Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

SUSANNA KUBE, MARCUS WEBER

Identification of Metastabilities in Monomolecular Conformation Kinetics

ZIB-Report 06-01 (January 2006)

Identification of Metastabilities in Monomolecular Conformation Kinetics

Susanna Kube and Marcus Weber*

January 4, 2006

Abstract

The identification of metastable conformations of molecules plays an important role in computational drug design. One main difficulty is the fact that the underlying dynamic processes take place in high dimensional spaces. Although the restriction of degrees of freedom to a few dihedral angles significantly reduces the complexity of the problem, the existing algorithms are time-consuming. They are mainly based on the approximation of a transfer operator by an extensive sampling of states according to the Boltzmann distribution and short-time Hamiltonian dynamics simulations. We present a method which can identify metastable conformations without sampling the complete distribution. Our algorithm is based on local transition rates and uses only pointwise information about the potential energy surface. In order to apply the cluster algorithm PCCA+, we compute a few eigenvectors of the rate matrix by the Jacobi-Davidson method. Interpolation techniques are applied to approximate the thermodynamical weights of the clusters. The concluding example illustrates our approach for epigallocatechine, a molecule which can be described by seven dihedral angles.

1 Introduction

In computational drug design, we examine the binding capacity of different ligands to certain target molecules. Our goal is to find out if structural similarities of the ligands imply functional similarities. The results can be used to decide which new ligands are worth to be tested in laboratory experiments. Since the behavior of a single ligand molecule depends on its structure, we want to determine those conformations which are suitable for the docking process [5][6]. Therefore, the first step consists in the identification of all metastable conformations and their thermodynamical weights.

Conformation dynamics based on Hybrid Monte Carlo techniques [20] and the transfer operator approach [21] are powerful methods for conformation analysis. The software package ZIBgridfree [17] [28] [18] combines these ideas with

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

mesh free methods. It does not only figure out metastable conformations and their thermodynamical weights, but also the correct transition probabilities. However, this algorithm is very expensive because it is based on a correct sampling of the Boltzmann distribution even in transition regions which are occupied very seldom. Alternative methods are considered at the time, for example replica exchange and jump techniques. These techniques deliver the correct distribution which is essential for the computation of observables. Clustering methods like successive PCCA [9] or PCCA+ [10] can be used to identify metastable conformations, but further calculations would be necessary to approximate transition probabilities between these conformations.

These are the reasons why we were looking for a method which can identify metastable conformations and approximate thermodynamical weights without a sampling of the complete Boltzmann distribution. We developed an algorithm based on local transition rates which uses only pointwise information about the potential energy surface. This idea was motivated by the work of I.L. Hofacker et.al. [29] who examine folding dynamics of RNA secondary structures. In their approach, transitions between secondary structures correspond to the opening and closing of base pairs. They identify macro states as basins of attraction of one or more local minima of the energy function. To avoid the course of dimensionality, they restrict their examination to a few minima with lowest energy.

Our approach is more general and can be applied to different classes of molecules. The molecules are represented as points in a space Ω spanned by the essential dihedral angles. We allow transitions between points which have a small distance $d < \varepsilon$ in Ω. Hence, our method requires an arbitrary set of points in Ω which sufficiently represent the area of interest. The application of PCCA+ to the transition rate matrix identifies those basins which belong to a metastable conformation, independent from the depth of the minima. Furthermore, we use interpolation methods to compute the thermodynamic weights. This illustrates the main difference to [29], in that we respect entropic effects.

2 Point Concept

The topology of a molecule is uniquely defined by the coordinates q of its atoms in position space $\mathcal{Q}$. They also define the potential energy $V(q)$. Throughout the paper, we consider the (n, v, T)-ensemble (constant number of particles, constant volume, constant temperature) where the probability of being in state $q \in \mathcal{Q}$ is given by the Boltzmann distribution

$$\pi(q) = \frac{1}{\mathcal{Z}_{\mathcal{Q}}} \exp(-\beta V(q)).$$

$Z_{\mathcal{Q}} = \int_{\mathcal{Q}} \exp(-\beta V(q)\, dq)$ is the unknown spatial partition function, and β denotes the inverse temperature

$$\beta = \frac{1}{k_b T},$$

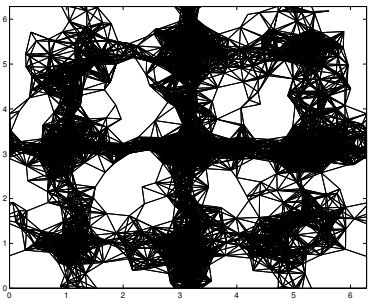

Figure 1: Set of possible transition pathways for pentane

where T is the temperature measured in Kelvin, and k_B is the Boltzmann constant.

