Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

CH. SCHÜTTE[†], W. HUISINGA[†‡], P. DEUFLHARD[†]

Transfer Operator Approach to Conformational Dynamics in Biomolecular Systems

[†]Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB), Germany. Internet: http://www.zib.de/MDgroup

[‡]supported by the Deutsche Forschungsgemeinschaft (DFG) under Grant De 293

Preprint SC 99-36 (October 1999)

Transfer Operator Approach to Conformational Dynamics in Biomolecular Systems

Ch. Schütte, W. Huisinga† and P. Deuflhard

Konrad–Zuse–Zentrum Berlin (ZIB)
Takustr. 7, 14195 Berlin, Germany

Abstract

The article surveys the development of novel mathematical concepts and algorithmic approaches based thereon in view of their possible applicability to biomolecular design. Both a first deterministic approach, based on the Frobenius-Perron operator corresponding to the flow of the Hamiltonian dynamics, and later stochastic approaches, based on a spatial Markov operator or on Langevin dynamics, can be subsumed under the unified mathematical roof of the transfer operator approach to effective dynamics of molecular systems. The key idea of constructing specific transfer operators especially taylored for the purpose of conformational dynamics appears as the red line throughout the paper. Different steps of the algorithm are exemplified by a trinucleotide molecular system as a small representative of possible RNA drug molecules.

Keywords. Transfer operator, Markov process, Markov chain, molecular dynamics, biomolecular conformations, canonical ensemble, transition probability, Hamiltonian dynamics, Langevin dynamics, nearly degenerate eigenvalues, Perron cluster

Mathematics subject classification. 65U05, 60J25, 60J60, 47B15

1 Introduction

In recent years, biomolecular design has attracted considerable attention both in the scientific and in the economic world. A few years ago, a research group at ZIB, partly supported by the DFG research program described in this volume, has started to work in this field. The problem of biomolecular design exhibits a huge discrepancy of time scales: those relevant from the pharmaceutical point of view are in the seconds, whereas present computations reach into the nanosecond regime at most. The reason for this is twofold: First, all available numerical integrators allow stepsizes of at most some femtoseconds only [38, 33]. Second, trajectory-oriented simulations are ill–conditioned after, say, a few thousand integration steps [1]. As a consequence, whenever *dynamical* informations (and

†supported by the Deutsche Forschungsgemeinschaft (DFG) under Grant De 293

not only *averages* of physical observables) are wanted—which is actually the case in biomolecular design—then only short term trajectories should be exploited. This message seems to be in direct contradiction to the desired aim of long term prediction in biomolecular design.

Aware of this seemingly contradiction, the ZIB group got inspired by work of DELLNITZ and co-workers [10, 9] on *almost invariant sets* of dynamical systems—within the same DFG research program. As documented in [11], the key idea was to interpret almost invariant sets in phase space as *chemical conformations*. Within chemistry, the latter term describes *metastable global states* of a molecule wherein the *large scale geometric structure* is conserved over long time spans. As it turned out, the chemists' dominant interest was anyway just in these conformations, their life spans, and their patterns of *conformational changes*. Therefore, our first approach [11] followed the line of the original paper by DELLNITZ AND JUNGE [10]: chemical conformations were identified via eigenmodes corresponding to an eigenvalue cluster of the *Frobenius-Perron operator* associated with the deterministic flow of the Hamiltonian system. However, upon keeping a clear orientation towards the design of biomolecular systems, the computational techniques based on this first approach appeared to be unsatisfactory for reasons of both lack of theoretical clarity and sheer computational complexity: The theoretical justification of the approach requires the introduction of artificial stochastic perturbations of the dynamics [10] regardless of any (physical) interpretation. Moreover, the computational techniques from [10, 11] are suitable only, if the objects of interest are rather low-dimensional, whereas the search for conformations will have to include the entire high-dimensional phase space of the molecular dynamics. Therefore, an almost complete remodelling with special emphasis on both physical interpretation and dimensionality of the problem turned out to be necessary in view of biomolecular applications.

In order to define conformations as *experimentally* determinable objects, concepts of *Statistical Physics* needed to be included. In addition, the remodelling had to include the aspect that chemical conformations are purely spatial objects determined via molecular geometry. These insights gave rise to the study of "spatial" *Markov operators* beyond the Frobenius-Perron operator as well as the associated Markov chains replacing the Hamiltonian dynamics [35, 34]. The thus arising special Markov operator was shown to exhibit all the desirable theoretical properties needed as a basis for efficient algorithms. Moreover, a Galerkin approximation of this Markov operator in a weighted L^2-space naturally led to the replacement of the original (expensive) subdivision techniques [9] by newly developed (cheap) Hybrid Monte Carlo (HMC) methods called reweighted adaptive temperature HMC, or short ATHMC [16]. On the basis of suggestions by AMADEI ET AL. [2], an algorithm for identifying the essential molecular degrees of freedom has been worked out that drastically reduces the eigenvalue cluster problem even in larger molecular systems see [22]. With these algorithmic improvements the applicability of our approach to realistic biomolecules came into reach. By applying the above ideas to the stochastic Langevin model of molecular dynamics [36], we succeeded to show that the fruitful coupling between the concepts of Statistical Physics and the transfer operator approach to effective dynamics can be exploited in a much more general framework.

The purpose of the present article is to survey what has been achieved, and to gain further insight from that. As will be shown subsequently, we are now able to subsume both our first deterministic approach [11] and the different stochastic approaches [35, 36] under the unified roof of transfer operators preserving the key idea of conformation analysis. In order to return to the original problem of biomolecular design, we illustrate the different steps of our present algorithm when applied to a small RNA molecular system.

2 Molecular Dynamics

In order to introduce our mathematical frame, we need to fix some notation. Consider a probability space $(\mathbf{X}, \mathcal{A}, \mu)$, where $X \subset \mathbf{R}^m$ for some $m \in \mathbf{N}$ denotes the state space, $\mathcal{A}$ the Borel σ–algebra on $\mathbf{X}$ and μ a probability measure on $\mathcal{A}$.

We will see below that in classical molecular dynamics the evolution of a single molecular system with initial data $x_0 \in \mathbf{X}$ is in general described by a homogeneous Markov process $\{X_t\}_{t \in \mathbf{M}}$ with $\mathbf{M} = \mathbf{R}_0^+$ or $\mathbf{M} = \mathbf{N}_0$ in continuous or discrete time, respectively. We assume that X_t is measurable and non–singular with respect to μ, i.e., $\mu({\mathbf{X}_t}^{-1}(A)) = 0$ for all $A \in \mathcal{A}$ with $\mu(A) = 0$. Furthermore, we assume that the process satisfies the semigroup properties: $X_{t=0} = Id$ and $X_{t+s} = X_t \circ X_s$ for all $t, s \in \mathbf{M}$. Then, the evolution of a single system starting in $x(0) = x_0$ is given by $x(t; x_0) = X_t(x_0)$ for all $t \in \mathbf{M}$. We choose this more general framework to describe molecular dynamics, since it is suitable both for the deterministic case and for the stochastic situation.

Markov processes may be defined in terms of stochastic transition kernels. A function $p : \mathbf{M} \times \mathbf{X} \times \mathcal{A} \to [0, 1]$ is called a *stochastic transition kernel* [6, 29], if

1. $p(t, x, \cdot)$ is a probability measure on $\mathcal{A}$ for every $t \in \mathbf{M}$, $x \in \mathbf{X}$ and furthermore, $p(0, x, X \setminus \{x\}) = 0$ for every $x \in \mathbf{X}$.

2. $p(t, \cdot, A)$ is measurable for every $t \in \mathbf{M}$, $A \in \mathcal{A}$.

3. $p(\cdot, x, A)$ satisfies the Chapman–Kolmogorov equation [19, 29]

$$p(t+s, x, A) \quad = \quad \int_{\mathbf{X}} p(t, x, \mathrm{d}y)\, p(s, y, A) \tag{1}$$

   for all $t, s \in \mathbf{M}$, $x \in \mathbf{X}$ and $A \in \mathcal{A}$.

The family $\{X_t\}_{t \in \mathbf{M}}$ is called a homogeneous Markov process, if [6, 29]

$$\mathbf{P}[X_t \in A \,|X_0 = x] \quad = \quad p(t, x, A) \tag{2}$$

for all $t \in \mathbf{M}$ and $A \in \mathcal{A}$. Thus $p(t, x, C)$ is the probability that the Markov process started in x stays in A after the time span t.

2.1 Modelling Molecular Motion

Classical models for molecular motion describe the molecular system under consideration via coupled equations of motion for the N atoms in the system (cf. textbook [1]). The most popular class of equations of motion can be written in the following general form:

$$\begin{aligned} \dot{q} &= M^{-1} p, \qquad (3) \\ \dot{p} &= -\nabla_q V(q) \,-\, \gamma(q, p)\, p \,+\, F_{\text{ext}}, \end{aligned}$$

where q and p are the atomic positions and momenta, respectively, M the diagonal mass matrix and $V = V(q)$ a differentiable potential energy function describing all the interactions between the atoms. The function $\gamma = \gamma(q,p)$ denotes the friction constant and F_{ext} the external forces acting on the molecular system. The state space of the system is $\Gamma \subset \mathbf{R}^{6N}$ and the solution (q_t, p_t) of (3) describes the dynamics of a *single molecular system*. In the notation introduced above, we hence have $\mathbf{X} = \Gamma$ and $X_t(q_0, p_0) = (q_t, p_t)$.

The Hamiltonian function

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

denotes the *internal energy* of the system in state $x = (q,p)$. In the following we assume $M = \mathrm{Id}_{\mathbf{R}^{3N}}$ for simplicity. In most cases, the phase space is simply given by $\Gamma = \Omega \times \mathbf{R}^{3N}$ for some $\Omega \subset \mathbf{R}^{3N}$. We will call Ω the *position space* of the system and distinguish between two fundamentally different cases:

(B) *Bounded system*: The position space Ω is unbounded, typically $\Omega = \mathbf{R}^{3N}$, and the potential energy function is smooth, bounded from below, and satisfies $V \to \infty$ for $|q| \to \infty$. Such systems are called bounded, since the energy surfaces $\{x : H(x) = E\}$ are bounded subsets of Γ.

(P) *Periodic systems*: The position space Ω is some $3N$ dimensional torus and V is continuous on Ω and thus bounded. There is an intensive discussion concerning the question of whether V can also be assumed to be smooth on Ω as we will do herein, see Sec. 2 of [34] for details.

Both cases are typical for molecular dynamics applications. Case (P) includes the assumption of periodic boundaries which is the by far the most popular modelling assumption for biomolecular systems. Subsequently, we will refer to these assumptions by referring to systems of type (B) or type (P).

Deterministic Hamiltonian Dynamics. Whenever $\gamma \equiv 0$ and $F_{\text{ext}} = 0$, equation (3) reduces to the classical Newtonian equations of motion:

$$\dot{q} \,=\, p, \qquad \dot{p} \,=\, -\nabla_q V(q). \tag{5}$$

The flow Φ^t associated with the Hamiltonian H from (4) aloows to denote the solution process of (5), i.e., $x(t; x_0) = X_t(x_0) = \Phi^t x_0$ and the transition kernel is given by

$$p(t,x,C) \,=\, \chi_C\left(\Phi^t x\right), \tag{6}$$

where χ_C denotes the characteristic function of the set $C \subset \Gamma$. In this deterministic case, the equations of motion (5) model an *energetically closed system*, i.e., the Hamiltonian denotes the *total energy* of the system, which is preserved by the dynamics.

