Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

SUSANNA KUBE, MARCUS WEBER

Coarse Grained Molecular Kinetics

ZIB-Report 06-35 (June 2006)

Coarse Grained Molecular Kinetics

Susanna Kube and Marcus Weber*

June 29, 2006

Abstract

The dynamic behavior of molecules can often be described by Markov processes. From computational molecular simulations one can derive transition rates or transition probabilities between subsets of the discretized conformational space. On the basis of this dynamic information, the spatial subsets are combined into a small number of so-called metastable molecular conformations. This is done by clustering methods like the Robust Perron Cluster Analysis (PCCA+). Up to now it is an open question how this coarse graining in space can be transformed to a coarse graining of the Markov chain while preserving the essential dynamic information. In the following article we aim at a consistent coarse graining of transition probabilities or rates on the basis of metastable conformations such that important physical and mathematical relations are preserved. This approach is new because PCCA+ computes molecular conformations as linear combinations of the dominant eigenvectors of the transition matrix which does not hold for other clustering methods.

AMS MSC 2000: 65C40, 65P99, 65L99, 65F15, 80A30

Keywords: Markov chains, soft clustering, coarse graining, master equation, molecular kinetics

1 Introduction

The understanding of transition pathways between different conformations of a molecule is an important issue in structural biology. Although the restriction of degrees of freedom to a few dihedral angles significantly reduces the complexity of the problem, it is still very difficult to identify conformations and their transition

*Zuse Institute Berlin (ZIB), Takustraße 7, D-14195 Berlin, Germany

probabilities or transition rates. Often, scientists are interested in single pathways, for example those over lowest energy barriers [2]. On the other hand, it is well known that molecular kinetics is not purely deterministic. All kinds of trajectories could appear, some with higher probability than others. Therefore it seems natural to consider population densities. Starting with a given probability density in position space, we are interested in the evolution of the density to figure out intermediate states.

A description of molecular dynamics based on single positions in phase space is infeasible for large molecules. Therefore we work with a set concept based on metastable conformations as introduced in [6, 5]. From our point of view, conformations are not only single molecular geometries but are formed by sets of several geometries. These sets are characterized by the property that the large scale geometric structure is conserved during the molecular dynamical process over a long period of time before a transition to another conformation occurs. The behavior of such high dimensional dynamical systems can be described by continuous-time Markov chains. To identify the metastable sets, we first reduce the dihedral space to a number of N states represented by basis functions [23] or boxes [20, 8]. By applying Markov chain–Monte Carlo sampling techniques, we construct a transition probability matrix $P \in \mathbb{R}^{N\times N}$ or a transition rate matrix $Q \in \mathbb{R}^{N\times N}$ and cluster states into metastable conformations by applying PCCA+. In other words, we find out which states of our discretization belong to the same metastable conformation. Thus we are able to reduce our model not only to a set of basis functions whose number can be very large, but to the few metastable sets which contain all important information about the system. Thus, the *essential* dynamic behavior can be described by transition probabilities or transition rates between these sets.

Transition rates provide chemical information concerning transition pathways between different geometrical conformations. Given an initial weighting x_A of the states, one can compute the corresponding weights and the spatial configuration density at any time by the master equation $\dot{x} = Q^\top x$. This is the desired dynamic in configuration space, which is not based upon single molecules but upon ensembles. The entry $Q(i,j), i \neq j$, can be considered as the reaction rate of the monomolecular reaction

$$x_i \quad \rightharpoonup \quad x_j$$

where x_i stands representatively for the weight or "concentration" of state i.

The extraction of the essential dynamic behavior is achieved by a reduction of the corresponding transition matrices P or Q to low-dimensional matrices P_c or Q_c as illustrated in Figure 1. If P is embeddable [4], i. e. there exists a unique rate matrix Q with $P(t) = \exp(tQ)$ where $\exp(\cdot)$ denotes the matrix exponential function [11], then Q can be obtained directly from the transition probability matrix P by $Q = \log(P)/t$. However, in most cases $P(t)$ stemming from discrete-time observations is not embeddable. There exist several techniques to find an approximate generator [15, 1, 4] which are very useful if solely the transition probability matrix $P(t)$ or a discrete timeseries is given. If information about the underlying

$$\begin{tikzcd}
P(t) \arrow[r, leftrightarrow] \arrow[d, leftrightarrow] & Q \arrow[d, leftrightarrow] \\
P_c(t) \arrow[r, leftrightarrow] & Q_C
\end{tikzcd}$$

Figure 1: Coarse graining scheme. The goal is to reduce the transition probability matrix P or its generator Q to low-dimensional coarse-grained matrices P_c or Q_c such that they reflect the correct essential dynamic behavior of the system.

continuous-time process is available, we recommend to approximate the rate matrix directly from the simulation data, as we will do in our example. Thus, the embedding problem is circumvented. However, our goal is to develop a coarse-graining method which covers the ideal case of an embeddable matrix P. Therefore the coarse-grained matrices are required to satisfy several properties.

- P_c is a stochastic matrix which contains the transition probabilities between the metastable conformations.
- P_c reflects the correct dynamic behavior of the system.
- Q_c is a rate matrix which contains the transition rates between the metastable conformations.
- Q_c reflects the correct dynamic behavior of the system.
- If Q is the generator of P, Q_c should be the generator of P_c.

However, it will turn out that in general it is not possible to meet all requirements at the same time. From our point of view, the most important fact is the conservation of the dynamic behavior. In the following we present different coarse-graining schemes and explain their qualities w. r. t. the above mentioned items.

2 Clustering of States

Biological systems can often be described by a continuous-time Markov chain $\{X(t) : t \geq 0\}$, represented by a discrete-time sample path $\{X_n\}_{n\in\mathbb{N}}$ in a finite state space $E = \{1, \ldots, N\}$ [12, 3, 16, 10]. Assume we have discretized our space of interest Ω into N states and counted the transition frequencies of this path resulting from a Markov chain–Monte Carlo simulation. Let $\overline{P}$ denote the matrix of relative transition frequencies with

$$\sum_{i,j=1}^{N} \overline{P}(i,j) = 1.$$

The transition probability matrix P is obtained by scaling $\overline{P}$ to row sum one. This corresponds to a multiplication of $\overline{P}$ with a diagonal matrix D

$$P = D^{-1}\overline{P}.$$

Let e denote the vector which has the entry 1 in each component. Then, the entries of D are given by

$$D = \text{diag}(\pi), \quad \pi = \overline{P}e.$$

Note that $\sum_i \pi(i) = 1$ and $\pi(i) \geq 0$ by construction.

