Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

PETER DEUFLHARD

From Molecular Dynamics to Conformational Dynamics in Drug Design

ZIB-Report 02–20 (April 2002)

From Molecular Dynamics to Conformational Dynamics in Drug Design

Peter Deuflhard

Dedicated to Good Bill Hunting,
Chief of Mount Highdle tribe,
on the occasion of his 60th birthday

Abstract

Computational drug design studies molecular recognition in the *virtual lab*. The arising Hamiltonian dynamics is known to be chaotic and ill-conditioned already after picoseconds ($= 10^{-12}$ seconds), whereas times of pharmaceutical interest are in the milliseconds ($= 10^{-3}$ seconds) up to minutes. Classical molecular dynamics with long term trajectory computation gives, at best, information about time and statistical ensemble averages. The present paper surveys a recent new modelling approach called *conformational dynamics*, which is due to Deuflhard and Schütte. This approach achieves information about the dynamics on longer time scales by telescoping a short term deterministic model with a statistical model. Examples of small biomolecules are included.

Keywords: molecular dynamics, conformational dynamics, drug design, Hamiltonian dynamics, almost invariant sets, transition operator, Markov chain, Perron cluster analysis, hybrid Monte Carlo

MSC (1991): 62P10, 65C05, 65F15, 65L05, 65L08

Introduction

The design of pharmaceuticals, briefly called *drug design*, is a pyramidal multistage process, from a broad basis to an extremely narrow tip:

- molecular recognition studies
- intracellular impact studies
- physiological investigations
- animal experiments
- clinical tests
- market introduction

The basis level "molecular recognition studies", in turn, consists of two parts: studies in the chemical lab and studies in the *virtual lab* by means of the computer, often named as *computational drug design*. The impact of this rather new scientific field cannot be overestimated: The cost of identifying a marketable drug out of a huge set of promising chemical substances is commonly estimated as 500 million Euro. If, at the basis level, the number of promising drug candidates could be halved, then the cost per successful marketable pharmaceutical would also roughly be halved, not to mention the reduction of "time to market".

In computational biotechnology, algorithms from discrete mathematics or computer science already play a publicly visible role – for example, multiple alignment in the decoding of the human genome. These approaches primarily aim at a clarification of the *geometric form* of molecular systems. In view of the *biological function*, however, the *dynamics* of molecular systems need to be studied in detail. Here the situation is characterized by the fact that real times of pharmaceutical interest are in the region of $msec$ up to min, whereas simulation times are presently in the region of $psec$ up to $nsec$ with $fsec$ timesteps. Therefore some computational scientists advocate that the available computer power is the essential limiting factor for gaining insight into the dynamics of molecular systems.

Even though the dynamics of molecules is well recognized in its importance, its mathematical treatment seems to be still at an early phase of involvement. Up to now, classical numerical analysis essentially only enters via

fast multipole methods (see Greengard and Rokhlin [**?**]) or via symplectic discretizations (cf. Sanz-Serna et al. [**?**]). However, the computation of molecular dynamics has a mathematical limitation, even stricter than the limitation by computer power: the arising trajectories are Hamiltonian and as such chaotic. Consequently, the traditional trajectory simulations give, at best, only information about time averages. Under some ergodic hypothesis, often carelessly a priori assumed, these averages are equivalent to statistical ensemble averages. Therefore an investigation of the dynamics of molecular systems over the time scales of interest will require a different mathematical approach.

In recent years the present author and Ch. Schütte have created some new mathematical model based on concepts of nonlinear dynamics (for early papers see, e.g., [**?**, **?**, **?**, **?**]). This approach, now called *conformational dynamics*, will be worked out here together with its algorithmic implications and its scientific perspectives.

1 Classical Molecular Dynamics

In classical molecular dynamics the simplifying assumption is made that the motion of atoms and molecules can be described by Newtonian differential equations just as in classical mechanics, replacing mechanical potentials by special molecular potentials. Such an assumption obviously ignores the role of quantum mechanics, which actually provides the correct physical framework for these microscopic processes. Some part of the quantum-mechanical effects, at least, are introduced into the classical formalism via a parametrization of the potentials.

Hamiltonian differential equations. Let N atoms of a molecular system be specified in terms of their spatial coordinates (position variables) $q_j \in \mathbb{R}^3$, $j = 1, \ldots, N$, and their corresponding N generalized moments (momenta variables) $p_j \in \mathbb{R}^3$. Then the Hamilton function H has the form

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

The first, quadratic term, involving the symmetric, positive definite mass matrix M, is the kinetic energy, the second term is the potential energy or just potential, which is often highly nonlinear in the molecular context. From

given H, the Hamiltonian differential equations are defined as

$$q_i' = \frac{\partial H}{\partial p_i}, \quad p_i' = -\frac{\partial H}{\partial q_i}, \qquad i = 1, \ldots, N.$$

Of course, the quality of any molecular dynamics calculation is strongly dependent on the quality of the available potential data (we mostly use MMFF due to [**?**]). These potentials have the general form

$$\begin{aligned} V(q) = &\sum_{k,l} V_{\text{bond}}(q_k, q_l) + \sum_{k,l,j} V_{\text{angle}}(q_k, q_l, q_j) \\ &+ \sum_{k,l,j,m} V_{\text{out-of-plane}}(q_k, q_l, q_j, q_m) + \sum_{k,l,j,m} V_{\text{dihedral}}(q_k, q_l, q_j, q_m) \\ &+ \sum_{k,l} V_{\text{Lennard-Jones}}(q_k, q_l) + \sum_{k,l} V_{\text{Coulomb}}(q_k, q_l) \end{aligned}$$

or, in abbreviation,

$$V = V_B + V_A + V_T + V_{LJ} + V_Q \ ,$$

where V_B describes the bond deformation, V_A the angle deformation, V_T the torsion angle deformation (two parts), V_{LJ} the van-der-Waals interaction in terms of the Lennard-Jones potential, and V_Q the electrostatic interaction in terms of Coulomb forces between charges Q.