Deterministic Thermostatted Dynamics. In general, the term $\gamma(q,p)\,p$ represents the effect of some "thermostat" on the system. In "thermostatted molecular dynamics", one designs deterministic descriptions of open but conservative molecular systems contained in a heat bath by choosing $\gamma \neq 0$ and (deterministic) forces $F_{\text{ext}} \neq 0$ such that the solution of (3) conserves either kinetic or total energy [14].

Stochastic Langevin Dynamics. The most popular model for an *open system* with stochastic interaction with its environment is the so-called Langevin model [32]:

$$\dot{q} = p, \qquad \dot{p} = -\nabla_q V(q) - \gamma p + \sigma \dot{W}. \tag{7}$$

It is a special case of (3) with some constant friction $\gamma(q,p) \equiv \gamma > 0$ and an external force $F_{\mathrm{ext}} = \sigma \dot{W}_t$ given by a $3N$-dimensional Wiener process W_t with zero mean $\langle W_t \rangle = 0$ and correlation $\langle W_t W_s \rangle = \delta(t-s)$. The external stochastic force models the influence of the Brownian motion of the heat bath surrounding the molecular system. In this case, the internal energy H is not preserved, but the interplay between stochastic excitation and damping equilibrates the internal energy as we will see in Section 3.2.

2.2 Long-Term Behavior and Conformations

In principle, a discretization of (3) permits a simulation of single system trajectories once the initial state is given. However, numerical analysis of present discretizations restricts the validity of such single system trajectory simulations to only short time spans and to comparatively small discretization steps. The reason for this is two-fold: First, numerical long-term simulation is an ill-posed problem for the Hamiltonian systems under consideration [1], and second, no numerical integrator is available that allows stepsizes larger than a few femtoseconds—neither for Hamiltonian nor for Langevin dynamics [38, 33].

On the smallest time scales of about one femtosecond molecular dynamics consists of fast oscillations or fluctuations around equilibrium positions. In contrast to these fast fluctuations the term conformations describes meta-stable global states of the molecule, in which the *large scale geometric structure* is understood to be conserved. *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. Thus, the effective conformational dynamics occurs on time scales not accessible via long–term simulation. We thus have to abandon the trajectory-based approach of identifying conformations via long-term simulations. Instead, we use the dynamical properties of conformations to introduce a *set-oriented concept*:

Conformations are related to geometric structure given by the atomic positions. This means that conformations are subsets of the position space. Under additional consideration of the dynamical properties, we characterize *conformations as special "almost invariant" subsets in position space* in the following sense: An *invariant set* can never be left by the dynamical process under consideration. If conformations were *invariant sets* of the molecular dynamics, then transitions between different conformations would be *impossible.* Since transitions between conformations exist but are *rare*, we have to understand conformations as *almost invariant* sets of the molecular dynamics.

In [10], Dellnitz and Junge proposed to identify almost invariant subsets of discrete dynamical systems via specific eigenvectors of corresponding transfer operators. In order to make this intriguing idea applicable to the identification of conformations, we will introduce some notation, define transfer operators for molecular motion and link them to concepts of statistical mechanics.

3 Molecular Ensembles and Transfer Operators

We *in principle* always have to accept experimental measurement uncertainties when determining the initial state —all the positions and momenta— of some molecule. As a consequence, when modelling the physical reality, we have to propagate a *statistical ensemble* of molecular systems which represents the *distribution of possible initial states determined via the initial measurement.* The distribution may be described by some *time dependent probability density* $u = u(x,t)$ in phase space. In the following, the density u is always meant with respect to the measure μ; consequently, the probability within the ensemble to encounter a system $x \in \mathbf{X}$ in a subset $C \subset \mathcal{A}$ at time t is given by

$$\mathbf{P}_t[x \in C] \quad = \quad \int_C u(x,t)\,\mu(\mathrm{d}x). \tag{8}$$

Physical experiments allow for measuring relative frequencies in the ensemble, e.g., to determine the relative frequency of systems within the ensemble whose state lies in $C \subset \mathbf{X}$ at time t. The probability $\mathbf{P}_t[x \in C]$ corresponds to the relative frequency introduced above and is thus physically measurable— in contrast to the probability density $u(x,t)$. Whenever physicists use the phrase "probability density" they refer to the density from (8) *with respect to the Lebesgue measure* $\mathrm{d}x$. This means, whenever $u(x,t)$ is the density with respect to μ and, additionally, μ is absolutely continuous with respect to $\mathrm{d}x$ with density $d(x,t)$, then the *physical density* is $f(x,t) = u(x,t)d(x,t)$. Nevertheless, it is sometimes mathematically advantageous to consider densities with respect to specific measures particularly adapted to the Markov process under investigation.

3.1 Forward and Backward Transfer Operators

The evolution of a probability density $u = u(x,t)$ in state space $\mathbf{X}$ is governed by the (micro-) dynamics $\{X_t\}_{t\in\mathbf{M}}$ of each of the identically prepared molecular systems within the ensemble. We may describe the evolution by the *propagator* or *forward transfer operator*

$$P_t u(x) \quad = \quad u(x,t),$$

which maps the initial probability density $u(x) = u(x,0)$ to the density $u(x,t)$ at time t. Assume for the moment that the transition kernel of the process $\{X_t\}$ is absolutely continuous with respect to the probability measure μ, i.e., $p(t,x,\mathrm{d}y) = \int_C p(t,x,y)\,\mu(\mathrm{d}y)$. Since $p(t,x,y)$ denotes the "probability" of the process to move from x to y within the time t, the propagator should have the form

$$P_t u(y) \quad = \quad \int_{\mathbf{X}} p(t,x,y)\,u(x)\,\mu(\mathrm{d}x). \tag{9}$$

However, since in general the transition kernel will not be absolutely continuous, we proceed in a different way and define P_t via the well–known *backward transfer operator* [19]

$$T_t u(x) \quad = \quad \mathbf{E}_x[u(X_t)] = \int_{\mathbf{X}} u(y)\,p(t,x,\mathrm{d}y), \tag{10}$$

where $\mathbf{E}_x[u(X_t)]$ denotes the expectation of an observable $u : \mathbf{X} \to \mathbf{C}$ under the condition that the process $\{X_t\}$ has been started at $t = 0$ in x.

Consider T_t as an operator on $L^\infty_\mu(\mathbf{X})$ and P_t on $L^1_\mu(\mathbf{X})$, and let $\langle \cdot, \cdot \rangle_\mu$ denote the duality bracket between $L^\infty_\mu(\mathbf{X})$ and $L^1_\mu(\mathbf{X})$. Then, as a generalization of (9), the forward transfer operator P_t is *defined as the adjoint operator* $P_t = T_t^*$ of the backward transfer operator T_t [19], i.e.,

$$\langle T_t u, v \rangle_\mu \quad = \quad \langle u, P_t v \rangle_\mu, \quad \text{for all } u \in L^\infty_\mu(\mathbf{X}), v \in L^1_\mu(\mathbf{X}). \tag{11}$$

Since $p(t, x, \cdot)$ is a transition kernel, the thereby defined operator P_t is a *Markov operator* on $L^1_\mu(\mathbf{X})$. Furthermore, the semigroup property of the Markov process implies that $\{P_t\}_{t \in \mathbf{M}}$ is a semigroup of Markov operators.

In view of equations (9) and (10), the notion of "forward" and "backward" transfer operator becomes clearer. For the forward case, the state average with respect to u is taken over all initial states x, which are propagated forward in time, while for the backward case, the state average is taken over all final states y.

Invariant Measures and Stationary Densities A measure μ on $\mathbf{X}$ is called *invariant* with respect to the process $\{X_t\}$, if

$$\mu(C) \quad = \quad \int_X p(t, x, C)\, \mu(dx), \quad \text{for all } C \in \mathcal{A} \text{ and } t \in \mathbf{M}.$$

Due to the properties of the transition kernel and the definition of the backward transfer operator, we have—independent of the measure μ—for every $t \in \mathbf{M}$,

$$T_t \chi_{\mathbf{X}} \;=\; \chi_{\mathbf{X}}.$$

The above equality does in general not hold for the forward transfer operator, because P_t depends via (11) on the probability measure μ. However, if we assume μ to be invariant, we also get

$$P_t \chi_{\mathbf{X}} \;=\; \chi_{\mathbf{X}} \tag{12}$$

for all $t \in \mathbf{M}$. In other words, $\chi_{\mathbf{X}}$ is an invariant density of P_t, whenever μ is invariant.

Remark. Suppose additionally that μ admits a density d with respect to the Lebesgue measure. Let moreover the ensemble be distributed according to μ so that d is the stationary physical probability density of the ensemble. Then, $f(\cdot, 0) = \chi_C \cdot d$ denotes the physical density of the *subensemble* of all systems being in $C \subset \mathbf{X}$ at some time $t = 0$. Since P_t denotes the evolution of the ensemble in time t, the physical density of the subensemble at time t is given by $f(\cdot, t) = P_t \chi_C \cdot d$. In contrast to this, $T_t \chi_C = p(t, \cdot, C)$ denotes the probability density to *access* C at time t. This again emphasizes the difference in interpretation between P_t and T_t: P_t denotes the physically interpretable propagator of the ensemble and is defined with respect to some measure μ, while T_t denotes the transfer operator related to the Markov process (independent of the measure μ) as usually considered in stochastic theory.

3.2 Canonical Ensemble

Most experiments on molecular systems are performed under the conditions of constant temperature $\mathcal{T}$ and volume. The corresponding ensemble density (with respect to the Lebesgue measure on $\mathbf{X}$) is the *canonical density* f_{can} associated with the Hamiltonian H:

$$f_{\text{can}}(x) \;=\; \frac{1}{Z}\exp\left(-\beta\, H(x)\right), \qquad Z = \int_{\Gamma}\exp\left(-\beta\, H(x)\right) dx, \tag{13}$$

where $\beta = 1/k_B\mathcal{T}$ denotes the inverse temperature and k_B Boltzmann's constant. Since H was assumed to be separable, f_{can} factorizes in a product of two densities $\mathcal{P}$ and $\mathcal{Q}$:

$$f_{\text{can}}(x) \;=\; \underbrace{\frac{1}{Z_p}\exp\left(-\frac{\beta}{2}\,p^T M^{-1} p\right)}_{=\mathcal{P}(p)}\;\underbrace{\frac{1}{Z_q}\exp\left(-\beta\, V(q)\right)}_{=\mathcal{Q}(q)}. \tag{14}$$

Since we are interested in the canonical ensemble, we define the *canonical probability measure* induced by the canonical density:

$$\mu_{\text{can}}(\mathrm{d}x) \;=\; f_{\text{can}}(x)\,\mathrm{d}x.$$

It will turn out advantageous to consider transfer operators acting on weighted function spaces with respect to μ_{can}.

3.3 Transfer Operators and the Canonical Ensemble

In general, an equation of motion for the process $\{X_t\}$ implies an equation of motion for a probability density u. We will see below that the processes induced by both, the Hamiltonian dynamics and the Langevin dynamics, leave the canonical measure μ_{can} invariant. Since we are interested in describing fluctuations within the canonical ensemble, we thus define the forward transfer operator with respect to the canonical probability measure μ_{can}, i.e., acting on $L^1_{\mu_{\text{can}}}(\mathbf{X})$.