However, it has been shown that the essential properties of molecules can be described by a few dihedral angles $\phi_1, \ldots, \phi_d$ which significantly reduces the number of degrees of freedom. Since we want to identify metastable conformations as subsets of $\Omega \in \mathbb{R}^d$, we do not only need a map $f : \mathcal{Q} \mapsto \Omega$ which is a standard routine in molecular dynamics, but we also need an energy function $\hat{V} : \Omega \mapsto \mathbb{R}$. In general, one point in Ω corresponds to several points in $\mathcal{Q}$ with different energies, but $\hat{V}$ has to be unique. Let us be given a set of states

$$\mathcal{S} = \{p_i\}_{i=1}^N \subset \Omega.$$

For every $p \in \Omega$ we define the energy function

$$\hat{V}(p) = \min\{V(q) \,|\, q \in \mathcal{Q},\, f(q) = p\}.$$

The probability of being in state p in equilibrium is given by

$$\pi(p) = \frac{1}{\mathcal{Z}} \int_{q, f(q)=p} \exp(-\beta V(q))\, dq.$$

Since we want to circumvent the sampling, we approximate

$$\pi(p) \propto \exp(-\beta \hat{V}(p)). \tag{1}$$

Furthermore, we define a set of possible transition ways

$$\mathcal{W} = \{(i,j) \,|\, p_i, p_j \in \mathcal{S},\, d(p_i, p_j) \leq \varepsilon\}$$

The distance measure d corresponds to the Euclidian distance in $\mathbb{R}^d$. Figure 1 shows an example for pentane where Ω is spanned by two dihedral angles.

Now, we describe the dynamics of the molecule as a random walk on $\mathcal{S}$ along paths in $\mathcal{W}$. We denote by $x_i(t)$ the probability that the molecule is in state p_i at time t. Then, the dynamics is described by the master equation

$$\dot{x}_i(t) = \sum_{j,\,(i,j)\in\mathcal{W}} q_{ji}\, x_j(t), \quad q_{ii} = -\sum_{j,\,(i,j)\in\mathcal{W}} q_{ij},$$

where q_{ji} denotes the transition rate from state p_j to state p_i. In matrix notation, the equation reads

$$\dot{\mathbf{x}} = Q^\top \mathbf{x}.$$

The matrix Q can be considered as the infinitesimal generator of an underlying continuous-time Markov jump process [8], given by the transition probability matrix P with $P(t) = \exp(tQ)$ where t is the simulation length of the corresponding molecular dynamics.

2.1 Energy Minimization

Assume we are given a set of points $\{q_i\}_{i=1}^N \in \mathcal{Q}$ and their corresponding images $p_i = f(q_i) \in \Omega$ which cover the relevant part of the dihedral space. These points can be obtained, for example, by a presampling procedure at high temperature which overcomes energy barriers [28]. Ideally, we want to preserve the dihedral angles during the minimization process,

$$V(q) \;\rightarrow\; \min!, \quad \phi_1, \ldots, \phi_d = \text{constant}.$$

We use the following restraint objective function in order to solve the minimization problem,

$$V_{\text{mod}}(q) = V(q) + \mu W(q) \;\rightarrow\; \min!, \quad \mu > 0, \quad W(q) = 1 - \cos(\phi(q) - \theta)$$

where ϕ is the vector of the current dihedral angles, and θ denotes the target angles or the target position in Ω respectively. Now we calculate the derivatives

$$\begin{aligned} \frac{\partial V_{\text{mod}}}{\partial q} &= \frac{\partial V_{\text{mod}}}{\partial \cos(\phi - \theta)} \frac{\partial \cos(\phi - \theta)}{\partial \cos(\phi)} \frac{\partial \cos(\phi)}{\partial q} \\ &= -\mu \left(\cos(\theta) - \frac{\sin(\theta)\cos(\phi)}{\sin(\phi)} \right) \frac{\partial \cos(\phi)}{\partial q}. \end{aligned}$$

The expression for the last factor can be obtained from [26]. As minimization routine we apply conjugate gradients with line search by the method of golden section.

