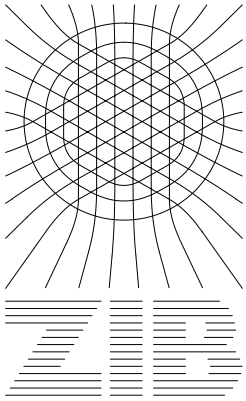

Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

PETER DEUFLHARD CHRISTOF SCHÜTTE

Molecular Conformation Dynamics and Computational Drug Design

ZIB-Report 03–20 (July 2003)

Molecular Conformation Dynamics and Computational Drug Design[1]

Peter Deuflhard[2][3][4] and Christof Schütte[4]

Abstract

The paper surveys recent progress in the mathematical modelling and simulation of essential molecular dynamics. Particular emphasis is put on computational drug design wherein time scales of *msec* up to *min* play the dominant role. Classical long-term molecular dynamics computations, however, would run into ill-conditioned initial value problems already after time spans of only $psec = 10^{-12} sec$. Therefore, in order to obtain results for times of pharmaceutical interest, a combined deterministic-stochastic model is needed.

The concept advocated in this paper is the direct identification of *metastable conformations* together with their life times and their transition patterns. It can be interpreted as a *transfer operator* approach corresponding to some underlying hybrid Monte Carlo process, wherein short-term trajectories enter. Once this operator has been discretized, which is a hard problem of its own, a stochastic matrix arises. This matrix is then treated by *Perron cluster analysis*, a recently developed cluster analysis method involving the numerical solution of an eigenproblem for a Perron cluster of eigenvalues. In order to avoid the 'curse of dimension', the construction of appropriate boxes for the spatial discretization of the Markov operator requires careful consideration. As a biomolecular example we present a rather recent SARS protease inhibitor.

AMS MSC 2000: 65C40, 65C05, 65P10

Keywords: conformation dynamics, Monte Carlo methods, transfer operators, Hamiltonian dynamics, Smoluchowski dynamics, metastable sets, Perron cluster analysis

[1]supported by DFG Research Center "Mathematics for Key Technologies" in Berlin

[2]Invited key note speaker, International Conference on Industrial and Applied Mathematics, July 2003, Sydney, Australia

[3]Zuse Institute Berlin (ZIB)

[4]Free University of Berlin, Dept. Mathematics and Computer Science

Contents

Introduction

In recent years, *prion* diseases, like the mad cow disease, but also *viral* diseases such as HIV or SARS, have attracted much public and political interest. Whenever any new such disease shows up, there is a highly competitive race for new drugs against them. This race typically starts in the computer.

For quite a while, algorithms from discrete mathematics or computer science have already played a publicly visible role – for example, in the decoding of the human genome. These approaches primarily aim at the geometry of the molecules under consideration, i.e., on the secondary or tertiary structure. A real understanding of *biological function*, however, requires detailed knowledge about biomolecular *dynamics*.

In dynamics, 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 $nsec = 10^{-9}sec$ with timesteps of less than $5fsec = 5 \cdot 10^{-15}sec$. The established 'molecular dynamics' approach (usually just called MD) realizes numerical integration of the *Hamiltonian dynamics* of the molecular systems – often limited by the available computer power. This kind of approach, however, has an even stricter mathematical limitation: the Hamiltonian trajectories to be computed are known to be asymptotically chaotic. Consequently, traditional long-term trajectory simulations may, at best, give information about time averages, which, under some ergodic hypothesis, are equivalent to statistical ensemble averages.

As a result of this insight, any investigation of the dynamics of molecular systems for time scales of interest in drug design will require a rather different mathematical approach. In the past few years, the present authors and their joint research group have created such a different approach based on concepts of nonlinear dynamics – for early papers see, e.g., [9, 39, 38, 13]. This approach, now called *conformation dynamics*, has already been surveyed in articles like [7, 40]. The present paper updates the state of the art in this fast moving research topic.

1 Transfer Operators and Metastable Conformations

Hamiltonian dynamics. We assume that the dynamics of the molecular system under consideration is characterized by a *separable* Hamilton function

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

where the first term, the kinetic energy, only depends on the generalized momenta variables p, while the second term, the potential energy, only depends

on the position variables q. From given H, the Hamiltonian differential equations for N atoms are defined as

$$q_i' = \frac{\partial H}{\partial p_i}, \quad p_i' = -\frac{\partial H}{\partial q_i}, \qquad i = 1, \ldots, N. \tag{1.1}$$

Of course, the quality of any molecular dynamics calculation is strongly dependent on the quality of the available potential data (we mostly use MMFF [27]). Details of the numerical integration of these ODEs are omitted here, they can be found, e.g., in Section 1.2. of the textbook [8].

The unique solution of this initial value problem can be written in terms of the flow Φ^t as

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

The sensitivity of the solution, i.e. the solution perturbation $\delta x(t)$ versus the initial perturbation δx_0, is characterized by the *condition number* κ. Following [8, Sect. 3.1.2], this quantity is defined (in first order perturbation analysis) as

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

As already discovered by H. Poincaré, Hamiltonian systems can be *chaotic*. In Numerical Analysis, we want to know the critical finite time, after which some kind of chaoticity (in the sense of almost complete loss of information about the initial state) occurs. In almost all molecular dynamics problems the condition number seems to grow exponentially such that almost all information concerning the initial state is lost after critical times t_{crit} no longer than a few *psec*. That is why the traditional MD with numerical long term integration can only interpreted as computing ensemble averages via time averages in the sense of the ergodic theorem – which need not hold in all cases.

On this basis, we are led to the following conclusion:

Instead of the point concept *of classical mechanics based on deterministic trajectories, with which it is only able to model short-term dynamics, we need to derive some* set concept *including stochastic elements to model long-term dynamics.*

Smoluchowski or Langevin dynamics. In the literature, several stochastic dynamical systems are discussed as alternative models for certain aspects of molecular motion in a heat bath. The most prominent of these are the Langevin or Smoluchowski dynamics. For medium to large molecular systems these models are believed to describe the effective dynamical behavior well enough. The Smoluchowski system models the dynamics in the position space only. It defines a *reversible* Markov process by means of the stochastic differential equation

$$\gamma \, \dot{q} = -\nabla_q V(q) + \, \sigma \, \dot{W}_t. \tag{1.2}$$

Here $\gamma > 0$ denotes some friction constant and $F_{\text{ext}} = \sigma \dot{W}_t$ the external forcing given by a $3N$-dimensional Brownian motion W_t. The external stochastic force is assumed to model the influence of the heat bath surrounding the molecular system. The stochastic differential equation (1.2) defines a continuous time Markov process Q_t on the state space Ω with invariant probability measure [36]

$$\mathcal{Q}(dq) \propto \exp(-\beta V(q))dq \; .$$

There is a long history of using it as a simple toolkit for investigation of dynamical behavior in complicated energy landscapes [4]. We will herein use it for the same purpose, i.e., we will concentrate on the stochastic reformulation of Hamiltonian motion (see next section) but use Smoluchowski dynamics for simplified illustration and comparison.

Biomolecular conformations and metastable sets. Today, the effective dynamics of many biomolecules is understood to be governed by statistically rare transitions between so-called conformations of the biomolecule (cf. [47]). In a conformation, the large scale geometric structure of the molecule is understood to be conserved, whereas on smaller scales the system may well rotate, oscillate or fluctuate. Furthermore, transitions between conformations are rare events or, in other words, a typical trajectory of a molecular system stays for long periods of time within the conformation, while exits are long-term events. Hence, the term conformation includes both *geometric* and *dynamical* aspects. From the geometrical point of view, conformations are understood to represent all molecules with the same large scale geometric structure and may thus be identified with a subset of the state space. From the dynamical point of view, a conformation typically persists for long periods of time (compared to the fastest molecular motions) such that the associated subset of the state space is *metastable* and the resulting *macroscopic dynamical behavior* can be described as a flipping process between the metastable subsets. Consequently, it is of utmost interest to decompose the state space of the molecular motion into some main metastable sets, evaluate the transition probabilities between them and perhaps learn about the transition pathways between these conformations.

The standard biophysical explanation for the existence of conformations is as follows: The *free energy landscape* of a molecular system, say a protein or peptide, decomposes into particularly deep wells each containing huge numbers of local minima. These wells are separated by relatively large barriers—as measured on the scale of the thermal energy ($\sim T$: temperature)—from each other and represent different metastable conformations. The hierarchy of barrier heights induces a hierarchy of conformations [17, 21, 20]. The corresponding hierarchy of time scales observed for conformational transitions seems to confirm the biophysical explanation for the existence of conformations [35]. However, this concept does *not* (at least not directly) refer to

dynamical aspects but describes conformation transitions in terms of a thermodynamic quantity, the free energy. In Section 2 below we will show that the Smoluchowski dynamics is an ideal setting to discuss similarities and differences between this thermodynamic concept and the dynamical concepts to be presented herein.

1.1 Transfer operators

The just mentioned set concept can be realized by virtue of some stochastic transfer operator (or Markov operator), which is discussed here to necessary detail.

