Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

CHRISTOF SCHÜTTE ALEXANDER FISCHER
WILHELM HUISINGA PETER DEUFLHARD

A Direct Approach to Conformational Dynamics based on Hybrid Monte Carlo

Preprint SC 98-45 (December 1998)

A Direct Approach to Conformational Dynamics based on Hybrid Monte Carlo

Ch. Schütte[1,2], A. Fischer[1], W. Huisinga[1], and P. Deuflhard[1,2]

[1] Konrad Zuse Zentrum Berlin (ZIB), Takustr. 7, 14195 Berlin, Germany
[2] Freie Universität Berlin, Fachbereich Mathematik und Informatik, Arnimallee 2–6, 14195 Berlin, Germany

Abstract. Recently, a novel concept for the computation of *essential* features of the dynamics of Hamiltonian systems (such as molecular dynamics) has been proposed [9]. The realization of this concept had been based on subdivision techniques applied to the Frobenius–Perron operator for the dynamical system. The present paper suggests an alternative but related concept that merges the conceptual advantages of the dynamical systems approach with the appropriate statistical physics framework. This approach allows to define the phrase "conformation" in terms of the dynamical behavior of the molecular system and to characterize the dynamical stability of conformations. In a first step, the frequency of conformational changes is characterized in statistical terms leading to the definition of some Markov operator T that describes the corresponding transition probabilities within the canonical ensemble. In a second step, a discretization of T via specific hybrid Monte Carlo techniques is shown to lead to a *stochastic* matrix P. With these theoretical preparations, an identification algorithm for conformations (to be presented in [11]) is applicable. It is demonstrated that the discretization of T can be restricted to few essential degrees of freedom so that the combinatorial explosion of discretization boxes is prevented and biomolecular systems can be attacked. Numerical results for the n-pentane molecule and the triribonucleotide adenylyl*(3'-5')*cytidylyl*(3'-5')*cytidin are given and interpreted.

Key words. conformation, conformational dynamics, hybrid Monte Carlo, reweighting, essential degrees of freedom, transition probabilities, Markov operator, transition operator.

Mathematics subject classification. 47A75, 47B38, 58F05, 60J20, 60J35, 65C05, 65U05.

1 Introduction

The classical microscopic description of molecular processes leads to a mathematical model in terms of Hamiltonian differential equations. In principle, the discretization of such systems permits a simulation of the dynamics. However, direct simulation is even today restricted to relatively short time spans and to comparatively small discretization steps. Fortunately, most questions of chemical relevance just require the computation of *averages* of physical observables, of *stable conformations*, or of *conformational changes*. In a conformation, the *large scale geometric structure* of the molecule is understood to be conserved, whereas on smaller scales the system may well rotate, oscillate or fluctuate. The computational characterization of a conformation via direct simulation thus often requires inaccessibly long time spans.

Therefore, most approaches to the identification of conformations neglect the dynamical aspect: they are interested only in finding clusters of molecular configurations with significantly different large scale geometric structure and realize this by a straightforward statistical analysis of some appropriate set of sampling data, compare [22,27]. Unlike these approaches, we herein advocate to *directly* attack the determination of conformations *together* with the computation of their stability time spans and the rate of transitions between them. Therefore, it is suggested to define the phrase "conformation" in terms of statistical mechanics and *not* in terms of molecular geometry: a *conformation* is understood as some *almost invariant* subset in the position space – a notion which means that the fraction of systems in the molecular ensemble, that leave this subset during some fixed observation time, is "small". The algorithm to be presented allows to decomposed the position space into such dynamically defined conformational subsets and to compute the corresponding transition probabilities. This approach distinctly differs from other approaches to the characterization of conformational transitions, e.g., via artificial acceleration of molecular processes (cf. [24,23,43]).

The key idea of the algorithmic realization of the new approach goes back to the work of M. Dellnitz and coworkers on the approximation of almost invariant sets in dynamical systems [8]. Therein, it had been suggested to compute almost invariants subsets in phase space via the discretized eigenvalue problem for the Frobenius–Perron operator, an operator which describes the propagation of probability within the system. This "dynamical systems" approach has been realized for molecular dynamics [9], but, even though the numerical results were intriguing, this approach suffers both from a (yet) unclear theoretical justification and from the so-called "curse of dimension" of the proposed subdivision algorithm.

Herein, we will propose an alternative strategy that merges the conceptual advantages of the dynamical systems approach with the appropriate statistical physics framework. The key step of its derivation is the replacement of the Frobenius–Perron operator by the statistically correct spatial transition operator. The conceptual background of this replacement and its algorithmic consequences are first outlined in Section 2 and subsequently discussed in more detail in Secs. 3 and 4. The single steps of the resulting algorithm are illustrated by numerical results for the rather simple n-pentane molecule (Sec. 5). Its applicability to biologically relevant systems —in particular the circumvention of the curse of dimension— is exemplified at a small ribonucleotide.

2 Outline of the Method

Before we go into the technical details of this paper, we want to give some "bird's eye view" of the new approach as a whole.

Theoretical Framework As usual in molecular dynamics, we assume that we are dealing with an ensemble of molecular systems that is described by some (stationary) density f_0 in the phase space Γ of the molecular systems under consideration. Moreover, we suppose that the dynamical behavior of a single molecular system starting at time $t = 0$ in state $x_0 \in \Gamma$ can be described by the formal solution $x(t) = \Phi^t x_0$ of certain Hamiltonian equations of motion (compare Sec. 3 for details). Then the *transition probability* between two subset $S_1, S_2 \subset \Gamma$ is given by

$$w(S_1, S_2, \tau) = \frac{1}{\int_{S_1} f_0(x)\, dx} \int\limits_{S_1} \chi_{S_2}(\Phi^\tau x)\, f_0(x)\, dx \tag{1}$$

with χ_S denoting the characteristic function of the set $S \subset \Gamma$, i.e., $\chi_S(x) = 1$ iff $x \in S$ and $\chi_S(x) = 0$ otherwise. We are interested in *almost invariant* subsets, i.e., in sets $S \subset \Gamma$ with large probabilities to stay within, which, for the time being, can be expressed as $w(S, S, \tau) \approx 1$. In [9], chemical conformations had been understood as such almost invariant subsets in phase space Γ. However, they are usually understood to be objects in *position space*. Therefore, we herein characterize *conformational subsets* as *spatial* subsets B of positions $q \in B$. If we allow for arbitrary momenta p, we are naturally led to the *phase space fiber*

$$\Gamma(B) = \{(q, p) \in \Gamma, \quad q \in B\} \tag{2}$$

associated with B. Consequently, the spatial subset B is said to be a conformational subsets whenever the phase space fiber $\Gamma(B)$ is almost invariant in the sense that $w(\Gamma(B), \Gamma(B), \tau) \approx 1$.

The crucial step towards the algorithmic identification of such *conformational subsets* is the derivation of some Markov operator T in Sec. 3.3, which describes the *probability of position fluctuations* within the canonical ensemble. Consequently, the Markov chain $\{q_k\}_{k=0,1,\ldots}$ generated by T allows to simulate the spatial transitions in the ensemble. The chain takes values in the position space Ω and has the following basic properties: First, its stationary probability to be within a spatial subset $B \subset \Omega$, denoted by $\pi(B)$, is given via the ensemble density f_0, i.e., $\pi(B) = \int_{\Gamma(B)} f_0(x)\, dx$, and, second, its one-step transition probabilities $P(q_1 \in C | q_0 \in B)$ between subsets $B, C \subset \Omega$ are given by the transition probabilities within the ensemble between the corresponding spatial fibers

$$\frac{P(q_1 \in C | q_0 \in B)}{\pi(B)} = w(\Gamma(B), \Gamma(C), \tau). \tag{3}$$

This illustrates, that the generator T of the chain is the statistically correct *spatial transition operator* of the ensemble. Following [8,9], our algorithmic strategy is to identify conformational subsets via eigenmodes of the dominant eigenvalues of T (see Sec. 3.3).

Algorithmic Realization In order to compute these eigenmodes (and thus the conformations), we will have to discretize the corresponding eigenvalue problem. We realize this by means of a Galerkin procedure (Sec. 4.1) based on a box covering $B_1, \ldots, B_n \subset \Omega$ of the position space. This discretization step results in a reversible stochastic transition matrix whose entries are just the transition probabilities $w(\Gamma(B_k), \Gamma(B_l), \tau)$ between the discretization boxes.

Due to (3), we may compute these entries of the transition matrix via simulation of the Markov chain associated with T. The approximation of this chain naturally leads to standard hybrid Monte Carlo (HMC) sampling techniques (Sec. 4.2). By construction, the transition probabilities of the resulting HMC chain are similar to that of the original chain whose probability to leave some conformational subset is extremely small. Consequently, the same *trapping problem* occurs for the HMC chain, which leads to the rather unsatisfactory convergence properties of HMC when applied to biomolecules, as reported in the literature [34]. In order to circumvent this problem, a novel approach combining HMC with the reweighting technique [13,5] has been presented in [14]. This HMC variant, called adaptive temperature hybrid Monte Carlo (ATHMC), facilitates the transitions by repeatedly switching to an increased temperature in order to cross crucial energy barriers followed by a correction of this momentary overheating via reweighting to the ensemble of the original temperature (cf. Sec. 4.2). Application of this technique allows us to compute the entries $w(\Gamma(B_k), \Gamma(B_l), \tau)$ of the transition matrix, even for larger molecules.