Langevin Dynamics. The process induced by the Langevin equation (7) leaves the canonical measure μ_{can} corresponding the the inverse temperature β invariant, if the noise and damping constants satisfy [32]:

$$\beta \;=\; \frac{2\gamma}{\sigma^2}. \tag{15}$$

The evolution of $u = u(x,t)$ with respect to μ_{can} (compare introduction to Section 3) is governed by the well–known *Fokker–Planck equation* [32]:

$$\partial_t u \;=\; \left(\underbrace{\frac{\sigma^2}{2}\Delta_p \;-\; p\cdot\nabla_q \;+\; \nabla_q V\cdot\nabla_p \;-\; \gamma p\cdot\nabla_p}_{=A}\right) u. \tag{16}$$

As a consequence, the Fokker–Planck operator A is the infinitesimal generator of the semigroup of forward transition operators $\{P_t\}_{t\in\mathbf{R}_0^+}$ acting on $L^1_{\mu_{\text{can}}}(\mathbf{X})$ with

$$P_t u \;=\; \exp(tA)u \tag{17}$$

and, since $\mu_{\rm can}$ is invariant, we have $P_t \chi_\Gamma = \chi_\Gamma$.

Moreover, under certain conditions on the potential V (systems of type (B) with potential $V \in C^\infty(\mathbf{X})$), this is the *unique stationary density* and the semigroup $\{P_t\}_{t\in\mathbf{R}_0^+}$ is asymptotically stable [23], i.e., $P_t u \to \chi_\Gamma$ for $t \to \infty$ and every density $u \in L^1_{\mu_{\rm can}}(\Gamma)$. Due to this property, the Langevin equation is the most prominent stochastic model for a heat bath driven relaxation of molecular ensembles to the canonical ensemble.

Hamiltonian Dynamics. The Hamiltonian equations of motion are the deterministic analogue of the Langevin equations with $\gamma = 0$ and $\sigma = 0$. As for the Langevin dynamics, the canonical probability measure $\mu_{\rm can}$ is invariant under the dynamics. Using $\gamma = \sigma = 0$, the equation of motion (16) for the probability density u reduces to the *Liouville equation* corresponding to the Hamiltonian H:

$$\partial_t u \;=\; \Bigg(\underbrace{-p\cdot\nabla_q \,+\, \nabla_q V\cdot\nabla_p}_{=i\mathcal{L}}\Bigg)\, u \tag{18}$$

where $\mathcal{L}$ denotes the well–known Liouville operator [25]. The solution of (18) satisfies $u(x, t+s) = u(\Phi^{-t}x, s)$ for all $t, s \in \mathbf{R}_0^+$. Using (17), the forward transfer operator acting on $L^1_{\mu_{\rm can}}(\Gamma)$ is given by

$$P_t u(x) \;=\; \exp(it\mathcal{L})u(x) \;=\; u(x,t) \;=\; u\left(\Phi^{-t}x\right), \tag{19}$$

which is just the definition of the *Frobenius–Perron operator* corresponding to the Hamiltonian flow Φ^t [26]. Additionally, inserting the transition kernel (6) in the definition (10) of the backward transfer operator yields

$$T_t u(x) \;=\; u\left(\Phi^t x\right), \tag{20}$$

which is simply the *Koopman operator* corresponding to Φ^t [26]. Equations (19) and (20) illustrate that P_t is the adjoint operator of T_t as discussed above.

As we have seen, the canonical density $f_{\rm can}$ induces the *invariant measure* $\mu_{\rm can}$ of the deterministic Hamiltonian dynamics. However, there are infinitely many other invariant measures induced by densities of the form $f(x) = \mathcal{F}(H(x))$ for some smooth function $\mathcal{F} : \mathbf{R} \to [0,1]$ of the Hamiltonian. Due to this ambiguity, pure Hamiltonian dynamics is not appropriate for modelling the relaxation of molecular ensembles to one specific ensemble, in our case the canonical ensemble. This observation corresponds to the fact that, for solving the Liouville equation, we have to specify an initial density $u(\cdot, t = 0)$. Physically, the specification of an initial density corresponds to an *initial experimental preparation* of the ensemble due to (8). Thus, selecting one of the possible invariant densities means the specific initial preparation of a stationary ensemble.

4 Almost Invariant Sets of Molecular Ensembles

Assume in this section that the molecular motion is described by a Markov process $\{X_t\}_{t\in\mathbf{M}}$ that leaves the probability measure μ invariant. Moreover, assume that the Markov process is initially distributed according to μ, i.e., the probability to find the process at time $t = 0$ in a subset $C \in \mathcal{A}$ is given by

$$\mathbf{P}[X_0 \in C] \quad = \quad \mu(C)$$

(see introduction to Section 3).

4.1 Ensemble Transition Probabilities

The *transition probability* $p(s, C, D)$ within the ensemble from $C \in \mathcal{A}$ to $D \in \mathcal{A}$ within the time span s is defined as the conditional probability

$$p(s, C, D) \; = \; \mathbf{P}[X_s \in D \,|X_0 \in C] \; = \; \frac{\mathbf{P}[X_s \in D \text{ and } X_0 \in C]}{\mathbf{P}[X_0 \in C]}. \tag{21}$$

The similar symbols for both the transition probability $p(s, C, D)$ and for the transition kernel $p(s, x, C)$ corresponding to the process emphasizes the strong relation to (2), which, in addition to the above assumption, allows to rewrite the transition probability as

$$p(s, C, D) \quad = \quad \frac{1}{\mu(C)} \int_C p(s, x, D)\, \mu(\mathrm{d}x). \tag{22}$$

The transition probabilities quantify the *dynamical fluctuations within the stationary ensemble*. Using the duality bracket $\langle \cdot, \cdot \rangle_\mu$ between $L^\infty_\mu(\mathbf{X})$ and $L^1_\mu(\mathbf{X})$, the definitions of the transfer operators T_t and P_t yield

$$p(s, C, D) \quad = \quad \frac{\langle T_s \chi_D, \chi_C \rangle_\mu}{\langle \chi_C, \chi_C \rangle_\mu} \; = \; \frac{\langle \chi_D, P_s \chi_C \rangle_\mu}{\langle \chi_C, \chi_C \rangle_\mu}. \tag{23}$$

The above defined transition probabilities can be *measured* via the following two–step experiment on the ensemble:

1. *Pre-Selection:* Select from the ensemble all such systems with states $x \in C \in \mathcal{A}$. This selection prepares a new ensemble, which is described by the probability measure

   $$\mu_C(D) \quad = \quad \frac{1}{\mu(C)}\, \mu(C \cap D), \quad D \in \mathcal{A}.$$

2. *Transition-Counting:* After the time span s, determine the *relative frequency* of systems in the ensemble μ_C with states in C. Since all systems evolve due to the process $\{X_t\}_{t \in \mathbf{M}}$, this relative frequency is equal to

   $$\int_{\mathbf{X}} p(s, x, C)\, \mu_C(dx) \quad = \quad p(s, C, C).$$

4.2 Conformations as Almost Invariant Subsets

We now aim at a dynamical characterization of conformations within the ensemble; this characterization will be based on the notion of almost invariance. As already mentioned, we have to define almost invariance in terms of ensemble dynamics rather than in terms of the duration of stay of a single system.

Following [10], We call some subset $C \in \mathcal{A}$ *almost invariant*, whenever the fraction of systems within the ensemble that stay in C after some characteristic time span $s \in \mathbf{M}$ is close to 1:

$$C \text{ almost invariant} \iff p(s, C, C) \approx 1.$$

This definition of almost invariance guarantees that its "degree" $p(s, C, C)$ can be *measured* via the two-step experiment introduced above.

Almost invariance may equivalently be characterized by $p(s, C, \mathbf{X} \setminus C) \approx 0$, which allows to relate it to the semigroup of forward transfer operators $\{P_t\}_{t \in \mathbf{M}}$ by the following general identity [37]:

$$\left\| P_s \frac{1}{\mu(C)} \chi_C - \frac{1}{\mu(C)} \chi_C \right\|_1 = 2\, p(s, C, \mathbf{X} \setminus C). \tag{24}$$

4.3 Identification Strategy

By definition, P_s is a Markov operator and consequently, its $L^1_\mu(\mathbf{X})$–spectrum is contained in the unit ball $\{\lambda \in \mathbf{C} : |\lambda| \leq 1\}$. Every invariant density $u \in L^1_\mu(\mathbf{X})$ of P_s satisfies $P_s u = u$ and therefore is an eigenvector of P_s corresponding to the eigenvalue $\lambda = 1$, the so–called *Perron root*. Since μ is assumed to be invariant, in particular $u = \chi_X$ is an invariant density.

Whenever a proper subset C of $\mathbf{X}$ is invariant under the Markov process, i.e., $p(t, x, \mathbf{X} \setminus C) = 0$ for all $x \in C$, the density $u = \chi_C / \mu(C)$ is an eigenvector corresponding to $\lambda = 1$.

Due to our above characterisation, the set $C \in \mathcal{A}$ is almost invariant if $p(\tau, C, \mathbf{X} \setminus C) \approx 0$, which via formula (24) implies that $\chi_C / \mu(C)$ is an approximate invariant density, i.e., an approximate normalized eigenvector associated with an eigenvalue close to the Perron root $\lambda = 1$. This motivates the following *algorithmic strategy*:

> *Invariant* sets can be identified via eigenvectors corresponding to the Perron root $\lambda = 1$, while *almost invariant* sets may be identified via eigenvectors corresponding to eigenvalues $|\lambda| < 1$ close to the Perron root $\lambda = 1$.

This strategy has first been proposed by Dellnitz and Junge [10] for discrete dynamical systems with weak random perturbations and has been successfully applied to molecular dynamics in different contexts [35, 36, 34]. It will be justified in more detail in Section 5.4 below, where more information about the properties of the transfer operators of interest will be available.

It is important to notice that almost invariance is defined herein with respect to some physically selected invariant probability measure μ that describes the stationary ensemble under consideration. Assume that the process $\{X_t\}$ admits another invariant measure ν, which, for the sake of simplicity, is absolutely continuous with respect to μ with density $d \in L^1_\mu(\mathbf{X})$. Then, the density $u = \chi_{\mathbf{X}} d$ is an eigenvector of P_s corresponding to $\lambda = 1$. As a consequence, one will not be able to decide in general whether some eigenvector corresponding to an eigenvalue $|\lambda| < 1$ close to the Perron root is related to an almost invariant subset of the ensemble represented by μ or rather by ν. Thus, the above algorithmic strategy requires uniqueness of the invariant measure. For its numerical realization via an eigenvalue problem we moreover need that the remaining spectrum of P_s is strictly bounded away from the Perron root, i.e., $\lambda = 1$ must be an isolated, simple eigenvalue of P_s. Additionally, the physical interpretation of the ensemble excludes other eigenvalues than $\lambda = 1$ *on* the unit circle or, equivalently, we exclude asymptotic periodicity of P_s.

We introduce the following two *fundamental conditions* on the forward transfer operator P_s that are sufficient to guarantee the desired properties:

(C1) P_s is asymptotically stable, i.e., $(P_s)^n u \to \chi_{\mathbf{X}}$ in $L^1_\mu(\mathbf{X})$ for every density $u \in L^1_\mu(\mathbf{X})$ as $n \to \infty$.