Perron–Frobenius operator. Starting point for the new approach was the pioneering work of M. Dellnitz and co-workers [6] on the numerical approximation of invariant measures $\bar{\mu}$ and their corresponding invariant sets $\bar{B}$ via the (unitary) Perron–Frobenius operator $\mathbf{U}$. In terms of this operator, $\bar{\mu}$ and $\bar{B}$ are characterized by the eigenvalue problem

$$\mathbf{U}\bar{\mu} = \bar{\mu}, \quad \Phi^{-t}(\bar{B}) \subset \bar{B}, \quad \forall t \geq 0 \tag{1.3}$$

for the Perron eigenvalue $\lambda = 1$. Moreover, eigenvalues $\lambda \neq 1$ close to the Perron eigenvalue seemed to have an interpretation in terms of *almost invariant sets* of the dynamical system.

The success of that approach was intimately linked to dynamical systems that asymptotically collapse to some dynamics on a low-dimensional manifold. This is definitely not the case in Hamiltonian dynamics, so that a generalization to molecular dynamics is all but trivial. A first attempt in this direction has been published in [9]. However, the subdivision technique applied there caused some *curse of dimension* that restricted the applicability of the method to a domain far away from realistic molecules.

Self-adjoint transfer operator. In [39, 38] a new stochastic operator $\mathbf{T}$ has been constructed, which embeds $\mathbf{U}$ into a canonical distribution

$$f_0(q,p) = \frac{1}{Z}\exp\left(-\beta(p^T M^{-1} p/2 + V(q))\right)$$

with Z as normalization factor and β proportional to the inverse temperature. For separable Hamiltonian H this distribution may be factorized according to

$$f_0 = \mathcal{P}\mathcal{Q}\,, \quad Z = Z_p Z_q\,, \qquad \int \mathcal{P}(p)dp = \int \mathcal{Q}(q)dq = 1\,, \tag{1.4}$$

where

$$\mathcal{P}(p) = \frac{1}{Z_p}\exp(-\frac{\beta}{2}p^T M^{-1} p)\,, \quad \mathcal{Q}(q) = \frac{1}{Z_q}\exp(-\beta V(q))\,.$$

At this point recall that metastable conformations are understood to be objects in *position space* $q \in \Omega$ rather than in the whole phase space Γ. Let $A, B \subset \Omega$ be subsets in position space and define cylinders

$$\Gamma(A) \;=\; \{(q,p) :\; q \in A\}.$$

The required transfer operator may then be constructed integrating the Perron-Frobenius operator $\mathbf{U}$ over the cylinders $\Gamma(\cdot)$ – thus achieving an operator $\mathbf{T}^\tau$ that acts on functions in position space:

$$\mathbf{T}^\tau u(q) \;=\; \int_{\mathbb{R}^d} u(\Pi_q \Phi^{-\tau}(q,\xi))\, \mathcal{P}(\xi) d\xi, \tag{1.5}$$

where Π denotes the projection $\Pi(q,p) = q$ onto the position space. In the sequel we will often omit the superindex τ, if the time scale τ that has been chosen is clear and does not change.

As has been shown in [38], $\mathbf{T}^\tau$ can be interpreted as the transfer operator associated with the Markov chain, to be called *Hamiltonian system with randomized momenta*,

$$q_{k+1} = \Pi\, \Phi^\tau(q_k,\, p_k)\;, \qquad p_k :\quad \mathcal{P} - \text{distributed}\;. \tag{1.6}$$

For a schematic representation see Fig. 1. This Markov chain 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 canonical distribution, which is just a Gaussian distribution due to the quadratic kinetic energy – see (1.4). For a discussion of the physical meaning of this stochastic model of the dynamics visit [41].

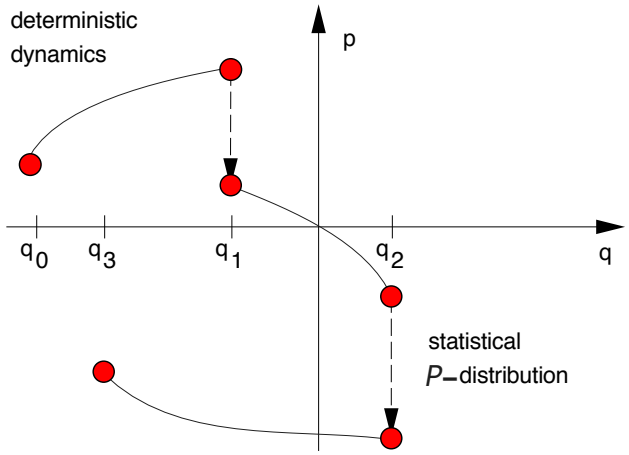

Figure 1: Markov chain (1.6).

The operator $\mathbf{T}^\tau$ is defined over the weighted spaces

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

The Hilbert space $L^2_{\mathcal{Q}}(\Omega)$ is naturally associated with the weighted inner product

$$\langle u, v\rangle_{\mathcal{Q}} = \int_\Omega u(q)v(q)\mathcal{Q}(q)dq \tag{1.7}$$

Among the properties of $\mathbf{T}^\tau$ for all τ we mention (from [38]):

- $\mathbf{T}^\tau$ is a Markov operator on $L^1_{\mathcal{Q}}(\Omega)$.
- $\mathbf{T}^\tau$ is *self-adjoint* in $L^2_{\mathcal{Q}}(\Omega)$.

Hence, its spectrum satisfies $\sigma(\mathbf{T}^\tau) \subset [-1, 1]$. Moreover, under certain quite general conditions the existence of metastable sets is deeply related to a cluster of eigenvalues close to the Perron eigenvalue $\lambda = 1$, called the *Perron cluster*, which is well-separated from the remaining (continuous) part of the spectrum (see Theorem 1.1 for details). Discretization of this operator (to be studied in Section 4 below) generates a stochastic sparse matrix T, which inherits the self-adjointness of the operator as symmetry with respect to a discrete analog of the weighted inner product $\langle\cdot,\cdot\rangle_{\mathcal{Q}}$.

With these preparations, we are ready to express all relevant information about the dynamical system. Let $\chi(A)$ denote the characteristic function of A, a set function that is 1 inside A and 0 outside. Then we obtain:

- The probability for the dynamical system to *be* within A is

$$\pi(A) = \int_{\Gamma(A)} f_0(p, q)dq\ dp = \int_A \mathcal{Q}(q)dq = \langle\chi_A, \chi_A\rangle_{\mathcal{Q}}\ . \tag{1.8}$$

- The conditional probability for the system, once it is in A, to *move* from A to B during time τ can be defined by virtue of $\mathbf{T}^\tau$ as

$$w(A, B, \tau) = \frac{\langle\chi_A, \mathbf{T}^\tau\chi_B\rangle_{\mathcal{Q}}}{\langle\chi_A, \chi_A\rangle_{\mathcal{Q}}}\ . \tag{1.9}$$

- The probability for the system, once it is in A, to *stay* in A during time τ (more exactly: to be found in A at time $t = \tau$ after being in A at time $t = 0$) comes out as

$$w(A, A, \tau) = \frac{\langle\chi_A, \mathbf{T}^\tau\chi_A\rangle_{\mathcal{Q}}}{\langle\chi_A, \chi_A\rangle_{\mathcal{Q}}}\ . \tag{1.10}$$

Given open sets A and B, we could compute these probabilities by means of long-term iteration of the Markov chain (1.6) associated with $\mathbf{T}^\tau$. Any realization would yield an sequence of positions $\{q_k\}$ that can be proved to be distributed according to $\mathcal{Q}$ asymptotically [38]. The relative frequency of transitions from A to B in this sequence asymptotically approximates $w(A, B, \tau)$ (see Section 4 for algorithmic consequences and difficulties). In addition we get a *sequence of τ-sub-trajectories* of the Hamiltonian system under consideration. If long enough this sequence will explore the state space entirely and contain all necessary information about the dynamical features of the system.

Transfer operator for Smoluchowski dynamics. The transfer operator describes the evolution of probability densities under the dynamics in question. For the Smoluchowski system (1.2) the evolution of probability densities f (w.r.t. the Lebesgue measure) is governed by the Fokker-Planck equation

$$\partial_t f \;=\; \left(\frac{\sigma^2}{2\gamma^2} \Delta_q + \frac{1}{\gamma} (\nabla_q V(q) \cdot \nabla_q \;+\; D^2 V(q)) \right) u.$$

Upon introducing the probability distribution $v = u/\mathcal{Q}$, this evolution equation reads

$$\partial_t v \;=\; \mathbf{A}_{\mathrm{Smo}}\, v \;=\; \left(\frac{\sigma^2}{2\gamma^2} \Delta_q - \frac{1}{\gamma} (\nabla_q V(q) \cdot \nabla_q) \right) v.$$

Thus, the associated transfer operators $\mathbf{T}^t_{\mathrm{Smo}}$ form a semigroup. For twice continuously differentiable $u \in L^r_{\mathcal{Q}}(\Omega)$ with $1 \leq r < \infty$, this semigroup admits $\mathbf{A}_{\mathrm{Smo}}$ as a strong generator such that in this case

$$\mathbf{T}^t_{\mathrm{Smo}} \;=\; \exp(t\mathbf{A}_{\mathrm{Smo}})\;.$$

For details on $\mathbf{A}_{\mathrm{Smo}}$ see the theory of Fokker-Planck equations and Kolmogoroff forward and backward equations [36, 42, 28].