Throughout the paper, we assume that the transition frequency matrix $\overline{P}$ is symmetric. Such Markov chains are called *reversible*. Then, the vector $\pi \in \mathbb{R}^N$ is the stationary distribution of P,

$$\pi^\top P = e^\top \overline{P} = e^\top \overline{P}^\top = (\overline{P}e)^\top = \pi^\top,$$

and the Markov chain meets the *detailed balance condition*,

$$DP = P^\top D \quad \leftrightarrow \quad \pi(i)P(i,j) = \pi(j)P(j,i). \tag{1}$$

If, moreover, P is irreducible, then π is the unique stationary distribution.

States, which belong to a metastable set, are characterized by large transition probabilities among each other but small transition probabilities to other states. In the case of completely decoupled stable sets, the matrix could be rearranged to block-diagonal structure. Hence, the identification of metastable conformations corresponds to the locating of a hidden block structure in the matrix.

2.1 Clustering Methods

Standard clustering methods would result in N_c clusters characterized by membership vectors $\chi_k \in \{0,1\}^N$, $k = 1, \ldots, N_c$, which have the entry 1 if the corresponding state belongs to cluster k and else the entry 0. These vectors satisfy

$$P\chi_k \approx \chi_k$$

because transition probabilities between states of different clusters are nearly zero. In this context, we speak of a *crisp* clustering. This approach works well if the states can be assigned uniquely to a cluster. However, in the discretization process it could happen that a state is located in a transition region (e. g. a saddle point of the potential energy surface) such that its assignment to a certain cluster would be wrong. To circumvent such difficulties, we focus on *soft* clustering methods. In contrast to the crisp clustering, the states are now assigned to all clusters with certain weights given by the entries of the *soft* membership vectors $\chi_k \in [0,1]^N$ with values *between* 0 and 1. As an important property, the membership vectors form a *partition of unity*

$$\sum_{k=1}^{N_c} \chi_k(i) = 1, \quad i = 1, \ldots, N.$$

In this context, conformations can be considered as fuzzy sets.

The Robust Perron Cluster Analysis PCCA+ [7] generates the soft membership vectors as a linear transformation of the dominant eigenvectors of the transition probability matrix P,

$$PX = X\Theta, \quad \Theta = \mathrm{diag}(\theta_i),\, \theta_i \approx 1, \quad X^\top DX = id,$$

where id denotes the identity matrix. This results in soft membership vectors χ,

$$\chi = XA, \quad A \text{ regular.} \tag{2}$$

The number N_c of clusters equals the number of eigenvalues of P near the Perron root $\theta_1 = 1$. A detailed perturbation analysis for the PCCA+ approach based on Markov chain theory provides robustness of this method [7]. Figure 2 shows such transformation for the eigenvectors stemming from a discretization of the butane dynamics.

The algorithm is based on the following ideas. A completely decoupled transition matrix $\widetilde{P}$ can be rearranged to block diagonal structure and has the N_c-fold eigenvalue $\theta = 1$. The corresponding eigenvectors are constant for states belonging to the same block. For nearly decoupled matrices P, the Perron root degenerates to a cluster of eigenvalues near $\theta = 1$. The corresponding eigenvectors are not piecewise constant any longer, but if they are considered row-wise, they nearly form a simplex[1] in $\mathbb{R}^{N_c-1}$ [23]. The complete algorithm is based on a constrained optimization method which minimizes the overlap of the resulting clusters and delivers membership vectors χ_k which are non-negative. The initial guess for this algorithm is constructed without regarding the non-negativity constraint. However, the initial guess is optimal w.r.t. the objective function. This led to a simplified version of PCCA+, in which the smallest entry of the corresponding membership vectors, the so called *minChi*-indicator [22], measures the feasibility of the initial guess. In most cases it suffices to take this initial guess as final solution because *minChi* is small enough[2]. In order to meet the non-negativity constraint a slight shift and rescaling of the initial membership vectors can be performed, such that $\chi = X\mathcal{A}$ still holds for an appropriate regular transformation matrix $\mathcal{A}$.

To sum up, metastable sets are convex linear combinations of all states from our discretization where the linear factors are given by the values of the membership vectors χ_k. These values can be considered as probabilities that a certain state

[1] One dimension can be skipped because the first eigenvector is constant.

[2] Or one determines N_c such that minChi is almost zero.

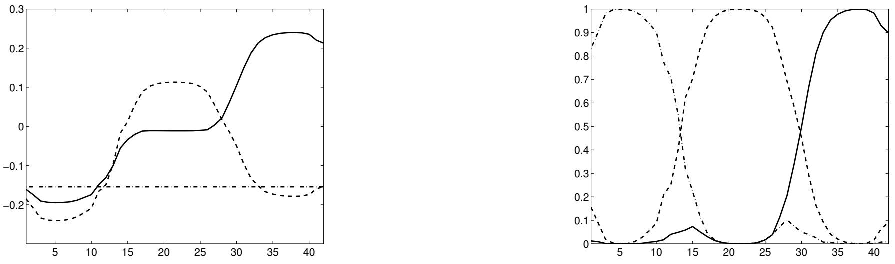

Figure 2: Eigenvectors and membership vectors of a transition matrix resulting from the discretization of butane.

belongs to a certain cluster. For Theorem 3.6 it is very important to use PCCA+ for the construction of the membership vectors, because $\chi = X\mathcal{A}$ is a basic assumption in the corresponding proof.

2.2 Coarse Graining

After having identified metastable sets, we are interested in transition probabilities between these sets. For this purpose, we first consider the transition frequencies between two clusters C_k and C_l, $k, l = 1, \ldots, N_c$. For a crisp clustering, these frequencies are obtained by just adding the rows and columns of the transition frequency matrix $\overline{P}$ which belong to the same cluster. This results in a coarse frequency matrix $\overline{P}_c \in \mathbb{R}^{N_c \times N_c}$,

$$\overline{P}_c(k,l) = \sum_{i \in C_k, j \in C_l} \overline{P}(i,j).$$

In terms of membership vectors this reads

$$\overline{P}_c = \chi^\top \overline{P} \chi = \chi^\top D P \chi.$$

If χ and D are positive, $\overline{P}_c$ is positive as well. The coarse stochastic matrix P_c is obtained from rescaling $\overline{P}_c$ to row sum one,

$$P_c = \tilde{D}^{-1} \overline{P}_c = \tilde{D}^{-1} \chi^\top D P \chi. \tag{3}$$