The numerical solution of the initial value problem for these differential equations first requires the selection of an efficient nonstiff discretization scheme – consult, e.g., the specialized textbook of Sanz-Serna [**?**] or Section 4.3.4 in the more recent textbook [**?**, **?**]. In the context of numerical integration an efficient evaluation of the right sides is needed. The above potential terms V_B, V_A, V_T, and V_{LJ} contribute a cost of order $O(N)$ operations. The direct evaluation of the long-range Coulomb potential V_Q appears to require $O(N^2)$ operations and hence constitutes a problem of its own, at least for realistic molecules. An efficient algorithm requiring only $O(N)$ operations is the *fast multipole method* of L. Greengard and V. Rokhlin [**?**].

In order to speed up the numerical computations, T. Schlick and followers suggested to skip the adaptive control of the numerical integrators and just run them with step sizes at the border of stability of the numerical schemes. Such an approach has an interpretation only in terms of some sampling based on the ergodic theorem – see, e.g., [**?**].

Condition of molecular initial value problems. Formally speaking, the above solution of the initial value problem is *unique*, which can be written in terms of the flow Φ as

$$x(t) = (q(t), p(t)) = \Phi^t x_0 \ .$$

For the purpose of numerical analysis, we additionally have to study the corresponding *condition number* κ, which characterizes the sensitivity of the unique solution under perturbation of the initial values. By virtue of first order perturbation theory such a quantity can be defined as (cf. Section 3.1.2 in [**?**, **?**])

$$\|\delta x(t)\| \dot{\leq} \kappa(t) \|\delta x_0\| \ , \quad \kappa(t) = \|\partial \Phi^t / \partial x_0\| \ .$$

As already discovered by H. Poincaré, Hamiltonian systems are *chaotic*. In general mathematical terms, this means a characterization of the asymptotic behavior – in the present notation $\kappa(\infty) = \infty$. In the context of numerical analysis, this means that an ever so slight perturbation of the initial values will induce a resulting perturbed trajectory deviating markedly from the unperturbed trajectory after some characteristic critical time. The question is: How long is that "critical time"? Detailed examination shows that for the subclass of *integrable* Hamiltonian systems (such as the popular Kepler problem) the condition number grows linearly – see, e.g. V. I. Arnold [**?**]. In real life molecular dynamics problems, however, the growth is exponential, i.e.

$$\kappa(t) \sim \exp(t/t_{\text{crit}}) \ , \tag{1.1}$$

where the critical times t_{crit} are typically no longer than a few ps.

Example: Trinucleotide ACC. We illustrate the effect for the small biomolecule ACC – compare Section 1.2 in [**?**, **?**]. This molecule is a short RNA segment consisting of 94 atoms; the genetic letters in its acronym stand for adenine (A) and cytosine (C). Figure **??** shows simulation snapshots at the times $t = 0.0$ ps, $t = 0.5$ ps, and $t = 20$ ps (picoseconds: 1 ps $= 10^{-12}$ *sec*).

As can be seen, the two molecular configurations are almost identical at the start, but differ completely after only 20 ps. The resulting configurations (left a spherical shape, right a stretched shape) remain essentially the same over quite long time spans. They are therefore called *metastable conformations*. These mathematical objects typically occur in nearly all molecular systems and should be directly computed as such.

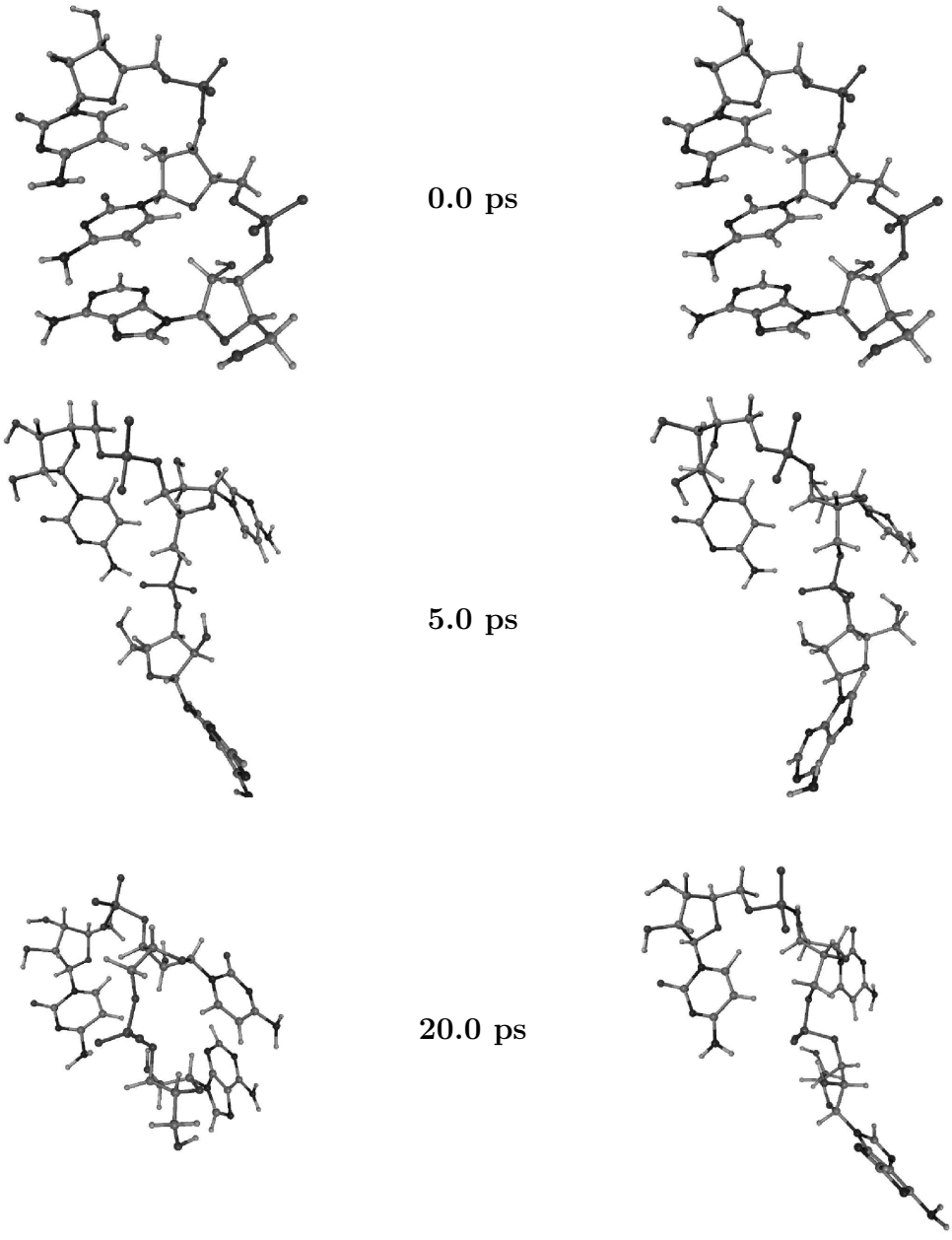

Figure 1: ACC molecule: Development of distinct conformations from nearly identical initial configurations