However, even if we can compute arbitrary transition probabilities, any discretization of the transition operator T will suffer from the "curse of dimension" whenever it were based on the decomposition 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— only relatively few *conformational* or *essential degrees of freedom* are needed to describe the conformational transitions [2]. Different techniques are available for identifying these essential degrees of freedom based on reliable simulation data (see Sec. 4.3). We herein suggest to apply these techniques to an ATHMC sampling. Having completed this identification process, we can avoid discretization of by far the most degrees of freedom of the molecular system under investigation; only the low-dimensional essential configuration space has to be discretized which leads to a tremendous reduction of dimension.

Once the entries of the corresponding transition matrix have been computed based on ATHMC sampling data, we have to determine the eigenvectors of its dominant eigenvalues. That is, only an approximation of the dominant eigenelements of the transition matrix is required, *not* its full diagonalization. Thus, actual evaluation of the required eigenvectors is efficiently possible using subspace oriented iterative techniques, even if the number of discretization boxes may be about 100.000 or larger (depending on the spectral properties of the matrix, see Sec. 4.3). The final step, the determination of the conformational subsets from these eigenvectors, is realized by means of a specific identification algorithm presented in [11].

The whole algorithmic scheme of the direct conformational dynamics approach is illustrated in Fig. 1.

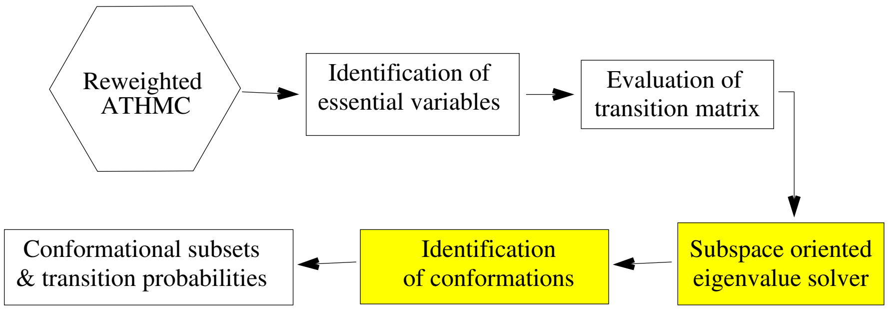

Fig. 1. Basic scheme of the algorithm. Gray boxes are presented in [11].

3 Conformations as Almost Invariant Sets

In classical MD (cf. textbook [1]) a molecule is modeled by a Hamiltonian function

$$H(q,p) \;=\; \tfrac{1}{2}\, p^T M^{-1} p \;+\; V(q), \tag{4}$$

where q and p are the corresponding positions and momenta of the atoms, M the diagonal mass matrix, and V a differentiable potential. The Hamiltonian H is defined on the phase space $\Gamma \subset \mathbb{R}^{6N}$. The corresponding canonical equations of motion

$$\dot{q} \;=\; M^{-1}p, \qquad \dot{p} \;=\; -\mathrm{grad}\, V \tag{5}$$

describe the dynamics of the molecule. The formal solution of (5) with initial state $x_0 = (q(0), p(0))$ is given by $x(t) = (q(t), p(t)) = \Phi^t x_0$, where Φ^t denotes the flow.

On the smallest time scales (say, 1 femtosecond) the dynamics described by the flow Φ^t consists of fast oscillations around equilibrium positions (bond length or bond angle vibrations). In contrast to these fast fluctuations the phrase "conformations" describes meta-stable global configurations of the molecule. *Conformational changes* are therefore rare events, which will show up only in long term simulations of the dynamics (e.g. on a nano- or millisecond time scale). From a mathematical point of view, conformations are special "almost invariant" subsets in position space: *Invariant sets* correspond to infinite durations of stay (or relaxation times). If the conformations were *invariant sets* of the flow of the Hamiltonian system, then transitions between different conformations would be *impossible*. Since such transitions exist but are *rare*, we must understand every conformation to be an *almost invariant* subset of the Hamiltonian flow.

3.1 Dynamical Systems Approach

In what follows, the concept of almost invariant sets and their algorithmic identification, which has been studied for rather general but low-dimensional dynamical systems, will shortly be reviewed:

Some subset $S \subset \Gamma$ is called *invariant* under the flow Φ^t iff, for all $t > 0$,

$$\Phi^t(S) = S \quad \text{and, thus,} \quad \Phi^{-t}(S) = S.$$

We now aim at a precise mathematical understanding of "almost invariance" of a subset $S \subset \Gamma$. Therefore, we have to introduce a measure for describing the fraction $S \cap \Phi^\tau(S)$ that remains in S under the action of the flow Φ^τ. The degree of invariance of S with respect to a certain probability measure μ is given by the corresponding conditional probability

$$\delta(S,\tau) = \frac{\mu\left(S \cap \Phi^\tau(S)\right)}{\mu(S)} \leq 1, \qquad S \ \mu\text{-measurable.} \tag{6}$$

In particular, if S is invariant, then $\delta(S,\tau) = 1$ independent of the choice of μ. We are interested in subsets S with $\delta(S,\tau)$ sufficiently close to $\delta = 1$, to be denoted as *almost invariant* subsets. The so-defined notion of almost invariance obviously depends on the choice of the time span τ. However, we will see in Sec. 3.3, that (at least for systems of chemical interest) the influence of τ on the identification of almost invariant subsets can be neglected.

Upon fixing a suitable time span τ, we have reduced the continuous dynamical system (5) to a discrete dynamical system

$$x_{k+1} = \Phi^\tau x_k, \quad k = 0, 1, 2, \ldots. \tag{7}$$

The long term behavior of this system is described by so-called *invariant measures*: a probability measure μ is invariant, iff $\mu(\Phi^\tau(S)) = \mu(S)$ for all measurable subsets $S \subset \Gamma$. Thus, $\mu(S)$ may be interpreted as the probability of finding the molecular system in S at an arbitrary instant $t = k\tau$, $k \in \mathbb{Z}$. Thus, invariant measures are the natural probability measures to be used in (6) for quantifying almost invariance. Consequently, uniqueness of the invariant measure is a desirable property since it guarantees that almost invariance is well-defined.

The numerical computation of invariant measures is equivalent to the solution of an *eigenvalue problem* for the so-called *Frobenius–Perron operator* U. Invariant measures correspond to eigenmodes of U for its largest eigenvalue $\lambda = 1$. It has been discovered in [8], that for many discrete dynamical systems

almost invariant sets are related to eigenmodes of the Frobenius–Perron operator for eigenvalues $\lambda \approx 1$ *inside* the unit circle ($|\lambda| < 1$). (8)

One strategy for identification of almost invariant sets is to discretize the Frobenius–Perron operator in order to approximate these eigenvalues $\lambda \approx 1$. In a sequence of articles (cf. [7,8]), M. Dellnitz and coworkers established numerical techniques realizing this strategy for different non-Hamiltonian systems. The Frobenius–Perron operator is discretized via a multi-level subdivision process, which generates a box covering of the system's relative global attractor. Recently, this approach has been extended to Hamiltonian systems with intriguing numerical results [9].

This "dynamical systems approach", however, has two crucial difficulties. First, this approach turns out to be useful only for small molecular systems, since it suffers from *combinatorial explosion* of the necessary number of discretization boxes already for moderate size molecules. Second, the approach has some deep-lying conceptual problems that are related to the properties of the Frobenius-Perron operator for Hamiltonian systems: To understand these problems, one has to discuss the physical meaning of the Frobenius–Perron operator U in the context of statistical mechanics. This will help us to draw the appropriate consequences for the molecular ensembles to be considered herein and, finally, to transform the key ideas of the dynamical systems approach into an algorithmic concept being applicable to the identification of biomolecular conformations.

3.2 Reformulation in Terms of Statistical Mechanics

In order to understand the physical meaning of the Frobenius–Perron operator for Hamiltonian systems, we recall the basic equations of motion in statistical mechanics. The evolution of a *statistical ensemble* of identically prepared systems is described by a time dependent probability density $f = f(x,t)$ in phase space. The propagation of the probability density is described by the Liouville equation for the Hamiltonian H:

$$\partial_t f = i\mathcal{L} f = \{H, f\}, \qquad f(t=0) = f_0, \tag{9}$$

where $\{\cdot,\cdot\}$ denotes the well–known Poisson bracket and $\mathcal{L} = -i\{H,\cdot\}$ the associated Liouville operator (cf. [30]). The density f_0 describes the initial probability distribution in the statistical ensemble, i.e., $f_0(x)$ is interpreted as the relative frequency in the ensemble of systems in state x at time $t = 0$. Therefore, the density must be *defined* in accordance with the *initial experimental preparation* of the ensemble.