The entries of the diagonal matrix $\tilde{D}$ are given by

$$\tilde{D}(i,i) = \sum_{j=1}^{N} \sum_{l=1}^{N} \sum_{k=1}^{N} \chi_i(k) \pi(k) P(k,l) \chi_j(l) = \sum_{k=1}^{N} \chi_i(k) \pi(k).$$

Lemma 2.1. *The diagonal elements of $\tilde{D}$ form the vector $\tilde{\pi}$ which is the stationary distribution of P_c,*

$$P_c^\top \tilde{\pi} = \tilde{\pi}, \quad \sum_{k=1}^{N_c} \tilde{\pi}(k) = 1.$$

Proof.

$$\sum_{k=1}^{N_c} \tilde{\pi}(k) = \sum_{k=1}^{N_c}\sum_{j=1}^{N} \pi(j)\chi_k(j) = \sum_{j=1}^{N} \pi(j) = 1,$$

$$\tilde{\pi}^\top P_c = e^\top \chi^\top DP\chi = e^\top DP\chi = \pi^\top P\chi = \pi^\top \chi = \tilde{\pi}^\top.$$

□

In the above derivation, we only used the fact that the rows of χ and P sum to one. Therefore, the same method can be applied for a soft clustering.

Definition 2.2. *Let $P \in \mathbb{R}^{N\times N}$ be a stochastic transition probability matrix with stationary distribution vector $\pi \in \mathbb{R}^N$. Let $\chi = [\chi_1, \dots, \chi_{N_c}] \in \mathbb{R}^{N\times N_c}$ denote the matrix of soft membership vectors and $D = diag(\pi)$ the matrix with π on its diagonal. With $\tilde{D} = diag(\chi^\top diag(D))$ we define the* **restriction** *and* **interpolation** *operators:*

$$R : \mathbb{R}^N \mapsto \mathbb{R}^{N_c}, \quad R = \chi^\top$$
$$I : \mathbb{R}^{N_c} \mapsto \mathbb{R}^N, \quad I = D\chi\tilde{D}^{-1}$$

Lemma 2.3. *The restriction and interpolation operator have column sum one. Furthermore, for crisp membership vectors $\chi_k \in \{0,1\}^{N_c}$ we have $RI = id$. In any case, the matrix RI is regular.*

Proof. The first statement is satisfied by construction. The second one results from

$$\begin{aligned} (RI)(i,j) &= \sum_k \chi_i(k)\pi(k)\chi_j(k)\tilde{\pi}(j)^{-1} \\ &= \delta_{ij}\sum_k \pi(k)\chi_j^2(k)\tilde{\pi}(j)^{-1} \\ &= \delta_{ij}\sum_k \pi(k)\chi_j(k)\tilde{\pi}(j)^{-1} \\ &= \delta_{ij}. \end{aligned}$$

From (2) we obtain

$$RI = \chi^\top D\chi\tilde{D}^{-1} = A^\top X^\top DXA\tilde{D}^{-1} = A^\top A\tilde{D}^{-1}.$$

Since A is regular, the same yields for RI. □

Densities with respect to the *fine* stochastic matrix $P \in \mathbb{R}^N$ are given as vectors $x^f \in \mathbb{R}^N$ with $\sum_i x^f(i) = 1$, $x(i) \geq 0$. In the following they are called *fine densities.* As a *coarse density*, we consider a vector $x^c \in \mathbb{R}^{N_c}$, $\sum_i x^c(i) = 1$, $x^c(i) \geq 0$. These densities represent the weights of the clusters. The above defined operators are used to transform the densities to each other. The coarse density is obtained by a projection of a fine density onto the clusters, i.e.

$$x^c = Rx^f. \tag{4}$$

Indeed, multiplication with R preserves positivity and the sum of the vector elements. Furthermore, $\tilde{\pi}$ is obtained from π by (4). The fine density can be obtained from a coarse density by interpolation,

$$x^f = Ix^c. \tag{5}$$

To make this transformation feasible, note that $\tilde{\pi}$ is transformed to π,

$$(I\tilde{\pi})(i) = \sum_{j=1}^{N_c} \pi(i)\chi_j(i)\frac{1}{\tilde{\pi}(j)}\tilde{\pi}(j) = \pi(i)\sum_{j=1}^{N_c} \chi_j(i) = \pi(i).$$

Furthermore, the transformation (5) preserves positivity and the sum of the vector elements.

The question we are concerned with is the following: How does the coarse stochastic matrix P_c represent the dynamic behavior described by the fine stochastic matrix P? For time dependent dynamical systems, the transition probabilities are actually time dependent. $P(t)(i,j)$ denotes the probability that the system starting in state i is in state j after the time span t. The distribution of states i in a finite state space E at time t is a vector $x(t) = \{x_t(i)\}_{i\in E}$ which is obtained from the initial distribution via

$$x(t) = P^\top(t)x(0). \tag{6}$$

The corresponding coarse equation reads

$$y(t) = P_c^\top y(0), \quad P_c \in \mathbb{R}^{N_c \times N_c}. \tag{7}$$

Ideally, there exists a connection between $x(t)$ and $y(t)$ by the previously defined restriction and interpolation operators. However, such a relationship can only be derived for special initial conditions.

Lemma 2.4. *Let be given a reversible stochastic matrix $P(t)$ and an initial distribution $x(0) = Iy(0)$, $y(0) \in \mathbb{R}^{N_c}$. Then the distribution $x(t) = P^\top(t)x(0)$ restricted to the space of conformations is equivalent to $y(t)$ computed from the coarse equation (7).*

Proof.

$$Rx(t) = RP^\top x(0) = RP^\top Iy(0) = P_c^\top y(0) = y(t).$$

□

In fact, the previous lemma is just a special case of the more general invariance of the space spanned by the weighted essential eigenvectors,

$$\mathcal{V} = \text{span}\{DX\} = \text{span}\{DX(:,1),\ldots,DX(:,N_c)\} = \text{span}\{D\chi\}.$$

Indeed, $x(0) = Iy(0) = DX\tilde{D}^{-1}y(0) \in \mathcal{V}$. Consider equation (6) and a density $z \in \mathbb{R}^{N_c}$. Due to the eigen-equation $P^\top DXz = DX\Theta(t)z$, the elements of $\mathcal{V}$ are invariant under the action of $P^\top$ and the following statement is satisfied.