2 Metastable Conformations as Almost Invariant Sets

The observations of the preceding section force severe changes in the mathematical modelling of molecular dynamics. Instead of the *point concept* of classical mechanics based on deterministic trajectories we need to derive a *set concept* based on the above mentioned metastable conformations. This is the key idea of conformational analysis to be presented here.

Perron–Frobenius operator. Starting point for the new approach was the pioneering work of Dellnitz and co-workers [**?**, **?**] based on the Perron–Frobenius operator U. This operator (dating back to Ulam) is defined via measures in phase space $x = (p, q) \in \Gamma \subset \mathbb{R}^{6N}$ as

$$U\mu(B) = \mu(\Phi^{-\tau}(B)) \, , B \subset \Gamma$$

An invariant measure $\bar{\mu}$ and the corresponding invariant set $\bar{B}$ are characterized by

$$\bar{\mu}(B) = \bar{\mu}(\Phi^{-\tau}(B)) \, , \quad \bar{B} = \Phi^{-\tau}(\bar{B}) \, ,$$

which lead to the eigenvalue problem

$$U\bar{\mu}(\bar{B}) = \bar{\mu}(\bar{B}) \tag{2.1}$$

for the Perron eigenvalue $\lambda = 1$. On this basis, these authors computed (relatively) global attractors by some adaptive multilevel box discretization. Moreover they found that (a) eigenvalues $\lambda \neq 1$ on the unit circle permit an interpretation in terms of cyclic dynamics, and (b) eigenvalues close to the Perron eigenvalue inside the unit cicle (due to discretization effects) seem to have an interpretation in terms of *almost invariant sets*.

The success of that approach was intimately linked to *hyperbolic* dynamics which is known to collapse asymptotically to some dynamics on a low-dimensional manifold. Being well aware of this restriction, the present author nevertheless risked to extend that basic scheme to *Hamiltonian* dynamics known not to collapse, but to remain on some high-dimensional energy surface. A first attempt in this direction, as published in [**?**], suffered from two important disadvantages. First, for a *deterministic* Hamiltonian system, the operator U is *unitary* in $L^2(\Gamma)$ so that real eigenvalues inside the unit circle cannot exist. But such eigenvalues had been computed and could be interpreted in detail within the model! The reason for that has been that the

discretization had allowed for *stochastic* perturbations of the deterministic system so that such eigenvalues could, in fact, occur and did contain information about almost invariant sets. Second, the subdivision technique caused some *curse of dimension* that restricted the applicability of the method to a domain far from realistic molecules.

Stochastic transition operator. In the above situation Ch. Schütte [**?**, **?**] constructed a new *self-adjoint* stochastic operator T. Starting point of his construction is the fact that in a chemical lab with constant temperature and constant volume the deterministic model should be embedded into a canonical or Boltzmann distribution f_0. With β the inverse temperature and for separable Hamiltonian $H = \frac{1}{2}p^T M^{-1} p + V(q)$ we may factorize this distribution according to

$$\begin{aligned} f_0 &= \frac{1}{Z}\ exp(-\beta H)\ ,\ Z = \int exp(-\beta H) dq\ dp \\ &= \frac{1}{Z_p}\ exp(-\frac{\beta}{2} p^T M^{-1} p)\ \frac{1}{Z_q} exp(-V(q)) \end{aligned} \tag{2.2}$$

$$f_0 = \mathcal{P}\mathcal{Q}, \quad Z = Z_p Z_q, \quad \int \mathcal{P}(p)dp = \int \mathcal{Q}(q)dq = 1$$

The key idea is now that the mathematical objects of interest, the metastable conformations, are objects in *position space* $q \in \Omega \subset \mathbb{R}^{3N}$ rather than in the whole phase space $\Gamma = \Omega \times \mathbb{R}^{3N}$. Let $A, B \subset \Omega$ be subsets in the position space and define cylinders $\Gamma(A) := A \times \mathbb{R}^{3N}$ – see Fig. **??**. Let $\chi(A)$ denote the characteristic function of a set A (a function which is 1 inside A and 0 outside). In this setting the probability for the dynamical system to *be* within A can be written as

$$\pi(A) = \int\limits_{\Gamma(A)} f_0(p,q) dq\ dp = \int\limits_A \mathcal{Q}(q) dq = \int\limits_\Omega \chi_A^2 \mathcal{Q}(q) dq =: \langle \chi_A, \chi_A \rangle_{\mathcal{Q}}\ , \tag{2.3}$$

where we introduced some inner product with weighting $\mathcal{Q}$.

The operator T is then constructed as the restriction of the Perron-Frobenius operator U to position space via averaging over the momentum part of the canonical distribution, which means integrating U over the cylinders $\Gamma(\cdot)$.