On one hand, the solution of (9) is given by the flow as

$$f(x,t) \;=\; f_0(\Phi^{-t}x),$$

on the other hand, it can be denoted using the semi-group generated by $\mathcal{L}$ on the Hilbert space $L^2(\Gamma)$:

$$f(\cdot,t) \;=\; \exp(it\mathcal{L})\, f_0. \tag{10}$$

Frobenius–Perron Operator in Statistical Mechanics For the Hamiltonian system (7), the Frobenius–Perron operator U of the dynamical systems approach is identical with the statistical propagator in (10), that is,

$$U \;=\; \exp(i\tau\mathcal{L}), \quad \text{yielding} \quad Uf \;=\; f\circ\Phi^{-\tau}, \tag{11}$$

acting on $L^2(\Gamma) = \{f : \int_\Gamma |f(x)|^2 dx < \infty\}$, for details see [31,41]. Since $\mathcal{L}$ is self-adjoint [29], U is unitary in $L^2(\Gamma)$. Thus, the spectrum of U in $L^2(\Gamma)$ lies *on* the unit circle and there simply are no eigenvalues $\lambda < 1$ allowing for the identification of almost invariant sets. (The same is true in $L^1(\Gamma)$, see [31], Prop. 3.1.2; or [41]).

Moreover, all stationary solutions of the Liouville equations are invariant densities of U, i.e., eigenvectors for the eigenvalue $\lambda = 1$. In particular, for *arbitrary* smooth functions $F : \mathbb{R} \to [0,1]$, the associated densities $f(x) = F(H(x))$ are stationary solutions of the Liouville equation. Consequently, there are infinitely many invariant densities (and associated invariant measures) for U.

As a consequence of our considerations, one has to replace the Frobenius–Perron operator by an alternative stochastic operator that represents the restriction to the stationary ensemble density under consideration and —since the conformation are purely *spatial* objects— describes spatial fluctuation within this ensemble. After introducing the appropriate notation in the subsequent paragraph, we will see in Sec. 3.3 that this can in fact be realized.

Spatial Fluctuations in the Canonical Ensemble Most experiments on molecular systems are performed under the conditions of constant temperature and volume. The corresponding stationary density is the *canonical density* associated with the Hamiltonian H

$$f_0(x) \;=\; \frac{1}{Z}\exp\left(-\beta\, H(x)\right), \quad \text{with} \quad Z = \int_\Gamma \exp\left(-\beta\, H(x)\right) dx,$$

where $\beta = 1/k_B\mathcal{T}$, with $\mathcal{T}$ being the system's temperature $\mathcal{T}$ and k_B Boltzmann's constant. Since H was assumed to be separable, f_0 is a product

$$f_0(x) \;=\; \underbrace{\frac{1}{Z_p}\exp\left(-\frac{\beta}{2}p^T M^{-1} p\right)}_{=\mathcal{P}(p)}\;\underbrace{\frac{1}{Z_q}\exp\left(-\beta\, V(q)\right)}_{=\mathcal{Q}(q)}, \tag{12}$$

where we normalize $\mathcal{P}$ and $\mathcal{Q}$ such that

$$\int \mathcal{P}(p)dp = \int \mathcal{Q}(q)dq = 1.$$

In the following we always consider this *canonical ensemble*, i.e., f_0 will always be given by (12).

We are interested in particular almost invariant subsets of the canonical ensemble f_0. Thus, the probability measure μ in the basic definition (6) of almost invariance is now given by the density f_0. Then, the definition (1) of the statistical transition probabilities allows to rewrite the degree $\delta(S,\tau)$ of invariance of some subset $S \subset \Gamma$ as $\delta(S,\tau) = w(S,S,\tau)$. Thus, $S \subset \Gamma$ is almost invariant if $w(S,S,\tau) \approx 1$.

As already discuss above, conformations are related to subsets of the *position space* $\Omega \subset \mathbb{R}^{3N}$ (the spatial component of the phase space $\Gamma = \Omega \times \mathbb{R}^{3N}$): conformational subsets are subsets $B \subset \Omega$ such that the corresponding phase space fiber $\Gamma(B)$ is almost invariant, i.e., such that

$$w(\Gamma(B), \Gamma(B), \tau) \;\approx\; 1,$$

where, as a consequence of (1) and (12),

$$w(\Gamma(B), \Gamma(C), \tau) \;=\; \frac{1}{\int_B \mathcal{Q}(q)\,dq} \int\limits_C \left\{ \int_{\mathbb{R}^{3N}} \chi_B(\xi_1 \Phi^\tau(q,p))\,\mathcal{P}(p)\,dp \right\} \mathcal{Q}(q)\,dq,$$

with ξ_1 denoting the projection onto the position component, i.e., $\xi_1(q,p) = q$. From now on, we are interested only in subsets of this form and denote the probability to be within $B \subset \Omega$ by

$$\pi(B) \;=\; \int\limits_B \mathcal{Q}(q)\,dq \;=\; \int\limits_{\Gamma(B)} f_0(x)\,dx. \tag{13}$$

3.3 Definition of the Spatial Transition Operator

As will turn out subsequently, an appropriate choice for a stochastic operator is the *spatial transition operator* T defined via momentum weighting due to

$$Tu(q) \;=\; \int u(\xi_1 \Phi^{-\tau}(q,p))\,\mathcal{P}(p)\,dp, \tag{14}$$

where $u = u(q)$ is a function $u : \Omega \to \mathbb{C}$ and $u(\xi_1 \Phi^{-\tau}(q,p))$ means $u(q_1)$ if $(q_1, p_1) = \Phi^{-\tau}(q,p)$ due to the definition of ξ_1. In comparison with (11), one may interpret T as the restriction of the Frobenius–Perron operator to the position coordinates via an appropriate averaging with respect to the canonical momentum distribution.

We consider T as an operator on the weighted spaces

$$L^p_{\mathcal{Q}}(\Omega) = \{u : \Omega \to \mathbb{C}, \quad \int_\Omega |u(q)|^p \mathcal{Q}(q)\, dq < \infty\}, \qquad p = 1, 2.$$

Obviously, $L^2_{\mathcal{Q}}(\Omega)$ is a Hilbert space with scalar product

$$\langle u, v\rangle_{\mathcal{Q}} = \int_\Omega u^*(q)\, v(q)\, \mathcal{Q}(q)\, dq$$

and induced norm $\|u\|^2_{\mathcal{Q}} = \langle u, u\rangle_{\mathcal{Q}}$. With respect to these spaces, the important properties of T are the following (cf. [41]):

1. T is a Markov operator on $L^1_{\mathcal{Q}}(\Omega)$.
2. T is bounded: $\|Tu\|_{\mathcal{Q}} \leq \|u\|_{\mathcal{Q}}$.
3. In $L^2_{\mathcal{Q}}(\Omega)$, T is *selfadjoint*, since Φ^τ is *reversible*. Hence, the spectrum $\sigma(T)$ of T is real-valued and bounded: $\sigma(T) \subset [-1, 1]$.
4. For subsets $B, C \subset \Omega$ we find:
$$\langle T\chi_B, \chi_C\rangle_{\mathcal{Q}} = \int\limits_{\Gamma(B)} \chi_{\Gamma(C)}(\Phi^\tau x)\, f_0(x)\, dx, \tag{15}$$
showing that T represents the transition probabilities of our interest.
5. T is asymptotically stable in $L^1(\Omega)$, i.e., the eigenvalue $\lambda = 1$ is dominant and simple in $L^1(\Omega)$ and $L^2(\Omega)$ (this holds for all systems of chemical interest).

The last property shows that T has a unique invariant density so that "almost invariance" is well-defined via (6). Thus, T has all necessary properties to replace the Frobenius–Perron operator such that, in analogy to (8), we may identify the conformational subsets via the eigenmodes of T for eigenvalues near $\lambda = 1$.

In contrast to the properties 1-4 which generally hold for Hamiltonian systems, the last property is only valid for systems satisfying a certain mixing condition: for every position $q \in \Omega$, the map $y_q(p) = \xi_1 \Phi^\tau(q, p)$ must have sufficiently strong mixing properties (e.g., y_q must not map all possible momenta p to a single position $q' \in \Omega$). This mixing condition is satisfied, e.g., for all molecular systems with periodic boundary condition [41]. It, however, excludes certain "degenerate" systems such as strictly harmonic systems with period τ (where $y_q(p) = q$ for every momentum p).

Moreover, for systems satisfying the above condition for every $\tau > 0$, the dominant eigenmodes of T —and, thus, the almost invariant sets— are rather insensitive to changes in τ [41]. In contrast to this insensitivity, the transition probabilities do crucially depend on τ. The time span τ appears to be a temperature-like parameter (increases in τ effect a kind of melting process of the fluctuation-induced mixing in position space, compare [41] for details).

For the systems of interest, the cluster of eigenvalues near $\lambda = 1$ is separated from the remaining part of the spectrum $\sigma(T)$ by some significant spectral gap (cf. [41], Sec. 3.2): $\sigma(T)$ can be decomposed into this so-called Perron cluster $\{\lambda_1 = 1, \lambda_2, \ldots, \lambda_k\}$ of isolated eigenvalues $\lambda_k \leq \ldots \leq \lambda_2 < 1$, and the remainder $\sigma_R(T) \subset [-\kappa, \kappa]$ with some value $0 < \kappa < \lambda_k$ such that (in most cases of interest) the gap $g = \lambda_k - \kappa$ is significantly larger than the distances between the eigenvalues within the Perron cluster (for examples see Sec. 5).

