chi_k \rangle_\mu} \, \chi_k.$$

Application of the Galerkin projection to $P^\tau v = \lambda v$ yields an eigenvalue problem for the discretized propagator $\Pi_n P^\tau \Pi_n$ acting on the finite-dimensional space $\mathcal{S}_n$. The matrix representation of this finite dimensional operator is given by the $n \times n$ *transition matrix* $\mathcal{T} = (\mathcal{T}_{kl})$, whose entries are given by

$$\mathcal{T}_{kl} \;=\; \frac{\langle P^\tau \chi_k, \chi_l \rangle_\mu}{\langle \chi_k, \chi_k \rangle_\mu}. \tag{7}$$

The transition matrix inherits the main properties of the transfer operator: it is a stochastic matrix with invariant measure given by the invariant measure μ of P^τ, it is reversible if P^τ is self-adjoint, and (if the discretization is fine enough) it also exhibits a Perron cluster of eigenvalues that approximates the corresponding Perron cluster of P^τ, and with eigenvectors that approximate the dominant eigenvectors of P^τ [39]. It thus allows to compute the metastable sets of interest by computation of the dominant eigenvectors of $\mathcal{T}$ and by realization of the identification strategy of page 6 based on these (discrete) eigenvectors. This has led to the construction of an aggregation technique called "Perron Cluster Cluster Analysis" (PCCA) [13, 14].

The entries of $\mathcal{T}$ can be computed from realizations of the underlying Markov process X_t. We have

$$\mathcal{T}_{kl} \;=\; \frac{1}{\langle \chi_k, \chi_k \rangle_\mu} \int_\mathbf{X} \chi_k(x) \mathbf{E}_x[\chi_l(X_\tau)] \, \mu(\mathrm{d}x).$$

If $x_0, \ldots, x_N$ denote a time series obtained from a realization of the Markov process with time stepping τ, then the entries of $\mathcal{T}$ can be approximated from the relative transition rates computed by means of this time series:

$$\mathcal{T}_{kl} \;\approx\; \mathcal{T}_{kl}^{(N)} \;=\; \frac{\sum_{j=1}^{N} \chi_k(x_j) \cdot \chi_l(x_{j+1})}{\sum_{j=1}^{N} \chi_k(x_j)^2}. \tag{8}$$

For a time series of whatever length, but with a high dimensional configuration variable, practical evaluation of the formula (8) may become problematic. There are two main reasons for potential difficulties.

Trapping problem. The *rate of convergence* of $\mathcal{T}_{kl}^{(N)} \to \mathcal{T}_{kl}$ depends on the smoothness of the partition functions χ_k as well as on the mixing properties of the Markov process [30]. The latter property is crucial here: The convergence is geometric with a rate constant $\lambda_1 - \lambda_2 = 1 - \lambda_2$ where λ_2 denotes the second largest eigenvalue (in modulus). In the case of metastability with λ_2 being very close to $\lambda_1 = 1$, we will have dramatically slow convergence. This is of no surprise because closeness of dominant eigenvalues is typically the main difficulty in all approaches to biomolecular dynamics and statistics, and it is also a bottleneck of the transfer operator approach. Much of the literature aims to tackle this *trapping problem* [4, 17]. In our largest examples evaluation of (8) is not practical. However we will *not* go into the depth of the discussion on overcoming the trapping problem, because we propose alternative approaches. We will simply assume in all of the following that we have already generated or can directly generate a time series that is "long enough" in the sense that it contains statistically significant information about more than one –if not all– interesting metastable states of the system under consideration. We will discuss later whether this is the case for our poly(AT) DNA time series.

Curse of Dimension. Any discretization of the transfer operator will suffer from the curse of dimension whenever it is based on a uniform partition of all of the hundreds or thousands of degrees of freedom in a typical biomolecular system. Fortunately, chemical observations reveal that—even for larger biomolecules—the curse of dimensionality can be circumvented by exploiting the hierarchical structure of the dynamical and statistical properties of biomolecular systems: only relatively few *essential degrees of freedom* are needed to describe the conformational transitions (see next section); furthermore, the canonical density has a rich spatial multiscale structure induced by the rich structure of the potential energy landscape. This structure induces a hierarchical cluster structure of the sampling data that can be identified and used to define a multilevel discretization adapted to the structures of the statistical data (see [39] or subsequent examples).

3.4 Illustrative Example

For simplicity we consider the Markov process given by so-called high-friction Langevin equation which is the limit of high friction of the famous Langevin equation, see [35, 38]. The high-friction Langevin equation is stated in the position space only and is given by the equation

$$\dot{x} = -\nabla_x V(x) + \sigma \dot{W}_t, \tag{9}$$

with $x(t) \in \mathbf{R}^d$ being the position vector of the system, W_t denoting d-dimensional standard Brownian motion, and σ the noise intensity parameter. The stochastic differential equation (9) defines a continuous time Markov process X_t with invariant probability measure $\mu(\mathrm{d}x) \propto \exp(-\beta V(x))\mathrm{d}x$ with $\beta = 2/\sigma^2$ [35]. There is a long history of using it as a simple toolkit for investigation of dynamical behavior in complicated energy landscapes [10]. It is known that under weak conditions on the potential function V the Markov process is reversible [25].

The associated semigroup (P^t) of propagators admits a strong generator $\mathcal{A}$ such that the semigroup can be written as $P^t = \exp(t\mathcal{A})$, respectively. For twice continuously differentiable $u \in L^2(\mu)$ we have the identity

$$\mathcal{A}\, u \;=\; \left(\frac{\sigma^2}{2}\Delta_x - \nabla_x V(x) \cdot \nabla_x\right) u.$$

For details on $\mathcal{A}$ see the theory of Fokker-Planck equations and Kolmogoroff forward and backward equations [35, 40]. Under appropriate conditions (the Perron cluster is a discrete part of the spectrum) we can compute the dominant eigenvectors of P^t via those of $\mathcal{A}$.

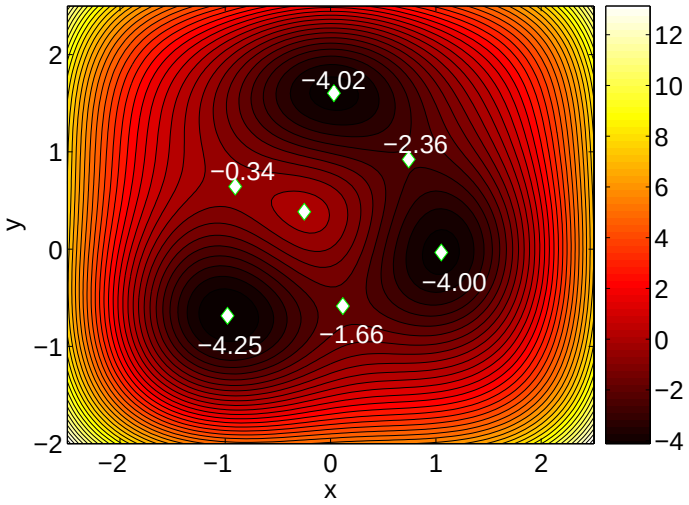

Figure 3: Potential V used for illustrative example. We observe three wells in the potential landscape (see colorbar). The tags indicate the minima and saddle points of the potential; the numbers give the value of the potential at these points. We observe that the leftmost minimum is the deepest well separated by the most pronounced energy barrier from the two other ones.

For illustrative means we use the potential V illustrated in Fig. 3 (thus setting $d = 2$). Fig. 4 shows typical realizations of the high friction Langevin Markov process associated with this potential (setting $\sigma = 0.131$). We observe that the vicinity of the wells in the potential energy landscape can approximately be identified with the metastable sets of the process; it is well-known from large deviation theory that in fact, for small enough noise intensity, the vicinity of the wells of the potential energy landscape are the metastable sets of high-friction Langevin processes (at least such wells that are separated from each other by significant energy barriers) [27].

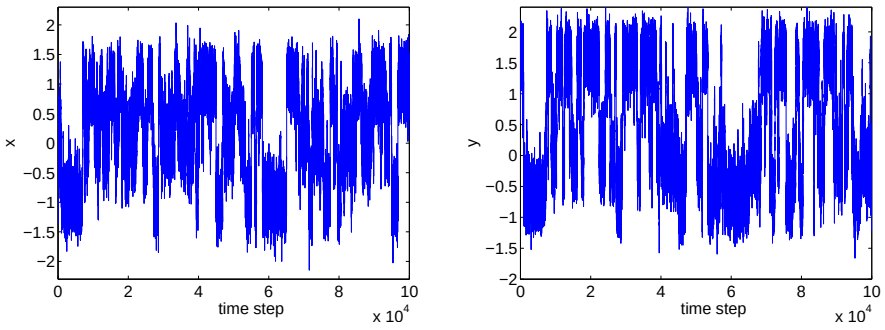

Figure 4: Typical realization of the high-friction Langevin dynamics (both components (left/right) of the state versus time) in the potential energy landscape V shown in Fig 3 for $\sigma = 0.131$.

Next, we discretized the transfer operator of the process (fine grid with 100×100 discretization boxes in discretization domain $[-3, 3] \times [-3, 3]$) for different values of τ which results in the dominant eigenvalues listed in Table 1.

$\sigma(P^\tau)$	λ_1	λ_2	λ_3	λ_4
$\tau = 0.01$	1.000	0.999	0.997	0.959
$\tau = 0.10$	1.000	0.994	0.975	0.656
$\tau = 1.00$	1.000	0.937	0.776	0.015

Table 1: Leading four eigenvalues of transfer operator P^τ for different values of τ for high-friction Langevin motion with potential and parameters as described in the text.