Hence the Smoluchowski case gives us the opportunity to study the relation between dominant eigenvectors of the transfer operator and metastable sets by means of partial differential operators. The fact that the Smoluchowski system has at least some relation to the Hamiltonian case is reflected in the following relation between the transfer operator $\mathbf{T}^\tau$ of the Hamiltonian system with randomized momenta and the Smoluchowski generator:

$$\mathbf{T}^\tau \;=\; \mathrm{Id} \;+\; \tau^2 \mathbf{A}_{\mathrm{Smo}} \;+\; \mathcal{O}(\tau^4).$$

For $u \in L^2_{\mathcal{Q}}(\Omega)$ the reversibility of the Smoluchowski dynamics implies that $\mathbf{A}_{\mathrm{Smo}}$ is self-adjoint.

1.2 Dominant Spectra and Metastability

There are several recent articles on the relation between metastability and dominant eigenmodes of the transfer operator associated with the considered dynamical system [41, 30, 28, 6, 39]. Within these approaches, metastability is a *set-wise* notion and conceptually defined in the following way: some dynamical system is said to exhibit metastability or to have a *metastable decomposition*, if its state space can be decomposed into a finite (hopefully small) number of disjoint sets such that the *probability of exit* from each of these sets is extremely small [41, 6]. There are basically two different concepts of probability of exit: (a) the probability of exit from a set is defined via an ensemble of systems and measures the fraction of systems that exit from the set during some fixed time interval [41, 39], (b) in case of a stochastic process the probability of exit is measured from the distribution of exit times from the set, i.e., the probability of exit is the smaller the larger the expected exit time is [3], or, equivalently, the slower the decay of the distribution of exit times is [30]. However, both concepts (a) and (b) are related to the dominant eigenvectors of the transfer operator. Accordingly, the basic insight of the transfer operator approach to metastability is [41]:

Identification of metastable decompositions. *Metastable decompositions can be detected via the discrete eigenvalues of the transfer operator* $\mathbf{T}^\tau$ *close to its maximal eigenvalue* $\lambda = 1$*; they can be identified by exploiting structural properties of the corresponding eigenfunctions. In doing so, the number of sets in the metastable decomposition is equal to the number of eigenvalues close to* 1*, including* $\lambda = 1$ *and counting multiplicity.*

We will later learn about the identification algorithm constructed based on this idea. Furthermore, we will present illustrating examples in Section 2. In the final paragraphs of this section however, we will present one of several mathematical statements supporting this idea. To this end, recall the formula for the probability to remain within some set A during time span τ:

$$w(A, A, \tau) = \frac{\langle \chi_A, \mathbf{T}^\tau \chi_A \rangle_\mathcal{Q}}{\langle \chi_A, \chi_A \rangle_\mathcal{Q}} \ .$$

The metastability of a set A may be measured by $w(A, A, \tau)$.

Definition: Metastability of a decomposition. For an arbitrary decomposition $\mathcal{D} = \{A_1, \ldots, A_m\}$ of the state space into m disjoint sets A_k, we define

$$w_m(\tau) = \sum_{i=1}^{m} w(A_i, A_i, \tau) \tag{1.11}$$

as the corresponding metastability.

The following crucial result is due to [31]; a specialized version for two subsets has been published by Huisinga in his thesis [28].

Theorem 1.1. *Let $T^\tau : L^2_{\mathcal{Q}}(\Omega) \to L^2_{\mathcal{Q}}(\Omega)$ denote a reversible transfer operator whose essential spectral radius is strictly less than 1 and for which the eigenvalue $\lambda = 1$ is simple. Then $\mathbf{T}^\tau$ is self–adjoint and the spectrum has the form*

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

with $-1 < a \leq b < \lambda_m \leq \ldots \leq \lambda_1 = 1$ and isolated, not necessarily simple eigenvalues of finite multiplicity that are counted according to multiplicity. Denote by $v_m, \ldots, v_1$ the corresponding eigenfunctions, normalized to $\|v_k\|_{L^2_{\mathcal{Q}}(\Omega)} = 1$. Let Q be the orthogonal projection of $L^2_{\mathcal{Q}}(\Omega)$ onto $\mathrm{span}\{\chi_{A_1}, \ldots, \chi_{A_m}\}$. *Then the following bounds hold:*

$$1 + \kappa_2\lambda_2 + \ldots + \kappa_m\lambda_m + c \leq w_m(\tau) \leq 1 + \lambda_2 + \ldots + \lambda_m,$$

where $\kappa_j = \|Qv_j\|^2_{L^2_{\mathcal{Q}}(\Omega)} \leq 1$, $j = 1, \ldots, m$, and $c = |a|\,(1-\kappa_2)\ldots(1-\kappa_m) < 1$.

This theorem obviously holds for the transfer operator of the Hamiltonian system with randomized momenta as well as for the one related to Smoluchowski dynamics. Whenever the dominant eigenfunctions $v_2, \ldots, v_m$ are almost constant on the metastable subsets $A_1, \ldots, A_m$ – which then implies that $\kappa_j \approx 1$ and thus $c \approx 0$ – then the above lower and upper bound are close. Moreover, Huisinga et al. [31] have even shown that both bounds are sharp and asymptotically exact. The idea to exploit almost constancy and sign changes of the dominant eigenmodes lies exactly at the heart of the algorithm to be presented that identifies a metastable decomposition into m sets via the m dominant eigenmodes, see Section 3 for details.

2 A Complete Picture in a Simplified Setting

In principle, the transfer operator approach as presented so far allows us to identify an almost optimal metastable decomposition of the state space. In terms of the biochemical background this gives us the main conformations of the molecular system under consideration. However, this solves "only" one of the four most important biophysical problems: One may want to (a) identify the *dominant conformations*, (b) characterize the *geometric and dynamical flexibility* of the molecule within one of its main conformations, (c) estimate the *transition probabilities* between conformations or the *exit rates* from a single one, and (d) characterize the *transition regions* and *pathways* on which the transitions between conformations will occur most probably.

The characterization of the internal flexibility and (roughly) of the transition regions are automatic by-products of the algorithmic realization of the transfer operator approach via sequences of sub-trajectories exploring state space (see Section 4 below). Furthermore this algorithmic realization will permit the direct computation of the transition probabilities w.r.t. some prescribed time span τ by means of formula (1.9).

In order to illustrate the relation between the different concepts (transfer operator approach, exit rates, transition pathways) we now work out details at a rather simple test case from *Smoluchoswki dynamics*.

Test system (2D). We consider the two-dimensional system given by the potential

$$\begin{aligned} V(x,y) &= 3\exp(-x^2-(y-1/3)^2) - 3\exp(-x^2-(y-5/3)^2) \\ &\quad -5\exp(-(x-1)^2-y^2) - 5\exp(-(x+1)^2-y^2); \end{aligned}$$

The potential is illustrated in Fig. 2. We observe that there are two equally important deep wells with minima at $(x,y) = (1,0)$ and $(x,y) = (-1,0)$, and a not so important one around $(x,y) = (0,5/3)$. The barrier between the two dominant wells is substantially higher than the barrier between each of the dominant ones and the less important wells. The inverse temperature $\beta = 2\gamma/\sigma^2$ is set to $\beta = 3$ (with $\gamma = 1$) such that crossing the barriers in the potential certainly will be a rare event.

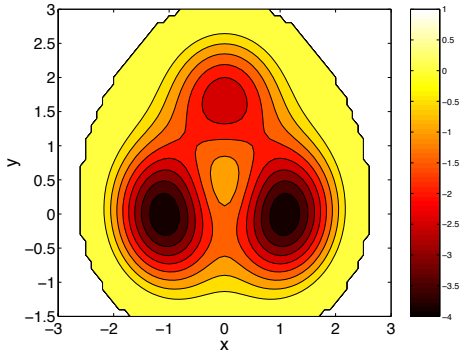

Figure 2: *Smoluchowski test problem:* Contour lines of potential V. Regions between the contour lines are shaded according to average value of potential.

Metastable decomposition. The eigenvalue problem of the generator $\mathbf{A}_{\text{Smo}}$ of the transfer operator $\mathbf{T}^t_{\text{Smo}}$ can be solved numerically by means of finite element eigenvalue solvers for elliptic problems. This leads to the following numerical results for the dominant eigenvalues of $\mathbf{A}_{\text{Smo}}$ in $L^2_{\mathcal{Q}}(\Omega)$:

λ_1	λ_2	λ_3	λ_4	...
0.000	-0.002	-0.144	-2.330	...

The eigenfunction associated with $\lambda_1 = 0$ obviously is the constant function. The eigenfunctions associated with λ_2 and λ_3 are shown in Fig. 3.