4 Transition Probabilities and Associated Markov Chains

Since the transition operator T is a Markov operator in $L^1(\Omega)$ satisfying $T\chi_\Omega = \chi_\Omega$, it generates a Markov chain $\{q_k\}_{k=0,1,\ldots}$ with values in the position space Ω via the transition function

$$P(q_1 \in B | q_0 = q) \;=\; P(q, B) \;=\; T\chi_B(q), \qquad \text{for all measurable } B \subset \Omega.$$

This chain can be realized via the *discrete stochastic dynamical system* [41]

$$q_{k+1} \;=\; \xi_1 \Phi^\tau(q_k, p_k), \quad k = 0, 1, \ldots, \tag{16}$$

with p_k being randomly chosen from the momentum distribution $\mathcal{P}$ in each step. For systems of chemical interest, the chain has been shown to be irreducible and aperiodic with unique stationary density $\mathcal{Q}$ [41]. Moreover, any simulation of the chain via (16) would allow to compute the desired transition probabilities in the ensemble, since the definition of its transition function implies

$$P(q_k \in C | q_0 \in B) \;=\; \langle \chi_C, T^k \chi_B \rangle_{\mathcal{Q}}, \tag{17}$$

which in particular yields (3) for the one-step transition probabilities.

Thus, the replacement of the Frobenius–Perron operator U by the spatial transition operator T induces an associated change in the dynamical description: the discrete deterministic dynamical system (7) associated with U is replaced by the stochastically perturbed dynamical system (16) associated with T. In other words, the restriction to *spatial* fluctuations via averaging with respect to the canonical momentum distribution may be interpreted as a specific *coarse graining* of the dynamical description.

In order to compute the conformational subsets via the eigenvalue problem for T, we will now proceed to the (spatial) discretization of T. We will see that this finally also leads to a certain discretization of the Markov chain $\{q_k\}_{k=0,1,\ldots}$ generated by T.

4.1 Spatial Discretization

If we restrict our attention to the weighted Hilbert space $L^2_Q(\Omega)$, we can (as in [8,9]) naturally derive a special Galerkin procedure to discretize the eigenvalue problem $Tu = \lambda u$. Let $B_1, \ldots, B_n \subset \Omega$ be a covering of Ω so that $B_k \cap B_l = \emptyset$ for $k \neq l$ and $\cup_{k=1}^n B_k = \Omega$. Then, the sets $\Gamma(B_k)$, $k = 1, \ldots, n$, are a covering of Γ. Our finite dimensional ansatz space $\mathcal{V}_n = \text{span}\{\chi_1, \ldots, \chi_n\}$ is spanned by the associated characteristic functions $\chi_k = \chi_{B_k}$. The Galerkin projection $\Pi_n : L^2_Q(\Omega) \to \mathcal{V}_n$ of $u \in L^2_Q(\Omega)$ is defined by

$$\Pi_n u = \sum_{k=1}^n \frac{1}{\pi(B_k)} \langle \chi_k, u \rangle_Q \, \chi_k .$$

The resulting discretized transition operator $\Pi_n T \Pi_n$ induces the approximate eigenvalue problem $\Pi_n T \Pi_n u = \lambda u$ in $\mathcal{V}_n$. Let λ be one of the corresponding eigenvalues and let the related eigenvector be $u = \sum_{k=1}^n \alpha_k \chi_k$. Then, the discretized eigenvalue problem has the form

$$\sum_{l=1}^n \langle T\chi_k, \chi_l \rangle_Q \; \alpha_l \;=\; \lambda \, \pi(B_k) \, \alpha_k, \qquad \forall k = 1, \ldots, n.$$

After division by $\pi(B_k)$ (known to be positive), we end up with the convenient form

$$P\alpha \;=\; \lambda\alpha \quad \text{with} \quad \alpha = (\alpha_1, \ldots, \alpha_n),$$

where in fact the entries of the $n \times n$ matrix P are given by the spatial transition probabilities from B_k to B_l:

$$P_{kl} \;=\; \frac{\langle T\chi_k, \chi_l \rangle_Q}{\pi(B_k)} \;=\; w(\Gamma(B_k), \Gamma(B_l), \tau). \tag{18}$$

This result finally confirms that (14) was the correct choice of a transition operator in the statistical context.

Since T is a Markov operator, its Galerkin discretization P is a (row) stochastic matrix, i.e., $P_{kl} \geq 0$ and $\sum_{l=1}^n P_{kl} = 1$ for all $k = 1, \ldots, n$ (for details about stochastic matrices see [4]). Hence, all its eigenvalues λ satisfy $|\lambda| \leq 1$. Moreover, we have the following four important properties (cf. [41]):

1. The row vector $\pi = (\pi_1, \ldots, \pi_n)$, $\pi_k = \pi(B_k)$ denotes the discretized invariant density. Simple calculus reveals that π is a left eigenvector to the eigenvalue $\lambda = 1$, i.e., that $\pi P = \pi$.
2. P is *irreducible and aperiodic*, which implies, that the eigenvalue $\lambda = 1$ is *simple*. Hence, the discretized invariant density π is the *unique* stationary distribution of P.
3. P is *reversible*, since T is self-adjoint. In other words, P fulfills the condition of *detailed balance*:

$$\pi_k \, P_{kl} \;=\; \pi_l \, P_{lk}, \qquad \forall k, l \in \{1, \ldots, n\}.$$

Therefore, all eigenvalues of P are real-valued: $\sigma(P) \subset [-1, 1]$.

4. Whenever the discretization is fine enough, the dominant eigenvalues of P are good approximations of the dominant eigenvalues of T. In this case, P also has a Perron cluster of eigenvalues near $\lambda = 1$ which is separated from the remainder of the spectrum by a significant gap (cf. Sec. 3, last paragraph).

This means that, for arbitrary coverings $B_1, \ldots, B_n \subset \Omega$, the discretization matrices P are inheriting the most important properties of the operator T.

As any stochastic matrix, our discretization matrix P also defines a *discrete Markov chain*, i.e., the stochastic (random) walk of a single system through phase space. The associated statistical interpretation is as follows: If at instance $j \in \mathbb{N}$ the system is in B_k, the probability of finding the system in B_l at instance $j+1$ is $P_{kl} = w(\Gamma(B_k), \Gamma(B_l), \tau)$. With $j \to \infty$ the system visits all subset B_k with the probability π_k, the value given by the stationary distribution of P.

According to our definition of "almost invariance", we are interested in such unions $B = \cup_{k \in I} B_k$ of our "discretization boxes" B_k, for which the probability $w(\Gamma(B), \Gamma(B), \tau)$ to stay within is sufficiently close to $\delta = 1$. In other words, we are looking for a nontrivial index set $I \subset \{1, \ldots, n\}$ so that the discrete system almost certainly stays within $B = \cup_{k \in I} B_k$ within one single step $j \to j+1$. As derived in [11], such index sets ("almost invariant aggregates") can be identified via the right eigenvectors of P for eigenvalues close to $\lambda = 1$. Once a conformational subset B has been identified, the probability $\delta(B, \tau) = w(\Gamma(B), \Gamma(B), \tau)$ to *stay within* B can easily be computed by virtue of the relation:

$$\delta(B, \tau) \;=\; \frac{1}{\sum\limits_{k \in I} \pi_k} \sum_{k,l \in I} \pi_k \, P_{kl}. \tag{19}$$

4.2 Realization via Hybrid Monte Carlo (HMC)

Up to now, the remaining question is how to compute the matrix P for given boxes B_k. According to (18) we have to determine the transition probabilities between the B_k. This task includes two subproblems:

1. "Sampling of the canonical density": That is, we have to generate a sequence of states $S = \{x_k, \quad k = 1, \ldots, M\} \subset \Gamma$ that is approximately distributed according to f_0.
2. Approximation of the transition probabilities: We will see below that this reduces to counting all such $x_j \in S$ for which $x_j \in \Gamma(B_k)$ and $\Phi^\tau x_j \in \Gamma(B_l)$. For checking the last condition, sufficient approximations $\tilde{x}_j \approx \Phi^\tau x_j$ of all M subtrajectories starting from S are needed.

The typical approach to sampling the canonical density is via Monte Carlo (MC) techniques. The literature on this topic is extremely rich and varied [6,42]. The reader might notice that we need not give particular merits to any special MC variant since *every* converging MC method would allow to realize the subproblem 1 from above. In addition, one may also apply MD-based techniques, e.g., constant temperature sampling of the canonical density [38,3].