Obviously, the molecule with minimized energy is represented by a different point $\tilde{p} \in \Omega$ than the original molecule. The parameter μ determines the flexibility of the molecule that we allow during the minimization process. If μ is chosen too large, the method fails to find the lowest energy in transition regions between minima (e.g. saddle points), i.e. we loose information about barriers on the energy surface. On the other hand, with a small μ we often get trapped in local minima which are irrelevant for the overall dynamic process. An automatic or maybe adaptive choice of μ is still ongoing work. At the time, we usually choose a large μ, but afterwards we use a selection routine which automatically eliminates those points in Ω with too large energy. Additionally, we apply the minimization algorithm to some points from the presampling trajectory with $\mu = 0$ in order to make sure that we find all relevant minima of $\hat{V}$.

2.2 Approximation of Transition Rates

It is a well known fact that the potential energy surface in $\mathcal{Q}$ forms a rough landscape with a number of local minima, especially for larger molecules. However, it turned out that $\hat{V}$ is much smoother because the fast degrees of freedom were eliminated. This allows the assumption that there is no barrier in the potential energy surface between points in Ω which are close to each other. Consequently, the transition rate between two points p_i and p_j only depends on the energy difference $\hat{V}(p_i) - \hat{V}(p_j)$ which can be related to the fraction $\frac{\pi_i}{\pi_j}$ by (1).

Since reversibility of Q implies reversibility of P [15], which is a natural and desired property of Markov jump processes, we want Q to meet the detailed balance condition

$$\pi_i q_{ij} = \pi_j q_{ji}, \quad \forall\, i, j \in \{1, \ldots, N\}.$$

This is a sufficient condition for reversibility.

It is important to note that validity of the detailed balance equation implies the existence of the stationary distribution π [8],

$$\pi^\top Q = 0.$$

Every expression

$$q_{ij} = \frac{1}{\pi_i} s_{ij}$$

with a symmetric function s meets the detailed balance condition. We cannot directly compute π due to the unknown partition function, but, as mentioned above, we operate with the fraction $R_{ij} = \frac{\pi_j}{\pi_i}$. This offers several possibilities for the choice of q_{ij} [8] [7].

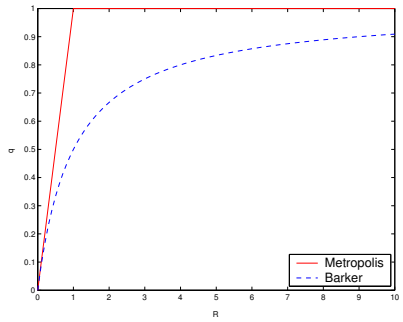

Figure 2: Illustration of transition rate q_{ij} for Metropolis and Barker kinetics dependent on the fraction R_{ij}.

Metropolis Dynamics The classic Metropolis choice is

$$q_{ij} = \min(1, R_{ij}) = \min(1, \frac{\pi_j}{\pi_i}) = \frac{1}{\pi_i} \min(\pi_i, \pi_j).$$

For irregularly distributed points in Ω, we take into account the number N_i of neighbors of a certain point p_i,

$$q_{ij} = \frac{1}{\pi_i} \min(\frac{\pi_i}{N_i}, \frac{\pi_j}{N_j}).$$

This accounts for the fact that the jump rate should be independent of the number of neighbors.

Barker Dynamics Chemists often use Barker's dynamic where q_{ij} is given by

$$q_{ij} = \frac{R_{ij}}{1 + R_{ij}} = \frac{1}{\frac{\pi_i}{\pi_j} + 1} = \frac{\pi_j}{\pi_i + \pi_j} = \frac{1}{\pi_i} \frac{1}{\frac{1}{\pi_i} + \frac{1}{\pi_j}},$$

or

$$q_{ij} = \frac{1}{\pi_i} \frac{1}{\frac{N_i}{\pi_i} + \frac{N_j}{\pi_j}},$$

respectively. The second factor can be interpreted as the total drag of two parallel resistors or the harmonic average, respectively.