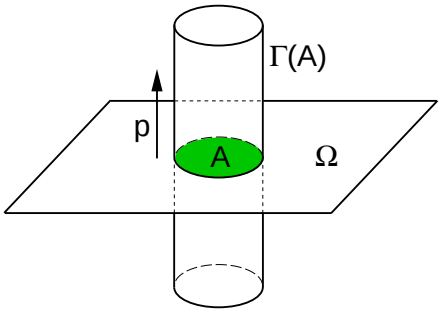

Figure 2: Position space fibre (here: cylinder) $\Gamma(A)$ in phase space

The conditional probability for the system to *move* during time τ from A to B during time τ can then be defined by virtue of the new operator T as

$$w(A, B, \tau) = \frac{\langle \chi_A, \mathrm{T}\chi_B \rangle_{\mathcal{Q}}}{\langle \chi_A, \chi_A \rangle_{\mathcal{Q}}} \ . \tag{2.4}$$

In the same manner, the probability for the system to *stay* in A during time τ comes out as

$$w(A, A, \tau) = \frac{\langle \chi_A, \mathrm{T}\chi_A \rangle_{\mathcal{Q}}}{\langle \chi_A, \chi_A \rangle_{\mathcal{Q}}} \ . \tag{2.5}$$

The operator T is defined over the weighted spaces

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

Obviously, the Hilbert space $L^2_{\mathcal{Q}}(\Omega)$ is associated with the above introduced weighted inner product $\langle \cdot, \cdot \rangle_{\mathcal{Q}}$. With this notation, the properties of T can be listed as follows (due to Schütte [**?**]):

1. T is bounded in $L^r_{\mathcal{Q}}(\Omega)$: $\|\mathrm{T}u\|_{\mathcal{Q}} \leq \|u\|_{\mathcal{Q}}$, for $r = 1, 2$.
2. T is a Markov operator on $L^1_{\mathcal{Q}}(\Omega)$.
3. T is *self-adjoint* in $L^2_{\mathcal{Q}}(\Omega)$. Hence, the spectrum $\sigma(\mathrm{T})$ is real-valued and bounded: $\sigma(\mathrm{T}) \subset [-1, 1]$.
4. There exists a cluster of eigenvalues close to the Perron eigenvalue well-separated from the remaining (continuous) part of the spectrum. We call it the *Perron cluster*.

In summary, the operator T arises as the transition operator of a *reversible* Markov chain. We will use this basic structure for the discretization of the operator – see the subsequent Section **??**. As a result of this kind of discretization we will obtain a stochastic and sparse matrix T which, due to the reversibility of the Markov chain, is also symmetric in a generalized sense.

Perron cluster analysis (PCCA). The newly introduced name "Perron cluster analysis" characterizes a cluster analysis technique based on some analysis of the arising Perron cluster of eigenvalues of the transition matrix of a Markov chain. For this reason it should more correctly be named **P**erron **C**luster **C**luster **A**nalysis, possibly abbreviated PCCA to distinguish it clear enough from the principal component analysis (PCA).

The PCCA method requires an input in terms of a stochastic (general symmetric) matrix $T = T_N$ of dimension N. The method analyzes the spectrum of such a matrix with respect to the possible existence of a Perron cluster of eigenvalues, say $\lambda_1 = 1, \lambda_2 \approx 1, \ldots, \lambda_k \approx 1$. The task is to identify k almost invariant sets corresponding to k metastable chemical conformations. Note that the number k is unknown in advance and must be identified as well. Here we will only sketch the main ideas behind the algorithm. For a broader introduction into the topic we refer to Section 5.5 in the recent editions of the textbook [**?**, **?**], for more details to the original paper [**?**].

Just as in (**??**), we here obtain the (discrete) eigenvalue problem

$$\pi^T T = \pi^T \ , \quad Te = e \ , \qquad \pi^T e = 1 \ , \tag{2.6}$$

where the left eigenvector $\pi^T = (\pi_1, \ldots, \pi_N)$ represents the discrete invariant measure and the right eigenvector $e^T = (1, \ldots, 1)$ is the discrete invariant set – each corresponding to the Perron eigenvalue $\lambda_1 = 1$. Assume now that the total index set $\mathcal{S} = \{1, 2, \ldots, N\}$ can be decomposed into k disjoint index subsets

$$\mathcal{S} = \mathcal{S}_1 \oplus \cdots \oplus \mathcal{S}_k$$

such that there exist k *uncoupled* Markov chains, each of which is running "infinitely long" within one of the index subsets. Then, for a reversible Markov chain, the total transition matrix T is strictly block diagonal with block submatrices $\{T_1, \ldots, T_k\}$ – see, e.g., [**?**]. Each of these submatrices is stochastic and gives rise to a single Perron eigenvalue $\lambda(T_i) = 1, \ i = 1, \ldots, k$. Let the submatrices be primitive. Then, due to the Perron-Frobenius theorem, each block T_i possesses a unique right eigenvector $e_i = (1, \ldots, 1)^T$ of length

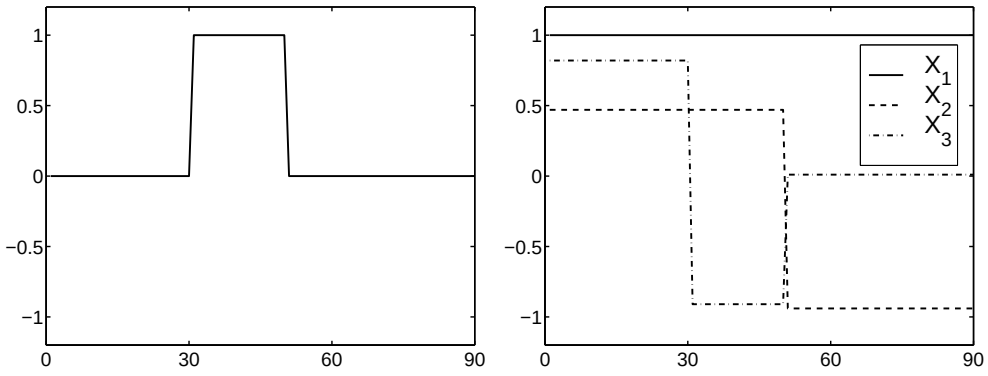

Figure 3: Uncoupled Markov chain over $k = 3$ disjoint index subsets. The state space $\mathcal{S} = \{s_1, \ldots, s_{90}\}$ divides into the subsets $\mathcal{S}_1 = \{s_1, \ldots, s_{29}\}$, $\mathcal{S}_2 = \{s_{30}, \ldots, s_{49}\}$ and $\mathcal{S}_3 = \{s_{50}, \ldots, s_{90}\}$. *Left*: Characteristic function $\chi_{\mathcal{S}_2}$. *Right*: Eigenbasis corresponding to the 3-fold eigenvalue $\lambda = 1$. Observe that each eigenvector is constant on each subset. The sign structure for state s_{69}, for example, is $(+, -, 0)$ in the sense of Lemma **??**.