While the eigenvector of the largest eigenvalue is constant, the corresponding second and third eigenvector of P^τ in $L^2(\mu)$ are shown in Figs. 5 (they are identical for all values of τ because of the semigroup property).

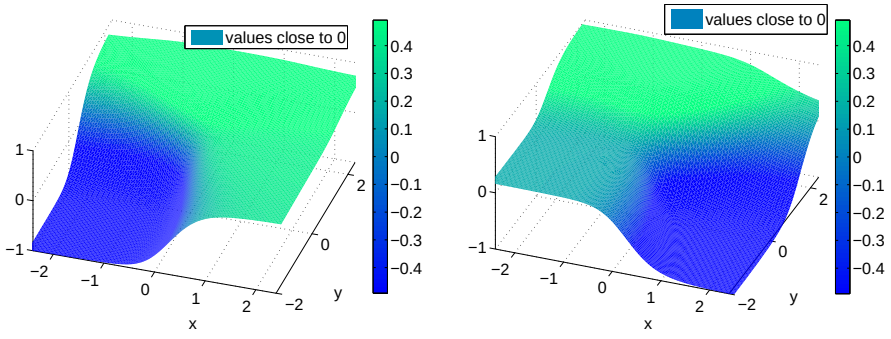

Figure 5: Second and third eigenvectors of the transfer operator for high-friction Langevin process in potential of Fig. 3 (details see text).

Having computed the dominant eigenvectors we can determine the optimal metastable decomposition by means of PCCA as introduced above. The results on the spectrum (see $\tau = 0.1$ for example) exhibit a hierarchy of metastability that is in perfect agreement with the general insight on metastability of high friction Langevin motion: We can apply PCCA to the first *two* eigenvectors of

the transfer operator; this results in the metastable decomposition that distinguishes between the vicinity of the deepest well and the remaining state space (see Fig. 6, left). When applying PCCA to the first *three* eigenvectors, however, the resulting metastable decomposition identifies the vicinities of all three well as the metastable regions of the system (see Fig. 6, right). This outcome is desirable and typical: metastable decomposition via spectral properties of the transfer operator are hierarchical in the sense that the process of including more and more leading eigenvalues uncovers finer and finer details of metastability within the system, see [39, 27].

So, what happens if we take the first four eigenvectors? This we can immediately understand by comparing the values of the functional meta for the optimal metastable decompositions $\mathbf{D}_m$ into $m = 1, 2, 3, 4$ sets ($\tau = 1$) as given in Table 2: Between $m = 3$ and $m = 4$ there is a significant drop in metastability indicating that it makes no real sense to speak of four metastable sets for the system under consideration.

m	1	2	3	4
$\mathrm{meta}(\mathbf{D}_m)$	1.000	0.967	0.899	0.613
$\frac{1}{k}\sum_{k=1}^{m}\lambda_k$	1.000	0.969	0.904	0.682

Table 2: Metastabilities of the optimal metastable decomposition $\mathbf{D}_m$ into $m = 1, 2, 3, 4$ sets (as computed by PCCA from the dominant eigenvectors) and its theoretical upper bound as of Theorem 3.1.

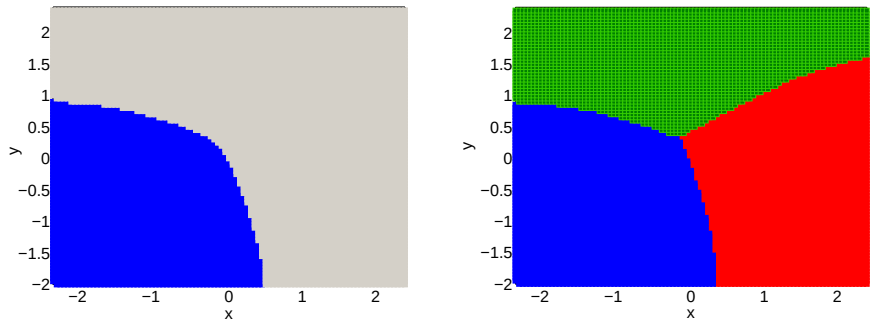

Figure 6: Optimal metastable decomposition resulting from PCCA based on the first two (left) and first three (right) eigenvectors of the transfer operator.

4 Algorithms

As already mentioned we assume herein that some "long enough" time series $(x_t)_{t=t_0,\ldots,t_N}$ of states of the system under consideration is already given. We are mainly interested in the case that the states x_{t_l} are from some high-dimensional state space $\mathbf{R}^d$. In this section we will often consider time series of some observables $(O_t)_{t=t_0,\ldots,t_N}$ computed from the time series (x_t), e.g., the time series of torsion angles or inter base pair parameters. We now want to discuss the question of *dimension reduction*. In the scope of this article this means the question of how to find/design observables such that the time series (O_t) on the one hand still allows to identify the main metastable states but on the other hand is significantly lower dimensional than the original state space.

Even if this dimension reduction is successful, the resulting time series (O_t) may *not* show clearly disjoint metastable states because of *overlapping* due to dimension reduction. For an example visit Fig. 4 again: While the metastable state of our illustrative example clearly are well disjoint in full state space (given by the vicinity of the wells of the energy landscape), the two one-dimensional time series shown in Fig. 4 clearly exhibit metastability but the metastable states are now overlapping (not clearly disjoint).

In the rest of this section we will, first, discuss methods for dimension reduction and, second, methods for identifying metastability despite overlaps between metastable states.

4.1 Essential Variables and Free Energy

One of the main ideas for dimension reduction is to find the *essential degrees of freedom* or *essential variables* of the system under consideration. In the (low dimensional) subspace of essential degrees of freedom most of the positional fluctuations occur, while in the remaining degrees of freedom the motion can be considered as "physically constrained".

Based on the available time series, we may determine essential degrees of freedom either in the position space according to Amadei et al. [1] or in the space of internal degrees of freedom, e.g., dihedral angles, by statistical analysis of circular data [26]. Either case is based on proper orthogonal decomposition (POD), also known as principal component analysis (PCA), for details see next paragraph. As shown in [26] and in Sec. 5 below, this procedure may results in an enormous dimension reduction.

4.1.1 Proper Orthogonal Decomposition (POD)

Proper orthogonal decomposition (POD), also called principal component analysis (PCA), provides a method for finding a best approximating subspace $S \subset \mathbf{R}^d$ to a given set of data [22] but *without* taking into account its temporal order. However, in our case the set of data always will be given by the time series $(y(t)) = (x_t)_{t=t_0,\ldots,t_N}$ of states or some time series $(y(t)) = (O_t)_{t=t_0,\ldots,t_N}$ of observations.

We may characterize the subspace S by a projection operator Π mapping $\mathbf{R}^d$ onto S. The task is to minimize

$$I(\Pi) = \sum_{l=1}^{N} \|y(t_l) - \Pi y(t_l)\|_R^2 \,,$$

wrt. to a given Riemannian metric $\|\cdot\|_R$ with associated scalar product $\langle\cdot,\cdot\rangle_R$. One then can easily show that the optimal m-dimensional subspace in $\mathbf{R}^d$ is spanned by the dominant eigenvectors Φ_k, $k = 1,\ldots,m$, of the covariance matrix C (for definition see below) of the data (y_t). That is, the Φ_k are the eigenvectors of the largest m eigenvalues ν_k of C (counting multiplicity), i.e., $C\,\Phi_k = \nu_k\Phi_k$. The subspace S is optimal in the sense that the energy norm on S

$$\sum_{l=1}^{N} \|\Pi y(t_l)\|_R = \sum_{j=1}^{m} \nu_j$$

is the maximum achieved by any m-plane in $\mathbf{R}^d$ (cf. [29]).

For the definition of the covariance matrix let us first consider the case that (y_t) is a time series in Euclidean space (i.e., no circular observables that may introduce problems with periodicity). Then,

$$C \;=\; \sum_{l=1}^{N} (y(t_l) - \bar{y}) \otimes (y(t_l) - \bar{y}), \tag{10}$$

where $\otimes$ denotes the tensor product wrt. $\langle\cdot,\cdot\rangle_R$, and $\bar{y}$ the mean value of y along the time series. Fig. 7 shows the spectrum of the covariance matrix of the time series of inter base pair parameters of the DNA segment introduced in Sec. 2.

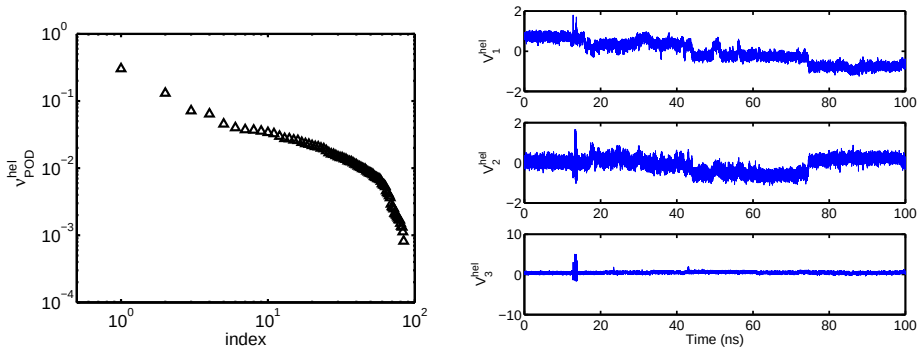

Figure 7: Spectrum of the covariance matrix (left, in double logarithmic representation) of the inter base pair parameter time series of the $GT(AT)_6C$ DNA segment and projection of the overall dynamics on first three principal components (right). We clearly observe at least two dominant POD modes ($\nu_1 = 0.30$, $\nu_2 = 0.13$, $\nu_3 = 0.07$, $\nu_4 = 0.06$).