3 Identification of Metastable Conformations

Subsets of Ω can be characterized by membership vectors $\chi : \mathcal{S} \mapsto [0, 1]$, i.e. $\chi(i)$ is the grade of membership of point $p_i \in \Omega$ to the conformation characterized by this vector. Similar to the analysis of almost invariant (or metastable) sets in conformation dynamics, we are looking for a partition of Ω into N_c subsets which remain almost invariant during the transition process. The membership vectors χ_k, $k = 1, \ldots N_c$, of these sets satisfy

$$Q\chi_k \approx 0, \quad \sum_{k=1}^{N_c} \chi_k = \mathbb{1}.$$

The first 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.

Proposition 3.1 *The eigenvalues of the rate matrix Q are located on the negative real axis in the interval $[-2 \max_i(|q_{ii}|), 0]$. Moreover, if Q is irreducible, the eigenvalue 0 is algebraically simple.*

Proof: Let D be the diagonal matrix with elements $d_{ii} = \pi_i$. Then, the reversibility constraint can be written in matrix notation,

$$DQ = Q^\top D.$$

Hence, $DQ = S$ is symmetric. Q can be reduced to a symmetric matrix Q^* by an orthogonal similarity transformation,

$$Q^* = D^{-1/2} S D^{-1/2} = D^{1/2} Q D^{-1/2}.$$

Consequently, Q has only real eigenvalues and is diagonalizable. Following Gershgorin's Theorem [12],

$$\lambda(Q) \subseteq \bigcup_{i=1}^{N} [q_{ii} - r_i, q_{ii} + r_i],$$

where $r_i = \min(\sum_{j,j\neq i} |q_{ij}|, \sum_{j,j\neq i} |q_{ji}|) = \min(|q_{ii}|, \sum_{j,j\neq i} q_{ji}) \leq |q_{ii}|$. Hence

$$\lambda(Q) \subseteq \bigcup_{i=1}^{N} [-2|q_{ii}|, 0],$$

and the proposition follows. Since Q has row sum zero, $e = (1, \ldots, 1)$ is an eigenvector of Q corresponding to the eigenvalue 0. Let $\theta = \max_i(|q_{ii}|)$ and consider the matrix $A = \theta I + Q$. The eigenvalues of A are given by $\lambda(A) = \theta + \lambda(Q) \subseteq [-\theta, \theta]$. The matrix A is non-negative and irreducible as long as Q is irreducible. According to the Perron-Frobenius-Theorem [4], there exists an eigenvalue $\lambda_0(A) > 0$ which is maximal in modulus among all the eigenvalues of A and algebraically simple. Due to the construction of A there is a one-to-one correspondence between the eigenvalues of A and Q. Since 0 is an eigenvalue of Q, θ is the largest eigenvalue of A. Since θ is algebraically simple, $\lambda_0(Q) = 0$ satisfies this proposition, too. □

Since Q is diagonalizable, the transition probability matrix $P = \exp(tQ)$ possesses the same eigenvectors as Q [15]. Motivated by our work on PCCA+ for such matrices [10], 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.

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

Since Q is ill-conditioned ($\sigma_{\max}/\sigma_{\min}$ is large), we suggest to compute the eigenpairs of $I - Q$ instead of Q. They are located in the interval $[1, 1 + |\lambda_{\max}(Q)|]$ with an eigenvalue cluster at 1.

3.1 The Jacobi-Davidson Method

We have to deal with a non-Hermitian eigenproblem

$$Ax = \lambda x.$$

Since Q is generalized symmetric, we could also solve the generalized eigenproblem

$$Sx = \lambda Dx$$

with the diagonal symmetric matrix D. However, D is not well conditioned which leads to highly distorted inner products. To circumvent this fact, we could use the so called QZ approach [11]. But it does not exploit symmetry of the involved matrix, so we do not gain anything.

The matrix A is large but sparse. Hence, a subspace oriented eigensolver is recommended. However, we do not know a priori how many eigenvalues of A are close to our target value. Therefore, it would be ineffective to work with a fixed subspace size. Instead of that, we decided to use the Jacobi-Davidson method [24] which delivers eigenvalues one after the other. We explain the main ingredients of this method following [11].