Lemma 2.5. *Consider equation (6) with initial condition $x(0) = DXy_0 \in \mathcal{V}$. Then the solution of (6) is given by $x(t) = DXy(t)$ with $y(t) = \Theta(t)y_0 \in \mathbb{R}^{N_c}$.*

Unfortunately, in general $y(t) = Rx(t) \notin \mathcal{V}$. Therefore, it is not possible to extend the result from the previous lemma to time steps nt, $n \in \mathbb{N}$, according to $x(nt) = P^{n\top}(t)x(0)$. On the other hand, note that $RP^\top = A^\top \Theta A^{-\top} R$. This leads to the idea of redefining the coarse transition probability matrix as

$$\hat{P}_c = (RI)^{-\top} I^\top P R^\top. \tag{8}$$

Lemma 2.6. *The matrix $\hat{P}_c$ defined in (8) is diagonalizable. Moreover, it has row sum 1 and the invariant density $\tilde{\pi}$.*

Proof. The first statements results from the fact that A is regular and

$$\hat{P}_c^\top = A^\top X^\top P^\top I(RI)^{-1} = A^\top \Theta X^\top I(RI)^{-1} = A^\top \Theta A^{-\top}.$$

Furthermore,

$$\hat{P}_c^\top \tilde{\pi} = A^\top \Theta A^{-\top} R\pi = A^\top \Theta X^\top \pi = A^\top X^\top P^\top \pi = \chi^\top \pi = \tilde{\pi}.$$

Since R and I have column sum 1, the matrices RI and $(RI)^{-1}$ have also column sum 1. Consequently, $(RI)^{-\top}$ has row sum 1. Multiplication with the row stochastic matrix $P_c = I^\top P R^\top$ preserves this property. □

However, besides the case of a crisp clustering, $\hat{P}_c(t)$ is not a stochastic matrix. Some entries might be negative. This is due to the definition of metastable sets because they are not disjoint sets but overlapping functions. Negative entries in the transition matrix are necessary to correct the global behavior. Nevertheless, the matrix can be used to calculate the propagation of densities by a coarse equation.

Theorem 2.7. *Let be given a reversible stochastic matrix $P(t)$ and an initial distribution $x(0) \in \mathbb{R}^N$. Then the distribution $x(t) = P^\top(t)x(0)$ restricted to the space of conformations is equivalent to $y(t) = \hat{P}_c^\top(t)y(0) \in \mathbb{R}^{N_c}$ with $y(0) = Rx(0)$.*

Proof.

$$Rx(t) = RP^\top x(0) = A^\top \Theta A^{-\top} Rx(0) = \hat{P}_c^\top y(0) = y(t).$$

□

This carries over to time steps nt because

$$RP^n(t) = A^\top \Theta^n A^{-\top} R = \hat{P}_c^n(t) R.$$

Thus, $\hat{P}_c(t)$ reflects the correct dynamic behavior w. r. t. the propagation of densities.

In the continuous-time case, the behavior of the dynamic system for arbitrary times t is governed by its infinitesimal generator. In the following section we explain this concept and examine how the coarse graining process carries over to the generator.

3 Conformation Kinetics

3.1 From Transition Probabilities to Transition Rates

Under certain conditions, the Markov chain is completely determined by a time-independent generator Q. It is defined for continuous-time Markov chains.

Definition 3.1. *[16, 3] Suppose that $\{P(t)\}_{t\geq 0}$ is a continuous transition semigroup on a countable state space E. Then, the limit*

$$Q = \lim_{t \mapsto 0+} \frac{P(t) - id}{t}$$

exists and defines the **infinitesimal generator** $Q = \{Q(i,j)\}_{i,j \in E}$ *with* $-\infty \leq Q(i,i) \leq 0 \leq Q(i,j) < \infty$.

Especially if E is finite, the semigroup is conservative,

$$-Q(i,i) = \sum_{j \neq i \in E} Q(i,j)$$

and therefore also stable

$$-Q(i,i) < \infty.$$

In the finite case, the transition semigroup $P(t)$ and its generator are related by the forward and backward Kolmogorov equation

$$\frac{dP(t)}{dt} = Q\,P(t) = P(t)\,Q.$$

Together with the initial condition $P(0) = id$, this gives the formal solution

$$P(t) = \exp(tQ) = \sum_{n=0}^{\infty} \frac{(tQ)^n}{n!}, \quad t \geq 0. \tag{9}$$

Differentiating equation (6) w. r. t. time and using Kolmogorow's equation, we obtain the *master equation*

$$\dot{x}(t) = Q^{\top} x(t). \tag{10}$$

For a finite state space, π is the stationary distribution of $P(t)$ iff $Q^{\top}\pi = 0$ [3]. Since $\sum_i Q(j,i) = 0$ (conservation), this is also equivalent to the balance equation

$$\sum_i \pi(i) Q(i,j) = \sum_i \pi(j) Q(j,i), \quad \forall\, j \in E.$$

However, in the following we even assume more. Q is supposed to be reversible, i. e. it meets the detailed balance equation

$$\pi(i) Q(i,j) = \pi(j) Q(j,i), \qquad \forall\, i,j \in E. \tag{11}$$

Lemma 3.2. *If Q satisfies the detailed balance condition, the transition probability matrix $P(t)$ is reversible.*

Proof. For a finite state space, pre- and post-multiplication with $D^{-1/2}$ preserves symmetry. Consequently, a finite Markov chain is reversible, if $D^{1/2}PD^{-1/2}$ is symmetric.
From (9) it follows that

$$D^{1/2}P(t)D^{-1/2} = \sum_{n=0}^{\infty} \frac{t^n}{n!} \left(D^{1/2}QD^{-1/2}\right)^n .$$

Since the right hand side is symmetric (Q meets detailed balance), the left hand side is symmetric, too. Hence, reversibility of Q implies reversibility of P. □

The reversibility requirement arises from the modeling point of view. Since we consider molecular dynamics in the equilibrium state, the system of Hamiltonian differential equations describing the molecular motions is reversible. This carries over to the discretization. For detailed explanations see [23].

Lemma 3.3. *The eigenvalues of a reversible rate matrix Q are located on the negative real axis in the interval $[-2\max_i(|Q(i,i)|), 0]$. Moreover, if Q is irreducible, the eigenvalue 0 is algebraically simple.*

Proof. see [17] □

Remark 3.4. *The proof shows that Q is diagonalizable and can be written as*

$$Q = Y\Sigma Y^\top D, \quad Y^\top DY = id,$$

where Y denotes the right eigenvectors, $Y^\top D$ the left eigenvectors, and Σ the diagonal eigenvalue matrix.