Despite this, we suggest to apply a certain *hybrid Monte Carlo* (HMC) technique, merely because it seems to be particularly appropriate for linking the above mentioned subproblems 1 and 2. In order to explain this advantage and the basic idea of HMC let us shortly recall that the transition probabilities may be computed via the Markov chain (16) associated with our transition operator T. Iterations of (16) realize sequences $\{q_k\}$ which are (asymptotically) distributed due to $\mathcal{Q}$ and allow to determine the relative frequency of transitions $q_k \in B_j \to q_{k+1} \in B_l$ for arbitrary box numbers j and l. The convergence guarantees that the relative frequencies approximate the desired transition probabilities in the sense that

$$\frac{\#\,(q_k \in B_j \wedge q_{k+1} \in B_l)}{\#\,(q_k \in B_j)} \;\to\; w(\Gamma(B_j), \Gamma(B_l), \tau). \tag{20}$$

Thus, we have to ask whether one can realize the iteration (16) by replacing the exact flow Φ^τ by an appropriate approximation. For answering this question, let $\Psi^{\Delta t}$ denote a reversible and volume-preserving one-step discretization of the flow Φ^t, i.e., of the Hamiltonian equations (5). The reader, who is not familiar with this notation, may think of $\Psi^{\Delta t}$ as denoting the well-known Verlet discretization [46,1] with stepsize Δt. The approximation of Φ^τ via m steps of this discretization yields the discrete flow

$$g \;=\; \left(\Psi^{\tau/m}\right)^m, \quad m \in \mathbb{N},$$

with m being large enough such that the stepsize τ/m is adequate. Unfortunately, the underlying stationary density f_0 is *not* invariant under the action of g, since g does not preserve the energy of the system. (There is no discretization which is symplectic and reversible and simultaneously preserves energy exactly [20]. We may reduce the energy error, produced by g, to an arbitrary small value by increasing m, but this would lead to a totally inefficient computation scheme.)

Standard Hybrid Monte Carlo (HMC) Hence, we have to look for a Markov chain, which allows to sample $\mathcal{Q}$ while containing only g and not the flow itself. This requirement naturally leads us to so-called "hybrid" Monte Carlo variants which to our knowledge have first been introduced in the late 80's (cf. [12]) and have in MD mostly been used for condensed matter and polymer-like systems (cf. [36,26,17]). HMC generates a sequence $(q_j) \subset \Omega$ in position space. The HMC update step $q_j \to q_{j+1}$ is based on the typical Metropolis Monte Carlo proposal/acceptance strategy: The first part of the HMC proposal step is to choose momenta p_j randomly from $\mathcal{P}$, gaining the state $x_j = (q_j, p_j)$. As the second part, compute the proposal state $\tilde{x}_j$ via a short approximate subtrajectory of the underlying Hamiltonian system, i.e., choose $\tilde{x}_j = g(x_j)$. Then, apply the standard Metropolis MC acceptance step to x_j and $\tilde{x}_j$, let the accepted state be x_{j+1}, and finally set $q_{j+1} = \pi_1 x_{j+1}$. In other words, HMC realizes an iteration of the Markov chain

$$q_{j+1} = \pi_1 a(q_j, p_j, r_j) \quad \text{with} \quad a(x,r) = \begin{cases} g(x), & \text{if } r \leq \alpha(x), \\ x & \text{otherwise,} \end{cases} \tag{21}$$

$$\begin{aligned} \text{setting} \quad & \alpha(x) = \min\{1, \exp(-\beta \Delta E(x))\}, \\ \text{with} \quad & \Delta E(x) \; = \; H(g(x)) - H(x), \end{aligned}$$

with p_j independently chosen randomly from $\mathcal{P}$ and r_j randomly from the equidistribution in $[0,1]$. In this form, HMC has to be understood as a *pure position sampling* of the spatial canonical distribution $\mathcal{Q}$ such that the resulting Markov chain $\{q_j\}$ allows to approximate the expectation values of appropriate *spatial* observables $\mathcal{A} : \Omega \to \mathbb{R}$ in the sense that we have asymptotically [41,37,44]

$$\Big| \frac{1}{M} \sum_{j=1}^{M} \mathcal{A}(q_j) \; - \; \int_{\Gamma} \mathcal{A}(q) \mathcal{Q}(q) \, dq \Big| \; \leq \; C \, M^{-1/2}, \tag{22}$$

with a constant C not explicitly depending on $\dim(\Gamma) = 6N$. Thus, we are able to approximate the desired transition probabilities $w(\Gamma(B_k), \Gamma(B_l), \tau)$ "simply" by counting according to (20). The main advantage of HMC in this context is obvious: we need approximations of $\Phi^\tau x_j$ and get them "for free" if we use $m\Delta t = \tau$ with sufficiently small Δt in the HMC iteration (21).

Theoretically, the transition matrix P is reversible. In order to reproduce this property for its approximation, we may simply count each transition from B_k to B_l as a transition $B_l \to B_k$, too (thus exploiting the reversibility of the discretization $\Psi^{\Delta t}$).

Reweighted Hybrid Monte Carlo (ATHMC) It is well-known that MC simulations for ensemble averages may suffer from possible "critical slowing down" [32]. This phenomenon occurs when the iteration $x_k \to x_{k+1}$ gets trapped near a local potential minimum due to high energy barriers so that a proper sampling of the phase space within reasonable computing times is prevented. Typically, this also happens to HMC applications to biomolecules [19,34]. Therefore, a novel approach combining HMC with the reweighting technique [13,5] has been developed [14]. This HMC variant generates the distribution of a mixed-canonical ensemble composed of two canonical ensembles at low and high temperature. Its analysis shows an efficient sampling of the canonical distribution at the low temperature, whereas the high temperature component facilitates crossing of the crucial energy barriers. We will call this variant "adaptive temperature HMC" (ATHMC) in the following. The sampling positions q_j generated by high temperature update steps have to be *reweighted* in order to guarantee overall convergence to the canonical position distribution to the low temperature. Moreover, we have to supply additional trajectories in order to guarantee that the initial momenta of the set of trajectories starting in one of the sampling position q_j are weighted according to the correct low temperature. For details of the ATHMC construction, the reader is referred to our article [14].

The necessity of introducing generalizations of HMC is caused by the *existence of almost invariant sets*: If there are almost invariant sets, denoted B and C, with small transition probability $w(\Gamma(B), \Gamma(C), \tau)$, then, both, the Markov chain (16) associated with the transition operator and the original HMC Markov chain need a huge number of iterations in order to produce sufficiently many of the rare transitions between B and C. This problem is circumvented by introducing the ATHMC chain which facilitates such transitions but has to be reweighted in order to yield samplings of the original canonical distribution.

The reader might also notice, that there are other Monte Carlo Markov chain techniques which allow to enforce barrier crossing (for example, the multicanonical algorithm [25], simulated tempering [35], J-walking [18], the fluctuating potential method [33] and other novel approaches [5].).

4.3 Essential Degrees of Freedom

Typical biomolecular systems contain hundreds or thousands of atoms. As a consequence, any direct spatial discretization of the transition operator T suffers from the curse of dimension, since the number of discretization boxes grows exponentially with the size of the molecular system under consideration. Our strategy to circumvent the curse of dimension is based on chemical observation. In the chemical literature conformations of biomolecules are mostly described in terms of few *essential degrees of freedom*. In the subspace of essential degrees of freedom anharmonic motion occurs that comprises most of the positional fluctuation, while in the remaining degrees of freedom the motion has a narrow Gaussian distribution and can be considered as "physically constrained". We may determine essential degrees of freedom either in the coordinate space according to AMADEI ET AL. [2] or in the space of internal degrees of freedom, e.g., torsion angles, by statistical analysis of circular data [15,16]. Both procedures result in a tremendous reduction of dimension (see Sec. 5.2).

After partitioning the chosen essential degrees of freedom resulting in discretization boxes $B_1, \dots, B_m$ we assemble the transition matrix P and solve the corresponding eigenvalue problem. Since we only need the Perron cluster of the largest eigenvalues near $\lambda = 1$, we apply subspace oriented iterative techniques (see, e.g., [40] or [10], Sec. 4.1) to solve the eigenvalue problem. It is important that the convergence rate only depends on the spectral gap between the Perron cluster and the remaining part of the spectrum (see Sec. 3.3) and is *independent of the size of the transition matrix* and thus of the number of discretization boxes. Therefore, neither the HMC sampling techniques nor the solution of the eigenvalue problem do scale exponentially with the size of the molecule.

5 Numerical Experiments

In this section, the performance of the above derived algorithm in application to n-pentane and to the triribonucleotide adenylyl*(3'-5')*cytidylyl*(3'-5')*cytidin are presented. The application to n-pentane allows to follow closely the single steps of the algorithm, while the case of the ribonucleotide exemplifies the performance of the algorithm when applied to biologically relevant systems.

5.1 Application to n-Pentane

Fig. 2 illustrates the chemically observed conformations of the n-pentane molecule $CH_3(CH_2)_3CH_3$.

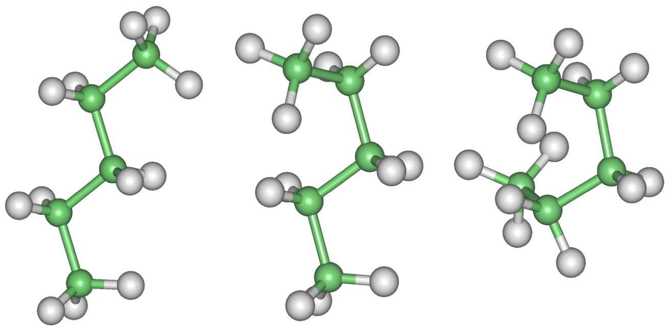

Fig. 2. Different conformations of n-pentane: From the left to the right: trans-trans, trans-gauche, gauche-gauche orientations.