Circular Observables. Whenever the observation time series $(y(t))$ contains circular coordinates the covariance matrix has to be redefined on the respective torus. Suppose the components of $y(t)$ are denoted $y_i(t)$ and that all y_i take values on $[0, 2\pi]$.

The covariance matrix C can be generalized as follows [18, 26]:

$$C_{kl} \;=\; \frac{r(y_k - y_l)^2 - r(y_k + y_l)^2}{\sqrt{(1 - r(y_k)^2)(1 - r(y_l)^2)}} \sqrt{4\log(r(y_k))\log(r(y_l))}, \tag{11}$$

where $r(y_k)$ is computed from the mean $\bar{e}_k = \sum_{l=1}^{N} \exp(iy_k(t_l))/N$ via

$$r(y_k)\exp(i\varphi(y_k)) \;=\; \bar{e}_k.$$

4.1.2 Free Energy

In statistical physics the concept of free energy is based on an m-dimensional manifold $\mathcal{M}$ in the configurational space of the system; this may, e.g., be the manifold spanned by the dominant POD modes. Suppose that the level sets of $c = c(q) \in \mathbf{R}^{3N-m}$ define this manifold of reaction coordinates with q denoting a configuration of the system in position space. The corresponding free energy is defined by

$$F(c') = -\frac{1}{\beta} \log Z(c'), \quad Z(c') = \int \exp(-\beta H(q,\xi))\delta(c' - c(q))\, dqd\xi.$$

Under very specific circumstances the free energy landscape is known to allow for an identification of the metastable sets of the system with its main wells [32]; however, this means that the reaction coordinates c already are the "conformational degrees of freedom" of the system, that is, they would allow for a perfect low-dimensional characterization of the metastable sets. This normally is not the case which, in turn, may make interpretation of the free energy landscape in terms of metastability and transitions misleading, or even wrong. We will observe a situation like that in the section on DNA dynamics.

Next we will explain how to approximate the free energy landscape based on given time series: Assume that we have access to a time series $(O_t)_{t=1,\dots,N}$ of the coordinates $O_t = c(q(t))$, and that the underlying time series $(q(t))_{t=1,\dots,T}$ is approximately distributed due to $\mu \propto \exp(-\beta H(q,\xi))$ in state space. Furthermore, let (O_t) be contained in $\mathcal{B} \subset \mathcal{M}$, and $(B_k)_{l=1,\dots,L}$ be a partition of $\mathcal{B}$ into "small", disjoint subsets ("bins"). The relative frequency of hits of box B_k along the time series is

$$r_{B_k} = \frac{\sum_{l=1}^{N} \chi_{B_k}(O_{t_l})}{N}.$$

Therefore, the histogramm of (O_t) wrt. (B_k) is $\sum_{k=1}^{L} r_{B_k}\chi_{B_k}$. Then, the function Z can be approximated by the histogram of (O_t) wrt. (B_k), i.e., by the step-function

$$Z(c') \approx \sum_{k=1}^{L} \frac{\sum_{l=1}^{N} \chi_{B_k}(O_{t_l})}{N}\, \chi_{B_k}(c').$$

Consequently, the approximate free energy landscape is given by

$$F(c') \approx \hat{F}(c') \;=\; -\frac{1}{\beta} \log\Big(\sum_{k=1}^{L} \frac{\sum_{l=1}^{N} \chi_{B_k}(O_{t_l})}{N}\, \chi_{B_k}(c') \Big) \tag{12}$$

There is a bunch of other (far better) approaches to the computation of the free energy landscape [42, 16, 21]. However, most of them require an entirely different approach to MD simulation such that these techniques are not applicable if some time series resulting from standard MD simulation are already given.

4.2 Algorithms based on Hidden Markov Models

Suppose the system under consideration has a metastable decomposition. Then, at any time t the system will be exactly in one of the associated metastable sets $B_q \subset \mathbf{X}$, $q = 1, \dots, m$. Therefore, at each time t there is "metastable state" $q(t)$

given by the number of the presently visited metastable set. Whenever a time series of values of observables (or "observations") $O = (O_t)_{t=t_0,\dots,t_N}$ is given, we want to identify the time series of metastable states $q = (q_t)_{t=t_0,\dots,t_N}$ associated with it. However, while the time series (O_t) is observed, i.e, known, the series (q_t) is *hidden* within the data.

Suppose that the observed data (O_t) is given with constant time stepping τ, i.e., $t_k = t_{k-1} + \tau$ for all $k = 1, \dots, N$. Setting $t_0 = 0$ we have $t_k = k\tau$ and especially $T = t_N = N\tau$. For the sake of simplicity of notation we thus may simply write $t = 0, ..., T$.

The probability to go from one metastable/hidden state q to another one q' is given by $\mathcal{T}_{qq'} = p(\tau, B_q, B_{q'})$. That is, the sequence (q_t) should be seen as the realization of a Markov chain with M states with transition matrix $\mathcal{T}$.

The observations O_t somehow result from the respective hidden state q_t by apriori unknown rules. For given time series of observations $O = (O_t)_{t=t_0,\dots,t_N}$ one is interested in finding the most probable series of metastable/hidden states.

Models like the one coarsely described above are well-known as *hidden Markov models* (HMMs). A HMM is a stochastic process with hidden and observable states; the hidden states of a HMM form Markov chain while the observable states are understood as output that is distributed according to a certain conditional distribution (conditioned to the hidden state).

To describe the whole system, we need to know the number M of hidden states, the transition matrix $\mathcal{T}$ between them, an initial distribution, and for each state a certain rule governing the probability distribution for the observation.

Stationary output. In standard HMMs the output distributions result from iid random variables, i.e., consecutive output states are statistically independent. That is, conditioned on the hidden state, the output is simply randomly chosen from a stationary distribution. In application to the analysis of data produced in the context of molecular dynamics this means that the system reaches the thermodynamical equilibrium distribution immediately after each transition between metastable states; this then is related to abrupt jumps of some physical observables.

The most popular choice for such stationary distributions are (multivariate) normal distributions. However, in the case of circular data (like torsion angle positions in a molecular dynamics simulation) the use of normal distribution often induces crucial problems due to periodicity and thus have to be replace by von Mises distributions [31].

The problem of the statistical analysis of the time series in this case will be reduced to the identification of the Markov transition matrix and equilibrium statistical distributions (often specified in a parameterized form) [3, 33, 34]. This approach was recently successfully applied to analysis of torsion angles dynamics of trialanine molecule [19].

In case of the DNA time series described above the dynamics of the torsion angle data is well-acquainted with main assumptions of the HMM–model (instantaneous switches in the hidden Markov-chain, equilibrated statistical distributions associated with each of the states, see Fig. 1), which may be explained by the small mass of the involved fragments (torsion angel characterizes the mutual positions of the four involved atoms).

SDE Output. In contrast, the relaxation behavior of the base parameters describing the dynamics of DNA-basepairs (consisting of 30 atoms) can not be consistently described in terms of canonical HMMs (see Fig. 8). Instead of assigning some predefined stationary distribution to each of the hidden states, we can use a parameterized stochastic dynamical process Y_t (e.g., given in form of a SDE) with parameters depending on the actual state of the hidden Markov–chain (metastable state) Y_t. Herein, we consider processes of the form

$$\begin{aligned} \dot{Y}_t &= -DV^{(q)}(Y_t) + \sigma^{(q)}\dot{W}, & (13) \\ q(t) &: \mathbf{R}^1 \to \{1, 2, ..., k\}, & (14) \end{aligned}$$

where $q(t)$ are the realizations of the hidden Markov chain, W is standard "white noise", and $\{V^{(q)}, \sigma^{(q)}\}$ is a set of the state-specific model parameters with harmonic potential $V^{(q)}$ of the form

$$V^{(q)}(Y) = \frac{1}{2}(Y - \mu^{(q)})^T D^{(q)}(Y - \mu^{(q)}) + V_0^{(q)}, \quad (15)$$

where $\mu^{(q)}$ and $D^{(q)}$ are *equilibrium position* and *Hesse-matrix* of the conformation q, respectively. Each of the SDEs has a unique invariant Gaussian distribution proportional to

$$\exp\Big(- (Y - \mu^{(q)})^T D^{(q)}(Y - \mu^{(q)})/{\sigma^{(q)}}^2 \Big). \quad (16)$$

In contrast to the standard HMM approach, such stochastic dynamical systems do not reach a new equilibrium distribution immediately after each jump of the Markov chain but relaxes into the new equilibrium state and reaches the invariant distribution ρ^i only after some characteristic *relaxation time* which can help to estimate the unknown parameters of the SDEs (14). This idea was recently exploited in the context of model reduction for complex systems exhibiting metastable behavior [23]. We will often abbreviate "HMMs with SDE output" by HmmSde .

Optimal Parametrization. Both types of HMMs, whether with stationary output distribution or with SDE output, contain sets of parameters (the entries of the transition matrix $\mathcal{T}$, the initial output distribution v, as well as the parameters of SDEs or output distributions, respectively), herein denoted by θ. We now want to identify optimal parameters for *given* observation data $O = (O_t)_{t=1,...,T}$. We have to define the likelihood functional with respect to which we then will have to determine the optimal parameters θ and the sequence of hidden states $q = (q_t)_{t=1,...,T}$.