In the following, denote by Λ the diagonal matrix comprising the first N_c eigenvalues of Q closest to zero, and by X the corresponding first N_c eigenvectors, i. e.

$$QX = X\Lambda, \quad X^\top DX = id. \tag{12}$$

Since Q is diagonalizable, the transition probability matrix $P(t) = \exp(tQ)$ possesses the same eigenvectors as Q [17]. The eigenvalue cluster of $P(t)$ at 1 transfers to an eigenvalue cluster of Q at 0,

$$\Theta(t) = \exp(t\Lambda).$$

Motivated by our work on PCCA+ for such matrices [7], we follow the same idea and try to find χ as a linear combination of eigenvectors of Q corresponding to eigenvalues $\lambda \approx 0$, i.e.

$$\chi = X\mathcal{A}, \quad QX = X\Lambda, \quad \Lambda = \operatorname{diag}(\lambda_i)_{i=1}^{N_c}, \ \lambda_i \approx 0.$$

The eigen-equation can be interpreted as follows. The rate matrix Q represents a closed system with mass conservation, indicated by the row sum zero property. For strictly characteristic vectors χ with values in $\{0,1\}$, χ characterizes the subsystems of Q for which this property holds, too.

3.2 Coarse Grained Kinetics

We are not interested in the evolution of the density $x(t) \in \mathbb{R}^N$ but in the evolution of the density $y(t) \in \mathbb{R}^{N_c}$ w. r. t. metastable sets. Since densities in the space of conformations are obtained by (4), our goal is to define a **coarse master equation**

$$\dot{y}(t) = Q_c^\top y(t), \quad Q_c \in \mathbb{R}^{N_c \times N_c}, \tag{13}$$

such that $||Rx(t) - y(t)||$ is small for all $t > 0$. Similar to (8), we define a coarse matrix

$$\hat{Q}_c^\top = RQ^\top I(RI)^{-1}. \tag{14}$$

Lemma 3.5. *The matrix $\hat{Q}_c$ defined in (8) is diagonalizable. Moreover, it has row sum 0 and the stationary distribution $\tilde{\pi}$.*

Proof. The proof is analog to the proof of Lemma 2.6. From the definition (14) we obtain

$$\hat{Q}_c^\top = A^\top X^\top Q^\top I(RI)^{-1} = A^\top \Lambda X^\top I(RI)^{-1} = A^\top \Lambda A^{-\top}.$$

The row sum 0 of Q is preserved due to the properties of R and I. Moreover,

$$\hat{Q}_c^\top \tilde{\pi} = A^\top \Lambda A^{-\top} R\pi = A^\top \Lambda X^\top \pi = A^\top X^\top Q^\top \pi = 0.$$

□

Note that $\hat{Q}_c$ can have some negative off-diagonal elements. The interpretation is the same as for $\hat{P}_c$ in section 2.2. However, in the following we still speak of a coarse rate matrix.

Theorem 3.6. *Let be given a reversible rate matrix Q with master equation $\dot{x}(t) = Q^\top x(t)$, $x(0) = x_0 \in \mathbb{R}^N$, and the corresponding restriction and interpolation operators. The result of the coarse master equation $\dot{y}(t) = \hat{Q}_c^\top y(t)$, $y(0) = Rx_0$, with $\hat{Q}_c^\top$ defined by (14) satisfies*

$$||Rx(t) - y(t)|| = 0, \, \forall \, t \in [0, \infty).$$

Proof. Note that

$$x(t) = \exp(tQ^\top)x(0), \qquad y(t) = \exp(t\hat{Q}_c)y(0).$$

Consequently

$$||Rx(t) - y(t)|| = ||(R\exp(tQ^\top) - \exp(t\hat{Q}_c^\top)R)x_0||.$$

Moreover, the following equation is satisfied,

$$\begin{aligned} R\exp(tQ^\top) &= A^\top X^\top \exp(tQ^\top) \\ &= A^\top \exp(t\Lambda) X^\top \qquad \text{(eq. (12))} \\ &= A^\top \exp(t\Lambda) A^{-\top} R. \end{aligned}$$

$$\begin{tikzcd}
P(t) \arrow[r, leftrightarrow] \arrow[d] & Q \arrow[d, dashed] & & P(t) \arrow[r, leftrightarrow] \arrow[d] & Q \arrow[d] \\
P_c(t) \arrow[r, leftrightarrow, dashed] & {}_{Q_c^{(2)}} \diagdown {}^{Q_c^{(1)}} & & \hat{P}_c(t) \arrow[r, leftrightarrow] & \hat{Q}_c
\end{tikzcd}$$

Figure 3: Two possible coarse graining schemes. The left one preserves the matrix properties but results in a wrong dynamic behavior. Moreover, the derivation of Q_c is not unique. The right scheme yields the desired dynamic behavior but the matrices have no physical meaning. The arrows indicate the direction of computation from input to output.

From Lemma 3.5 we obtain

$$\exp(t\hat{Q}_c^\top) = \exp(tA^\top \Lambda A^{-\top}) = A^\top \exp(t\Lambda) A^{-\top}.$$

This yields the proposition. □

If the coarse matrices are defined analogous to the crisp clustering, in general it holds

$$Q_c = I^\top Q R^\top \neq \frac{1}{t}\log(P_c(t)).$$

Thus, it is not clear how to derive a coarse rate matrix in this case. Furthermore, the solution of the coarse equation cannot be related directly to the solution of the corresponding fine equation.

On the other hand, the modified coarse graining method is consistent. Comparing (8) and (14) it can be seen that

$$\hat{P}_c = \exp(t\hat{Q}_c).$$

Moreover, by Theorems 2.7 and 3.6, the essential dynamic behavior is preserved. Note, that even if the matrix $P(t)$ has no unique logarithm due to negative or complex eigenvalues in the lower part of the spectrum, the logarithm of the coarse matrix $\hat{P}_c$ is unique, if the dominant eigenvalues are real and non-defective. Thus it may be possible to derive a generator for the coarse grained dynamic process although the generator for the fine process might not exist.

Figure 3 illustrates the two concepts. The coarse graining according to the crisp clustering (left picture) does not maintain the dynamic properties and leads to inconsistencies in the definition of Q_c. However, the desired dynamic is obtained, if the usual matrix properties are abandoned (right picture).

Since $N_c \ll N$, we can save a considerably amount of work by using the coarse model. A similar approach was previously investigated by the group of I.L. Hofacker at the University of Vienna [24]. However, they set up the coarse rate matrix in advance. They identify macro states as basins of attraction of local minima of the energy function by extensive search strategies. In contrast, in our method the macro states are identified automatically by clustering the fine matrix, and its number will be much smaller than the number of local minima. The reason is that metastable sets generally include several local minima because we also take into account entropic effects.

3.3 Extraction of Kinetic Information

We are not necessarily interested in solving (13), but in the information we gain from $\hat{Q}_c$. In the following, assume that the sample paths of $\{X(t)\}$ are right-continuous step functions. Such processes are also called Markov jump processes. Then the entries of the generator Q can be related to the mean holding times within the states and transition rates between different states.