For the n-pentane Hamiltonian, we use the united atom model (cf. Fig. 3) with the typical bond length and bond angle potentials, and a Lennard-Jones potential modelling the interaction between the first and the last of the united "atoms". The dihedral angle potentials are chosen according to [39], cf. Fig. 3. The form of the dihedral angle potential shows three different minima corresponding to the trans and gauche orientations of the angles. The vibrational frequencies induced by these potentials are considerably smaller than those induced by the bond interactions. Consequently, in this simple example, the dihedral angles can be selected as the essential degrees of freedom mentioned above in Sec. 4.3.

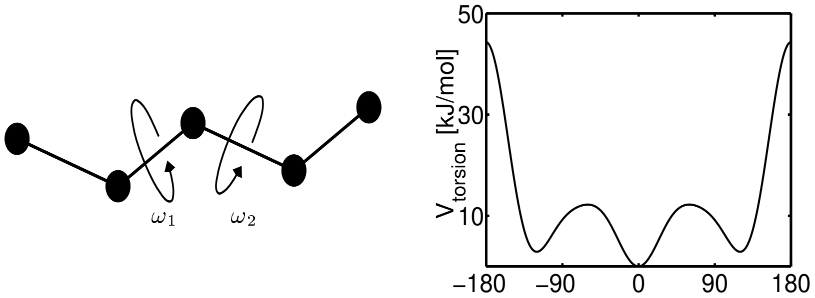

Fig. 3. United atom model of n-pentane with the two dihedral angles ω_1 and ω_2. On the left: Dihedral angle potential due to [39]. The main minimum corresponds to the trans orientation of the angle, the two side minima to the $\pm$gauche orientations.

Figures 4 to 7 below illustrate the performance of the algorithm for the temperature $\mathcal{T} = 300$K. The discretization boxes are constructed via uniform decomposition of the possible values $[0, 2\pi] \times [0, 2\pi]$ of the two dihedral angles ω_1 and ω_2 in $n = 20 \times 20 = 400$ boxes. The HMC sampling has been realized using the Verlet time discretization with a subtrajectory length of $\tau = 160$fs. Fig. 4 shows the resulting sequences of HMC steps in terms of the dihedral angles.

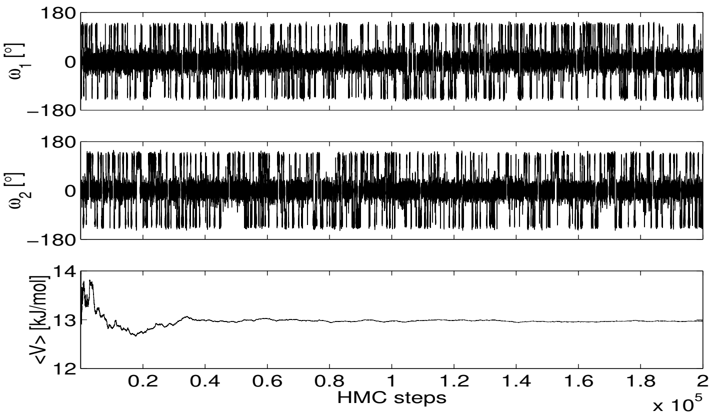

Fig. 4. HMC simulation of n-pentane for $\mathcal{T} = 300$K. From top to bottom: The two dihedral angles versus the step number and the convergence of the potential energy expectation $\langle V \rangle$.

We observe frequent transitions between the different "trans" and "gauche" orientations of both angles. This observation illustrates that it is not sufficient to know the probability to *be within* a particular orientation of the angles but that the essential dynamical information is given by the probability to *stay within* it until a transition into another orientation occurs.

Based on such a HMC sampling with $M = 200.000$ steps, the *transition matrix* P is assembled by the procedure explained in Sec. 4.2. Within this sampling length, the HMC method produces a sufficient sampling of the canonical density (see the equilibration diagram on bottom of Fig. 4). That is, in this case, we observe no serious trapping problems and application of ATHMC is not absolutely necessary. When switching to lower temperatures (as, e.g., for the simulation underlying Fig. 8 below), the rate of convergence of the HMC sampling slows down significantly and an application of ATHMC allows to decrease sampling lengths for more than an order of magnitude (cf. [14]).

From Sec. 4.1 we know that the discrete invariant density $(\pi(B_k))_{k=1,\ldots,n}$ is given by the left eigenvector of P for the largest eigenvalue $\lambda_1 = 1$. The result is given in Fig. 5. As expected, the invariant density shows distinct local maxima at the minima of the dihedral angle potentials.

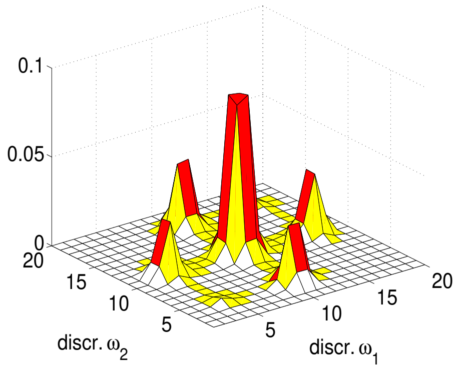

Fig. 5. Discrete canonical distribution for n-pentane versus the indices of the discretization boxes of the two dihedral angles ω_1 and ω_2. $\mathcal{T} = 300$K.

Conformations. Following [11], the chemical conformations are analyzed via the right eigenvectors corresponding to an eigenvalue cluster near $\lambda = 1$. A presentation of the derivation of the algorithmic procedure would be beyond the scope of the present paper. We herein only give a sketch of the construction principle: In a first step, determine the eigenvalue cluster near $\lambda = 1$, which is separated from the remaining part of the spectrum by a significant spectral gap – in our case, these are the seven largest eigenvalues. Fig. 6 shows a schematic plot of the corresponding right eigenvectors. We observe that we may decompose the discretization domain into disjoint regions by distinguishing between different positive, negative, and almost zero values of these eigenvectors. The details of the algorithmic realization are nontrivial, because it has to include an iterative procedure to decide what is "almost zero".

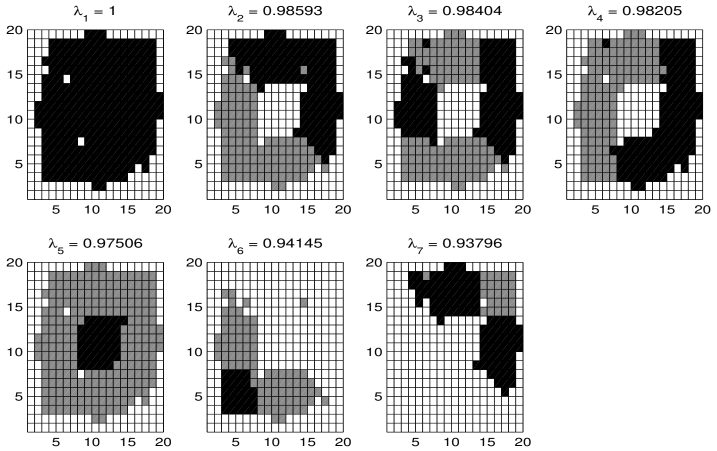

Fig. 6. Schematic plot of the right eigenvectors corresponding to the seven largest eigenvalues $\lambda_1, \ldots, \lambda_7$ of P versus the indices $(1, \ldots, 20) \times (1, \ldots, 20)$ of the discretization boxes of the two dihedral angles ω_1 and ω_2. Positive entries of the eigenvectors are indicated by black boxes, negative entries by gray boxes and white boxes indicate almost zero entries. $\mathcal{T} = 300$K.

By analyzing the eigenvectors as illustrated, the algorithm from [11] identifies the conformational subsets shown in Fig. 7. As can be seen the automatic procedure in fact supplies the chemically expected information. After identifying the conformations, the corresponding probabilities to stay within each conformational subset can be computed due to equation (19). The resulting values p are also given in Fig. 7. We observe that the trans/trans conformation is slightly more stable than the different trans/gauche and gauche/trans conformations. As expected, the two gauche/gauche conformations are clearly less stable.

As already emphasized above, the probabilities to *stay within* should *not* be confused with the probability to *be within* a conformation, which is already given by the invariant density (cf. Fig. 5). In the table below, these two different probabilities are listed for each of the conformational subsets shown in Fig. 7 (±g and t denote the ±gauche and trans orientations):

conformation	-g/t	t/+g	-g/-g	t/t	t/-g	+g/t	+g/+g
prob. to be within	0.120	0.132	0.012	0.473	0.117	0.132	0.013
prob. to stay within	0.976	0.980	0.910	0.982	0.979	0.970	0.865

The slight differences between the probabilities to be within the ±g/t and t/±g orientations may be used as an error indicator for the sampling. The probability to be within the +gauche/-gauche or -gauche/+gauche orientations is less than 0.0005, showing that they are irrelevant in this context.