$\dim(T_i)$ corresponding to its Perron root. Therefore, in terms of the total transition matrix T, the eigenvalue $\lambda = 1$ is k–fold and the corresponding eigenspace is spanned by the vectors

$$\chi_{\mathcal{S}_i} = (0, \ldots, 0, e_i^T, 0, \ldots, 0)^T, \qquad i = 1, \ldots, k \; .$$

In view of the identification problem to be treated, our notation deliberately emphasizes that these eigenvectors can be interpreted as *characteristic functions* of the invariant index subsets (see Fig. **??**, left). In general, any basis $\{X_i\}_{i=1,\ldots,k}$ of the eigenspace corresponding to $\lambda = 1$ can be written as a linear combination of the characteristic functions $\chi_{\mathcal{S}_i}$ with coefficients $\alpha_{ij} \in \mathbb{R}$ such that

$$X_i = \sum_{j=1}^{k} \alpha_{ij} \, \chi_{\mathcal{S}_j}, \qquad i = 1, \ldots k \; .$$

As a consequence, eigenvectors corresponding to $\lambda = 1$ are *constant on each index subset* (see Fig. **??**, right).

In reality, the block diagonal form will not be apparent due to unknown index permutations. We therefore need some elementwise criterion that is independent of any index permutation.

Lemma 2.1 *[?] Given a block-diagonal transition matrix T consisting of reversible, primitive blocks, a left eigenvector $\pi > 0$ and a basis $\{X_i\}_{i=1,\ldots,k}$ of its eigenspace corresponding to $\lambda = 1$. Associate with every state s_i its sign structure*

$$s_i \longmapsto (\mathrm{sign}((X_1)_i), \ldots, \mathrm{sign}((X_k)_i)).$$

Then

1. *invariant index subsets are collections of states with common sign structure,*
2. *different index subsets exhibit different sign structures.*

Next suppose that we have k *nearly uncoupled* Markov chains, each of which is staying "for a long time" in one of the index subsets $\mathcal{S}_i$. For the transition probabilities (**??**) and (**??**) this means that

$$w(\mathcal{S}_i, \mathcal{S}_i, \tau) = 1 - O(\epsilon), \qquad w(\mathcal{S}_i, \mathcal{S}_j, \tau) = O(\epsilon), \quad i \neq j \tag{2.7}$$

in terms of some not further specified perturbation parameter that indicates the *metastability* of the index subsets. In this case the transition matrix T is (after some unknown permutation) block diagonally *dominant*. Moreover, a *Perron cluster*

$$\lambda_1 = 1, \quad \lambda_2 = 1 - O(\epsilon), \ \ldots \ , \lambda_k = 1 - O(\epsilon)$$

arises as a perturbation of the k-fold Perron root in the uncoupled case $\epsilon = 0$. Upon applying Kato's perturbation theory [**?**] we obtain the following results for the corresponding eigenvectors:

Theorem 2.2 *[?] Let $T(\epsilon)$ be a family of matrices satisfying certain regularity conditions not specified here (for details see [?]). Let Π_j denote the projection on the eigenspace spanned by the eigenvector X_j of the unperturbed transition matrix $T(0)$. Then, for real ϵ, there exist π–orthonormal eigenvectors $X_1(\epsilon), \ldots, X_k(\epsilon)$ of the following form:*

(i) An eigenvector corresponding to the Perron root $\lambda_1(\epsilon) \equiv 1$ given by

$$X_1(\epsilon) \equiv e,$$

(ii) A set of $k-1$ eigenvectors corresponding to the eigenvalue cluster $\lambda_2(\epsilon), \ldots, \lambda_k(\epsilon)$ close to $\lambda = 1$ of the form

$$X_i(\epsilon) = \sum_{j=1}^{k} (\alpha_{ij} + \epsilon\beta_{ij})\, \chi_{\mathcal{S}_j} + \epsilon \sum_{j=k+1}^{n} \frac{1}{1-\lambda_j} \Pi_j T^{(1)} X_i + \mathcal{O}(\epsilon^2)$$

for appropriate coefficients $\alpha_{ij}, \beta_{ij} \in \mathbb{R}$ and index subsets $\mathcal{S}_1, \ldots, \mathcal{S}_k$ corresponding to the block-diagonal form of $T(0)$.

The theorem nicely indicates that we can essentially use the tools from the unperturbed case also for the perturbed case. As an illustration, see Fig. **??** where the locally constant pattern over each of the index subsets is still visible even under perturbation. Upon applying Lemma **??** and carefully observing perturbations of the strict zero, we again have an elementwise criterion independent of any permutation.

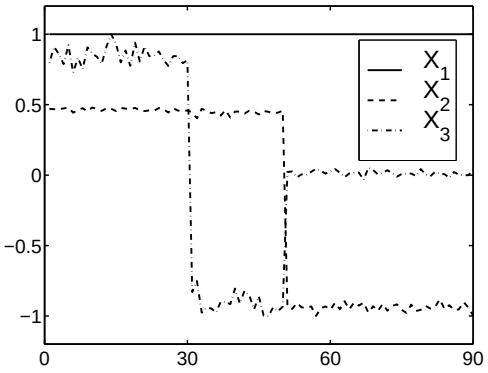

Figure 4: Eigenbasis X_1, X_2, X_3 corresponding to Perron cluster $\lambda = 1, 0.75, 0.52$ of the transition matrix associated with $k = 3$ nearly uncoupled Markov chains. Observe the nearly constant level pattern on each of the index subsets $\mathcal{S}_1, \mathcal{S}_2$ and $\mathcal{S}_3$ – to be compared with Fig. **??** for the uncoupled case.

Summarizing, we finally have the desired *k metastable chemical conformations* (in spatial box discretization) as the k almost invariant subsets $\mathcal{S}_1, \ldots, \mathcal{S}_k$. In the true spirit of scientific computing these objects must be appropriately visualized – a scientific topic of its own right, which, however, cannot be touched upon here. For these conformations the algorithm supplies the following information:

- the probabilities $\pi(\mathcal{S}_i)$ for the system to *be* within the subset $\mathcal{S}_i$, as defined in (**??**),
- the probabilities $w_{ii} = w(\mathcal{S}_i, \mathcal{S}_i, \tau)$ for the system to *stay* during time τ in the subset $\mathcal{S}_i$, as defined in (**??**), and
- the probabilities $w_{ij} = w(\mathcal{S}_i, \mathcal{S}_j, \tau)$, $i \neq j$, for the system to *move* from the subset $\mathcal{S}_i$ to the subset $\mathcal{S}_j$, as defined in (**??**).