Definition 3.7. *For a stochastic process $\{X(t)\}$ the random variable*

$$T_i(t) = \inf(s \geq 0 : X(t+s) \neq i, X(t) = i)$$

is called holding time in state i.

If $X(0) = i$, $X(T_i)$ represents the state the Markov chain visits just after leaving state i. It can be shown [16] that T_i decays exponentially in t, i.e.

$$\mathbb{P}[T_i > s] = \exp(-h(i)s), \qquad \forall\, s > 0.$$

$h(i)$ is called the *jump rate* associated with state $i \in E$. The average life time of state i is given by

$$\mathbb{E}[T_i] = \frac{1}{h(i)}.$$

Lemma 3.8. *For a homogeneous finite Markov chain $X(t)$ in continuous time with infinitesimal generator Q, $\sum_{j \neq i} Q(i,j) > 0$, the following equations are satisfied:*

1. $$-\frac{Q(i,j)}{Q(i,i)} = \mathbb{P}[X(T_i) = j | X(0) = i],$$

2. $$Q(i,i) = -h(i).$$

Proof. See [16, 17]. □

Thus, the generator Q my be represented by the inverse average life times $H = \mathrm{diag}\{h(i)\}$ and some transition matrix K,

$$Q = H(K - id),$$

where $K(i,j) = Q(i,j)/h(i)$ is the conditional probability of a transition from state i to state j given that the process starts in i. K describes the *embedded* Markov chain. This characterization forms a basis for a numerical simulation of the Markov jump process.

The other way round, given $\hat{Q}_c$ one can reconstruct $\hat{H}_c$ and $\hat{K}_c$ which contain the information about mean holding times and conditioned transition probabilities for the coarse states or conformations. However, this method cannot be applied if $\hat{Q}_c$ has negative off-diagonal entries. Since these entries are mostly very small, we correct the matrix by

$$\tilde{Q}(i,j) = \max(\hat{Q}_c(i,j), 0), \quad i \neq j, \quad \tilde{Q}(i,i) = -\sum_{j \neq i} \tilde{Q}(i,j). \tag{15}$$

3.4 Steered Kinetics

Equation (10) is not very interesting because the process simply converges towards the equilibrium distribution π. Assume we are interested in a simulation of a transition from metastable conformation C_k to a metastable conformation C_l and the corresponding transition behavior. If all species are in one conformation C_k, the corresponding distribution in the coarse setting is given by the unity vector $z_0 = e_k \in \mathbb{R}^{N_c}$. In order to assure that $Rx_0 = z_0$ we set $x_0 = I(RI)^{-1}z_0$. The same is applied to the end state $x_e = I(RI)^{-1}z_e$ where $z_e = e_l \in \mathbb{R}^{N_c}$. Then (10) has to be solved as an initial value problem with initial distribution x_0 and an absorbing end state given by the distribution x_e. Chemically, one would permanently eliminate conformation C_l out of the ensemble in order to push the reaction into the direction of this product. This corresponds to *Le Chatelier's Principle.* Mathematically this can be done by a projection of $x(t)$ onto the orthogonal complement of the desired end point x_e before applying Q. Thus, the absorbing kinetics equation is

$$\dot{x}(t) = Q_\Pi^\top x(t), \quad Q_\Pi = \Pi Q, \quad x(0) = x_0, \tag{16}$$

with

$$\Pi = id - \frac{x_e x_e^\top}{x_e^\top x_e}. \tag{17}$$

Obviously, the underlying dynamic is not reversible. Even though $x_e^\top Q_\Pi = 0$, it is not clear if Q_Π is a rate matrix. If this was the case we could proceed as in the previous subsection. It can be shown that Q_Π has row sum zero and negative diagonal elements. However, numerical experiments have shown that Q_Π can have small negative off-diagonal elements which we correct by (15). An alternative approach was recently presented by Crommelin and Vanden-Eijnden [4]. They solve

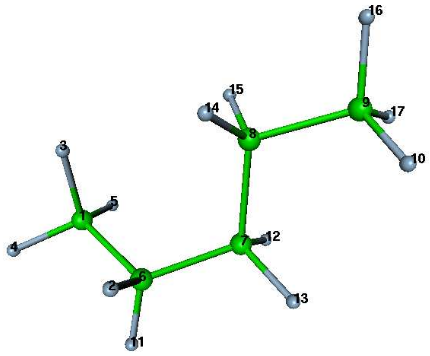

Figure 4: Balls-and-stick-representation of pentane with atom indices.

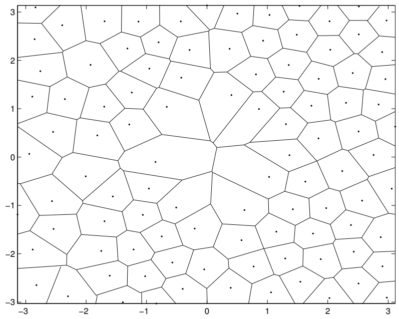

Figure 5: Discretization of the 2d space of torsion angles by Voronoi cells.

a minimization problem to find the generator whose eigenspectrum resembles the required one (in our case the one of Q_{Π}) as closely as possible. This assures to maintain the important features of the dynamics. In our opinion this method could be applied here. We especially aim to preserve the left eigenvector x_e as well as the right eigenvectors and eigenvalues which are used to construct the membership vectors χ.

4 Application to n-Pentane

We present the application to the n-pentane molecule $CH_3(CH_2)_3CH_3$ which was modeled with Merck Molecular Force Field [13, 14] at a temperature of $300K$. The rate matrix Q resulted from a conformation dynamics simulation with ZIBgridfree, a program package based on mesh-free methods which was developed at Zuse Insti-

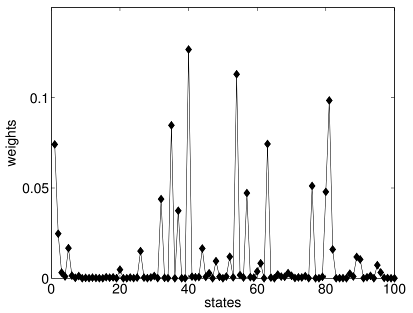

Figure 6: Invariant density of the 100 × 100 rate matrix.