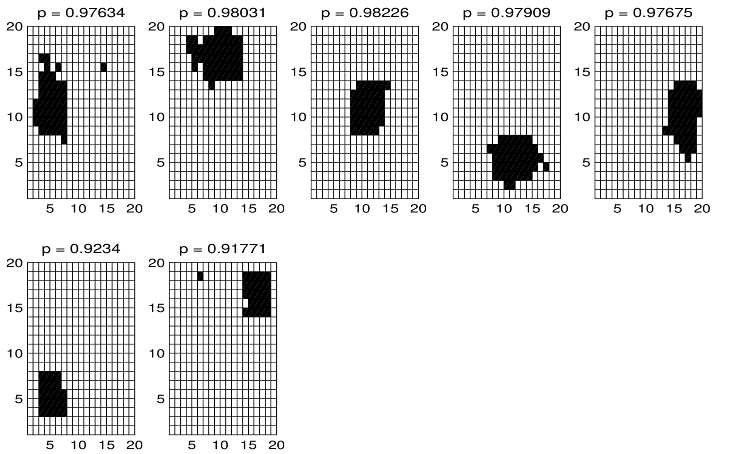

Fig. 7. Almost invariant sets for $\mathcal{T} = 300$K. The numbers p on top of each figure are the probabilities to stay within the corresponding subsets during the time span τ. From the left hand side on top to the right hand side below we see the -gauche/trans, trans/+gauche, -gauche/-gauche, trans/trans, trans/-gauche, +gauche/trans, and +gauche/+gauche conformations (cf. Fig. 2).

Parameter Sensitivity. The results presented herein surely depend on a number of crucial parameters, some of them being of physical nature (e.g., the temperature $\mathcal{T}$), others being introduced by the algorithm (e.g., the number n of discretization boxes or the length M of the HMC sampling). We want to emphasize that the algorithm as it stands now is far from being perfectly tuned. We thus can only present some experiences from numerical experiments for the n-pentane molecule and some other comparably small systems.

At first, let us consider the dependence of the conformations on the temperature $\mathcal{T}$. Varying the temperature between $\mathcal{T} = 200$K and $\mathcal{T} = 600$K we do not observe an influence on the identified conformations. But, as to be expected, the probabilities to stay within these conformations are decreasing with increasing $\mathcal{T}$: Fig. 8 shows the corresponding decrease of the nine largest eigenvalues of the transition matrices $P = P(\mathcal{T})$. It also illustrates that in all cases tested so far there exists a distinct *spectral gap* between the seven largest eigenvalues used to identify the conformational subsets, and the remaining part of the spectrum.

Obviously, the quality of the results depends crucially on the length M of the HMC sampling. If, for fixed temperature and spatial discretization, the number of steps is decreased from $M = 200.000$ down to $M = 50.000$, we observe that the approximation quality of the invariant density slowly deteriorates. This corresponds to a slowly increasing distortion of the approximate "conformational" subsets. Thus, it is of primary importance to check the quality of the Monte Carlo sampling via appropriate convergence indicators [21].

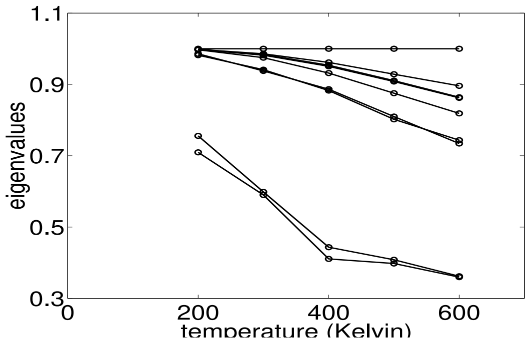

Fig. 8. Temperature dependence of the nine largest eigenvalues of the transition matrix P.

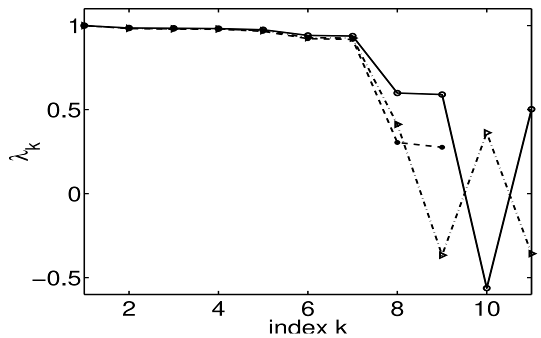

Fig. 9. Sensitivity of the *absolutely* largest eigenvalues of P for different uniform discretizations of $[0, 2\pi]^2$ with $n = 3 \times 3 = 9$ boxes (dashed line), $n = 9 \times 9 = 81$ (dashed-dotted), and $n = 20 \times 20 = 400$ boxes (dense line). Note that the seven largest eigenvalues – only these are used for the identification of the conformations – remain almost unperturbed if the grid gets coarser.

Dependence on Discretization. Finally, let us illustrate an extremely important property of the presented algorithm, the stability of the results even when significantly coarser discretizations are used. For the n-pentane molecule we indeed can reduce the decomposition of the discretization domain from $n = 20 \times 20$ boxes to $n = 3 \times 3$ boxes but the algorithm still identifies approximately the same conformations and nearly the same probabilities (both to stay and to be within). The reason for this is illustrated in Fig. 9: since the HMC procedure samples the phase space independent of the discretization, the seven largest eigenvalues of the transition matrix P are only insignificantly perturbed when the number of discretization boxes is reduced.

5.2 Application to a Ribonucleotide

In this section, the performance of the algorithm in application to the triribonucleotide adenylyl*(3'-5')*cytidylyl*(3'-5')*cytidin at temperature $\mathcal{T} = 295$K is presented. The trinucleotide molecule is modelled by means of the potential and masses of the extended atom representation of Gromos [45]. Solvent effects are neglected.

The numerical results to be presented are based on an ATHMC sampling of the canonical density using subtrajectories of length $\tau = 80$ femtoseconds computed by means of the Verlet discretization with stepsize $\Delta t = 2$ femtoseconds. For these parameters, HMC simulations typically require thousands of iterations only to leave the neighborhood of the initial configuration. Application of ATHMC (with adaptive temperatures between $\mathcal{T} = 295$K and $\mathcal{T}^+ = 400$K) circumvents the problem: one observes frequent transitions in the crucial torsion angles of the molecule (for details see [14]). The ATHMC simulation was terminated by the associated convergence indicator [21] after $M = 32.000$ steps, resulting in the sampling sequence $q_1, \ldots, q_M$, and corresponding reweighting factors. The sampling process was completed by the "transition sampling" by computing four subtrajectories $\Phi^\tau(q_k, p_{k,l})$ for each of the sampling positions q_k with initial momenta $p_{k,l}$ randomly chosen from $\mathcal{P}$.

Based on this ATHMC sampling, the essential degrees of freedom of the molecule were determined by applying an identification procedure based on statistical analysis of circular data [15,16] similar to that proposed by Amadei et al. [2] but using torsion angles instead of position information [28]. In this procedure *generalized angle coordinates* are introduced (linear combinations of the torsion angles defined by eigenvectors of the circular covariance matrix that measures correlations between the torsion angles). The distribution of the sampling sequence (q_k) with respect to these generalized coordinates has the form of some narrow Gaussian for most of the coordinates (indicating that they can be considered as "physically constrained"), while it is non-Gaussian for a small number of coordinates only (cf. Fig. 10). In our case, only four degrees of freedom showed such non-Gaussian distribution. The partitioning of the corresponding four-dimensional essential configuration space was chosen such that these distributions are decomposed into their single Gaussian-like parts (cf. Fig. 10). This process generated 36 discretization boxes.

For this partitioning, the transition matrix P (size 36×36) was assembled by counting the transitions between the discretization boxes based on the $4 \times 32.000 = 128.000$ subtrajectories of the transition sampling and weighting each transition due to its reweighting factor. Since every box had been hit by sufficiently many events the statistical sampling was accepted to be reliable. The computation of the dominant eigenvalues of P yielded a Perron cluster of 8 eigenvalues with a significant gap to the remaining part of the spectrum.

k	1	2	3	4	5	6	7	8	9	...
λ_k	1.000	0.999	0.989	0.974	0.963	0.946	0.933	0.904	0.805	...

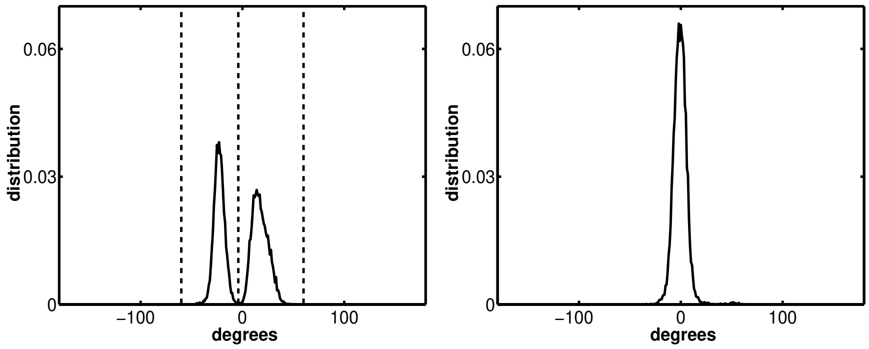

Fig. 10. Distribution of the sampling sequence $q_1, \ldots, q_M$ with respect to two of the generalized angle coordinates introduced in the text. Left: Distribution for an essential degree of freedom (possible decomposition illustrated by dashed lines). Right: Gaussian distribution for some nearly "physically constrained" degree of freedom.