(C2) The essential spectrum of P_s is strictly bounded away from $|\lambda| = 1$.

These conditions exclude some very prominent models for molecular motion. For example, in the pure Hamiltonian case the invariant density is *not* unique in $L^1(\mathbf{X})$, and, worse, the spectrum of the Frobenius–Perron operator P_s in $L^1(\mathbf{X})$ lies *on* the unit circle[1]. Despite these fundamental problems, Deuflhard et al. computed almost invariant subsets of Hamiltonian systems in the above sense with quite intriguing results [11]. However, they did not use the exact Hamiltonian flow Φ^t but added small, but significant perturbations originating from time discretization errors and the related energy fluctuations. It is a widely accepted approach to model such discretization effects by small random perturbations. Under appropriate conditions, the thereby resulting transfer operator is compact and may have a unique invariant measure (see [10]). In [11] another interpretation of this approach via a sequence of nested function spaces based on subsequent coverings of the energy cell is indicated.

There are other models that satisfy our conditions without additional artificial perturbations. An example is the Langevin model introduced above. For appropriate systems (see above), its unique invariant measure is the canonical measure. Hence, application of our algorithmic strategy to the Langevin model seems to allow to attack chemically interesting systems. However, there is another condition which has to be considered and prevents the Langevin model from being a good starting point: *Chemical conformations are usually understood to be objects in position space* Ω. Therefore, a proper model needs to yield a family of forward transition operators, which are defined on $\mathbf{X} = \Omega$ rather than on the entire phase space Γ of the molecular systems.

5 Conformational Dynamics in Position Space

Since conformations are objects in position space, this section is devoted to an adequate theory of ensemble dynamics in position space, including two examples. We introduce two (reduced) Markov processes in position space and define the corresponding transfer operators. Due to physical reasons and as a consequence of (23), we restrict ourselves to the semigroup of forward transfer operators or propagators $\{P_t\}_{t \in \mathbf{M}}$ for the canonical ensemble $\mathcal{Q}$.

[1] Here, $L^1(\mathbf{X})$ may be replaced by $L^1_\mu(\mathbf{X})$ where μ may stand for μ_{can} or for any other invariant measure of the Hamiltonian flow Φ^t that is absolutely continuous with respect to the Lebesgue measure on the phase space $\mathbf{X}$.

5.1 Positional Dynamics and Transfer Operators

Let $(\Omega, \mathcal{A}, \mu_{\mathcal{Q}})$ denote the positional probability space with $\mu_{\mathcal{Q}}(\mathrm{d}q) = \mathcal{Q}(q)\mathrm{d}q$ and refer by $L^r{}_{\mathcal{Q}}(\Omega)$ for $r = 1, 2, \ldots, \infty$ to the corresponding function spaces with respect to the canonical measure $\mu_{\mathcal{Q}}$. Note that $L^2{}_{\mathcal{Q}}(\Omega)$ is a Hilbert space with scalar product

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

and induced norm $\|u\|^2_{\mathcal{Q}} = \langle u, u\rangle_{\mathcal{Q}}$.

As a consequence of Subsection 4.3 we have to transform the *state space* dynamics into a pure *position space* dynamics. Assume that the transformed dynamics of a single system in Ω is described by a (homogeneous) Markov process $\{Q_t\}_{t\in\mathbf{M}}$ with stochastic transition kernel $p(t, q, C)$, invariant measure $\mu_{\mathcal{Q}}$ and initial distribution $\mathbf{P}[Q_0 \in C] = \mu_{\mathcal{Q}}(C)$. Then, the semigroup of forward transfer operators $\{P_t\}_{t\in\mathbf{M}}$ for the canonical ensemble is given by $P_t : L^r{}_{\mathcal{Q}}(\Omega) \to L^r{}_{\mathcal{Q}}(\Omega)$ such that for all $C \in \mathcal{A}$

$$\int_C P_t u(q)\mathcal{Q}(q)\, \mathrm{d}q \quad = \quad \int_\Omega u(q)p(t, q, C)\mathcal{Q}(q)\, \mathrm{d}q \tag{25}$$

under suitalbe conditions of the integrability of the transition kernel. In the following, we will consider P_t mainly as an operator acting on the Hilbert space $L^2{}_{\mathcal{Q}}(\Omega)$, since—as we will see below—the corresponding scalar product may reveal possible additional properties of P_t and allows to define Galerkin projections for the discretization procedure.

We conclude by stating all assumptions on the transfer operators, which result from the requirements of Subsection 4.3:

(C1) P_s is asymptotically stable, i.e., $(P_s)^n\, u \to \chi_\Omega$ in $L^1_{\mathcal{Q}}(\Omega)$ for $n \to \infty$ and every density $u \in L^1_{\mathcal{Q}}(\Omega)$. This implies that $\lambda = 1$ is an isolated, simple eigenvalue in $L^2{}_{\mathcal{Q}}(\Omega)$.

(C2) The essential spectrum of P_s in $L^2{}_{\mathcal{Q}}(\Omega)$ is strictly bounded away from $|\lambda| = 1$.

5.2 Discrete Time Markov Chain

The first example of a reduced positional dynamics is based on the Hamiltonian equation of motion within the canonical ensemble f_{can} (14) and a characterization of conformations as special almost invariant subsets. A subset $C \subset \Omega$ of the position space is called almost invariant, if the enlarged "cylindrical" subset $C \times \mathbf{R}^d \subset \Gamma$ of the state space is almost invariant with respect to the Hamiltonian dynamics.

Let $p_{\mathbf{X}}(t, x, A)$ denote the stochastic transition kernel of the Markov process in state space (see (6)). Fix an observation time span $\tau > 0$. Then, $C \subset \Omega$ is almost invariant (with respect to τ), if $p_{\mathbf{X}}(\tau, C \times \mathbf{R}^d, C \times \mathbf{R}^d) \approx 1$. For *fixed* τ, this definition can be used to derive a reduced dynamics in position space. For two subsets $C, D \subset \Omega$, we have due to (6) and (22):

$$
\begin{aligned}
& p_{\mathbf{X}}(\tau, C \times \mathbf{R}^d, D \times \mathbf{R}^d) \\
&\quad = \frac{1}{\int_{C\times\mathbf{R}^d} f_{\mathrm{can}}(x)\,\mathrm{d}x} \int_{C\times\mathbf{R}^d} \chi_{D\times\mathbf{R}^d}(\Phi^\tau(x)) f_{\mathrm{can}}(x)\,\mathrm{d}x \\
&\quad = \frac{1}{\int_C \mathcal{Q}(q)\,\mathrm{d}q} \int_C \underbrace{\int_{\mathbf{R}^d} \chi_D(\Pi_q \Phi^\tau(q,p))\mathcal{P}(p)\,\mathrm{d}p}_{=:p^\tau_\Omega(1,q,D)}\ \mathcal{Q}(q)\,\mathrm{d}q. \qquad (26) \\
&\quad = p^\tau_\Omega(1, C, D),
\end{aligned}
$$

where Π_q denote the projection onto the position space. A comment on the dependence of the one-step transition probability $p^\tau_\Omega(1, C, D)$ on the observation time span τ can be found in the remark below. It is easy to show, that $p^\tau_\Omega(1, q, D)$ is a transition kernel and thus defines a *discrete time* Markov process $\{Q_n\}_{n\in\mathbf{N}_0}$ on the position space Ω. Furthermore, $\{Q_n\}$ satisfies inductively for all $n \in \mathbf{N}_0$ the stochastic dynamical equation (SDE) [34]

$$
Q_{n+1} \quad = \quad \Pi_q \Phi^\tau(Q_n, P_n) \qquad (27)
$$

with P_n chosen randomly according to the momenta distribution $\mathcal{P}$. The SDE (27) is the reduced positional dynamics that we were looking for. In mathematical terms, it corresponds to a Hamiltonian motion with randomly chosen momenta at discrete (physical) times $\tau, 2\tau, \ldots$. As shown in [34], $\{Q_n\}$ leaves the canonical ensemble $\mathcal{Q}$ invariant.

Via Equation (25), the transition kernel also defines a discrete time semigroup of transition operators $\{P_n\}_{n\in\mathbf{N}_0}$ on $L^r{}_{\mathcal{Q}}(\Omega)$. Exploiting that Φ^τ is a reversible, symplectic and $\mu_{\mathcal{Q}}$ invariant mapping (see (19) and below, and [34]), we get

$$
P_1 u(q) \quad = \quad \int_{\mathbf{R}^d} u(\Pi_q \Phi^{-\tau}(q,p))\mathcal{P}(p)\,\mathrm{d}p \qquad (28)
$$

for $u \in L^r{}_{\mathcal{Q}}(\Omega)$. For all systems of type (P), P_t satisfies the requirements stated in Subsection 5.1 [34]; furthermore, it is self–adjoint on $L^2{}_{\mathcal{Q}}(\Omega)$ due to reversibility and symplecticness of the Hamiltonian flow [34]. As a consequence, the $L^2{}_{\mathcal{Q}}(\Omega)$–spectrum of P_t is real–valued, bounded and contained in the interval $(-1, 1]$; the essential spectrum is bounded away from 1.

Remark. In (26), we have defined the *one step* transition kernel $p^\tau_\Omega(1, q, D)$ for fixed τ. Changing the observation time to σ results in a new one step transition kernel $p^\sigma_\Omega(1, q, D)$. In contrast to that, the n–term transition kernel $p^\tau_\Omega(n, q, D)$ is defined recursively by the Chapman–Kolmogorov equation (1). In general, $p^{2\tau}_\Omega(1, q, D) \neq p^\tau_\Omega(2, q, D)$ and, consequently, $P_1^{2\tau} \neq P_2^\tau$, where the superscript indicates the corresponding observation time span (for an example, see [34, Sec. 3.7.1]). In terms of the SDE (27), this is not surprising, since $P_1^{2\tau}$ includes only one choice of momenta according to $\mathcal{P}$, while P_2^τ does include two.

5.3 High–Friction Langevin Dynamics

The second example of a reduced positional dynamics is based on the Langevin equation. We will see that in a specific high friction limit $\gamma \to \infty$ the Langevin equation acting on the state space reduces to the so–called high–friction Langevin equation acting only on the position space.

Consider the Langevin equation (7) written in second order form

$$\ddot{q} \;=\; -\nabla_q V(q) - \gamma\dot{q} + \sigma\dot{W}. \tag{29}$$

For the high friction case, let ϵ be a small positive parameter and consider the transformed friction γ/ϵ. In order to preserve the temperature $\mathcal{T}$ of the surrounding heat bath, we simultaneously have to scale the white noise constant $\sigma \mapsto \sigma/\sqrt{\epsilon}$ due to (15). This yields

$$\ddot{q} \;=\; -\nabla_q V(q) - \frac{\gamma}{\epsilon}\dot{q} + \frac{\sigma}{\sqrt{\epsilon}}\dot{W}.$$

After rescaling the time according to $t \mapsto \epsilon\, t$ one gets

$$\epsilon^2\ddot{q} \;=\; -\nabla_q V(q) - \gamma\dot{q} + \sigma\dot{W}.$$

For systems of type (B), for which the gradient of V satisfies a global Lipschitz condition, and $0 < \epsilon \ll 1$ one may neglect the ϵ^2–term [30, Thm. 10.1] and finally get the high–friction Langevin equation[2]

$$\dot{q} = -\frac{1}{\gamma}\nabla_q V(q) + \frac{\sigma}{\gamma}\dot{W} \tag{30}$$

modelling the high friction positional dynamics within the canonical ensemble. The stochastic differential system (30) defines a *continuous time* Markov process $\{Q_t\}_{t\in\mathbf{R}_0^+}$ on the position space Ω with corresponding transition kernel $p(t,q,C)$. The process leaves the canonical measure μ_Q invariant [32].