In other words: The Perron cluster analysis supplies the number, the life times, and the decay pattern of the metastable chemical conformations. As for the parameter ϵ used above without specification, we naturally arrive at the definition

$$\epsilon = \max_{i=1,\ldots,k} (1 - w_{ii}) = 1 - \min_{i=1,\ldots,k} w_{ii} \tag{2.8}$$

For each of the $\mathcal{S}_i$ the characteristic life times are roughly found to be

$$\tau_{\mathcal{S}_i} \approx \frac{\tau}{1 - w_{ii}} \ .$$

The blow-up from $\tau \ll t_{\text{crit}}$, the deterministic time scale as defined in (**??**), to the time scales $\tau_{\mathcal{S}_i}$ of the metastable conformations is significant. This relation documents in a nutshell the telescoping of the deterministic model, based on short term trajectories, and the statistical model, based on the eigenvalue problem for the (discretized) transition operator, to obtain a long term model.

Above all it is clear that the whole Perron cluster analysis will only work, if the stochastic transition operator T can be discretized avoiding the curse of dimension – which is the topic of the next section.

3 Approximation of the Transition Operator

The spatial stochastic transition operator T as discussed in Section **??** is associated with an underlying Markov chain. Upon introducing the projection π on the position variables via $\pi(q, p) = q$, we may write this Markov chain as

$$q_{k+1} = \pi \, \Phi^\tau(q_k, \, p_k) \ , \qquad p_k : \quad \mathcal{P} - \text{distributed} \ . \tag{3.1}$$

As shown schematically in Fig. **??**, it combines a short term deterministic model, characterized by the flow Φ^τ, with a statistical model, characterized

by the $\mathcal{P}$-distribution, the momentum part of the Boltzmann distribution – see (**??**).

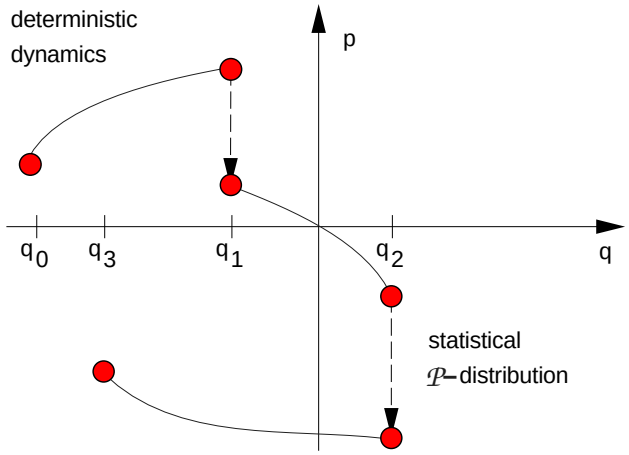

Figure 5: Hybrid Monte Carlo process and Markov chain (**??**)

Hybrid Monte Carlo method. Given a discretization of the position space Ω in terms of boxes $\{B_1, \ldots, B_N\}$, the elements of the transition matrix $T = (T_{ij})$ can be computed by virtue of

$$T_{ij} = \frac{\#\{q_{k+1} \in B_j \ \wedge \ q_k \in B_i\}}{\#\{q_k \in B_i\}} \qquad i, j = 1, \ldots N \ .$$

By construction, the evaluation of the matrix elements thus leads to some *hybrid Monte Carlo* process – see again Fig. **??**. If we run M samples within such a process, then we obtain an approximation $T^{(M)}$ with an approximation error

$$|T - T^{(M)}| \leq \gamma/\sqrt{M} \ .$$

As in all Monte Carlo type processes, *trapping* within local minima will occur, unless we take special precautions. In particular, if the spectral gap at the Perron root approaches 0, then the above constant γ blows up to ∞. However, this is just the case treated here, since we want to analyze Perron clusters! In this situation a technique of temperature embedding has been developed, which circumvents critical slowing down of the MC process in the case under consideration. Unlike simulated annealing this method can "heat" the momenta of the system separately in a nonphysical fashion – compare the

factorization in (**??**). First results have been published in the early paper [**?**] by A. Fischer et al., an improvement in the direction of a hierarchical coupling-uncoupling method can be found in [**?**].

Spatial box discretization. The number N of spatial boxes is also the dimension of the arising transition matrix T. In order to avoid the *curse of dimension* we must assure that N remains of moderate size even for larger molecular systems.

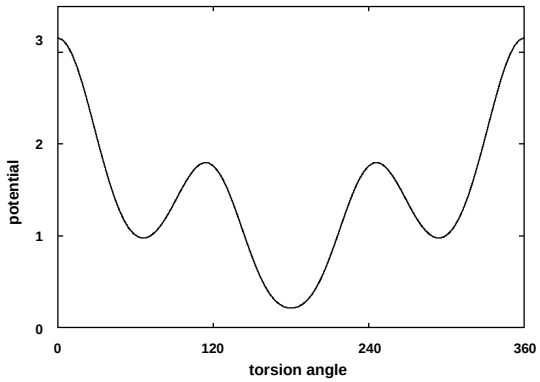

Figure 6: Molecular torsion potential with triple well ($s = 3$)

From chemical insight into the problem, different conformations are caused by the double or triple well structure in the torsion potentials – see Fig. **??**. Let s be the number of minima in the torsion potential ($s = 2$ or $s = 3$) and m the number of torsion angles ($m \approx 7$ per nucleotide), then we obtain a number

$$N \approx s^m$$

of boxes. For the above example molecule ACC we have $m = 37$ and would therefore arrive at some $N > 10^{11}$ – which is certainly intolerable for such a small system!