For given parameters θ, the likelihood $\mathcal{L}(\theta|O_t, q_t)$ has to be the probability of output O_t under the condition of being in metastable state q_{t_j} for given parameters θ:

$$\mathcal{L}(\theta|O_t, q_t) = p(O, q|\theta) = v(q_0)\rho(O_0|q_0)\prod_{t=1}^{T}\mathcal{T}(q_{t-1}, q_t)\rho(O_t|q_t, O_{t-1}), \quad (17)$$

where $\rho(\cdot|q, O_{t-1})$ denotes the output distribution at time t under the condition that the system is in hidden state q_t. In the case of SDE output this distribution also depends on the previous output state O_{t-1} which is not the case for

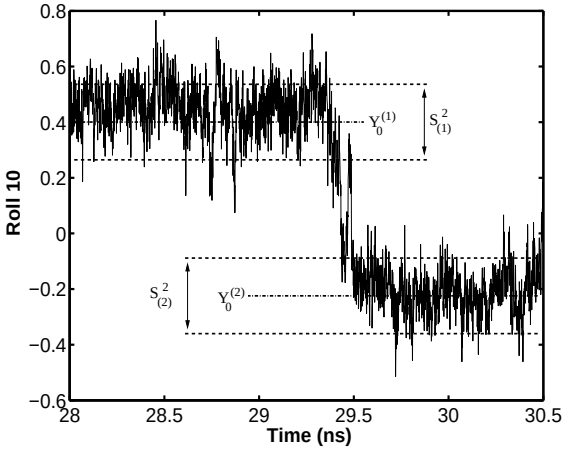

Figure 8: Fragment of one of the inter base pair parameter time series showing relaxation behavior. Statistics of the time series in each of the states and relation (16) can be used to calculate $\sigma_{(i)} = \sqrt{2D_{(i)}}S_{(i)}$.

stationary output distributions. Stationary output means that the formula for the output distribution is explicitly given; as demonstrated in detail in [23, 24] there also is an explicit formula for the output distribution $\rho(\cdot|q, O_{t-1})$ for SDE output as long as the potentials in (14) are harmonic.

The next task now will be to construct algorithms that

(1) determine the optimal parameters $(\mathcal{T}, \mu^q, D^q, \sigma^q)_{q=1,\ldots,M}$ by maximizing the likelihood $\mathcal{L}(\theta|O, q)$; this is a nonlinear global optimization problem,

(2) determine the optimal sequence of hidden metastable states (q_t) for given optimal parameters, and

(3) determine the number of important metastable states; up to now we also simply assumed that the number M of hidden states is a priori given - how can we determine the appropriate number?

To solve problem (1) we will use the *expectation-maximization* (EM) algorithm. The EM algorithm is a learning algorithm: it alternately iterates two steps, the Expectation step and the Maximization step. Starting with some initial parameter set θ_0 the steps iteratively refine the parameter set, i.e., in step k the present parameter set θ_k is refined to θ_{k+1}. We will work out the details of the EM algorithm for the problem under investigation by following the general framework given in [6]:

The key object of the EM algorithm is the expectation

$$Q(\theta, \theta_k) \;=\; \mathbf{E}\Big(\log p(O, q|\theta) \;|\, O, \theta_k \;\Big) \tag{18}$$

of the complete-data likelihood $\mathcal{L}(\theta|O, q) = p(O, q|\theta)$ (in our case given by (17)) wrt. the hidden sequence q given the observation sequence and the current parameter estimate θ_k. One step of the EM algorithm then realizes the following two steps:

- Expectation-step: This step evaluates the expectation value Q based on the given parameter estimate θ_k.
- Maximization-step: This step determines the refined parameter set θ_{k+1} by maximizing the expectation:

$$\theta_{k+1} = \operatorname*{argmax}_{\theta} Q(\theta, \theta_k). \tag{19}$$

The maximization guarantees that $\mathcal{L}(\theta_{k+1}) \geq \mathcal{L}(\theta_k)$.

Algorithmic realizations of these two steps are standard for stationary Gaussian and von Mises output [19]; for SDE output the algorithmic details can be found in [23, 24]. In both cases the necessary computational effort for one step of the EM algorithm scales linearly in the length of the observation sequence and quadratically in the number of hidden states.

Optimal sequence of hidden states. Based on the results of the EM algorithm, Problem (2) from above can be solved by applying the standard Viterbi algorithm [43]. For given θ and O this algorithm computes the most probable hidden path $q^* = (q_0^*, \ldots, q_T^*)$. This path is called the *Viterbi path*. For an efficient computation we define the highest probability along a single path, for the first t observations, ending in the hidden state S_i at time t,

$$\delta_t(i) = \max_{q_0, q_1, \ldots q_{t-1}} P(q_0, q_1 \ldots q_t = S_i, O_0, O_1 \ldots O_t | \theta).$$

This quantity is given by induction as

$$\delta_t(j) = \max_{1 \leq i \leq M} [\delta_{t-1}(i) \mathcal{T}_{ij}] \rho(O_t | q_t, O_{t-1}). \tag{20}$$

In addition, the argument i that maximizes (20) is stored in ψ in order to actually retrieve the hidden state sequence. These quantities are calculated for each t and j, and then the Viterbi path will be given by the sequence of the arguments in ψ obtained from backtracking. For more details see [19].

Number of metastable states. In the setup of all HMM techniques for a given observation sequence one is confronted with the task to select *in advance* the number M of hidden states. There are no general solutions to this problem, and the best way to handle this problem often is a mixture of insight and preliminary analysis. However, since our goal is to identify metastable states we can proceed as suggested in [19]: Start the EM algorithm with some sufficient number of hidden states, say M, that should be greater than the expected number of metastable states. After termination of the EM algorithm, take the resulting transition matrix A and aggregate the M hidden states into $\mathcal{M} \leq M$ metastable states by means of PCCA. The resulting conformation states will then allow an interpretation of the results in terms of metastable states.

In all numerical experiments in the following the initial parameter guesses are based on the same procedure: The initial $M \times M$ transition matrix is chosen to be a stochastic matrix with offdiagonal entries 0.001 and identical diagonal entries. The initial values of the model parameters were obtained by the respective re-estimation formulas of the EM algorithm based on randomized determination of the probabilities $P(O_t | q_t, O_{t-1})$ (they were chosen uniformly distributed on $[0, 1]$).

Combination of results from different projections. Assume that we already applied HMMSDE or some HMM-based method to several low-dimensional observation time series of the system under consideration, but to each one independently. Suppose that the different time series simply are resulting from different projections of the full time series in state space; for example, think of the different time series given by each single torsion angle of system, or of the time series given by each single of the leading POD modes. In this situation one may be interested in combining the hidden states from each of the single projections into "higher dimensional" metastable states of the system. This can be done by analyzing the Viterbi paths derived from the single low-dimensional observation time series: Suppose we are concerned with J low-dimensional time series and therefore J Viterbi paths. The J Viterbi paths can be understood as a J-dimensional discrete time series. Every state of this time series lies in the discrete state space consisting of all possible combinations of the metastable states of the single low-dimensional time series. We obviously can take this time series, compute its transfer matrix by counting transitions between its discrete states, determine the dominant eigenmodes of this transfer matrix, and again apply PCCA to identify metastable decompositions of the discrete state space. The sets in such a metastable decomposition have to be interpreted as aggregates of the metastable states from the low-dimensional time series where the aggregation is done based on additional insight coming from the combination of all of the low-dimensional information. This concept leads to the following algorithm:

1. Determine model parameters and Viterbi paths for each low-dimensional observation time series;

2. combine the Viterbi paths and compute the transfer matrix in the discrete state space of combined metastable states;

3. determine metastable decompositions via PCCA.

4.3 Illustrative Example Revisited

We now assume a time series $(x(t))_{t=t_0,\ldots,t_N}$ with $t_k - t_{k-1} = \tau = 0.01$ and $N = 10^5$ being given of the test system introduced in Sec. 3.4. For given $t = t_0, ..., t_N$ let $x(t) \in \mathbf{R}^2$ be the full state of the system.

For this choice of τ the transfer operator $P^\tau = \exp(\tau\mathcal{A})$ of the high-friction Langevin motion considered in Sec. 3.4 has the following dominant eigenvalues

$$\sigma(P^t) = \{1.000, 0.999, 0.997, 0.959, \ldots\}.$$

Let us consider the two observation time series $(O_t^{(j)})_{t=t_0,\ldots,t_N}$, $j = 1, 2$, with $O_t^{(j)} = x_j(t)$ (the first and second components of the state of system).

We first apply HMMSDE to observation time series $(O_t^{(1)})$ (see Fig. 9 for illustration) and set $M = 3$. Eleven iterations of the EM algorithm result in the following transition matrix

$$\mathcal{T} = \begin{pmatrix} 0.9983 & 0.0013 & 0.0004 \\ 0.0017 & 0.9983 & 0.0000 \\ 0.0008 & 0.0000 & 0.9992 \end{pmatrix}.$$

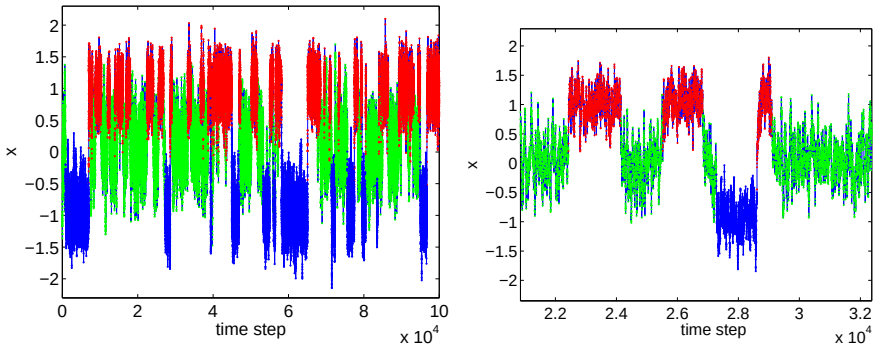

Figure 9: Observation time series ($O_t^{(1)}$). Left: Entire time axis. Right: Magnification clearly exhibiting metastability and overlapping. Color/grey scale due to Viterbi path (see text below).

that has the spectrum

$$\sigma(\mathcal{T}) = \{1.000, 0.999, 0.997\},$$

which perfectly agrees with the results of the transfer operator approach (that is based on the full two-dimensional information instead of on the reduced observation time series). The HmmSde results for the parameters of the potential and the noise intensities are given in the table below and are in very good agreement with the results to be expected.

parameter	$j=1$	$j=2$	$j=3$
$\mu^{(j)}$	0.0552	1.0169	-0.9584
$\sigma^{(j)^2}$	0.1325	0.1321	0.1302
$D^{(j)}$	0.5589	1.0507	0.9324

Table 3: Parameters of HmmSde for training with ($O_t^{(1)}$).