For the reason of numerical stability, it is advantageous to work with an orthonormal basis. Therefore, we compute a partial Schur form

$$AQ_k = Q_k R_k,$$

with an $(n \times k)$ orthonormal matrix Q_k and an upper triangular $(k \times k)$ matrix R_k. The diagonal entries of R_k represent the eigenvalues $\{\lambda_i\}_{i=1}^k$ of A. The pairs (q_i, λ_i) are called Schur pairs. Given the eigen-decomposition of R_k,

$$R_k Y = Y\Lambda,$$

the eigenpairs of A are obtained by

$$AQ_kY = Q_kR_kY = Q_kY\Lambda \quad \rightarrow \quad AX = X\Lambda \quad \text{with} \quad X = Q_kY.$$

To avoid that the same eigenpair is computed twice, we use a deflation technique [3]. Suppose that we have already detected $k-1$ Schur pairs, i.e. $AQ_{k-1} = Q_{k-1}R_{k-1}$. The following Schur vector q_k is required to satisfy

$$q_k \perp Q_{k-1}, \quad A[Q_{k-1}\, q_k] = [Q_{k-1}\, q_k] \begin{bmatrix} R_{k-1} & a \\ 0 & \lambda_k \end{bmatrix}$$

Hence, (q_k, λ_k) meets

$$q_k \perp Q_{k-1}, \quad Aq_k = Q_{k-1}a + q_k\lambda_k.$$

Denote by Q_{k-1}^* the conjugate transpose of Q_{k-1}. Since Q_{k-1} is orthonormal, a satisfies

$$a = Q_{k-1}^*(A - \lambda_k I)q_k.$$

This leads to

$$q_k \perp Q_{k-1}, \quad (I - Q_{k-1}Q_{k-1}^*)(A - \lambda_k I)q_k = 0.$$

Consequently, (q_k, λ_k) is an eigenpair of the deflated matrix

$$A_d = (I - Q_{k-1}Q_{k-1}^*)A(I - Q_{k-1}Q_{k-1}^*). \tag{2}$$

This eigenproblem is solved with the Jacobi-Davidson method. It is based on two main principles.

First, it is a subspace method, i.e. we look for an approximate eigenvector q in a search space $V \perp Q_{k-1}$. For this purpose, we compute the Schur form of the matrix

$$M = V^* A_d V = V^* A V, \quad MS = ST,$$

where S and T are ordered such that $\lambda = T(1,1)$ is closest to some target value $\tau \in \mathbb{C}$. The pair $(q, \lambda) = (VS(:,1), T(1,1))$ is an approximation for a wanted eigenpair of A_d. Observe that $\lambda = T(1,1) = S(:,1)^* MS(:,1) = q^* A_d q = q^* A q$.

The second idea of the Jacobi-Davidson method includes the expansion of the search space V with the solution of the Jacobi correction equation. Assume we are given an approximate eigenpair (q, λ), $||q||_2 = 1$ of the matrix A_d with residual

$$r = (A_d - \lambda I)q.$$

Note that $q \perp r$ because $\lambda = q^* A_d q$. We are looking for a correction $(v, \Delta\lambda)$ such that

$$A_d(q+v) = (\lambda + \Delta\lambda)(q+v), \quad v \perp q, \quad v \perp Q_{k-1}.$$

By ignoring the 2nd order terms we obtain the approximation

$$(A_d - \lambda I)v - \Delta\lambda\, q \approx -r.$$

Multiplication with the projection

$$P = I - qq^*$$

leads to the Jacobi correction equation

$$P(A_d - \lambda I)Pv = -r. \tag{3}$$

Afterwards, the search subspace $\mathrm{span}\{V\}$ is expanded with the orthonormal complement of v w. r. t. V. Observe that

$$\begin{aligned} r &= (A_d - \lambda I)q \\ &= (I - Q_{k-1}Q_{k-1}^*)A(I - Q_{k-1}Q_{k-1}^*)q - \lambda q \\ &= (I - Q_{k-1}Q_{k-1}^*)A(I - Q_{k-1}Q_{k-1}^*)q - \lambda(I - Q_{k-1}Q_{k-1}^*)(I - Q_{k-1}Q_{k-1}^*)q \\ &= (I - Q_{k-1}Q_{k-1}^*)(A - \lambda I)(I - Q_{k-1}Q_{k-1}^*)q. \end{aligned}$$