As a first remedy we adopted the technique of identification of *essential degrees of freedom* originally suggested by Berendsen et al. [**?**]. Generally speaking, this method is based on a principal component analysis (PCA) of fluctuations of the time series obtained by molecular dynamics calculations. We modified the method such that it only works on the torsion angles, i.e. on a preassigned subset of the variables – see [**?**]. Note that in this case the cylinder $\Gamma(A)$ reduces to a fibre associated with this subset – see Fig.

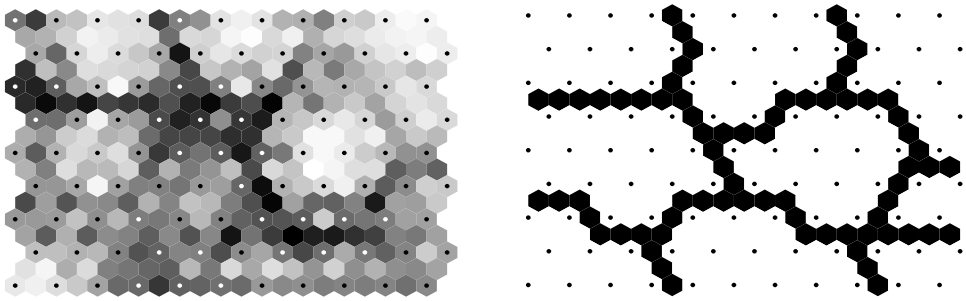

Figure 7: Results of cluster analysis. *Left*: SOM. *Right*: SOM combined with new Perron cluster analysis

??. For the ACC molecule, which was just mentioned above, the method suggests only $m_{\text{ess}} = 4$ generalized torsion coordinates and a number of

$$N_{\text{ess}} = 36$$

of boxes. For a while we were quite content with this approach, until we found out experimentally that it is also not efficient enough for larger molecules.

In a further step of the development the MD group at ZIB recurred to *neural networks*, especially to self-organizing maps (SOM) as suggested by Kohonen [**?**]. Upon combining SOM with the Perron cluster analysis as discussed in Section **??**, T. Galliat et al. managed to develop some much more efficient tool for box discretization – see [**?**]. In Fig. **??** we illustrate the improvement achieved by the addition of the Perron cluster analysis to SOM using a typical SOM representation in terms of hexagonal topology. The result on the right in the figure was obtained a lot faster than the result on the left (a quarter of an hour on a work station as compared to about a week). In [**?**, **?**] the idea has been further developed toward an adaptive multilevel box discretization called *self-organizing box maps* (SOBM) extending techniques from numerical partial differential equations to neural networks.

Example: Tri–nucleotide ACC. This example has been used several times before for illustration purposes. From the neural network approach to box discretization we obtain $N = 54$ boxes. In Fig. **??** the sparse pattern of the associated $(54, 54)$-matrix is given, representing the discretization of the stochastic transition operator over the given 54 boxes.

In Table **??** we list the first eigenvalues of the transition matrix (ordered

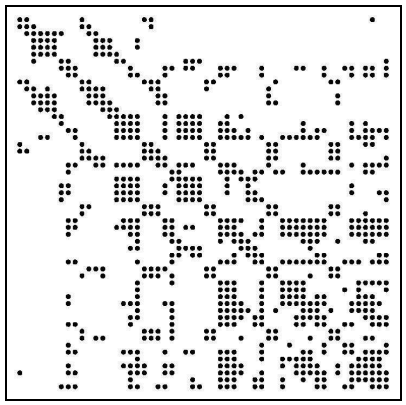

Figure 8: ACC: sparse transition matrix of dimension $N = 54$

k	1	2	3	4	5	6	7	8	9	$\cdots$
λ_k	1.000	0.999	0.995	0.993	0.980	0.972	0.961	0.930	0.874	$\cdots$

Table 1: ACC: eigenvalues of transition matrix

according to modulus). As can be observed, there are gaps at $k = 2$ and at $k = 8$, which can both be analyzed. Note that by construction via Lemma **??** a larger value of k just leads to some substructuring of the conformations: the extended sign structure of the eigenvectors just adds more sign information to the already existing one. In Table **??** we list the computed probabilities $\pi(\mathcal{S}_i)$ to *be* within and $w_{ii} = w(\mathcal{S}_i, \mathcal{S}_i, \tau)$ to *stay* for $\tau = 50$ fsec within one of the conformations $\mathcal{S}_i$ for $i = 1, \ldots, 8$. All elements w_{ii} in the second row are close below 1, which indicates that the computed conformations are in

conformations	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$	$\mathcal{S}_5$	$\mathcal{S}_6$	$\mathcal{S}_7$	$\mathcal{S}_8$
$\pi(\mathcal{S}_i)$	**0.325**	0.097	0.009	0.037	0.107	0.105	**0.273**	0.046
w_{ii}	0.995	0.992	0.919	0.966	0.964	0.991	0.987	0.969

Table 2: ACC: probabilities for metastable conformations

fact *metastable*. Moreover, the numbers in the first row clearly indicate that the conformations $\mathcal{S}_1$ and $\mathcal{S}_7$ dominate the dynamics, which explains the first eigenvalue gap at $k = 2$. From the numbers w_{ii} and definition (**??**) we here obtain the perturbation parameter $\epsilon = 0.081$.

Example: HIV protease inhibitor VX-478. This molecule is the basis for the anti-AIDS drug Agenerase distributed by Glaxo Wellcome. Generally speaking, the HIV is hard to attack directly by drugs, since it is a so-called retrovirus that mutates faster than any molecular recognition can take place. As a consequence, any HIV pharmaceutical will attack the supporting enzymes. One of them is the HIV protease, which regulates the passage of HIV through the cell membrane. The here selected molecule has been exactly designed (by Vertex) to inhibit this passage. The molecular data were taken from the public domain Protein Data Bank (PDB).

We started the conformational analysis at a virtual temperature of $1400K$ (to avoid trapping in the HMC process, see above). At this level there arose $k = 3$ metastable conformations. At the next lower level ($1000K$), these conformations could be analyzed in terms of substructures. In Fig. **??** two out of these substructures are shown. In view of drug design it is important to understand which of the conformations (of the same molecule!) actually exhibits the desired pharmaceutical effect. Questions of this kind can be studied in terms of the probabilities as exemplified above in the tables for the ACC molecule – assuming, of course, that the input potentials give a reliable description of the physics of the molecule.