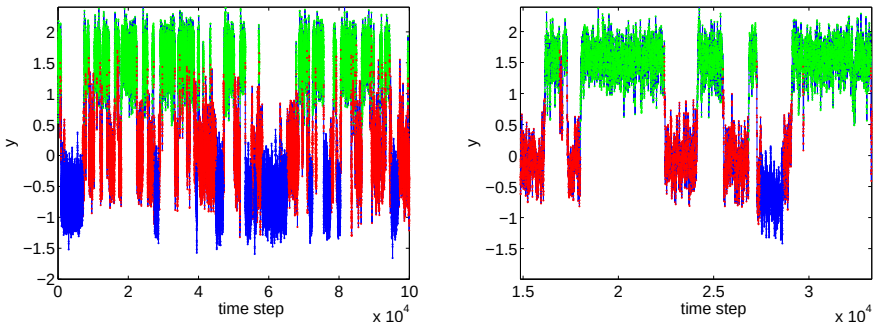

Figure 10: Observation time series ($O_t^{(2)}$). Left: Entire time axis. Right: Magnification clearly exhibiting metastability and overlapping. Color/grey scale due to Viterbi path (see text below).

Next we apply HmmSde to observation time series ($O_t^{(2)}$) (see Fig. 9) and set $M = 3$. Nine iterations of the EM algorithm result in the following transition

matrix

$$\mathcal{T} = \begin{pmatrix} 0.9987 & 0.0013 & 0.0000 \\ 0.0014 & 0.9981 & 0.0005 \\ 0.0000 & 0.0007 & 0.9993 \end{pmatrix}.$$

with spectrum

$$\sigma(P^t) = \{1.000, 0.999, 0.997\}.$$

The HMMSDE results now are again in good agreement with the results to be

parameter	$j = 1$	$j = 2$	$j = 3$
$\mu^{(j)}$	1.5526	-0.0084	-0.6693
$\sigma^{(j)^2}$	0.1318	0.1347	0.1343
$D^{(j)}$	1.0607	0.5018	1.1037

Table 4: Parameters of HMMSDE for training with $(O_t^{(2)})$.

expected.

Next we compute the Viterbi paths for the two HMMSDE results based on $(O_t^{(1)})$ and $(O_t^{(2)})$, respectively. This yields the assignment to metastable states as illustrated in Figs. 9 and 10, and in a two-dimensional representation in Fig. 11. We observe that the agreement of the assignment with the metastable states resulting from the transfer operator approach (see Fig. 5) is good. However, as it can be seen from the picture, the assignment of the points in the transition regions gets ambiguous. The algorithm for combining the results of our two different projections (as of page 20) yields the results shown in Fig. 12, where all points which are not clearly assigned to any of the metastable states are identified as belonging to some "transition state".

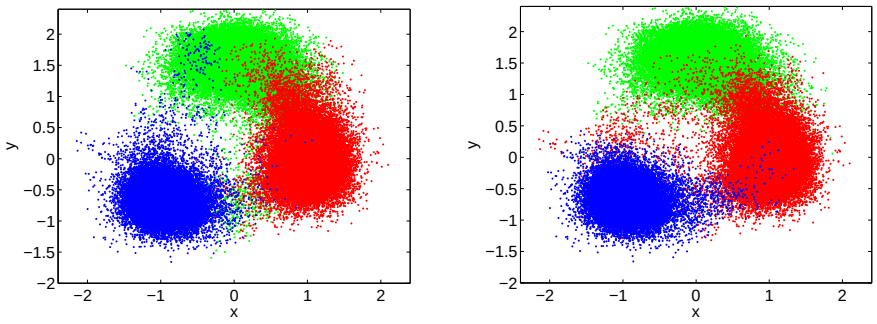

Figure 11: Visualization of the assignment of states to the three metastable states as resulting from the Viterbi paths computed via HMMSDE based on $(O_t^{(1)})$ (left) and $(O_t^{(2)})$ (right).

5 Numerical Results

In this section we will apply the techniques of Sec. 4 to different time series that result from the 100 ns MD simulation of $GT(AT)_6C$ DNA described in Sec. 2.2. We will particularly consider two differnet types of time series: the times series of the backbone angles $(O_t^{(1)})_{t=1,\ldots,100000}$ with $O_t^{(1)} \in \mathbf{R}^{84}$, and the time series

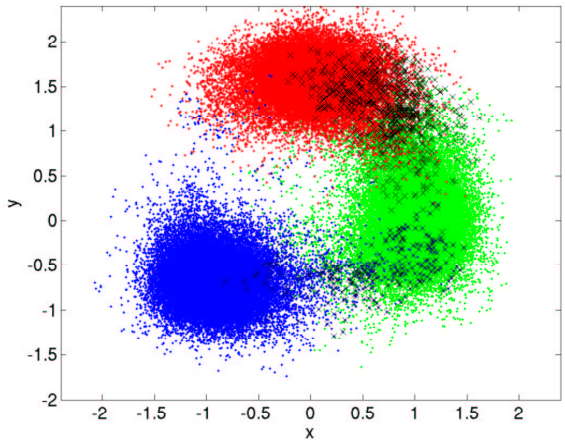

Figure 12: Visualization of the assignment of states to the three metastable states (points of three different grey tones) and transition states (black crosses) as resulting from the clustering of both one-dimensional Viterbi paths computed according to the transfer operator approach.

of the inter base pair parameters $(O_t^{(2)})_{t=1,\ldots,100000}$ with $O_t^{(2)} \in \mathbf{R}^{84}$. After a detailed separate analysis of these families of data we will finally compare the outcome.

5.1 Analysis of Backbone Torsion Angle Dynamics

Our first step is to apply POD to the time series $(O_t^{(1)})$. Fig. 13 shows the results of the POD approach when using the Fisher covariance matrix. We immediately observe that two eigenvalues dominate the spectrum, that some others may have significant contributions, and that by far most of the eigenvalues exhibit insignificant contributions (to the variance of the data).

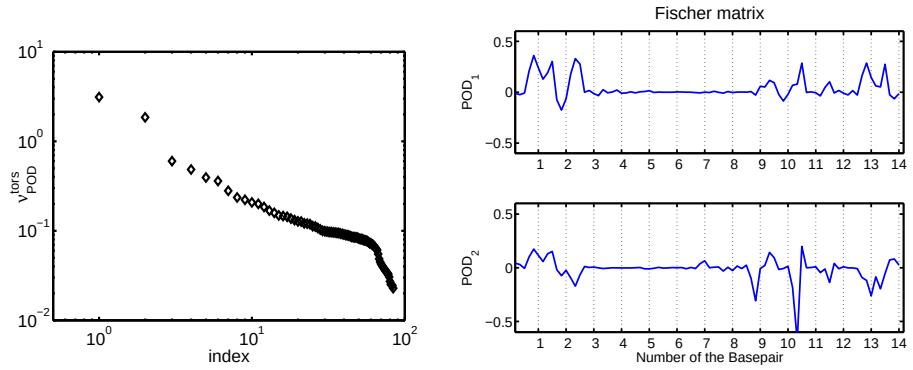

Figure 13: Spectrum of the torsion angles covariance matrix due to (11) (dominant eigenvalues $\nu_1 = 3.12, \nu_2 = 1.85, \nu_3 = 0.60, \nu_4 = 0.48$)(left) and two dominant eignvectors as functions of basepair torsion angles (right).

Suppose that $u_j \in \mathbf{R}^{84}$, $j = 1, \ldots, 4$, denote the associated four leading POD eigenmodes, and that the u_j are normalized, i.e., the Euclidean scalar product of u_j with itself is $u_j^T \cdot u_j = 1$. We now can project the time series $(O_t^{(1)})$ onto

the first POD components, i.e., we can compute the time series

$$O_t^{(1,j)} = u_j^T \cdot O_t^{(1)}, \quad t = 1, \ldots, 100000.$$

The first four time series are shown in Fig. 18 below, and obviously exhibit metastability. But before going into details we will analyze the time series $O_t^{(1,1)}$ associated with the first POD mode. This time series is shown in Fig. 14 together with the approximate free energy landscape. We observe that this free energy landscape exhibits at least six wells.

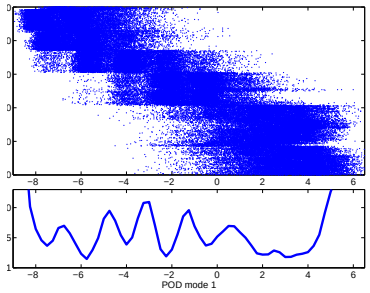

Figure 14: Projection $O_t^{(1,1)}$ of the overall DNA dynamics onto the first POD mode (top), and the approximate free energy in terms of the first POD mode as computed from time series $O_t^{(1,1)}$ (bottom).

We therefore first apply HmmSde to the time series $O_t^{(1,1)}$ with $M = 8$ hidden states (we could have chosen a larger number but this would make the following illustrations much more difficult). The results are shown in Fig. 15. We observe that the 8 different harmonic potentials identified by HmmSde approximately coincide with the wells of the free energy landscape with one exception in the region around POD mode $1 = -1.8$. However, the associated Viterbi path exhibits frequent flips between distinguished pairs of hidden states; it seems to show that the metastable behavior should be expressed in terms of four states only.