As for the general Langevin dynamics (16) in state space, the continuous time semigroup of forward transition operators $\{P_t\}_{t\in\mathbf{R}_0^+}$ may be defined in terms of its infinitesimal generator [19]:

$$A \;=\; \frac{\sigma^2}{2\gamma^2}\Delta_q - \frac{1}{\gamma}\nabla_q V(q)\cdot\nabla_q \tag{31}$$

acting on a suitable subspace of $L^r{}_Q(\Omega)$. As a consequence, one gets

$$\begin{aligned} P_t : L^r{}_Q(\Omega) \;&\to\; L^r{}_Q(\Omega) \\ u \;&\mapsto\; P_t u \;=\; \exp(tA)u. \end{aligned} \tag{32}$$

Thus, every probability density $u = u(q,t)$ with respect to μ_Q evolves according to the Fokker–Planck equation $\partial_t u = Au$ and its solution is formally given by (32).

It is shown in [36, 4] that for systems of type (B) the semigroup of forward transition operators satisfies the requirements of Subsection 5.1. Furthermore, $\{P_t\}_{t\in\mathbf{R}_0^+}$ is a self–adjoint semigroup in $L^2_Q(\Omega)$, since the infinitesimal generator A is self–adjoint with respect to $\langle\cdot,\cdot\rangle_Q$ [36].

[2] In contrast to the usual quasistatic approximation in mechanics, we cannot simply assume that the accelaration $\ddot{q}$ is bounded since the white noise process is unbounded. However, the investigation in [30] shows that the Langevin solution $q^\epsilon_{\mathrm{Lan}}(t;q_0,p_0)$ and the solution $q_{\mathrm{fric}}(t;q_0)$ of (30) satisfy for all p_0, with probability one: $\lim_{\epsilon\to 0}|q_{\mathrm{fric}}(t) - q^\epsilon_{\mathrm{Lan}}(t)| = 0$ uniformly for t in compact subintervals of $[0,\infty)$.

Remark. The physical density $f(q,t) = u(q,t)\mathcal{Q}(q)$ (see introduction to Section 3) evolves according to the Fokker–Planck equation $\partial_t f = A^* f$, where A^* denotes the formal adjoint of A, i.e., $A^* = \sigma^2/(2\gamma^2)\Delta_q + 1/\gamma\nabla_q V(q)\cdot\nabla_q + 1/\gamma\Delta_q V(q)$ (see [19]).

Almost Invariance and First Exit Times. Due to experimental requirements, almost invariance of conformations is defined at *discrete points in time* (see Eq. (21)):

$$p_{\mathrm{discr}}(t,C,C) \quad = \quad \frac{\mathbf{P}[Q_s \in C : s = 0 \text{ and } s = t]}{\mathbf{P}[Q_0 \in C]}. \tag{33}$$

This definition also holds for the continuous time Markov processes. However, one could alternatively want to characterize almost invariance of conformations based on *continuous time observations*:

$$p_{\mathrm{cont}}(t,C,C) \quad = \quad \frac{\mathbf{P}[Q_s \in C : \text{ for all } s \in [0,t]]}{\mathbf{P}[Q_0 \in C]}. \tag{34}$$

Obviously, the two definitions will in general produce different result, since the former definition does not take into account fluctuations in between the two instances. However, in contrast to the latter definition, the former one can be realized by the two-step experiment from Section 4.1.

Mathematically, both characterizations are closely related by Fokker–Planck equations on appropriate function spaces. Let τ_C^q denote the *first exit time* of the Markov process $\{Q_t\}_{t\in\mathbf{R}_0^+}$, started at time zero in $q \in C$, from an open subset $C \subset \Omega$,

$$\tau_C^q \quad = \quad \inf\{t \geq 0 : Q_t(q) \notin C\}. \tag{35}$$

For open, bounded subsets C with sufficiently smooth boundary ∂C the distribution of exit times $v_C(q,t) = \mathbf{P}[\tau_C^q > t] = \mathbf{P}[Q_s(q) \in C : \text{ for all } s \in [0,t]]$ for $q \in C$ is given by the Fokker–Planck equation on $C \cup \partial C$ with *Dirichlet boundary conditions*:

$$\partial_t v = Av, \qquad v(\cdot,0) = \chi_C \ \text{ and } \ v(\cdot,t) = 0 \text{ for all } t \geq 0.$$

In contrast, $u_C(q,t) = \mathbf{P}[Q_s(q) \in C : s = 0 \text{ and } s = t]$ satisfies the Fokker–Planck *Cauchy problem* on Ω:

$$\partial_t u = Au, \qquad u(\cdot,0) = \chi_C$$

(with implicit "transparent boundary conditions"). With respect to the above two characterization of almost invariance, we finally get

$$p_{\mathrm{discr}}(t,C,C) \quad = \quad \frac{1}{\mu_{\mathcal{Q}}(C)}\int_C u_C(q,t)\mathcal{Q}(q)\,\mathrm{d}q$$

and

$$p_{\mathrm{cont}}(t,C,C) \quad = \quad \frac{1}{\mu_{\mathcal{Q}}(C)}\int_C v_C(q,t)\mathcal{Q}(q)\,\mathrm{d}q.$$

5.4 Justification of the Algorithmic Strategy

Here, we want to pick up the algorithmic strategy presented in Section 4.3 and state more precisely how one can use eigenvectors corresponding to eigenvalues near the Perron root 1 in order to identify almost invariant subsets. In the following, we fix a time $s \in \mathbf{M}$ and—in accordance with the properties of the above two examples—we assume that the transition operator P_s is self–adjoint in $L^2_{\mathcal{Q}}(\Omega)$. Moreover, for the sake of simplicity, we restrict our considerations to the case that the Perron root is "nearly two–fold degenerate", i.e., we assume that the spectrum of P_s has the form

$$\sigma(P_s) \subset [-r, r] \cup \{\lambda_2\} \cup \{1\},$$

with $0 < r < \lambda_2 < \lambda_1 = 1$; furthermore, we assume that λ_1 and λ_2 are simple eigenvalues. The eigenvector corresponding to $\lambda_1 = 1$ is χ_Ω, while we denote the eigenvector corresponding to λ_2 by $\phi \in L^2_{\mathcal{Q}}(\Omega)$ with normalization $\|\phi\|_{\mathcal{Q}} = 1$. Note that $\langle \phi, \chi_\Omega \rangle_{\mathcal{Q}} = 0$.

Nonrigorous Approach. One intuitive idea is to interpret almost invariance as "perturbed invariance". Therefore, we assume that the above transition operator results from a *continuous* perturbation of some self-adjoint Markov operator $\tilde{P}$ with degenerate, two-fold Perron root and invariant measure μ. If the degeneracy of the Perron root is caused by the existence of two disjoint *invariant* sets, say C and $C^c = \Omega \setminus C$, the eigenspace E_1 of the Perron root is spanned by the eigenvectors χ_C and χ_{C^c}. Neither C nor C^c are invariant sets of P_s, however, $\chi_C/\mu(C)$ and $\chi_{C^c}/\mu(C^c)$ remain to be "approximative" invariant densities of P_s, in the sense that (compare Section 4.3)

$$\left\| P_s \frac{1}{\mu(C)} \chi_C - \frac{1}{\mu(C)} \chi_C \right\|_1 \approx 0.$$

By means of the *general formula* (24), this implies that C as well as C^c are *almost invariant* sets of P_s. Since χ_Ω is a common eigenvector of $\tilde{P}$ and P_s, we choose another orthonormal basis of $E_1 = \mathrm{span}\{\chi_\Omega, u_C\}$ with

$$u_C = \sqrt{\frac{\mu(C^c)}{\mu(C)}} \chi_C - \sqrt{\frac{\mu(C)}{\mu(C^c)}} \chi_{C^c}. \tag{36}$$

Since P_s is assumed to be a continuous perturbation of $\tilde{P}$, we have to expect that the so-defined u_C is an approximation of the eigenvector ϕ of P_s corresponding to λ_2. This motivates the algorithmic strategy to identify the almost invariant sets via the second eigenvector ϕ (or $-\phi$) according to

$$C \approx \{q : \ \phi(q) > 0\} \quad \text{and} \quad C^c \approx \{q : \ \phi(q) \leq 0\}. \tag{37}$$

For more details concerning the *identification algorithm* for the more general case see [13].

Rigorous Approach. Although the perturbation analysis yields an intuitive understanding of the form of the second eigenvector of P_s, we subsequently will *not* assume any kind of perturbation embedding of P_s but rather proceed in another way towards a rigorous justification of the following "equivalence":

$$
\boxed{\begin{array}{ccc} \text{Decomposition into almost} & & \text{Eigenvalue cluster } \{1,\lambda_2\} \text{ separated} \\ \text{invariant subsets } \Omega = C \cup C^c : & & \text{from remaining spectrum:} \\ p(s, C, C^c) \approx 0, & \Longleftrightarrow & \\ C \approx \{q: \ \phi(q) > 0\} & & \epsilon = \frac{1-\lambda_2}{1-r} \ll 1 \end{array}} \tag{38}
$$

The following rigorous statements are closely similar to the results of E.B. Davies [7]. To simplify reference to his results, let us denote by η_2 and ρ the positive values with

$$
\lambda_2 \;=\; \exp(-s\eta_2) \quad \text{and} \quad r \,=\, \exp(-s\rho),
$$

where s denotes the initially fixed time span. For the "$\Leftarrow$"-direction in (38), we assume that $\epsilon = (1-\lambda_2)/(1-r)$ is small enough, and introduce $c = \|\phi\|_\infty$ satisfying $c \geq 1$. Due to [7],[3] there exists $C \in \mathcal{A}$ given by $C = \{q : \phi(q) > 0\}$ such that $\frac{1}{2c^2} \leq \mu(C) \leq 1 - \frac{1}{2c^2}$ and

$$
\|\phi \,-\, u_C\|_2 \;\leq\; 4c\sqrt{\epsilon}.
$$

Furthermore, the subset C is almost invariant with

$$
p(ns, C, C^c) \,\leq\, K\,\epsilon\,(1 + \rho n s), \quad \text{for all } n \in \mathbf{N},
$$

where K depends on c and $\mu(C)$, and is independent of ϵ.

For the "$\Rightarrow$"-direction in (38), we assume that C is almost invariant with

$$
p(ns, C, C^c) \;\leq\; K\,\delta\,(1 + \rho n s), \quad \text{for all } n \in \mathbf{N},
$$

with $K = \frac{1-\mu(C)}{12}$ and sufficiently small $\delta > 0$. Then, we again get that u_C approximates the second eigenvector in the sense that $\|\phi - u_C\|_2 < \sqrt{2\delta}$, and that, due to Thm. 5 in [7], $0 < \eta_2/\rho < \delta$ implying

$$
\epsilon \;<\; \frac{1 - r^\delta}{1 - r}.
$$

Thus, the formal equivalence (38) can be taken seriously. The above statement can be generalized to the situation of more than one eigenvalue close to the Perron root, but bounded away from the remaining part of the spectrum (see [8]).