Perspectives

Conformational analysis opens the door to an understanding of molecular dynamics on time scales of pharmaceutical interest. Even though the essential structure of the mathematical model and its algorithmic realization seem to be quite clear at this time, further progress is needed to allow for the successful analysis of larger biomolecules. In the opinion of the author, the new mathematical concepts of conformational dynamics have a real chance to play an important role in drug design in the near future.

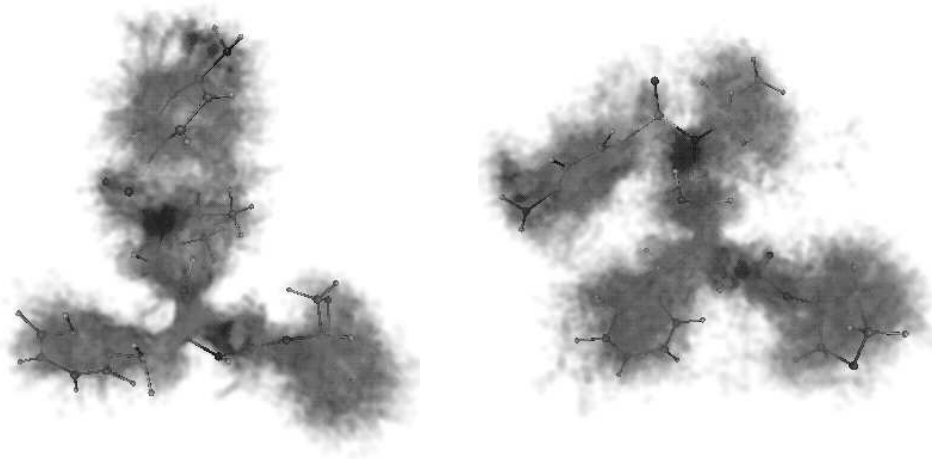

Figure 9: HIV protease inhibitor: T-bone and double T conformations

Acknowledgements. The author is greatly indebted to all members of the ZIB and FU molecular dynamics groups who have joined their efforts to improve the mathematical models and to develop efficient algorithms for computational drug design.

References

[1] A. Amadei, A.B.M. Linssen, and H.J.C. Berendsen. *Essential dynamics on proteins.* Proteins **17**, pp. 412-425 (1993).

[2] V. I. Arnold. *Mathematical Methods of Classical Mechanics.* Second edition. Springer, Heidelberg, New York (1989).

[3] M. Dellnitz and A. Hohmann. *A subdivision algorithm for the computation of unstable manifolds and global attractors*, Numer. Math. **75**, pp. 293–317 (1997).

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

[5] P. Deuflhard and F. Bornemann. *Numerische Mathematik II. Gewöhnliche Differentialgleichungen*. 2. Auflage. Walter de Gruyter, Berlin, New York (2002).

[6] P. Deuflhard and F. Bornemann. *Scientific Computing with Ordinary Differential Equations*. Springer, Berlin, Heidelberg, New York (2002).

[7] P. Deuflhard, M. Dellnitz, O. Junge, and Ch. Schütte. *Computation of essential molecular dynamics by subdivision techniques*. In [**?**], pp. 98–115 (1999).

[8] 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, Berlin, Heidelberg, New York (1999).

[9] P. Deuflhard and A. Hohmann. *Numerische Mathematik I. Eine algorithmisch orientierte Einführung*. 3. Auflage. Walter de Gruyter, Berlin, New York (2002).

[10] P. Deuflhard and A. Hohmann. *Introduction to Scientific Computing*. 2nd edition. Springer, Berlin, Heidelberg, New York (2002).

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

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

[13] A. Fischer, Ch. Schütte, P. Deuflhard, and F. Cordes. *Hierarchical uncoupling-coupling of metastable conformations*. In [**?**], (2002).

[14] T. Galliat. *Adaptive Multilevel Cluster Analysis by Self-Organizing Box Maps*. Submitted as PhD thesis, Department of Mathematics and Computer Ccience, Free University of Berlin, (March 2002).

[15] T. Galliat, P. Deuflhard, R. Roitzsch, and F. Cordes. *Automatic identification of metastable conformations via self-organized neural networks*. In [**?**], (2002).

[16] T. Galliat, W. Huisinga, and P. Deuflhard. *Self-organizing maps combined with eigenmode analysis for automated cluster identification.* In H. Bothe and R. Rojas, editors, *Proceedings of the 2nd International ICSC Symposium on Neural Computation*, Academic Press, pp. 227–232 (2000).

[17] L. Greengard, and V. Rokhlin. *On the evaluation of electrostatic interactions in molecular modeling.* Chem. Ser. **29A**, pp. 139–144 (1989).

[18] T.A. Halgren. *Merck molecular force field. I-V.* J. Comp. Chem., **17**, pp. 490–641 (1996).

[19] T. Kato. *Perturbation Theory for Linear Operators.* Springer, Berlin, Heidelberg, New York (1995).

[20] T. Kohonen. *Self-Organizing Maps.* Springer, Berlin, Heidelberg, New York, 3rd edition (2001).

[21] C. D. Meyer. *Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems.* SIAM Rev., **31**, pp. 240–272 (1989).

[22] J. Sanz-Serna, and M. Calvo. *Numerical Hamiltonian Problems.* Chapman and Hall, London, UK (1994).

[23] T. Schlick. *Some Failures and Successes of Long-Time Approaches to Biomolecular Simulations.* In [**?**], pp. 227–262 (1999).

[24] T. Schlick and H. H. Gan, editors. *Computational Methods for Macromolecules: Challenges and Applications — Proc. of the 3rd Intern. Workshop on Algorithms for Macromolecular Modelling, New York, 2000.* Springer, Berlin, Heidelberg, New York, 2002, in press.

[25] Ch. Schütte. *Conformational Dynamics: Modelling, Theory, Algorithm, and Application to Biomolecules.* Habilitation thesis, Department of Mathematics and Computer Science, Free University of Berlin, 1998. Available as ZIB-Report SC-99-18 via http://www.zib.de/bib/pub/pw/.

[26] Ch. 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**, pp. 146–168 (1999).