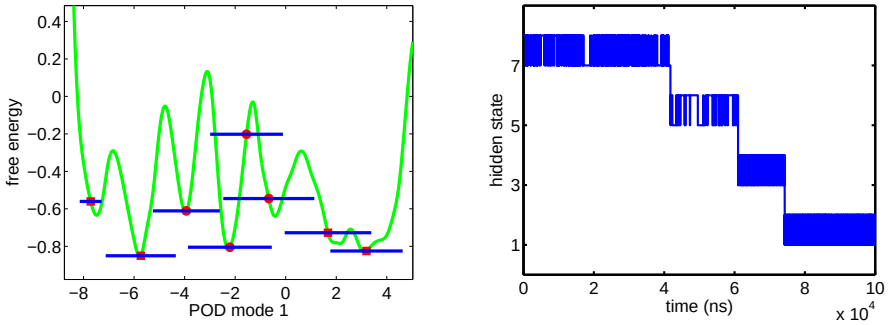

Figure 15: Results of application of HmmSde to time series $O_t^{(1,1)}$ with $M = 8$ hidden states. Left: Associated free energy landscape with additional markers indicating the location of equilibria $\mu^{(q)}$ of the harmonic potentials resulting from HmmSde and bars indicating the widths $s_q = \sigma^{(q)}/\sqrt{2D^{(q)}}$ of the invariant measures of the associated SDEs. Right: Viterbi path in terms of the 8 hidden state.

In fact, when analyzing the associated HmmSde transition matrix we find four clearly dominant eigenvalues, see Table 5, and some other eigenvalues that seem to correspond to not-so-dominant metastable features.

k	1	2	3	4	5	6	7	8
λ_k	1.00000	.99997	.99994	.99992	.99361	.98290	.92211	.12896

Table 5: Eigenvalues of the HmmSde transition matrix for $M = 8$ hidden states as computed from $O_t^{(1,1)}$.

The relation between the dominant and not-so-dominant features is exemplified in Fig. 16 which shows a fragment of the time series $O_t^{(1,1)}$ between times $t = 59.5$ ns and $t = 62.3$ ns and the HmmSde assignment of states to the hidden states within this fragment. We observe that HmmSde in fact identifies metastable phases correctly, but that the time span of stability of the states associated with the not-so-dominant features is of the order of 0.5 ns, while the dominant features are stable for several nanoseconds.

In order to judge the robustness of our observations we now can apply HmmSde to the time series $O_t^{(1,1)}$ with less states, e.g., with $M = 7$ hidden states. This results in a transition matrix with eigenvalues shown in table 6, and the Viterbi path shown in Fig. 17 .

k	1	2	3	4	5	6	7
λ_k	1.00000	.99997	.99994	.99992	.98170	.95462	.77414

Table 6: Eigenvalues of the HmmSde transition matrix for $M = 7$ hidden states as computed from $O_t^{(1,1)}$.

We observe that this again results in four dominant metastable states (which due to the two Viterbi paths are identical for $M = 7$ and $M = 8$). The only difference is that two (state 1 and 2) of the former 8 states obviously have been aggregated into one (now state 1).

These results seems to indicate that we should apply HmmSde with $M = 4$ hidden states. The result is shown in Fig. 18.

This figure also illustrates the outcome of HmmSde if applied to each of the four time series $O_t^{(1,j)}$, $j = 1, \ldots, 4$. In each case we chose $M = 4$. After computing the Viterbi path for each of the four time series (see Fig. 18, left), we cluster these paths according to the algorithm introduced in Sec. 4.2, page 20. This results in $m = 4$ metastable states; the resulting global Viterbi path is illustrated in Fig. 19.

Exploiting the time series of torsion angles we also can easily compute the approximate free energy landscape in terms of the first two POD modes. The result is illustrated in Fig. 20 together with a sketch of the Viterbi path projected onto the first two POD modes.

5.2 Analysis of Inter Base Pair Parameter Dynamics

Our second step is to apply POD to the time series $(O_t^{(2)})$. Fig. 21 shows the results of the POD approach. Again two eigenvalues dominate the spectrum, some others may have significant contributions, and by far most of the eigenvalues exhibit insignificant contributions (to the variance of the data). Suppose

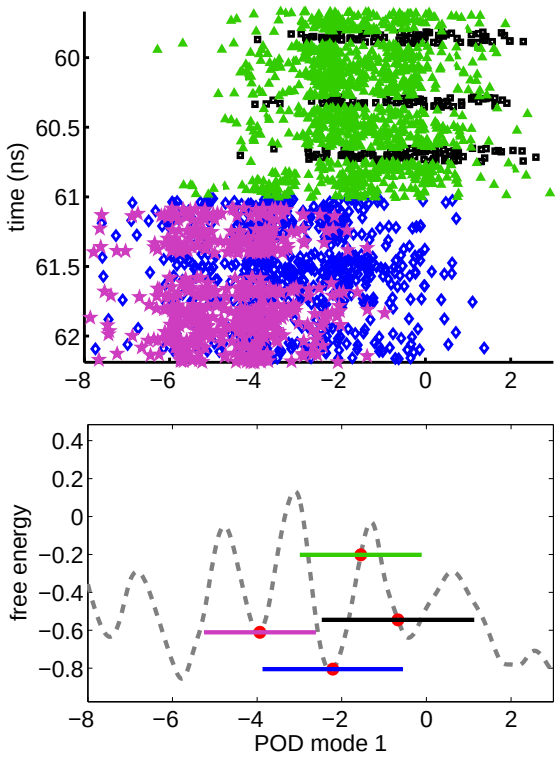

Figure 16: Results of application of HmmSde to time series $O_t^{(1,1)}$ with $M = 8$ hidden states. Top: Fragment of the time series $O_t^{(1,1)}$ between times $t = 59.5$ ns and $t = 62.3$ ns, and coloring of states according to the assignment to hidden states due to HmmSde Viterbi path; four hidden states are visible indicated by green/light grey triangles, black squares, blue/dark grey diamonds, and magenta/grey stars. Bottom: Associated free energy landscape with additional markers indicating the location of equilibria $\mu^{(q)}$ of the harmonic potentials and bars indicating the widths s_k of the associated invariant measures for the four hidden states visible in the fragment.

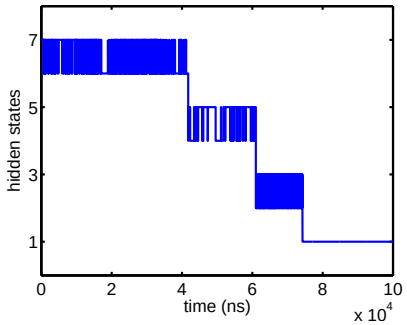

Figure 17: Viterbi path resulting from application of HmmSde to time series $O_t^{(1,1)}$ with $M = 8$ hidden states.

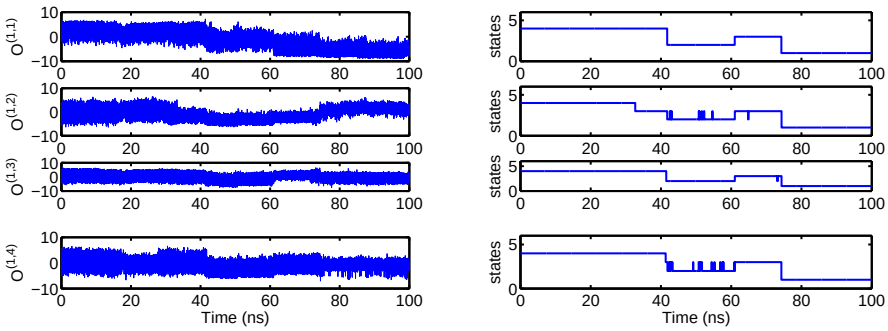

Figure 18: Projection of the overall DNA dynamics on first four POD modes (left) and the Viterbi paths resulting from HmmSde analysis of each of the projected time series.

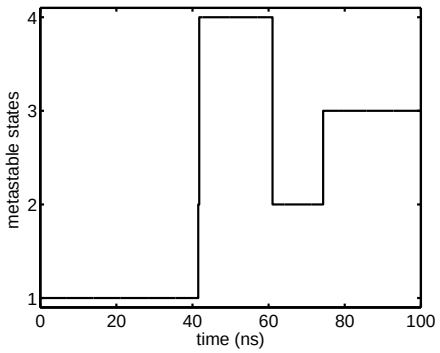

Figure 19: Viterbi-path of the dynamics in first four POD modes; 4 clusters were found

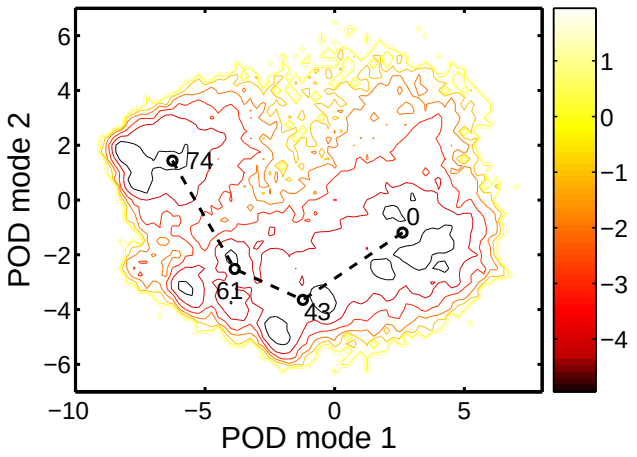

Figure 20: Free energy landscape revealed by two first POD modes of torsion data and the projection of the Viterbi-path (dashed) with arrival times in ns.

that $w_j \in \mathbf{R}^{84}$, $j = 1, 2, 3, 4$, denote the associated four leading POD normalized eigenmodes. Projection of $(O_t^{(2)})$ onto the first POD components is shown in Fig. 22.