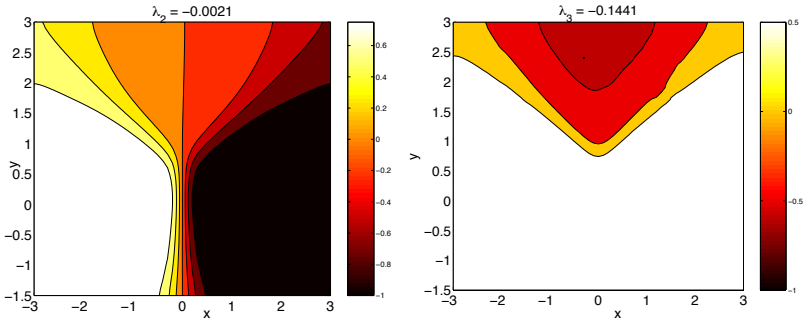

Figure 3: *Smoluchowski test problem:* second and third eigenfunction. Illustration via contour lines as explained in Fig. 2.

From Fig. 3 we observe: (a) the three metastable sets given by the three wells in the potential show up as regions of almost constancy of the three eigenfunctions, (b) the less important well (being coded into the third eigenfunction with a significantly larger eigenvalue) obviously exhibits less metastabilty than the other two ones.

In Section 3 below, we will present an algorithm for the identification of metastable decompositions as shown in Fig. 4.

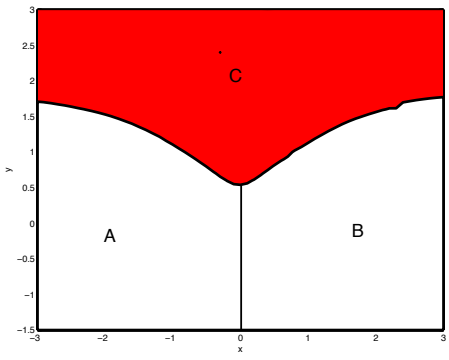

Figure 4: State space decomposition into sets A, B, C by identification algorithm as presented in Section 3.

Exit times and exit rates. The exit rate from a set A is defined as the decay rate of the exponential decay of the distribution of exit times from the set [30]. If A is the open set defined by one of the strictly positive or negative components of an eigenfunction of $\mathbf{A}_{\mathrm{Smo}}$, then the exit rate can be shown to equal the modulus of the eigenvalue associated with this eigenfunction [30]. If, for example, A is the left of the main wells of our example, the exit rate for $\beta = 3$ is given by $|\lambda_2| = 0.002$, i.e., most exit times will be in the hundreds of units. If β is asymptotically large, the expected exit time $\bar{\tau}$ is known to scale like

$$\bar{\tau} \;=\; C \exp(\beta \Delta V)$$

where ΔV denotes the smallest energy barrier via which the exit is possible. However, the preconstant C increases with the "narrowness" of the saddle point region through which the exit occurs. Situations like in our example are of utmost interest: there are two such regions, one that is a little bit wider but whose energy barrier is a little higher than that of the other one (which is more narrow). If β is not asymptotically large exits will occur in both regions. For real life applications this is the crucial problem of all algorithms designed to identify transition regions, pathways, or states: one always has to ask whether all important regions have been explored.

Transition pathways. Transition state theory tells us that the transition pathways between two (disjoint) wells W_1 and W_2 can be computed from some reaction coordinate $\xi : X \to \mathbb{R}$ where X may denote the important portion of the state space in which the wells $W_1 \subset X$ and $W_2 \subset X$ are dominant. IN the case of our test system, W_1, and W_2 should be left and right main wells of the potential energy landscape, i.e., the core sets of the metastable sets A and B of Fig. 4. In general, ξ is given by the following boundary value problem [43]:

$$\begin{aligned} \mathbf{A}_{\mathrm{Smo}}\,\xi &= 0, \quad \text{in } X \setminus \{W_1 \cup W_2\} \\ \xi|_{\partial W_1} &= 0 \\ \xi|_{\partial W_2} &= 1 \\ \partial_n \xi|_{\partial X} &= 0\,. \end{aligned} \tag{2.1}$$

Under certain circumstances the solution is closely related to the dominant eigenvalues: Let, e.g., μ denote the probability measure generated by the invariant density $\exp(-\beta V)$, and suppose that almost all weight is concentrated in W_1 and W_2, i.e., $\mu(W_1 \cup W_2) \approx 1$. Moreover, let there be only one negative eigenvalue λ_2 of $\mathbf{A}_{\mathrm{Smo}}$ very close to $\lambda_1 = 0$. Then we approximately have

$$\xi \approx \mu(W_1)\chi_X \;+\; \sqrt{\mu(W_1)\mu(W_2)}\, u_2,$$

where u_2 denotes the eigenfunction associated with λ_2.

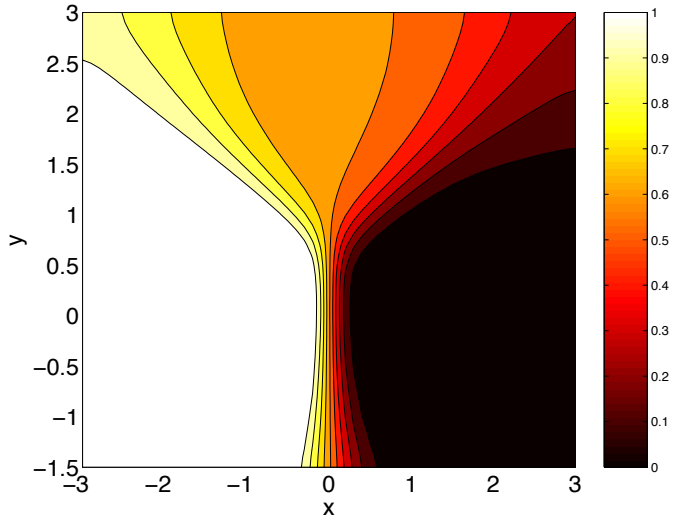

Figure 5: Reaction coordinate ξ for the system under consideration with W_1 being the left and W_2 the right main well. Illustration via contour lines in the same sense as explained in Fig. 2.

This is obviously the case in our test system: as we can see in Fig. 5, the solution of (2.1) for the case where W_1 and W_2 are identical to the left and right main well, respectively. The figure exhibits the *level set* $\xi = \xi_0$ for given ξ_0 of ξ. Transition state theory tells us that the transition pathways intersect the level sets of ξ perpendicularly [43].

From all possible transition paths only those are of importance that intersect the level sets where the restricted invariant distributions

$$\nu|_{\xi_0} = \frac{1}{Z(\xi_0)} \exp(-\beta V)|_{\xi=\xi_0}, \quad Z(\xi_0) = \int \delta(\xi - \xi_0) \exp(-\beta V(x,y))\, dx\, dy$$

is large enough.

Fig. 6 exhibits some of these restricted invariant distributions together with the level sets of the reaction coordinate ξ for a transition from the left to the right main well of our test system. We observed that there are at least two different transition regions, that contain different optional transition pathways. In situations like this the usual concept of free energy landscapes is not general enough; e.g., it is not clear over which variables the energy landscape has to be averaged in order to compute an useful free energy for both transition regions. However, it should be obvious that the identification of transition regions is closely related to the dominant eigenmodes of the transfer operator, and that a complete picture of the effective dynamical effects of the systems has to be based on the information coded in these dominant eigenmodes. An very promising direction of work that combines

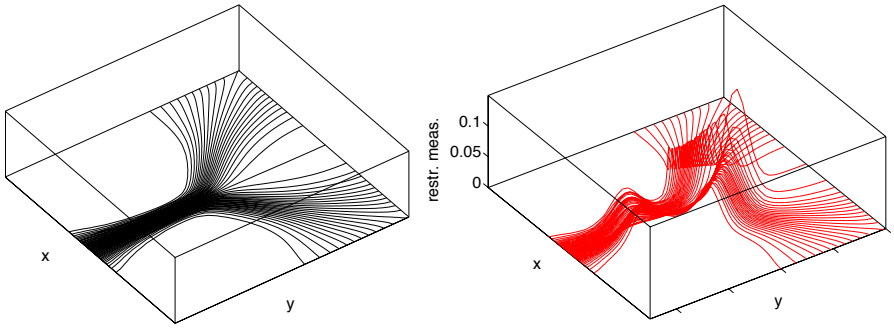

Figure 6: Level sets of ξ (left) and restricted invariant distribution $\nu|_{\xi_0}$ (right) on the level sets $\xi = \xi_0$ for some $\xi_0 = 0.1, \ldots, 0.9$ for the system under consideration.

aspects of the "global" transfer operator approach with a "localized", so-called *string method* [16] for the direct computation of transition pathways is presented in [43].