Hence $r \perp Q_{k-1}$. Since $q \perp Q_{k-1}$ and $(I - Q_{k-1}Q_{k-1}^*)^n = I - Q_{k-1}Q_{k-1}^*$, we obtain

$$\begin{aligned} (A_d - \lambda I)Pv &= A_d Pv - \lambda(I - Q_{k-1}Q_{k-1}^*)^2 Pv \\ &= (I - Q_{k-1}Q_{k-1}^*)(A - \lambda I)(I - Q_{k-1}Q_{k-1}^*)Pv. \end{aligned}$$

Now we can rewrite (3) as fully deflated Jacobi correction equation

$$\begin{aligned} &(I - qq^*)(I - Q_{k-1}Q_{k-1}^*)(A - \lambda I)(I - Q_{k-1}Q_{k-1}^*)(I - qq^*)v = -r. \\ \leftrightarrow \quad &(I - \tilde{Q}\tilde{Q}^*)(A - \lambda I)(I - \tilde{Q}\tilde{Q}^*)v = -r \quad \text{with} \quad \tilde{Q} = [Q_{k-1}\, q]. \end{aligned} \tag{4}$$

Since $\tilde{Q}$ is orthonormal and $r \perp \tilde{Q}$, this equation is equivalent to the following system of equations:

$$\begin{bmatrix} A - \lambda I & -\tilde{Q} \\ \tilde{Q}^* & 0 \end{bmatrix} \begin{bmatrix} v \\ \Delta\lambda \end{bmatrix} = \begin{bmatrix} -r \\ 0 \end{bmatrix} \tag{5}$$

Remark 3.2 *Consider the eigenvalue problem for $A = I - Q$. Since we already know the first eigenvector $\mathbf{e} = \frac{1}{\sqrt{N}}(1, \ldots, 1) \in \mathbb{R}^N$ according to $\lambda = 1$, we can start to search for eigenvectors $\mathbf{q}_i \perp \mathbf{e}$ by setting $R(1,1) = 1$ and $\mathbf{q}_1 = \mathbf{e}$. Then $\mathbf{e}_1 = (1, 0, \ldots, 0) \in \mathbb{R}^k$ will be an eigenvector of R_k corresponding to $\lambda_1 = 1$. Consequently we obtain $\mathbf{x}_1 = Q_k \mathbf{e}_1 = \mathbf{q}_1 = \mathbf{e}$ as desired.*

3.2 Solution of the Correction Equation

In general, (4) is solved by iterative methods. To achieve moderate convergence rates in the outer iteration, it suffices to solve (4) only approximately [24][23].

If (4) is to be solved by an iterative solver like GMRES, the rate of convergence can be improved considerably by using preconditioning. In [25] it was suggested to work with a preconditioner

$$\tilde{K} = (I - \tilde{Q}\tilde{Q}^*)K(I - \tilde{Q}\tilde{Q}^*),$$

where K is a preconditioner for $A - \tau I$ for one fixed value of τ which remains constant over some iterations. During the iteration process, we only have to solve

$$Kx = b.$$

An incomplete LU-factorization might be appropriate. However, if τ is close to an exact eigenvalue, then $A - \tau I$ is nearly singular. When using a large drop tolerance, the matrix factors are ill-conditioned in that some diagonal entries are nearly zero. There are several possibilities to circumvent these problems. For example, we can reorder the matrix elements to obtain a nearly block diagonal structure with well-conditioned leading blocks similar to the aggregation technique used for Markov chains [27]. Sleijpen and Wubs [22] have shown how such block-wise decompositions can be updated efficiently.

Instead of LU-decompositions, we could also use multilevel solvers. Since we are looking for eigenvalues close to zero, we choose $\tau = 0$, and the preconditioner for $A - \tau I$ is the same as for Q. Here, we can make use of the fact that $-Q$ is a singular M-matrix. For such matrices, algebraic multilevel solvers can be applied [1] but they did not turn out to be efficient for our problems.