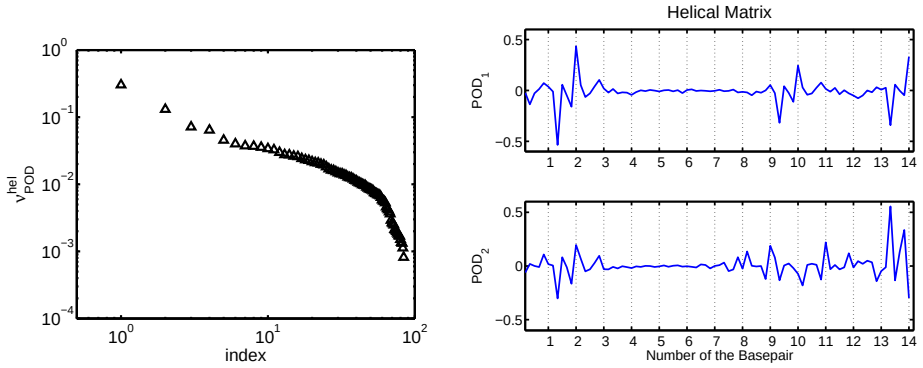

Figure 21: Spectrum of the inter base pair parameters covariance matrix (left) and eigenvectors corresponding to the two dominant POD modes ($\nu_1 = 0.30$, $\nu_2 = 0.13$, $\nu_3 = 0.07$, $\nu_4 = 0.06$)(right).

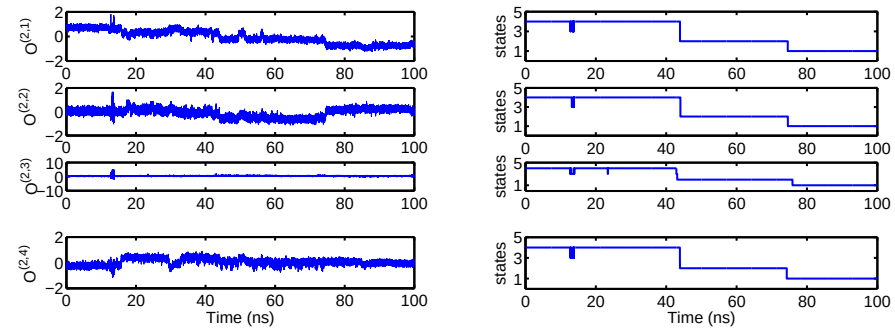

Figure 22: Projection of the overall DNA dynamics on first four POD modes (left) and the Viterbi paths resulting from HmmSde analysis of each of the projected time series.

We initially applied HmmSde to the first POD mode with $M = 8$ hidden states and observed again that the main features can be described in terms of only four dominant metastable states. Therefore, we herein apply HmmSde to each of the four time series $O_t^{(2,j)} = w_j \cdot O_t^{(2)}$, $j = 1, \ldots, 4$ with $M = 4$ in each case, and cluster the Viterbi paths (see Fig. 22) of each of the time series results in $m = 4$ metastable states; the resulting global Viterbi path exhibits 3 metastable sets and is illustrated in Fig. 23.

The approximate free energy landscape in terms of the first POD modes is shown in Fig. 24; we computed the free energy in two different subspaces: mode 2 versus mode 1 and mode 4 versus mode 1.

In order to understand the effect of both sequence and metastability on shape parameters we next consider the mean values along the time series within each metastable state (as identified by the Viterbi path cf. Fig. 23) of two dominant deformation parameters, namely twist and roll, for each of the TA and AT steps

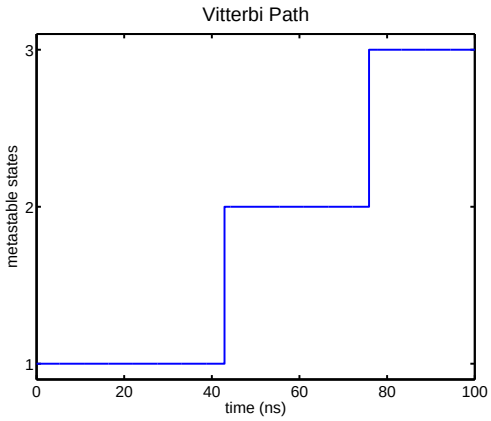

Figure 23: Viterbi-path of the dynamics in four first POD modes, 3 global clusters were found

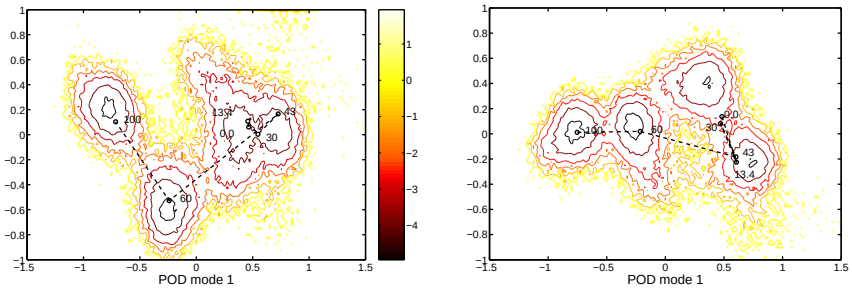

Figure 24: Free energy landscape revealed by first POD modes of inter base pair data and the projection of the Viterbi-path (dashed) with arrival times in ns. Left: POD mode 2 versus POD mode 1. Right: POD mode 4 versus mode 1.

along the oligomer. In Fig. 25 these values are superposed on the mean free-energy of the twist and roll for AT and TA steps computed via the average over all AT, respectively all TA, steps of the free energy at each step computed according to formula (12).

The first observation is that there is indeed a well-pronounced sequence dependence as the TA and AT steps do indeed have significantly different averages. Further for TA steps the mean free energy is markedly multi-well, and the mean values populating different wells are dependant upon their metastable set as identified by the Viterbi path. In contrast for AT steps, while the average values in each metastable set are still distinguishable and clustered, there is no striking multi-welled character in the mean free energy surface.

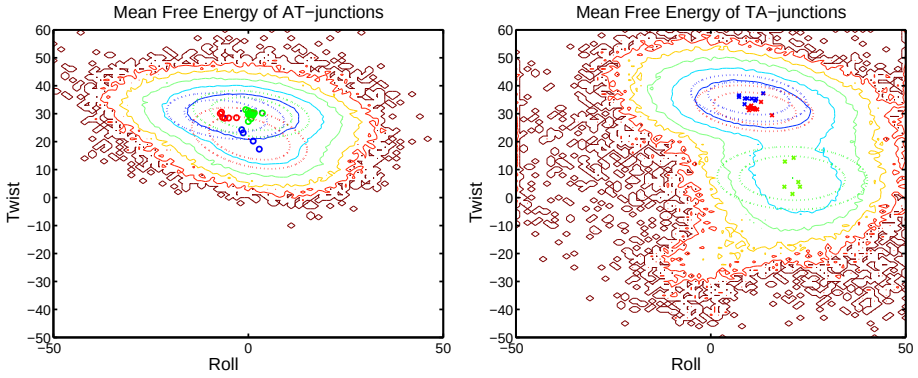

Figure 25: Plots of average values of AT and TA step twists and rolls evaluated at each step and in each metastable state, and superposed on the mean free energies over all AT, respectively TA, steps.

5.3 Comparison of Analysis based on Torsion Angles and Inter base pair Parameters

As it can bee seen from Fig. 26, despite of the difference in the number of identified metastable sets (four in the case of torsion angles and three for inter base pair parameters), there are some apparent similarities.

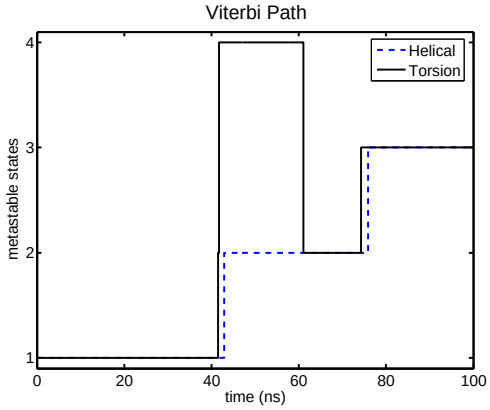

Figure 26: Comparison of inter base pair and torsional Viterbi-paths. Note the time-delays (1.5 ns) between the state changes in two representations.

First, the two switches between the states (at 42 and 75 ns) seem to happen almost simultaneously in both representations. Second, the delay times between torsional and inter base pair switches are both of around 1.5 ns. The possible explanation of these facts is that both Viterbi paths describe the same process of large-scale geometry change during the dynamics; while the inter base pair process is in some sense effected by the switch in the backbone. Fig. 27 shows the mean geometrical configurations (in terms of a string representation of the backbone) that are associated with the metastable sets of the torsion angle dynamics. The metastable states 2 and 4 (that are visited between 42 and 75 ns, see Fig. 26) of the torsion angle based analysis are geometrically close to each other (the difference is a difference in the position of the end-groups of the DNA helix). Closer inspection exhibits that they are not distinguishable in inter base pair picture.

6 Conclusion

We have presented a variety of algorithmic concepts for the identification of metastable states in dynamical systems. Possible strategies for application to very complex metastable systems, and the performance of the resulting algorithms have been demonstrated by analyzing a full-scale MD simulations of a poly-(AT) B-DNA oligomer.