3 Perron cluster analysis

Suppose we have already discretized the above transfer operator $\mathbf{T}$ – a topic postponed to the subsequent Section 4, since it requires techniques to be presented first. Then, in order to identify m almost invariant sets corresponding to m metastable chemical conformations, we need only deal with a stochastic (generalized symmetric) matrix T of dimension N. This is a problem of *cluster analysis*, where, in addition, m is unknown in advance. Comparable to (1.3), we start from the eigenvalue problem

$$\pi^T T = \pi^T \,, \quad Te = e \,, \qquad \pi^T e = 1 \,, \tag{3.2}$$

where the left eigenvector $\pi^T = (\pi_1, \ldots, \pi_N)$ represents the discrete invariant measure and the right eigenvector $e^T = (1, \ldots, 1)$ the characteristic function of the discrete invariant set – each corresponding to the Perron eigenvalue $\lambda_1 = 1$. The basic approach to be described requires an analysis of the *Perron cluster* of eigenvalues

$$\lambda_1 = 1, \lambda_2 \approx 1, \ldots, \lambda_m \approx 1$$

and their corresponding eigenvectors $V_m = [v_1, \ldots, v_m]$. For given $u, v \in \mathbb{R}^N$ we will use the special inner product and norm

$$\langle u, v \rangle_\pi = \sum_{l=1}^{N} u_l \pi_l v_l = u^T D^2 v \,, \quad \|v\|_\pi = \langle v, v \rangle_\pi^{1/2} \,, \tag{3.3}$$

where $D = \mathrm{diag}(\sqrt{\pi_1}, \ldots \sqrt{\pi_N})$ is a diagonal scaling matrix. Obviously, (3.3) is the discrete analog of the continuous inner product (1.7). Any reversible stochastic matrix T is symmetric under this inner product; as a consequence, for any right eigenvector $y = (y_1, \ldots, y_N)$ there exists a left eigenvector $z = (z_1, \ldots, z_N)$ with $z_l = \pi_l y_l$, or, equivalently,

$$z = D^2 y \; . \tag{3.4}$$

Algorithm PCCA. The first algorithm to tackle this problem has been the PCCA method (abbreviated from: **P**erron **C**luster **C**luster **A**nalysis), as worked out in detail in [13]; for a rather elementary introduction see also Section 5.5 of the latest edition of the undergraduate textbook [11]. We will also sketch the more robust variant called PCCA+, which has originally been suggested by M. Weber [44, 45] and will be further improved in a forthcoming paper [14].

Uncoupled Markov chains. Let $\mathcal{S} = \{1, 2, \ldots, N\}$ denote the total index set decomposed as

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

into m disjoint index subsets, which represent m uncoupled Markov chains, each of which is running "infinitely long" within the corresponding subset. Then the total transition matrix T is strictly block diagonal with block submatrices $\{T_1, \ldots, T_m\}$ – see, e.g., [33]. Each of these submatrices is stochastic and gives rise to a single Perron eigenvalue $\lambda(T_i) = 1$, $i = 1, \ldots, m$. Let the submatrices be primitive. Then, due to the Perron-Frobenius theorem, each block T_i possesses a unique right eigenvector $e_{\mathcal{S}_i} = (1, \ldots, 1)^T$ of length $\dim(T_i)$ having unit entries over the index subset $\mathcal{S}_i$. Therefore, in terms of the total transition matrix T, the eigenvalue $\lambda = 1$ has multiplicity m and the corresponding eigenspace is spanned by the vectors

$$\chi_i = (0, \ldots, 0, {e_{\mathcal{S}_i}}^T, 0, \ldots, 0)^T, \qquad i = 1, \ldots, m \; .$$

Our notation deliberately emphasizes that these eigenvectors can be interpreted as *characteristic functions* of the invariant index subsets (see Fig. 7, left).

In general, any Perron eigenbasis $V_m = \{v_1, \ldots, v_m\}$ can be written as a linear combination of the characteristic functions $\chi = [\chi_1, \ldots, \chi_m]$ such that

$$\chi = V_m \mathcal{A} \; , \quad V_m = \chi \mathcal{A}^{-1} \tag{3.5}$$

wherein the (m, m)-matrix $\mathcal{A} = (\alpha_{ij})$ is nonsingular (due to $\dim \ker(\mathcal{A}) = 0$) so that $\mathcal{A}^{-1} = (a_{ij})$ exists. In PCCA, each subset $\mathcal{S}_i$ for $i = 1, \ldots, m$ is identified by some componentwise sign structure of the eigenvectors V_m using the three values $\{+, 0, -\}$ for the sign function — compare [12].

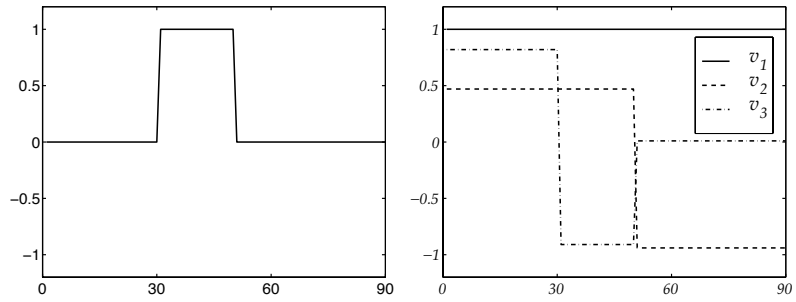

Figure 7: Uncoupled Markov chain over $m = 3$ disjoint index subsets. The state space $\mathcal{S} = \{s_1, \ldots, s_{90}\}$ divides into the index subsets $\mathcal{S}_1 = \{s_1, \ldots, s_{29}\}$, $\mathcal{S}_2 = \{s_{30}, \ldots, s_{49}\}$, and $\mathcal{S}_3 = \{s_{50}, \ldots, s_{90}\}$. *Left*: Characteristic function $\chi_2 = e_{\mathcal{S}_2}^T$. *Right*: Perron eigenbasis $V_3 = \{v_1, v_2, v_3\}$ corresponding to 3-fold Perron eigenvalue $\lambda = 1$.

Nearly uncoupled Markov chains. Suppose now we have m nearly uncoupled Markov chains, each of which is staying "for a long time" in one of the conformations i. For the transition probabilities (1.9) and (1.10) this means that

$$w(i, i, \tau) = 1 - O(\epsilon), \qquad w(i, j, \tau) = O(\epsilon), \quad i \neq j \ , \tag{3.6}$$

in terms of some perturbation parameter ϵ not further specified here. In this case the transition matrix T is (after some unknown permutation) block diagonally *dominant*. As a perturbation of the m-fold Perron root in the uncoupled case $\epsilon = 0$, the *Perron cluster*

$$\lambda_1 = 1, \quad \lambda_i = 1 - O(\epsilon), \quad i = 2, \ldots, m$$

arises. In the PCCA approach, the cluster identification is done exploiting the fact that, for $\epsilon = 0$, each cluster is clearly associated with the set of signs of the components of the eigenvectors $v_1, \ldots, v_m$, where $v_1 = e$ is set. Clearly, the signs of the components are preserved as long as the perturbation ϵ is 'small enough'; for 'too small' entries in an eigenvector $v_i, i > 1$, however, we will have to define some 'dirty zero' as a perturbation of the exact sign function value 0 – compare v_3 over the index subset $\mathcal{S}_3$ in Fig. 7, right. Therefore, in PCCA, the least squares requirement

$$\|\chi - V_m \mathcal{A}\|_\pi = \min \tag{3.7}$$

is imposed and solved iteratively by successive reduction of the 'dirty zero' parameter. In this way, some discontinuity enters into the algorithm, which leads to some lack of robustness of the PCCA approach as a whole.

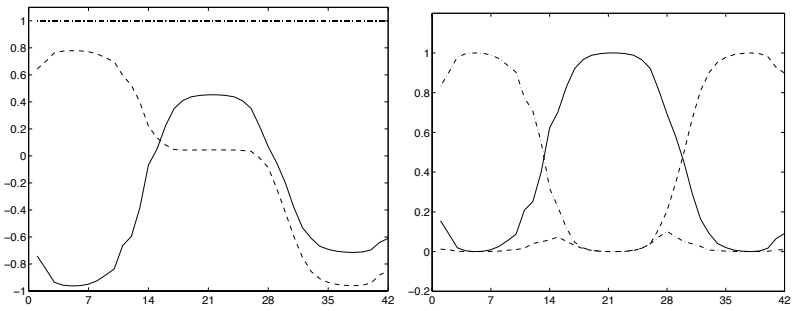

Figure 8: Perron cluster $\lambda = 1, 0.99, 0.98$ in butane molecule. *Left:* Eigenbasis v_1, v_2, v_3. *Right:* Soft characteristic functions.