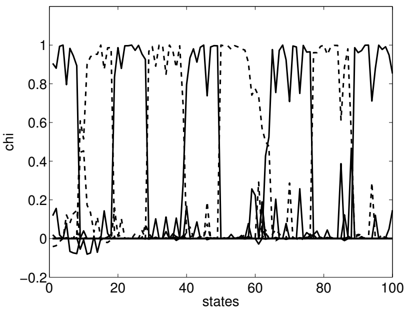

Figure 7: Membership functions χ_i of the 100 × 100 rate matrix. Note that the states on the right hand side are ordered according to the clusters such that the numbering of states is different compared to Figure 6.

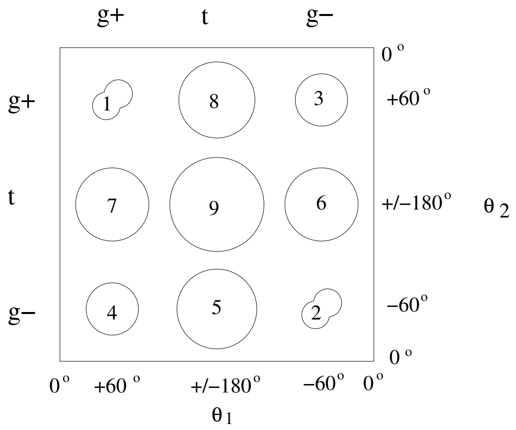

Figure 8: 2d schematic plot of the Boltzmann density of pentane w. r. t. the two dihedral angles θ_1 and θ_2 (left). The conformations are numbered from 1 to 9.

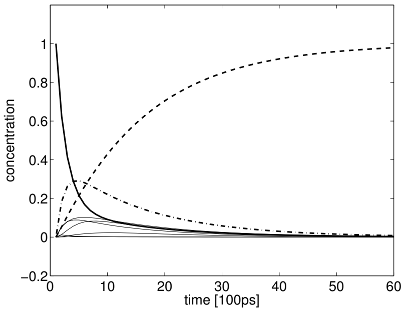

Figure 9: Matlab [9] plot of a conformation kinetics simulation. The 9 lines correspond to the weights of the 9 clusters. We selected the (g+/t)-conformation of pentane as start conformation (solid line) and the (t/g+)-conformation as end conformation (dashed line). The density is plotted over an interval of 6000ps at every 100ps.

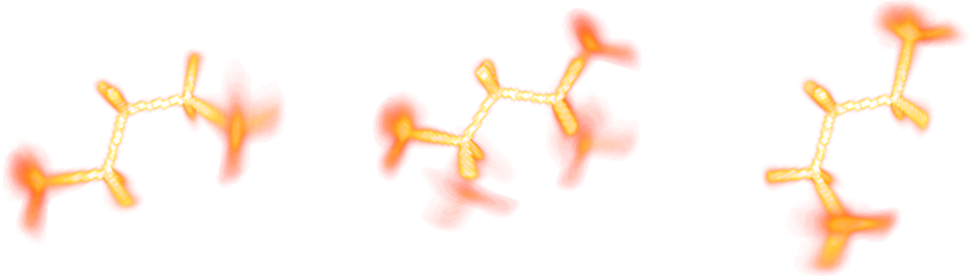

Figure 10: Volume rendering of the (g+/t)-conformations of pentane (left), the (t/g+)-conformation (right), and the corresponding transition macro-state (middle) in *amira/amiraMol* [21, 19].

tute Berlin [23, 18]. The dynamic behavior of pentane is described by two torsion angles θ_1 and θ_2. θ_1 is the torsion angle spanned by the atoms 6-7-8-9, θ_2 is spanned by the atoms 8-7-6-1, see Figure 4. The 2-dimensional space was discretized by 100 Voronoi cells, see Figure 5. The center of these cells were obtained by a selection of 100 sampling points from a high temperature pre-sampling at $1000K$. Within each cell, we generated nearly 3000 points according to the Boltzmann distribution by hybrid Monte-Carlo sampling with umbrella strategies and Gelman-Rubin convergence indicator. These points were propagated by molecular dynamics simulations until they left their starting cells. Average life times and conditioned transition probabilities were computed to set up the rate matrix Q.

To obtain the conformations, we applied PCCA+. We found 9 eigenvalues of Q close to 0,

$$\begin{aligned}\lambda &= \{2.6e-9,\ -2.2e-3,\ -3.0e-3,\ -3.7e-3,\ -4.7e-3,\\ &\quad -5.7e-3,\ -6.8e-3,\ -8.9e-2,\ -9.8e-2\},\end{aligned}$$

followed by a gap to the 10th eigenvalue $\lambda_{10} = -0.70$. This corresponds to 9 metastable conformations which can be distinguished according to the orientation of one of the two dihedral angles (g± and t denote the ± gauche and trans orientations), see Figure 8:

$$\text{conformations} = \{g+/g+, g-/g-, g+/g-, g-/g+, g-/t, t/g-, t/g+, g+/t, t/t\}$$

with

$$\text{weights} = \{0.0036, 0.0032, 0.0640, 0.0680, 0.1248, 0.1232, 0.1140, 0.1567, 0.3426\}.$$

The membership functions χ_i, $i = 1, \ldots, N_c$, are illustrated in Figure 7. The *minChi* value is 0.0808. From the coarse rate matrix, computed according to (14), we obtained the approximated mean holding times

$$h^{-1} = \{10.39, 11.38, 361.68, 334.35, 198.91, 227.31, 172.40, 235.74, 283.64\}ps,$$

and the transition probabilities of the embedded Markov chain

$$\hat{K}_c = \begin{pmatrix} 0 & 0.0000 & 0.0097 & 0.0105 & 0.0002 & -0.0002 & 0.5690 & 0.4045 & 0.0062 \\ 0.0000 & 0 & 0.0149 & 0.0098 & 0.4288 & 0.5403 & -0.0003 & 0.0009 & 0.0055 \\ 0.0172 & 0.0206 & 0 & 0.0000 & 0.0003 & 0.4502 & 0.0020 & 0.5089 & 0.0010 \\ 0.0148 & 0.0137 & -0.0001 & 0 & 0.5042 & 0.0020 & 0.4591 & 0.0048 & 0.0012 \\ 0.0003 & 0.1687 & -0.0001 & 0.1634 & 0 & -0.0062 & 0.0017 & 0.1881 & 0.4842 \\ 0.0000 & 0.3049 & 0.1462 & 0.0007 & -0.0337 & 0 & 0.1161 & 0.0023 & 0.4634 \\ 0.2898 & 0.0001 & 0.0003 & 0.1410 & 0.0011 & 0.0951 & 0 & -0.0164 & 0.4889 \\ 0.2184 & 0.0002 & 0.1355 & 0.0010 & 0.1778 & 0.0018 & -0.0243 & 0 & 0.4899 \\ 0.0018 & 0.0015 & 0.0000 & 0.0002 & 0.2518 & 0.2076 & 0.2675 & 0.2695 & 0 \end{pmatrix}.$$