In regard to the algorithmic aspects our conclusion is that for realistically large simulations of biomolecules it is not practical to compute explicitly the full discretized transfer operator in cartesian coordinates. This is due to the curse of dimensionality. Moreover it is still not practical to compute the full transfer operator after reduction to coarse grain variables (to dimension 84 in

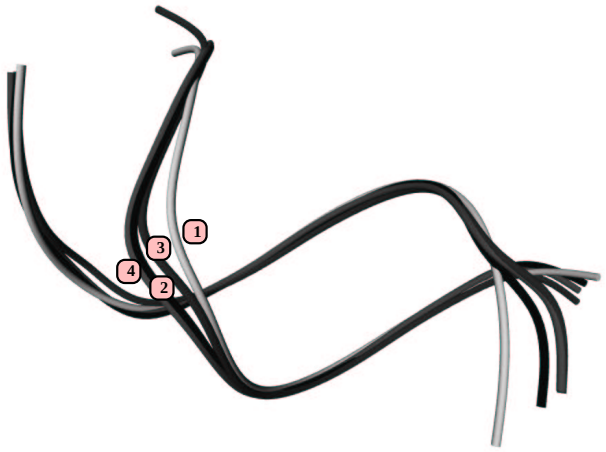

Figure 27: 3D-geometry visualized with AMIRA-software [2] of the four mean configurations defined by four metastable sets resulting from torsion angle based analysis. The 3D-geometry is illustrated by a string-like representation of the backbone. Numbers denote the indexes of metastable states as in Fig. 26

our DNA example), but now due to problems of slow convergence and overlapping metastable states in the projection. In contrast, standard HMM methods applied to a reduced time series with sharp transitions (the backbone time series in the DNA example) can identify the metastable sets. And even in a projection with less well-defined transitions (the base pair step parameter time series) the new method of HmmSde can still identify the metastable sets.

Our results show that the backbone dynamics of B-DNA exhibit metastable behavior (visible in both base pair and torsion angles representations of the dynamics) on nano- to microsecond time scales, and that this metastability might be sequence-dependent and of importance for macroscopic modelling of B-DNA elasticity and dynamics. Most specifically the average values of AT and TA base pair parameters are quantified and confirmed to be quite distinct. In addition the values of these averages are shown to depend upon the particular metastable set of the oligomer.

On a less positive note it is apparent that the simulation time scale of a few hundred nanoseconds is much too short to compute transition probabilities for the backbone accurately, i.e., to analyze quantitatively the possible sequence dependence of backbone conformation transitions. Most specifically the trajectory we have computed is demonstrably not ergodic.

References

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

[2] Amira—advanced visualization, data analysis and geometry reconstruction, user's guide and reference manual. Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB), Indeed–Visual Concepts GmbH and TGS Template Graphics Software Inc., 2000.

[3] L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3:1–8, 1972.

[4] B. J. Berne and J. E. Straub. Novel methods of sampling phase space in the simulation of biological systems. *Curr. Opinion in Struct. Biol.*, 7:181–189, 1997.

[5] D.L Beveridge, G. Barreiro, K.S Byun, D.A Case, T.E Cheatham III, S.B Dixit, E. Giudice, F. Lankas, R. Lavery, J.H. Maddocks, R. Osman, E. Seibert, H. Sklenar, G. Stoll, K.M. Thayer, P. Varnai and M.A. Young, Molecular Dynamics Simulations of the 136 Unique Tetranucleotide Sequences of DNA Oligonuleotides. I.Research Design and Results on d(CpG) Steps. Biophysical Journal, 87 (2004), 3799-3813

[6] J. A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and Hidden Markov Models. Technical report, International Computer Science Institute, Berkeley, 1998.

[7] S. D. Bond and B. B. L. Benedict J. Leimkuhler. The Nosé–Poincaré method for constant temperature molecular dynamics. *JCP*, 151(1):114–134, 1999.

[8] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein. Metastability in stochastic dynamics of disordered mean–field models. *Probab. Theor. Rel. Fields*, 119:99–161, 2001.

[9] D.A. Case, D.A. Pearlman, J.W. Caldwell, T.E. Cheatham III, J. Wang, W.S. Ross, C.L. Simmerling, T.A. Darden, K.M. Merz, R.V. Stanton, A.L. Cheng, J.J. Vincent, M. Crowley, V. Tsui, H. Gohlke, R.J. Radmer, Y. Duan, J. Pitera, I. Massova, G.L. Seibel, U.C. Singh, P.K. Weiner and P.A. Kollman AMBER 7, University of California, San Francisco, 2002.

[10] D. Chandler. Finding transition pathways: throwing ropes over rough montain passes, in the dark. In B. Berne, G. Ciccotti, and D. Coker, editors, *Classical and Quantum Dynamics in Condensed Phase Simulations*, pages 51–66. Singapure: World Scientific, 1998.

[11] E. B. Davies. Metastable states of symmetric Markov semigroups I. *Proc. London Math. Soc.*, 45(3):133–150, 1982.

[12] M. Dellnitz and O. Junge. On the approximation of complicated dynamical behavior. *SIAM J. Num. Anal.*, 36(2):491–515, 1999.

[13] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Lin. Alg. Appl.*, 315:39–59, 2000.

[14] P. Deuflhard and M. Weber. Robust Perron cluster analysis in conformation dynamics. ZIB-Report 03-19, Zuse Institute Berlin, 2003.

[15] J.L. Doob, *Stochastic Processes*. Wiley, New York, 1953

[16] W. E and E. Vanden-Eijnden. Metastability, Conformation Dynamics, and Transition Pathways in Complex Systems. *Multiscale, Modelling, and Simulation*. Part I, p. 35(34), Springer, Berlin, 2004.

[17] D. M. Ferguson, J. I. Siepmann, and D. G. Truhlar, editors. *Monte Carlo Methods in Chemical Physics*, volume 105 of *Advances in Chemical Physics*. Wiley, New York, 1999.

[18] A. Fischer, F. Cordes, and C. Schütte. Hybrid Monte Carlo with adaptive temperature in a mixed–canonical ensemble: Efficient conformational analysis of RNA. *J. Comput. Chem.*, 19:1689–1697, 1998.

[19] A. Fischer, S. Waldhausen, and C. Schütte. Identification of biomolecular conformations from incomplete torsion angle observations by Hidden Markov Models. *submitted to J. Comput. Phys.*, 2004.

[20] O. Gonzalez and J.H. Maddocks. Extracting parameters for base pair level models of DNA from molecular dynamics simulations. Theor. Chem. Acc. (2001) 106: 76-82

[21] C. Hartmann and Ch. Schütte A constrained hybrid Monte-Carlo algorithm and the problem of calculating the free energy in several variables *ZAMM*, submitted, 2005

[22] H. Heuser, J. Lumley and G. Berkooz *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge University Press, 1996.

[23] I. Horenko, E. Dittmer, A. Fischer and Ch. Schuette. Automated model reduction for complex systems. *MMS*, submitted, 2005.

[24] I. Horenko, E. Dittmer, and C. Schütte. Reduced stochastic models for complex molecular systems. *Submitted to CVS*, 2005.

[25] W. Huisinga. *Metastability of Markovian systems: A transfer operator approach in application to molecular dynamics*. PhD thesis, Free University Berlin, 2001.

[26] W. Huisinga, C. Best, R. Roitzsch, C. Schütte, and F. Cordes. From simulation data to conformational ensembles: Structure and dynamic based methods. *J. Comp. Chem.*, 20(16):1760–1774, 1999.

[27] W. Huisinga, S. Meyn, and C. Schütte. Phase transitions & metastability in Markovian and molecular systems. Submitted, 2002.

[28] W. Huisinga and B. Schmidt. Metastability and Dominant Eigenvalues of Transfer Operators, in preparation, 2002.

[29] S. Lall, J. E. Marsden and S. Glavaski. *A subspace approach to balanced truncation for model reduction of nonlinear control systems. Int.J.Robust Nonlin.Control.*, 12:519–535, 2002.

[30] P. Lezaud. Chernoff and Berry–Esséen inequalities for Markov processes. *ESIAM: P & S*, 5:183–201, 2001.

[31] K. V. Mardia. *Statistics of Directional Data*. Academic Press, New York, 1972.

[32] Y. Mu, P. H. Nguen, and G. Stock. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins*, 58:45–52, 2004.

[33] L. R. Rabiner. A tutorial on Hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.

[34] L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[35] H. Risken. *The Fokker-Planck Equation*. Springer, New York, 2nd edition, 1996.

[36] W. Saenger. Principles of Nucleic Acid Structure. Springer Verlag, New York, 1984.

[37] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys., Special Issue on Computational Biophysics*, 151:146–168, 1999.

[38] C. Schütte and W. Huisinga. On conformational dynamics induced by Langevin processes. In B. Fiedler, K. Gröger, and J. Sprekels, editors, *EQUADIFF 99 - International Conference on Differential Equations*, volume 2, pages 1247–1262, Singapore, 2000. World Scientific.

[39] C. Schütte and W. Huisinga. Biomolecular Conformations can be Identified as Metastable Sets of Molecular Dynamics. In P. G. Ciaret and J.-L. Lions, editors, *Handbook of Numerical Analysis*, volume on Computational Chemistry. North–Holland, 2003.

[40] C. Schütte, W. Huisinga, and P. Deuflhard. Transfer operator approach to conformational dynamics in biomolecular systems. In B. Fiedler, editor, *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems*, pages 191–223. Springer, 2001.

[41] G. Singleton. Asymptotically exact estimates for metatstable Markov semigroups. *Quart. J. Math. Oxford*, 35(2):321–329, 1984.

[42] M. Sprik and G. Ciccotti Free Energy from Constrained Molecular Dynamics J. Chem. Phys., 109(18):7737–7744, 1998

[43] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. on Inform. Theory*, IT-13:260–269, 1967.