Remark. We are aware of the fact that the above assumption $\epsilon \ll 1$ on the distribution of the eigenvalues is quite restrictive. However, we observed intriguing results of the identification strategy even for situations corresponding to ϵ-values close to 1 [13].

[3]The proofs of Thms. 3 and 5 of [7] have to be adapted to our situation. In the proof of Thm. 3, the arguments using the generator H have to be replaced by analogous arguments for $1 - P_s$.

6 Spectral Approximation of Transfer Operators

We are interested in fluctuation within the canonical ensemble for some fixed observation time span τ. As a result, we restrict our consideration to the time–s transition operator P_s with $s = \tau$ or $s = 1$ for the high–friction Langevin equation or the discrete time Markov chain with the same observation time in (26), respectively. Since both associated semigroups of transfer operators are self–adjoint, we assume in this section, that P_s is a self–adjoint operator acting on $L^2_{\mathcal{Q}}(\Omega)$.

6.1 Galerkin Discretization

In order to compute the conformational subsets exploiting certain eigenvectors of P_s, we will introduce a special Galerkin procedure to discretize the eigenvalue problem $P_s u = \lambda u$.

Let $B_1, \dots, B_n \subset \Omega$ be a partition of Ω such that $B_k \cap B_l = \emptyset$ for $k \neq l$ and $\cup_{k=1}^n B_k = \Omega$. Our finite dimensional ansatz space $\mathcal{V}_n = \text{span}\{\chi_1, \dots, \chi_n\}$ is spanned by the associated characteristic functions $\chi_k = \chi_{B_k}$. Then, the Galerkin projection $\Pi_n : L^2_{\mathcal{Q}}(\Omega) \to \mathcal{V}_n$ of $u \in L^2_{\mathcal{Q}}(\Omega)$ is defined by

$$\Pi_n u \;=\; \sum_{k=1}^{n} \frac{1}{\langle \chi_k, \chi_k \rangle_{\mathcal{Q}}} \langle \chi_k, u \rangle_{\mathcal{Q}}\, \chi_k.$$

Note that $\langle \chi_k, \chi_k \rangle_{\mathcal{Q}} = \int_{B_k} \mathcal{Q}(q)\, \mathrm{d}q$ is simply the weight of the subset B_k. The resulting discretized transition operator $\Pi_n P_s \Pi_n$ induces the approximate eigenvalue problem $\Pi_n P_s \Pi_n u = \lambda \Pi_n u$ in $\mathcal{V}_n$. Using $u = \sum_{k=1}^n \alpha_k \chi_k$, the discretized eigenvalue problems in coordinate representation reads

$$\sum_{l=1}^{n} \langle P_s \chi_k, \chi_l \rangle_{\mathcal{Q}}\, \alpha_l \;=\; \lambda\, \langle \chi_k, \chi_k \rangle_{\mathcal{Q}}\, \alpha_k, \qquad \forall k = 1, \dots, n.$$

After dividing by $\langle \chi_k, \chi_k \rangle_{\mathcal{Q}} > 0$, we end up with the convenient form

$$S\alpha \;=\; \lambda\alpha \quad \text{with} \quad \alpha = (\alpha_1, \dots, \alpha_n).$$

The entries of the $n \times n$ matrix S are given by the one step transition probabilities from B_k to B_l:

$$S_{kl} \;=\; \frac{\langle P_s \chi_k, \chi_l \rangle_{\mathcal{Q}}}{\langle \chi_k, \chi_k \rangle_{\mathcal{Q}}} \;=\; p(s, B_k, B_l). \tag{39}$$

Since P_s is a Markov operator, its Galerkin discretization S is a (row) stochastic matrix, i.e., $S_{kl} \geq 0$ and $\sum_{l=1}^n S_{kl} = 1$ for all $k = 1, \dots, n$. Hence, all its eigenvalues λ satisfy $|\lambda| \leq 1$. Moreover, we have the following four important properties [35, 34]:

1. The row vector $\pi = (\pi_1, \dots, \pi_n)$ with $\pi_k = \int_{B_k} \mathcal{Q}(q)\, \mathrm{d}q$, which represents the discretized invariant density $\mathcal{Q}$, is a left eigenvector corresponding to the eigenvalue $\lambda = 1$, i.e., $\pi S = \pi$.

2. S is *irreducible and aperiodic*, if P_s is asymptotically stable. As a consequence, the eigenvalue $\lambda = 1$ is *simple*. Hence, the discretized invariant density π is the *unique* stationary distribution of S.

3. The transition matrix S is *self–adjoint* with respect to the discrete scalar product $\langle u, v\rangle_\pi = \sum u_i v_i \pi_i$, since P_s is self–adjoint. Equivalently, S satisfies the condition of *detailed balance*:

$$\pi_k\, S_{kl} \quad = \quad \pi_l\, S_{lk}, \qquad \text{for all } k,l \in \{1,\dots,n\}.$$

Therefore, all eigenvalues of S are real–valued and, due to 2., contained in the interval $[-1, 1]$.

In other words, for an arbitrary covering $B_1, \dots, B_n \subset \Omega$, the discretization matrix S inherits the most important properties of the transition operator P_s.

6.2 Convergence of Discrete Eigenvalues

Denote by $\sigma(P_s)$ the $L^2_Q(\Omega)$–spectrum of P_s and by $\sigma_{\text{discr}}(P_s) \subset \sigma(P_s)$ the subset of all isolated eigenvalues of finite (algebraic) multiplicity. Then, $\sigma_{\text{discr}}(P_s)$ is called the discrete spectrum, while $\sigma_{\text{ess}}(P_s) = \sigma(P_s) \setminus \sigma_{\text{discr}}(P_s)$ is called the essential spectrum of P_s [24]. Assume that the essential spectrum is bounded away from 1 (condition C2 on page 12), i.e., there exists a constant $0 < \kappa < 1$ such that $\sigma_{\text{ess}}(P_s)$ is contained is the ball with radius κ centered a the origin.

We are interested in approximating a cluster of (real–valued) discrete eigenvalues $\lambda_c, \dots, \lambda_1 \in \sigma_{\text{discr}}(P_s)$ near 1 "outside" the essential spectrum:

$$\kappa < \lambda_c \le \cdots \le \lambda_2 < \lambda_1 = 1,$$

repeated according to multiplicity. The corresponding eigenvectors $u_c, \cdots, u_1$ are assumed to be orthogonal; this is always possible, since P_s is assumed to be self–adjoint.

Assume that the sequence of Galerkin ansatz spaces $\mathcal{V}_1 \subset \mathcal{V}_2 \subset \dots$ is dense in $L^2_Q(\Omega)$ and the corresponding partitions are getting finer and finer, $\max_{B\in\mathcal{V}_n} \text{diam}(B) \to 0$ as $n \to \infty$. Denote by $S_{\mathcal{V}_n}$ the transition matrix (39) associated with the ansatz space $\mathcal{V}_n$ and by $\lambda_{i,\mathcal{V}_n}$,$u_{i,\mathcal{V}_n}$ its eigenvalues and eigenvectors, respectively, ordered in decreasing magnitude and taken into account multiplicity.

Under the above stated assumptions, the dominant eigenvalues of $S_{\mathcal{V}_n}$ are good approximations of the dominant eigenvalues of P_s, whenever the discretization is fine enough; in this case, $P_{\mathcal{V}_n}$ also has a cluster of eigenvalues $\lambda_{c,\mathcal{V}_n} \le \dots \le \lambda_{2,\mathcal{V}_n} < \lambda_{1,\mathcal{V}_n} = 1$ near 1. More precisely, for every $i = 1, \dots, c$, we get [34]

$$\lambda_{i,\mathcal{V}_n} \longrightarrow \lambda_i \quad \text{and} \quad u_{i,\mathcal{V}_n} \longrightarrow u_i \quad \text{as } n \to \infty$$

in modulus and $L^2_Q(\Omega)$–norm, respectively.

7 Algorithmic Realization

In this section, we want to outline the basic steps for an algorithmic realization of identifying molecular conformations, their meta–stability and the transition rates between them. In doing so, we will exclusively focus on the discrete time Markov chain and the related transition operator defined in Section 5.2 due to the following two reasons. First, it is the common belief that the discrete time Markov chain approach based on Hamiltonian motion is more realistic for modelling conformational dynamics of biomolecules than the high friction Langevin approach. Second, we managed to prevent the numerical effort for solving the eigenvalue problem for the transition operator from exploding combinatorially with the number of atoms in the molecule: This was done by discretizing it by means of a special hybrid Monte Carlo method [16], such that the computational effort does not depend explicitly on the dimension of the system. The basic scheme of the resulting algorithm is illustrated in Fig. 1. We will explain the single algorithmic steps subsequently.

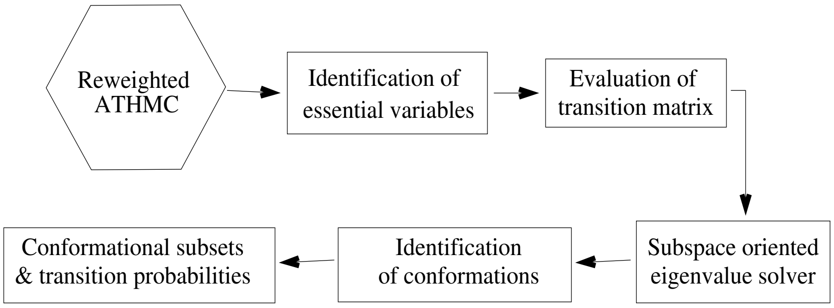

Figure 1: Basic scheme of the algorithm.

Each step of the algorithm is illustrated by application to the triribonucleotide adenylyl(3'-5')cytidylyl(3'-5')cytidin (r(ACC)) model system in vacuum (see Fig. 2). Its physical representation ($N = 70$ atoms) is based on the GROMOS96 extended atom force field [41].

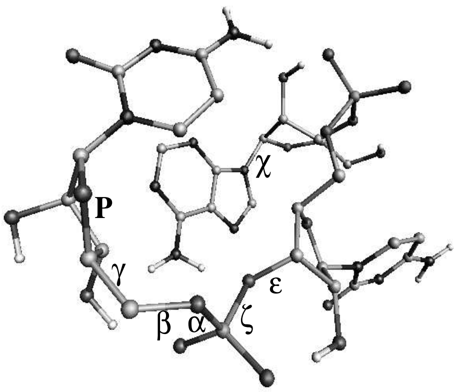

Figure 2: Configuration of the trinucleotide $r(ACC)$ in a ball–and–stick representation. The Greek symbols indicate some of the important dihedral angles of the molecule.

7.1 Evaluation of the Transition Matrix

In order to compute the discretization matrix S of the transition operator — called the *transition matrix* in the following— we have to be able to determine transition probabilities between subsets. This task includes three subproblems: Generation of an adequate box partition of the position space Ω, sampling of the canonical ensemble $\mathcal{Q}$ and approximation of the internal dynamics within the ensemble.

Sampling of the Canonical Density. The typical approach to sample the canonical density is via Monte Carlo (MC) techniques. There is an extremely rich and varied literature on this topic (see, e.g., [5, 39]) and *every* converging MC method would allow to realize this subproblem. In addition, one may also apply MD-based techniques, e.g., constant temperature sampling of the canonical density [31, 3].