Finally, the conformational subsets were computed based on the corresponding 8 right eigenvectors of P via the identification algorithm presented in [11]. The results turned out to be rather insensitive to further refinements of the partitioning. The corresponding probabilities to stay within and to be within these conformational subsets are listed in the following table:

conformation	1	2	3	4	5	6	7	8
prob. to be within	0.320	0.285	0.116	0.107	0.095	0.038	0.028	0.011
prob. to stay within	0.991	0.981	0.961	0.986	0.962	0.949	0.888	0.938

The resulting dynamical conformations are closely related to the conformations resulting from standard geometric identification algorithms, but the available dynamical information allows to gain further insight in the transitions between the conformational subsets (for a detailed comparison, see [28]).

Acknowledgments. It is a pleasure to thank Frank Cordes for many helpful discussions. W.H. was supported within the DFG–Schwerpunkt "Ergodentheorie, Analysis und effiziente Simulation dynamischer Systeme" under Grant De 293/2-1.

References

1. M.P. Allen and D.J. Tildesley. *Computer Simulations of Liquids.* Clarendon Press, Oxford, 1990.
2. A. Amadei, A.B.M. Linssen, and H.J.C. Berendsen. Essential dynamics of proteins. *Proteins*, 17, 1993.

3. H.C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, 72:2384–, 1980.
4. A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences.* Academic Press, New York, 1979. Reprinted by SIAM, Philadelphia, 1994.
5. B.J. Berne and J.E. Straub. Novel methods of sampling phase space in the simulation of biological systems. *Current Opinion in Structural Biology*, 7:181–189, 1997.
6. K. Binder. *The Monte Carlo method in condensed matter physics*, volume 71 of *Topics in applied physics.* Springer Verlag, Berlin, 1992.
7. M. Dellnitz and A. Hohmann. The computation of unstable manifolds using subdivision and continuation. In H.W. Broer, S.A. van Gils, I. Hoveijn, and F. Takens, editors, *Nonlinear Dynamical Systems and Chaos*, pages 449–459. Birkhäuser, *PNLDE* **19**, 1996.
8. M. Dellnitz and O. Junge. On the approximation of complicated dynamical behavior. *To appear in SIAM J. Num. Anal.*, 1998. http://www-math.uni-paderborn.de/ agdellnitz/publications/.
9. P. Deuflhard, M. Dellnitz, O. Junge, and Ch. Schütte, *Computation of Essential Molecular Dynamics by Subdivision Techniques.* In P. Deuflhard, J. Hermans, B. Leimkuhler, A. Mark, S. Reich, and B. Skeel, editors, *Computational Molecular Dynamics: Challenges, Methods, Ideas.*, Lecture Notes in Computational Science and Engineering, Vol. 4, pages 98–115. Springer Verlag, Berlin, 1998.
10. P. Deuflhard, T. Friese, and F. Schmidt. A nonlinear multigrid eigenproblem solver for the complex Helmholtz equation. Preprint SC 97-55, Konrad Zuse Zentrum, Berlin, 1997. Available via http://www.zib.de/bib/pub/pw/.
11. P. Deuflhard, W. Huisinga, A. Fischer, and Ch. Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. Preprint SC 98-03, Konrad Zuse Zentrum, Berlin, 1998. Available via http://www.zib.de/bib/pub/pw/.
12. S. Duane, A.D. Kennedy, B.J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Phys. Letters B*, 195(2):216–222, 1987.
13. A.M. Ferrenberg and R.H. Swendsen. New Monte Carlo technique for studying phase transitions. *Phys. Rev. Letters*, 61(23):2635–2638, 1988.
14. A. Fischer, F. Cordes, and Ch. Schütte. Hybrid Monte Carlo with adaptive temperature in mixed-canonical ensemble: Efficient conformational analysis of RNA. *J. Comput. Chem.*, 19:1689–1697, 1998.
15. Nick I. Fisher. *Statistical analysis of circular data.* Cambridge University Press, Cambridge, 1993.
16. Nick I. Fisher and A. J. Lee. A correlation coefficient for circular data. *Biometrica*, 70(2):327–332, 1983.
17. B.M. Forrest and U.W. Suter. Hybrid Monte Carlo simulations of dense polymer systems. *J. Chem. Phys.*, 101(3):2616–2629, 1994.
18. D.D. Frantz, D.L. Freeman, and J.D. Doll. Reducing quasi-ergodic behavior in Monte Carlo simulation by J-walking: Applications to atomic clusters. *J. Chem. Phys.*, 93:2769–2784, 1990.
19. D. Frenkel and B. Smit. *Understanding Molecular Dynamics.* Academic Press, New York, San Francisco, London, 1996.
20. Z. Ge and J.E. Marsden. Lie-Poisson integrators and Lie-Poisson Hamiltonian-Jacobi theory. *Phys. Lett. A*, 133:134–139, 1988.

21. A. Gelman. Inference and monitoring convergence. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 131–143. Chapman & Hall, London, 1995.
22. H.L. Gordon and R.L. Somorjai. Fuzzy cluster analysis of molecular dynamics trajectories. *PROTEINS: Structure, Function, and Genesis*, 14:249–264, 1992.
23. H. Grubmueller. Predicting slow structural transitions in macromolecular system: Conformational flooding. *Phys. Rev. E*, 52:2893–2906, 1995.
24. H. Grubmueller and P. Tavan. Molecular dynamics of conformational substates for a simplified protein model. *J. Chem. Phys.*, 101, 1994.
25. U.H.E. Hansmann and Y. Okamoto. Prediction of peptide conformation by multicanonical algorithm: New approach to the multiple-minima problem. *J. Comput. Chem.*, 14:1333–1338, 1993.
26. D.W. Heermann and L. Yixue. A global-update simulation method for polymer systems. *Makromol. Chem., Theory Simul.*, 2:299–308, 1993.
27. B. Hendrickson and R. Leland. An improved spectral graph partitioning algorithm for mapping parallel computations. *SIAM J. Sci. Comp.*, 16:452–469, 1995.
28. W. Huisinga, C. Best, F. Cordes, R. Roitzsch, and Ch. Schütte. From simulation data to essential conformations: A comparison of methods. Preprint in preparation, to appear Jan. 99, Konrad Zuse Zentrum, Berlin, 1999.
29. B.O. Koopman. Hamiltonian dynamics and transformations in Hilbert space. *Proc. Nat. Acad. Sci.*, 17, 1931.
30. R. Kurt. *Axiomatics of Classical Statistical Mechanics.* Pergamon Press, Oxford, New York, 1980.
31. A. Lasota and C. Mackey. *Chaos, Fractals and Noise.* Springer, New York, 1994.
32. E. Leontidis, B.M. Forrest, A.H. Widmann, and U.W. Suter. Monte Carlo algorithms for the atomistic simulation of condensed polymer phases. *J. Chem. Soc. Faraday Trans.*, 91(16):2355–2368, 1995.
33. Z. Liu and B.J. Berne. Methods for accelerating chain folding and mixing. *J. Chem. Phys.*, 99:6071–6077, 1993.
34. L.D. Loyens, B. Smit, and K. Esselink. Parallel Gibbs-ensemble simulations. *Mol. Phys.*, 86, 1995.
35. E. Marinari and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.*, 19:451–458, 1992.
36. B. Mehlig, D.W. Heermann, and B.M. Forrest. Hybrid Monte Carlo method for condensed-matter systems. *Phys. Review B*, 45(2):679–685, 1992.
37. S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability.* Springer, Berlin, Heidelberg, New York, Tokyo, 1993.
38. S. Nose. A molecular dynamics methods for simulations in the canonical ensemble. *Mol. Phys.*, 52, 1984.
39. J.-P. Ryckaert and A. Bellemans. Molecular dynamics of liquid n-butane near its boiling point. *Chem. Phys. Letters*, 30(1):123–125, 1975.
40. Y. Saad. *Numerical Methods for Large Eigenvalue Problems.* Manchester University Press, Manchester, 1992.
41. Ch. Schütte. Conformational dynamics: Modelling, theory, algorithm, and application to biomolecules. Habilitation thesis, manuscript available via email to schuette@zib.de, 1998.
42. A.D. Sokal. Monte Carlo methods in statistical mechanics. Lecture note, Department of Physics, New York University, 1989.

43. T.P. Straatsma and J.A. McCammon. Computational alchemy. *Annu. Rev. Phys. Chem.*, 43:407–435, 1992.
44. L. Tierney. Introduction to general state-space Markov chain theory. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *Markov chain Monte-Carlo in practice*, pages 59–74. Chapman and Hall, London, Glasgow, New York, Tokyo, 1997.
45. W.F. van Gunsteren, S.R Billeter, A.A. Eising, P.H. Hünenberger, P. Krüger, A.E. Mark, W.R.P. Scott, and I.G. Tironi. *Biomolecular Simulation: The GROMOS96 Manual and User Guide.* vdf Hochschulverlag AG, ETH Zürich, 1996.
46. L. Verlet. Computer experiments on classical fluids. Part I. *Phys. Rev.*, 159:98–103, 1967.