Algorithm PCCA+. In this approach, the linear least squares problem (3.7) is replaced by modifying the 'crisp' characteristic functions χ_i to certain 'soft characteristic functions' $\widetilde{\chi}_i(\epsilon)$ as represented schematically in Fig. 8. This may be interpreted as replacing the sets by 'fuzzy sets'. The soft characteristic functions are defined such that the relation (3.5) is modified according to

$$\widetilde{\chi} = V_m \widetilde{\mathcal{A}} \, . \tag{3.8}$$

Moreover, they are assumed to satisfy the *positivity* property

$$\widetilde{\chi}_i(l) \geq 0 \, , \quad i = 1, \ldots, m \, , \quad l = 1, 2, \ldots, N \tag{3.9}$$

and the *partition of unity* property

$$\sum_{i=1}^{m} \widetilde{\chi}_i(l) = 1 \, , \quad l = 1, 2, \ldots, N \, . \tag{3.10}$$

The actual computation of $\widetilde{\chi}$ is performed such that the *metastability* $w_m(\tau)$ as defined in (1.11) above is *maximized*, which is a well–known problem from discrete mathematics; the link to Theorem 1.1 is obvious. More details will be given in [14].

In view of the property (3.4), we may define

$$\widetilde{\pi}_i = D^2 \widetilde{\chi}_i = D^2 v_i \widetilde{\mathcal{A}}$$

via the left eigenvectors $D^2 v_i$ corresponding to the right eigenvectors v_i. In other words, we may interpret the soft characteristic functions $\widetilde{\chi}_i$ via the modified probabilities

$$\widetilde{\pi}_i = (\widetilde{\pi}_i(1), \ldots, \widetilde{\pi}_i(N) \,) = (\pi_1 \widetilde{\chi}_i(1), \ldots, \pi_N \widetilde{\chi}_i(N)) \tag{3.11}$$

associated with conformation i.

From this analysis we finally obtain the desired *m metastable chemical conformations* via the m soft characteristic functions $\widetilde{\chi}_1, \ldots, \widetilde{\chi}_m$. They may be interpreted as "mixed states" generated by perturbation of "pure states" $\chi_1, \ldots, \chi_m$. For these conformations the algorithm supplies the following information:

- the probabilities $\widetilde{\pi}_i$ for the system to *be* within state i as

$$\widetilde{\pi}_i = \pi^T \widetilde{\chi}_i = \langle \widetilde{\chi}_i, e \rangle_\pi \ , \tag{3.12}$$

  which is a variation of (1.8),

- the probabilities $w_{ii} = w(i, i, \tau)$ for the system, once it is in state i, to *stay* during time τ

$$w_{ii} = \frac{\langle \widetilde{\chi}_i, T\widetilde{\chi}_i \rangle_\pi}{\langle \widetilde{\chi}_i, e \rangle_\pi} = \frac{\langle \widetilde{\chi}_i, T\widetilde{\chi}_i \rangle_\pi}{\widetilde{\pi}_i} \ , \tag{3.13}$$

  which is a variation of (1.10), and

- the probabilities $w_{ij} = w(i, j, \tau)$, $i \neq j$, for the system, once it is in state i, to *move* to state j,

$$w_{ij} = \frac{\langle \widetilde{\chi}_i, T\widetilde{\chi}_j \rangle_\pi}{\langle \widetilde{\chi}_i, e \rangle_\pi} = \frac{\langle \widetilde{\chi}_i, T\widetilde{\chi}_j \rangle_\pi}{\widetilde{\pi}_i} \ , \tag{3.14}$$

  which is a variation of (1.9).

As for the parameter ϵ used above without specification, we quote the definition

$$\epsilon = \max_{i=1,\ldots,m} (1 - w_{ii}) = 1 - \min_{i=1,\ldots,m} w_{ii} \ , \tag{3.15}$$

which has been derived in [13].

Summarizing, we may state the following:

Given a sufficiently accurate approximation matrix T of the transfer operator $\mathbf{T}$*, the Perron cluster analysis supplies the number, the life times, and the decay pattern of the metastable chemical conformations.*

4 Approximation of Stochastic Operator

The whole Perron cluster analysis as described in Section 3 will only work, if the stochastic operator $\mathbf{T}$ can be approximated appropriately, which is the topic of this section. As has been shown in [38], $\mathbf{T}$ can be interpreted as a *transition operator* associated with the Markov chain (1.6).

Hybrid Monte Carlo method (HMC). First we want to briefly describe the mixed deterministic-stochastic process that directly mimics the Markov chain shown in Fig. 1. For details see references [15, 39].

In order to approximate the Hamiltonian flow Φ^τ in the definition of the transfer operator, we will have to discretize the Hamiltonian equations of motion (1.1). Suppose that this discretization with time step $h = \tau/k$ yields the discrete flow Ψ^h such that $\Phi^\tau x_0$ is approximated by

$$x_{j+1} = \Psi^h x_j \ , \quad j = 0, \ldots, k-1 \ .$$

All explicit discretizations with certain long-term stability properties, e.g., symplectic ones, do *not* exactly conserve the energy. Therefore, the chain

$$q_{k+1} = \pi \, (\Psi^h)^N (q_k, \, p_k) \ , \qquad p_k : \quad \mathcal{P} - \text{distributed} \ ,$$

will in general *not* sample the distribution $\mathcal{Q}$ of interest. In order to correct this, one has to use the Metropolis acceptance procedure. This yields the HMC chain, which leads to a chain of the same structure as the one shown in Fig. 1, has the correct invariant measure, and still contains good approximations of sub-trajectories of the Hamiltonian system.

Monte Carlo approximation of transition probabilities. Given a discretization of the position space Ω in terms of boxes $\{B_1, \ldots, B_N\}$, and a realization $\{q_1, \ldots, q_M\}$ of the HMC chain, the elements T_{ij} of the transition matrix T 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 means of this we obtain an approximation $T^{(M)}$ with an error like

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

where this estimate has to be understood in the sense of the central limit theorem for Markov chains (under special conditions there are much sharper convergence results [34]). As in all Monte Carlo type processes, however, *trapping* within local minima will occur, unless we take special precautions. In fact, the above constant γ exceeds any bound, if the spectral gap at the Perron root approaches 0. However, as we want to analyze Perron clusters, this is just the case treated here. Below we will present a temperature embedding technique especially designed to deal with this difficulty.

Spatial box discretizations. The number N of spatial boxes is also the dimension of the arising transition matrix T. Of course, we must assure that N remains of moderate size even for larger molecular systems. From chemical insight into the problem, different conformations occur corresponding to the double or triple well structure in the torsion angle potentials – see

Fig. 9. Let s be the number of minima in the torsion potential ($s = 2$ or $s = 3$) and n the number of torsion angles ($n \approx 7$ per nucleotide), then our first applied subdivision technique from [9] would have led to a number

$$N \approx s^n$$

of boxes. For the small RNA segment with 70 atoms and three genetic letters (ACC) given in [7], we have $n = 37$; this would have led to $N > 10^{11}$, which is, of course, intolerable! This combinatorial explosion is the well-known "curse of dimension".

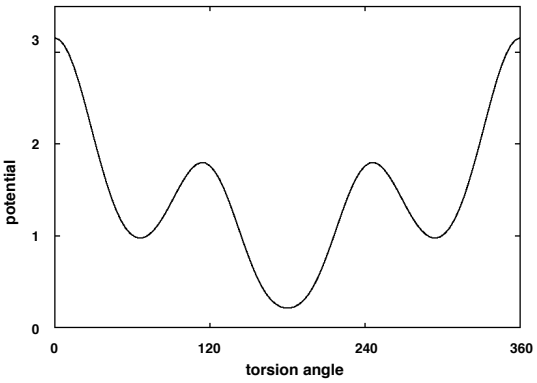

Figure 9: Molecular torsion potential with triple well ($s = 3$)

In order to overcome this undesirable effect, we have experimented with several heuristics. First, we adapted the method suggested by Amadei et al. [1] to circular coordinates [29, 39]; this method identifies "essential degrees of freedom" by principal component analysis (PCA) of dynamical fluctuations. This technique turned out to lack robustness already for quite small molecules. Next, we tried self-organizing maps (SOM) due to Kohonen [32] in combination with our PCCA: the speed-up of the combined cluster algorithm has been reported in [25]; an advanced multilevel version called self-organizing *box* maps (SOBM) has been developed in detail by Galliat et al. [22, 23, 24]. Our present favorite box discretization technique is a combination of the two heuristics to be described next.

Successive PCCA of dihedrals. This kind of box discretization heuristics is due to Cordes et al. [5]. It starts from the chemical insight that dihedrals (or torsion angles) are useful indicators for conformational changes. The principle of the algorithm is as follows: On the basis of a precomputed HMC series, we afford to construct rather fine discretizations for each of the dihedrals *separately*. This defines separate "dihedral transition matrices" T for each dihedral decomposition, which are analyzed in terms of PCCA+.

Among these matrices, the one with eigenvalue λ_2 closest to $\lambda_1 = 1$ is selected and subdivided according to the PCCA+ strategy. Upon applying this idea recursively to the remaining dihedral subspaces, a rather useful "coarse grid" is constructed, which is then taken as the box discretization for the final transition matrix to be analyzed as a whole. In Fig. 10, a few steps of this recursive scheme are schematically presented in a two-dimensional dihedral plane.