It is widely known, that MC simulations for ensemble averages in biomolecular systems may suffer from possible "trapping problems" [28]. As illustrated in [34], this phenomenon is related to the existence of almost invariant sets for the Monte-Carlo Markov chain.[4] We use a specific MC method, the *hybrid Monte Carlo method with adaptive choice of temperature* (ATHMC) [16], which was especially constructed to overcome this trapping problem. Moreover, ATHMC is particularly useful for linking the sampling technique with the ensemble dynamics. Future approaches will be based on a hierarchical sampling technique [15], which might be understood as a specific multilevel approach to ATHMC that merges its superior sampling properties with the identification of almost invariant sets.

The result of every converging MC method is a finite sampling $\Sigma \subset \Omega$ of positions that are distributed according to the canonical ensemble.

Application to r(ACC). The simulation data were generated by means of an ATHMC sampling of the canonical density at $\mathcal{T} = 300$K. The subtrajectories of length 80 femtoseconds were computed by means of the Verlet discretization with a stepsize of 2fs. For these parameters, standard MC simulations typically require thousands of iterations only to leave the neighborhood of the initial configuration. Application of ATHMC (with adaptive temperatures between 300K and 400K) circumvents this trapping problem: one observes frequent transitions in the crucial dihedral angles of the molecule (for details see [15]). The simulation was divided into 4 Markov chains, each starting with a different state chosen from a high temperature run at 500K, which allowed the molecule to move into different conformations. The sampling took about 12h on a workstation with MIPS R10.000 processor. It was terminated by a convergence indicator [21] associated with the potential energy and all 37 dihedral angles after 320.000 steps, resulting in the sampling sequence $\Sigma = \{q_1, \dots, q_{32.000}\}$, considering only every 10th step. Since the temperature can change during the ATHMC run, each configuration is connected with a reweighting factor with respect to the canonical ensemble at 300K.

[4] The trapping phenomenon occurs when the Monte-Carlo Markov chain gets trapped near a local potential energy minimum due to high energy barriers so that a proper sampling of the entire phase space within reasonable computing times is prevented.

Box Partition via Essential Degrees of Freedom. Typical biomolecular systems contain hundreds or thousands of atoms. As a consequence, the number of discretization boxes, and thus the dimension of the transition matrix S, would grow exponentially with the size of the molecular system, if we would generate a box decomposition for Ω by simply partitioning every degree of freedom.

Chemical insight allows to circumvent this "curse of dimension". In the chemical literature, conformations of biomolecules are mostly described in terms of a few *essential degrees of freedom*. In the subspace of essential degrees of freedom anharmonic motion occurs that comprises most of the positional fluctuations, while in the remaining degrees of freedom the motion has a narrow Gaussian distribution and can be considered as "physically constrained".

Based on the sampling of the canonical ensemble, we may determine essential degrees of freedom either in the position space according to AMADEI ET AL. [2] or in the space of internal degrees of freedom, e.g., dihedral angles, by statistical analysis of circular data [22]. These techniques are based on the following statistical analysis of the sampling data: The correlations between atomic motions within the simulation data are expressed by the covariance matrix C.[5] Since C is symmetric, it can always be diagonalized, i.e., there is an orthonormal matrix U such that $C = U^T \Lambda U$ with Λ being the diagonal matrix whose entries are the eigenvalues of C. The matrix U defines the transformation of the original coordinates (positions or internal degrees of freedom) into the uncorrelated coordinates. The matrix Λ is connected to the systems constraints in the following way [2]: Transformed coordinates corresponding to zero or nearly zero eigenvalues behave effectively as constraints; they have narrow Gaussian distributions and do not contribute significantly to the fluctuations. In contrast to that, transformed coordinates corresponding to large eigenvalues have large deviations from their mean position, i.e., they belong to important fluctuations. Mostly, only a few coordinates exhibit such important fluctuations; these are called essential degrees of freedom. Thus, this procedure results in a tremendous reduction of the number of degrees of freedom and, consequently, in a moderate number of partition boxes when discretizing the essential variables only. [22].

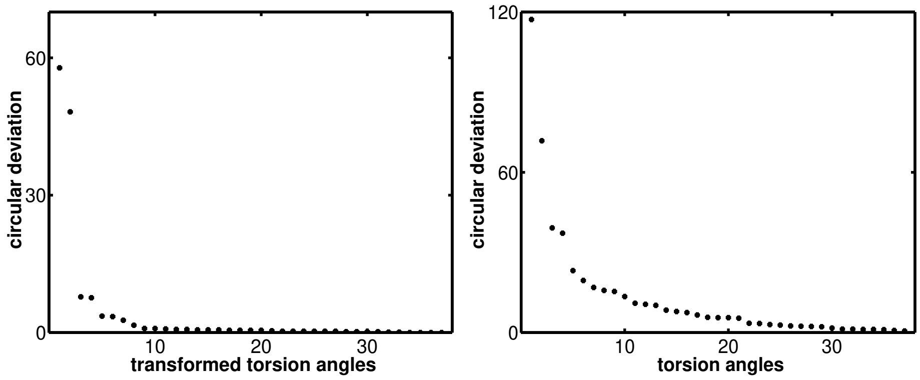

Figure 3: Top: circular deviation of the *transformed* dihedral angles ordered by magnitude (left) and circular deviation of the original dihedral angles (right).

[5]To analyze the simulation data in terms of the dihedral angles we have to apply statistical methods for circular data [17, 18]; see [22] for resulting definition of the covariance matrix.

Remark. As discussed in detail in [34], the transition operator can be restricted to the coordinate space spanned by the essential variables without loosing its desired spectral properties.

Application to r(ACC). Since essential degrees of freedom should solely reflect internal fluctuations of the molecule, we only consider the 37 dihedral angles of the r(ACC) molecule (see Fig. 2). The above explained transformation process based on the simulation data for r(ACC) is exemplified in Fig. 3 and Fig. 4. Figure 3 shows the circular deviations of the original and transformed dihedral angles in decreasing order of magnitude. Only the first four transformed dihedral angles have relevant circular deviation *and* are far from being Gaussian shaped (see Fig. 4), while the remaining transformed dihedral angles are Gaussian like.

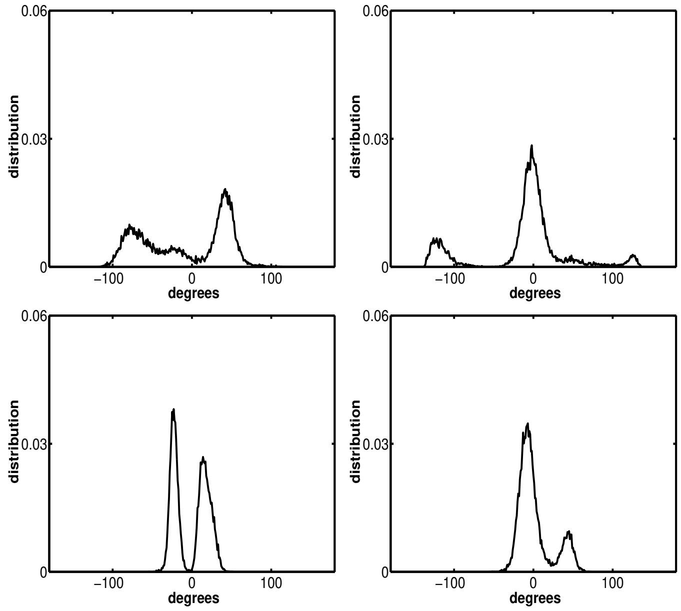

Figure 4: Distribution of the four essential dihedral angles. The distributions at the top allow to identify three maxima each, while there are two maxima for each distribution at the bottom.

The configurational space was discretized into "boxes" $B_1, \ldots, B_d$, by means of all four essential degrees of freedom (see Fig. 4) resulting in $d = 36$ discretization boxes.

Approximation of Internal Dynamics. Due to equations (22) and (39) the entries of the transition matrix S with respect to the boxes B_k are given by

$$\begin{aligned} S_{kl} &= p(1, B_k, B_l) \\ &= \frac{1}{\int_{B_k} \mathcal{Q}(q)\,\mathrm{d}q} \int_{B_k} p(1, q, B_l)\, \mathcal{Q}(q)\,\mathrm{d}q. \end{aligned} \tag{40}$$

Let now $q_1, q_2, \ldots$ denote an arbitrary sequence of positions generated by some ergodic Markov chain Monte Carlo method that is asymptotically distributed according to the canonical density $\mathcal{Q}$. Then, due to the law of large number for Markov chains [40], we may rewrite S_{kl} as

$$S_{kl} \;=\; \lim_{n\to\infty} \frac{\sum_{j=1}^{n} p(1,q_j,B_l)\chi_{B_k}(q_j)}{\sum_{j=1}^{n} \chi_{B_k}(q_j)}.$$

By using our particular MC sequence Σ, we thus get

$$S_{kl} \;\approx\; \frac{\sum_{q\in\Sigma} p(1,q,B_l)\chi_{B_k}(q)}{\sum_{q\in\Sigma} \chi_{B_k}(q)}.$$

Finally, we have to approximate the transition kernel $p(1,q,B_k)$ for all $q \in \Sigma$ and all B_i. For the discrete Markov chain, this can be done by applying some integration scheme to (26). A convergence analysis is presented in [34].

Application to r(ACC). The dynamical fluctuations within the canonical ensemble were approximated by integrating four short trajectories of length $\tau =$ 80fs starting from each sampling point $q \in \Sigma$. To facilitate transitions, analogous to the ATHMC sampling, the momenta were chosen according to the momenta distribution $\mathcal{P}$ for 4 different temperatures between $300K - 400$K and reweighted afterwards. This resulted in a total of $4 \times 32.000 = 128.000$ transitions. This calculation took less than 25% of the total computing time. Then the 36×36 transition matrix S was computed based on the 128.000 transitions taking the different weighting factors into account. Since every box had been hit by sufficiently many transitions, the statistical sampling was accepted to be reliable.

7.2 Solving the Eigenvalue Problem

Once the entries of the transition matrix have been computed, we have to determine the eigenvectors corresponding to a cluster of eigenvalues near the dominant eigenvalue 1. That is, only a small part of the spectrum of S is required, *not* its full diagonalization. Actual evaluation is efficiently possible using subspace oriented iterative techniques, even if the number of discretization boxes may be about 100.000 or larger [27, 20].

Application to r(ACC). The computation of the eigenvalues of S near 1 yielded a cluster of eight 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	...

7.3 Identification of Conformations

According to the definition of almost invariance, we are interested in unions $C = \cup_{k\in I} B_k$ of partition sets, for which $p(1,C,C) \approx 1$. In other words, we are looking for a nontrivial index set $I \subset \{1,\ldots,n\}$ such that the process Q_t almost certainly stays within $B = \cup_{k\in I} B_k$ after one step. Using the transition probabilities $p(1,B_k,B_l)$ between the partition sets, cluster algorithm can be used to identify almost invariant subsets [22]. We apply the identification strategy of Sec. 5.4 in its algorithmic realization due to [13], which exploits a certain almost constant level structure of eigenvectors corresponding to a cluster of eigenvalues near 1.

Application to r(ACC). Finally, the conformational subsets were computed based on the eigenvectors of S via the identification algorithm. This yielded eight conformations.