Observe that conformation 3 and 4 have large mean holding times even though they have small weights. This hints to the fact that there are large energy barriers to the other conformations. The rate matrix was used for a kinetics simulation with predefined start and end conformation. The results for a $(g+/t) \rightharpoonup (t/g+)$ transition of pentane are shown in Figure 9. We performed the coarse kinetics according to

(14) with $\hat{Q}_c \in \mathbb{R}^{9\times 9}$ instead of the fine kinetics (10) with $Q \in \mathbb{R}^{100\times 100}$. Figure 9 can be interpreted as follows. During the conformational change from (g+/t) to (t/g+)-pentane, other conformations are formed which can be seen as transition states. The transition is visualized in Figure 10. The left picture shows the start conformation (g+/t), the right one the end conformation (t/g+). At each step of the 6000ps kinetics simulation, a similar density plot can be computed. The picture in the middle shows the transition state at 1000ps simulation length. It is very similar to the t/t-conformation and can be considered as the intermediate distribution of states at this particular time.

5 Conclusion

The complexity of molecular kinetics can be reduced significantly by a restriction to metastable conformations which are almost invariant sets of molecular dynamical systems. The main goal is to describe the dynamics as a Markov process on these conformations. Ideally, the coarse graining process results in a small transition probability or rate matrix which has the correct stationary distribution and reflects the essential dynamic behavior. However, due to discretization errors or the lack of metastabilities in the dynamical system, it is mostly not possible to meet all requirements at the same time, as mentioned in the introduction. While previous articles aimed to construct coarse matrices with certain structural properties, we take a different point of view and construct the coarse matrices such that they reflect the correct dynamic behavior. This is possible because the clustering method PCCA+ characterizes the conformations as linear combinations of the eigenvectors of the unreduced transition matrix. On this basis we defined a restriction and interpolation operator which are used for density transformations. However, our coarse matrices do not always admit a physical interpretation due to negative entries. But if there really exists a hidden low-dimensional Markov process in the model, these entries are small and can be neglected. Thus, for example it is possible to derive mean holding times and conditioned transition probabilities between the clusters from the coarse grained rate matrix. Furthermore, we have shown how one can construct a rate matrix corresponding to a steered kinetics process.

Acknowledgment The authors like to thank P. Deuflhard for his suggestions concerning mesh free methods, as well as W. Huisinga for his illustrative lecture on Markov processes. Furthermore, thanks to J. Schmidt-Ehrenberg and A. Riemer for their assistance in programming in *amira*.

References

[1] M. Bladt and M. Sørensen. Statistical inference for discretely observed Markov jump processes. *J. R. Statist. Soc. B*, 67:395–410, 2005.

[2] P. G. Bolhuis, C. Dellago, P. L. Geissler, and D. Chandler. Transition path sampling: throwing ropes over mountains in the dark. *Journal of Physics: Condensed Matter*, 12:A147–A152, 2000.

[3] P. Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues.* Number Texts in Applied Mathematics in 31. Springer-Verlag New York, 1999.

[4] D. T. Crommelin and E. Vanden Eijnden. Fitting timeseries by continuous-time Markov chains: A quadratic programming approach. Technical report, Courant Institute of Mathematical Sciences, New York University, New York, USA, January 2006. submitted to J. Comp. Phys.

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

[6] 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.

[7] P. Deuflhard and M. Weber. Robust Perron Cluster Analysis in Conformation Dynamics. In M. Dellnitz, S. Kirkland, M. Neumann, and C. Schütte, editors, *Lin. Alg. App. – Special Issue on Matrices and Mathematical Biology*, volume 398C, pages 161–184. Elsevier Journals, 2005.

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

[9] TheMathWorks Inc. Germany. Matlab(R) 6.5.0, 1994–2005.

[10] D. T. Gillespie. *Markov Processes: An Introduction for Physical Scientists.* Academic Press, Inc., San Diego, 1992.

[11] G.H. Golub and C.F. van Loan. *Matrix Computations.* Johns Hopkins University Press, 3rd edition, 1996.

[12] G. Grimmett and D. Stirzaker. *Probability and Random Processes.* Oxford University Press, New York, 2004.

[13] T.A. Halgren. The representation of van der Waals (vdW) interactions in molecular mechanics force fields: potential form, combination rules, and vdW parameters. *J. Am. Chem. Soc.*, 114:7827–7843, 1992.

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

[15] R. B. Israel, J. S. Rosenthal, and J. Z. Wei. Finding generators for Markov chains via empirical transition matrices, with applications to credit ratings. *Mathematical Finance*, 11(2):245–265, April 2001.

[16] M. Kijima. *Markov Processes for Stochastic Modeling.* Stochastic Modeling Series. Chapman and Hall, 1997.

[17] S. Kube and M. Weber. Conformation kinetics as a reduced model for transition pathways. Technical Report ZIB 05-43, Zuse Institute Berlin, 2005.

[18] H. Meyer. Die Implementierung und Analyse von HuMfree–einer gitterfreien Methode zur Konformationsanalyse von Wirkstoffmolekülen. Master's thesis, Free University Berlin, February 2005.

[19] J. Schmidt-Ehrenberg, D. Baum, and H.-Ch. Hege. Visualizing dynamic molecular conformations. In *IEEE Visualization 2002*, pages 235–242. IEEE Computer Society Press, 2002.

[20] 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.

[21] D. Stalling, M. Westerhoff, and H.-Ch. Hege. Amira - a highly interactive system for visual data analysis. In Christopher R. Johnson and Charles D. Hansen, editors, *Visualization Handbook*. Academic Press, November 2004.

[22] M. Weber und W. Rungsarityotin und A. Schliep. An Indicator for the Number of Clusters Using a Linear Map to Simplex Structure. In M. Spiliopoulou und R. Kruse und C. Borgelt und A. Nürnberger und W. Gaul, editor, *From Data and Information Analysis to Knowledge Engineering, Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation e.V., Universität Magdeburg, März 2005*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 103–110. Springer, 2006.

[23] M. Weber. *Meshless Methods in Conformation Dynamics*. PhD thesis, Free University Berlin, 2006.

[24] M. T. Wolfinger, W. A. Svrcek-Seiler, Ch. Flamm, I. L. Hofacker, and P. F. Stadler. Efficient computation of RNA folding dynamics. *J. Phys. A: Math. Gen.*, 37:4731–4741, 2004.