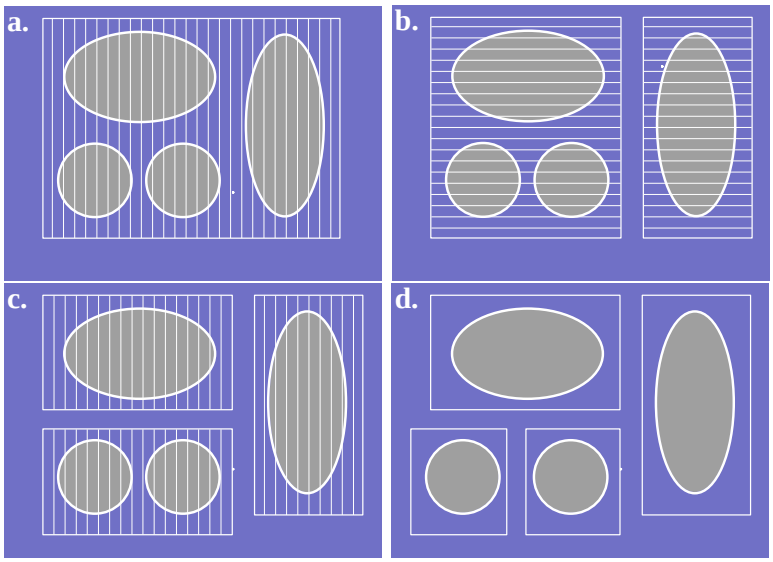

Figure 10: *Algorithmic scheme for successive PCCA of dihedrals:* Four metastable regions are drawn as ellipses in a 2-dimensional dihedral space. Thin lines show the successive fine discretizations of each dihedral. Figures a. to d. illustrate the alternation between fine discretization and coarse grid construction. The final coarse grid (Fig. d) consists of four spatial boxes.

This rather simple strategy is surprisingly robust and works well even for rather complex molecules. It will clearly fail whenever there is a coupling between torsion angles that have successively been selected for PCCA+. We are therefore planning to combine this technique with our former neural network strategy (SOM, see above) to avoid such a situation already at the level where it could occur. At present, such an occurrence is detected and corrected at some later stage of the UCMC strategy to be described next.

Uncoupling-Coupling Monte Carlo method (UCMC). This technique has been developed by A. Fischer et al. [19, 18]. From an abstract point of view, the algorithmic scheme is a Monte Carlo extension of aggregation/disaggregation techniques suggested in 1989 by C. D. Meyer [33]; there, however, the stationary distribution was the object of interest, which in our context is given as input.

As the starting point for an algorithmic realization of the transfer operator approach we need a sample of the state space distributed according to the canonical distribution $\mathcal{Q}^* \propto \exp(-\beta_* V)$ at inverse temperature β_*. Yet, a direct sampling of the state space via the associated HMC Markov chain (1.6) will result in *slow mixing* and, hence, poor convergence caused by the presence of metastabilities – which we actually want to compute.

In order to address this problem, an iterative scheme of alternating uncoupling and coupling is applied, which realizes the steps

(a) embedding $\mathcal{Q}^*$ in a series of canonical distributions of increasing temperatures – which decreases metastability,

(b) hierarchical decomposition of state space into metastable sets and restart of restricted Markov chains therein, applying a type of annealing strategy, and

(c) coupling the samples from restricted Markov chains for proper reweighing of the samples at $\mathcal{Q}^*$.

The sampling starts with one HMC Markov chain at the highest temperature level searching the whole state space. Step (b) already includes transfer operator techniques for the identification of metastable sets, but within the annealing strategy the state space is decomposed as soon as some metastability emerges. By construction, all restricted HMC Markov chains exhibit *rapid mixing*, which speeds up the computation and, at the same time, increases robustness of the overall algorithm. In coupling step (c) we set up a coupling matrix by computing quotients of normalizing constants between samples at neighboring temperatures with an overlapping domain in the hierarchy. Coupling factors connecting samples from different domains are then given by the entries of the stationary distribution of the coupling matrix. The situation is illustrated in Fig. 11.

As a result of the UCMC technique, we obtain a weighted sample, which is distributed according to $\mathcal{Q}^*$. Technical details of this quite complicated process can be found in [19, 18].

5 Example: SARS Protease Inhibitor

The here described bunch of new mathematical methods for the identification of metastable conformations has been published in a series of papers by the research group of the authors, among which the surveys [7, 40] also contain numerical results for interesting biomolecules, e.g., the green tea molecule epigallocatechine, a suspected anti-cancer drug, or an HIV protease inhibitor.

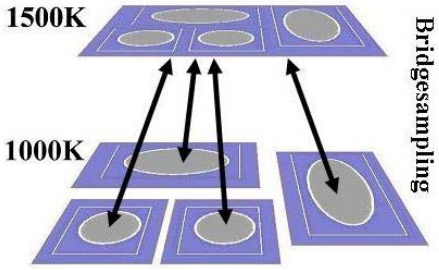

Figure 11: *Hierarchical simulation protocol for UCMC:* After decomposition, the metastable subsets of the conformational space are sampled independently at a lower temperature level. Two temperature levels are connected via bridge samplings.

In the present paper we restrict our attention to SARS (abbreviation for **S**evere **A**cute **R**espiratory **S**yndrome). The corresponding corona virus responsible for the sudden occurrence of the epidemics arose early this year, unknown until then. It is only since May 30, 2003, that the 3D structure of one of its enzymes, a protease, is available on the internet [46]; this molecule takes part in the viral metabolism by cutting larger proteins into smaller peptide strands. The underlying biochemical experiments have been published by the research group of Hilgenfeld [2]. In Fig. 12, we show the result obtained from a homology model on top of an X-ray analysis of a similar molecule, which seemed to reveal some active site of the SARS protease; the associated molecule in the active site has been observed to fit into the molecular pocket, but is not expected to be a drug against SARS. Instead the search race continues with high speed.

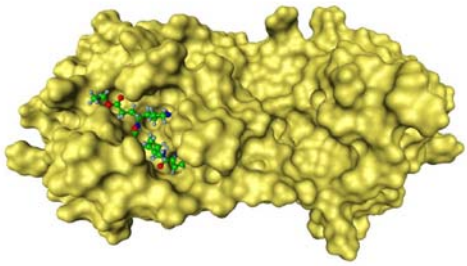

Figure 12: *SARS protease:* active site as suspected from X-ray analysis

Starting from the internet data, we investigated a molecule, the inhibitor AG7088, with our mathematical tools for conformation dynamics. Upon applying the UCMC technique for box discretization at the temperatures 1500K, 1000K, 600K, and 300K, we obtained the results arranged in Table 5.

T[K]	coarse spectrum	coupling matrix
1500	1.000 0.984 <u>0.975</u> 0.861	0.982 0.003 0.015 0.003 0.976 0.021 0.001 0.002 0.997
1000	1.000 0.994 0.987 <u>0.971</u> 0.955	0.992 0.001 0.005 0.002 0.000 0.966 0.024 0.010 0.001 0.014 0.982 0.003 0.001 0.006 0.003 0.990
1000	1.000 0.999 0.997 0.990 0.985 <u>0.982</u> 0.971	0.987 0.000 0.009 0.000 0.004 0.000 0.000 0.997 0.001 0.000 0.000 0.002 0.001 0.000 0.984 0.002 0.008 0.005 0.000 0.000 0.001 0.970 0.000 0.029 0.000 0.001 0.001 0.000 0.985 0.013 0.000 0.000 0.000 0.002 0.003 0.995
1000	1.000 0.995 0.992 0.990 <u>0.988</u> 0.982	0.978 0.002 0.016 0.004 0.000 0.002 0.976 0.000 0.014 0.008 0.003 0.000 0.987 0.009 0.001 0.000 0.002 0.005 0.986 0.007 0.000 0.001 0.001 0.017 0.981
600	1.000 0.998 0.994 0.988 <u>0.987</u> 0.979	0.959 0.032 0.001 0.000 0.008 0.008 0.981 0.003 0.002 0.006 0.001 0.012 0.980 0.002 0.005 0.000 0.001 0.000 0.966 0.033 0.000 0.001 0.000 0.006 0.993

Table 1: *SARS protease inhibitor:* hierarchical temperature sequence and coarse grid spectra, as obtained from UCMC and successive PCCA+. At 600K only the metastable conformation with highest thermodynamical weight has been selected, which then decomposes into 5 subsets at human body temperature 300K.

On each of the subsets we ran the fast mixing Markov chains based on HMC. The Perron clusters obtained from PCCA+ in connection with the box discretization technique of successive PCCA are included. As can be seen, we detected $m = 3$ metastable conformations at 1500K, which divide into 4, 6, and 5 conformations separately at 1000K. At 600K, only the metastable conformation with highest thermodynamical weight has been selected, which then decomposes into 5 subsets at room temperature or human body temperature 300K, respectively.

Of course, these data are mainly of interest for the drug designer. That is why, in Fig. 13, we additionally present an image of the molecule in the frame of conformation dynamics: there we combine a volume rendering visualization of the (discrete) invariant measure at 1500K together with one snapshot of the molecule in ball and stick representation.