The conformational subsets identified turned out to be rather insensitive to further refinements of the discretization. The weighting factors within the canonical ensemble and the meta–stability of the eight identified conformations are given in the following table:

conformations	D1c	D1t	D2c	D2t	D3c	D3t	D4c	D4t
weighting factor	0.107	0.011	0.116	0.028	0.320	0.038	0.285	0.095
meta–stability	0.986	0.938	0.961	0.888	0.991	0.949	0.981	0.962

The transition probabilities between the different conformations are visualized schematically in Fig. 5. The matrix allows to define a hierarchy between the conformations, which is inherent to the algorithm. On the top level, there are two conformations, D1&D2 and D3&D4 corresponding to the two 4×4 blocks on the diagonal of S. On the next level, each of these conformations split up into two subconformations yielding D1, ... ,D4. On the bottom level, each conformation is further divided into a c–part and a t–part (for interpretation see [22]). The evaluation of the transition matrix together with the execution of the identification algorithm took less than 2% of the computing time required for evaluation of the simulation data via ATHMC.

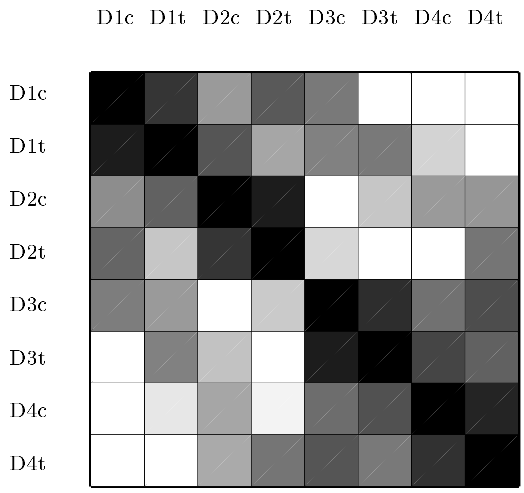

Figure 5: Visualization of the one–step transition probabilities $p(1, D_{\mathrm{from}}, D_{\mathrm{to}})$ between the conformation D_{from} (row) and D_{to} (column). The colors are chosen according to the logarithm of the corresponding entries; black: $p(1,\cdot,\cdot) \approx 1$, white: $p(1,\cdot,\cdot) \approx 0$.

Acknowledgment. One of us (W.H.) was supported within the DFG–Schwerpunkt "Ergodentheorie, Analysis und effiziente Simulation dynamischer Systeme" under Grant De 293/2-1.

References

[1] M. Allen and D. 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*, 17:412–425, 1993.

[3] H. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, 72:2384–, 1980.

[4] A. Arnold, P. Markowich, G. Toscani, and A. Unterreiter. On logarithmic Sobolev inequalities, Csiszar–Kullback inequalities, and the rate of convergence to equilibrium for Fokker-Planck type equations. *Submitted to Comm. PDE*, 1998.

[5] K. Binder[Ed.]. *The Monte Carlo method in condensed matter physics*, volume Bd. 71 of *Topics in applied physics.* Springer Verlag, Berlin, Heidelberg, New York, 1992.

[6] K. L. Chung. *Lectures from Markov Processes to Brownian Motion*, volume 249 of *A Series of Comprehensive Studies in Mathematics.* Springer, New York, Heidelberg, Berlin, 1982.

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

[8] E. B. Davies. Metastable states of symmetric Markov semigroups II. *J. London Math. Soc.*, 26(2):541–556, 1982.

[9] M. Dellnitz and A. Hohmann. A subdivision algorithm for the computation of unstable manifolds and global attractors. *Numerische Mathematik*, 75:293–317, 1997.

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

[11] P. Deuflhard, M. Dellnitz, O. Junge, and C. Schütte. Computation of essential molecular dynamics by subdivision techniques. In *[12]*, pages 98–115, 1999.

[12] P. Deuflhard, J. Hermans, B. Leimkuhler, A. E. Mark, S. Reich, and R. D. Skeel, editors. *Computational Molecular Dynamics: Challenges, Methods, Ideas*, volume 4 of *Lecture Notes in Computational Science and Engineering.* Springer, 1999.

[13] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte. Identification of almost invariant aggregates in nearly uncoupled Markov chains. Accepted in Lin. Alg. Appl., Available via http://www.zib.de/bib/pub/pw, 1998.

[14] D. Evans and G. Morriss. *Statistical Mechanics of Nonequilibrium Liquids.* Academic Press, London, 1990.

[15] A. Fischer. Coupling/uncoupling of reversible markov chains via subdivision monte carlo techniques. Report, Konrad–Zuse–Zentrum, Berlin. Available via http://www.zib.de/MDGroup, 1999.

[16] 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.

[17] N. I. Fisher. *Statistical Analysis of Circular Data.* University Press, Cambridge, 1993.

[18] N. I. Fisher and A. J. Lee. A correlation coefficient for circular data. *Biometrica*, 70(2):327–332, 1983.

[19] M. Freidlin and A. Wentzell. *Random perturbations of dynamical systems.* Springer, New York, London, 1984. Series in Comprehensive Studies in Mathematics.

[20] T. Friese, P. Deuflhard, and F. Schmidt. A multigrid method for the complex Helmholtz eigenvalue problem. In C.-H. Lai, P. E. Bjørstad, M. Cross, and O. B. Widlund, editors, *Eleventh International Conference on Domain Decomposition Methods*, 1998. Submitted.

[21] A. Gelman and D. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511, 1992.

[22] W. Huisinga, C. Best, R. Roitzsch, C. Schütte, and F. Cordes. From simulation data to conformational ensembles: Structure and dynamic based methods. To appear in J. Comp. Chem. 1999. Available via http://www.zib.de/bib/pub/pw, 1998.

[23] K. Ichihara and H. Kunita. A classification of the second order degenerate elliptic operators and its probabilistic characterization. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 30:235–254, 1974.

[24] T. Kato. *Perturbation Theory for Linear Operators.* Springer, 1995.

[25] R. Kurt. *Axiomatics of Classical Statistical Mechanics.* Pergamon Press, Oxford, New York, 1980.

[26] A. Lasota and M. C. Mackey. *Chaos, Fractals and Noise*, volume 97 of *Applied Mathematical Sciences.* Springer, New York, 2nd edition, 1994.

[27] R. B. Lehoucq, D. C. Sorensen, and C.Yang. *ARPACK User's Guide: Solution of Large Eigenvalue Problems by Implicit Restartet Arnoldi Methods.* Rice University Houston, 1998.

[28] E. Leontidis, B. Forrest, A. Widmann, and U. Suter. Monte Carlo algorithms for the atomistic simulation of condensed polymer phases. *J. Chem. Soc. Faraday Trans.*, 91(16):2355–2368, 1995.

[29] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability.* Springer, Berlin, Heidelberg, New York, Tokyo, 1993.

[30] E. Nelson. *Dynamical Theories of Brownian Motion.* Mathematical Notes. Princeton University Press, 1967.

[31] S. Nosé. A molecular dynamics methods for simulations in the canonical ensemble. *Mol. Phys.*, 52, 1984.

[32] H. Risken. *The Fokker-Planck equation.* Springer, New York, London, 2nd edition, 1996.

[33] T. Schlick. Some failures and successes of long–timestep approaches to biomolecular simulations. In *[12]*, pages 227–262, 1999.

[34] C. Schütte. *Conformational Dynamics: Modelling, Theory, Algorithm, and Application to Biomolecules.* Habilitation Thesis, Fachbereich Mathematik und Informatik, Freie Universität Berlin, 1998. Available as Report SC-99-18 via http://www.zib.de/bib/pub/pw/.

[35] 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.

[36] C. Schütte and W. Huisinga. On conformational dynamics induced by Langevin processes. Preprint SC-99-25, Konrad–Zuse–Zentrum, Berlin. Available via http://www.zib.de/MDGroup, 1999.

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

[38] R. Skeel and J. Izaguirre. The five femtosecond time step barrier. In *[12]*, pages 312–325, 1999.

[39] A. Sokal. Monte Carlo methods in statistical mechanics. Lecture note, Department of Physics, New York University, 1989.

[40] L. Tierney. Introduction to general state-space Markov chain theory. In W. Gilks, S. Richardson, and D. Spiegelhalter, editors, *Markov chain Monte-Carlo in practice*, pages 59–74. Chapman and Hall, London, Glasgow, New York, Tokyo, 1997.

[41] W. van Gunsteren, S. Billeter, A. Eising, P. Hünenberger, P. Krüger, A. Mark, W. Scott, and I. Tironi. *Biomolecular Simulation: The GROMOS96 Manual and User Guide.* vdf Hochschulverlag AG an der ETH Zürich, 1996.

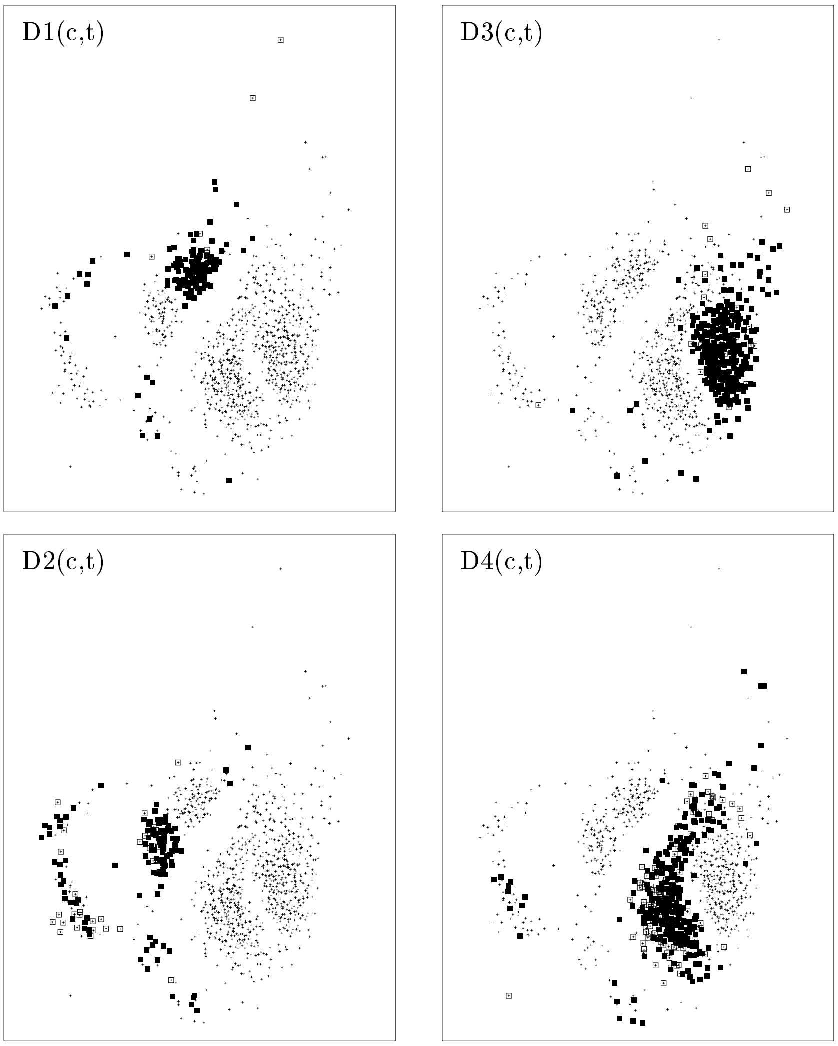

Figure 6: 2d plot of the four conformations D1, ... ,D4 (squares). The distinction between open and filled squares indicates a further splitting into eight conformations resulting from a partition into a c–conformation and a t–conformation.