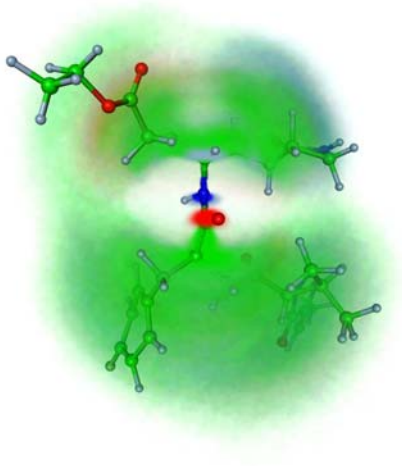

Figure 13: *SARS protease inhibitor:* volume rendering representation of invariant measure at temperature $T = 1500$K. Insertion of ball and stick representation of two dominant conformations

More insight into these conformations can be gained from the isosurfaces for the conformations as given in Fig. 14 for the dominant one and in Fig 15 for the subdominant one.

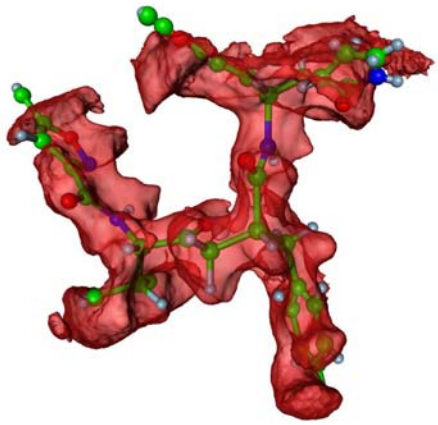

Figure 14: *SARS protease inhibitor:* isosurface representation of dominant conformation (probability ~ 56.5 % to *be* within)

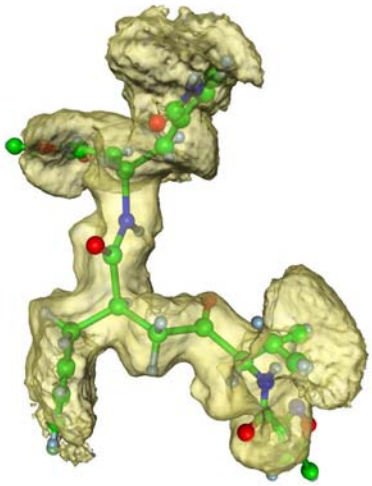

Figure 15: *SARS protease inhibitor:* isosurface representation of subdominant conformation (probability ~ 35.6 % to *be* within)

Remark. The authors are aware of the fact that in *prion* diseases (such as scrapie or the mad cow disease) rather rare conformations with high probability to *stay* within may nevertheless well play a decisive role – as has been pointed out by Griffith [26] already in 1967.

Acknowledgements. The authors want to thank all of their coworkers for their collaboration in this fascinating field, in particular Frank Cordes, Alexander Fischer, Wilhelm Huisinga, and Marcus Weber for invaluable groundwork to this article.

References

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

[2] K. Anand, J. Ziebuhr, P. Wadhani, J. R. Mesters, and R. Hilgenfeld. Coronavirus main proteinase (3clpro) structure: Basis for design of anti-sars drugs. *Science*, 300:1763, 2003.

[3] A. Bovier, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes II: Precise asymptotics for small eigenvalues. WIAS preprint, Sept. 2002.

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

[5] F. Cordes, M. Weber, and J. Schmidt-Ehrenberg. Metastable Conformations via successive Perron-Cluster Cluster analysis of dihedrals. Technical Report ZIB 02-40, Zuse Institute Berlin, 2002.

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

[7] P. Deuflhard. From molecular dynamics to conformational dynamics in drug design. In M. Kirkilionis, S. Krömker, R. Rannacher, and F. Tomi, editors, *Trends in Nonlinear Analysis*, pages 269–287. Springer, 2003.

[8] P. Deuflhard and F. Bornemann. *Scientific Computing with Ordinary Differential Equations*, volume 42 of *Texts in Applied Mathematics*. Springer, New York, 2002.

[9] P. Deuflhard, M. Dellnitz, O. Junge, and Ch. Schütte. Computation of essential molecular dynamics by subdivision techniques. In *[10]*, pages 98–115, 1999.

[10] 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-Verlag, 1999.

[11] P. Deuflhard and A. Hohmann. *Numerical Analysis in Modern Scientific Computing: An Introduction*, volume 43 of *Texts in Applied Mathematics*. Springer, New York, 2003.

[12] P. Deuflhard, W. Huisinga, A. Fischer, and Ch. Schütte. Identification of almost invariant aggregates in nearly uncoupled Markov chains. Accepted in Lin. Alg. Appl., Available via http://www.zib.de/MDGroup, 1999.

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

[14] P. Deuflhard and M. Weber. Robust Perron Cluster Analysis in Conformation Dynamics. Technical Report ZIB 03-19, Zuse Institute Berlin, 2003.

[15] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Phys. Lett. B*, 195(2):216–222, 1987.

[16] W. E, W. Ran, and E. Vanden-Eijnden. Probing multiscale energy landscapes using the string method. *Phys. Rev. Lett.*, to appear, 2002.

[17] R. Elber and M. Karplus. Multiple conformational states of proteins: A molecular dynamics analysis of Myoglobin. *Science*, 235:318–321, 1987.

[18] A. Fischer. *An Uncoupling–Coupling Method for Markov chain Monte Carlo simulations with an application to biomolecules*. PhD thesis, Free University Berlin, 2003.

[19] A. Fischer, Ch. Schütte, P. Deuflhard, and F. Cordes. Hierarchial uncoupling–coupling of metastable conformations. In *[37]*, pages 235–259, 2002.

[20] H. Frauenfelder and B. H. McMahon. Energy landscape and fluctuations in proteins. *Ann. Phys. (Leipzig)*, 9(9–10):655–667, 2000.

[21] H. Frauenfelder, P. J. Steinbach, and R. D. Young. Conformational relaxation in proteins. *Chem. Soc.*, 29A(145–150), 1989.

[22] T. Galliat. *Adaptive Multilevel Cluster Analysis by Self-Organizing Box Maps*. PhD thesis, FU Berlin, March 2002.

[23] T. Galliat and P. Deuflhard. Adaptive hierarchical cluster analysis by self-organizing box maps. Konrad–Zuse–Zentrum, Berlin. Report SC-00-13, 2000.

[24] T. Galliat, P. Deuflhard, R. Roitzsch, and F. Cordes. Automatic identification of metastable conformations via self–organized neural networks. In *[37]*, pages 260–284, 2002.

[25] 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, *Neural Computation*, pages 227–232. ICSC Academic Press, 2000.

[26] J. Griffith. Self-replication and scrapie. *Nature*, 215:1043–1044, 1967.

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

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

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

[30] W. Huisinga, S. Meyn, and Ch. Schütte. Phase transitions & metastability in Markovian and molecular systems. accepted in Ann. Appl. Probab., 2002.

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

[32] T. Kohonen. *Self–Organizing Maps*. Springer, Berlin, 2nd edition, 1997.

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

[34] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, Berlin, 1993.

[35] G. U. Nienhaus, J. R. Mourant, and H. Frauenfelder. Spectroscopic evidence for conformational relaxation in Myoglobin. *PNAS*, 89:2902–2906, 1992.

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

[37] 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*, Berlin, Heidelberg, New York, 2000. Springer.

[38] Ch. Schütte. *Conformational Dynamics: Modelling, Theory, Algorithm, and Application to Biomolecules*. Habilitation Thesis, Fachbereich Mathematik und Informatik, Freie Universität Berlin, 1999.

[39] 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*, 151:146–168, 1999.

[40] Ch. Schütte and W. Huisinga. Biomolecular conformations as metastable sets of Markov chains. In R. S. Sreenivas and D. L. Jones, editors, *Proceedings of the Thirty–Eight Annual Allerton Conference on Communication, Control, and Computing, Monticello, Illinois*, pages 1106–1115. University of Illinois at Urbana-Champaign, 2000.

[41] Ch. Schütte and W. Huisinga. Biomolecular conformations can be identified as metastable sets of molecular dynamics. In P. G. Ciarlet and J.-L. Lions, editors, *Handbook of Numerical Analysis*, volume Computational Chemistry. North–Holland, 2002. in press.

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

[43] E. Vanden-Eijnden. Metastability and effective dynamics in ergodic systems, 2003.

[44] M. Weber. Improved Perron Cluster Analysis. Technical Report ZIB 03-04, Zuse Institute Berlin, 2003.

[45] M. Weber and T. Galliat. Characterization of transition states in conformational dynamics using Fuzzy sets. Technical Report Report 02–12, Konrad–Zuse–Zentrum (ZIB), Berlin, March 2002.

[46] A. Wiley and Gh. Deslongchamps. Homology model of SARS-CoV Mpro protease, 2003.

[47] H. X. Zhou, S. T. Wlodec, and J. A. McCammon. Conformation gating as a mechanism for enzyme specificity. *Proc. Nat. Acad. Sci. USA*, 95(9280–9283), 1998.