However, in our special case we already know the first eigenvector. This leads to the idea to use a preconditioner for the matrix

$$\widetilde{M} = \begin{bmatrix} A - \lambda I & -q_1 \\ q_1^* & 0 \end{bmatrix}.$$

In contrast to $M = A - \lambda I$, $\widetilde{M}$ is regular and we obtain well-conditioned matrix factors. Let $\hat{Q} = [Q_{k-1}(:, 2:k-1) \quad q]$. Then we can split the system

$$\begin{bmatrix} M & -q_1 & -\hat{Q} \\ q_1^* & 0 & 0 \\ \hat{Q}^* & 0 & 0 \end{bmatrix} = \begin{bmatrix} \widetilde{M} & 0 \\ \hat{Q}^* & I \end{bmatrix} \begin{bmatrix} I & -\widetilde{M}^{-1}\hat{Q} \\ 0 & Z \end{bmatrix}$$

with $Z = \hat{Q}^* \widetilde{M}^{-1} \hat{Q}$. Thus, solving (5) just requires the action of the inverse of $\widetilde{M}$ or the action of the preconditioner in each iteration step, respectively. As new vectors are appended to the matrix Q, we must update $\widetilde{M}^{-1}\hat{Q}$ and Z. This is also necessary when the shift τ changes.

On the other hand, if we have a high-quality preconditioner available for $A - \lambda I$ the solution of (5) is already a good expansion vector for the search space and iterative methods need not be applied. This was pointed out by Sleijpen and Wubs in [22]. Furthermore, it turned out that it also suffices for our problem to solve (5) with constant shift $\lambda = \tau$ over all iterations. No updating of the preconditioner is needed. For this purpose we computed an incomplete LU-decomposition of $\widetilde{M}$ with small drop tolerance *droptol*. This must only be done once at the beginning of the outer iteration. Figure 3 illustrates the results for a matrix $Q \in \mathbb{R}^{1914 \times 1914}$ which was obtained from a discretization of the potential energy surface of epigallocatechine.

4 Computation of Observables

So far we have obtained a number of points and corresponding membership vectors. Now we would like to use this information in order to compute the weights of the conformations which are given by

$$\pi_i = \frac{1}{\mathcal{Z}_\mathcal{Q}} \int_\Omega \chi_i(q) \exp(-\beta V(q)) dq$$

and must be approximated by

$$\pi_i \approx \sum_j \chi_i(p_j) \exp(-\beta \hat{V}(p_j)) w(p_j).$$

Unfortunately, the points are not distributed according to the Boltzmann distribution and we do not know the correct weights $w(p_j)$. However, we can try to approximate these weights by interpolation. For this purpose, we used different interpolation techniques as described below. The computation of conformational weights is just an instance of the more general problem of computing observables

$$< f >= \frac{1}{\mathcal{Z}_\mathcal{Q}} \int_\Omega f(q) \exp(-\beta V(q)) dq.$$

In our case, $f = \chi_i$. Note that we only know the value of χ_i at the given points p_j.

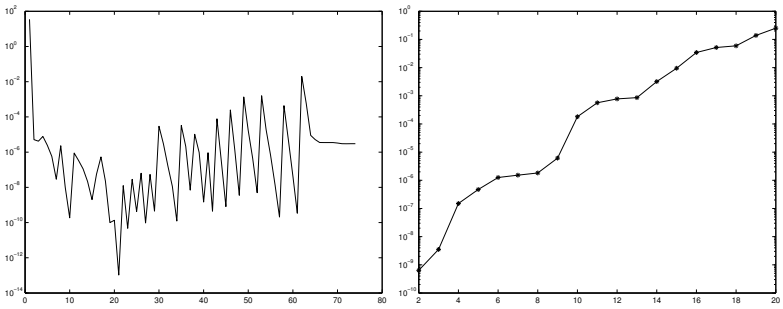

Figure 3: Convergence of eigenvalues of a rate matrix for epigallocatechine during the Jacobi-Davidson iteration. Computation of the first 20 eigenvalues close to zero. We chose $A = I - Q$ and kept the constant shift $\tau = 1$ during all iterations. $\widetilde{M}$ was approximated by an incomplete LU-decomposition with drop tolerance $1e - 10$. The figure on the left shows the convergence history, i.e. the $\log_{10}$ of the Euclidian norm of the residuals as a function of the iteration number. The jump in the curve corresponds to the detection of eigenpairs and a start for another eigenpair. Eigenpairs were accepted if the norm of the residual was less than $1e-8$. The method stagnated after the detection of 18 eigenvalues. The figure on the right shows the detected eigenvalues $1 - \lambda(A)$ plus the next two values that were obtained by a new start of the algorithm with a different shift $\tau = 1.3$. Note that the first eigenvalue equals zero.

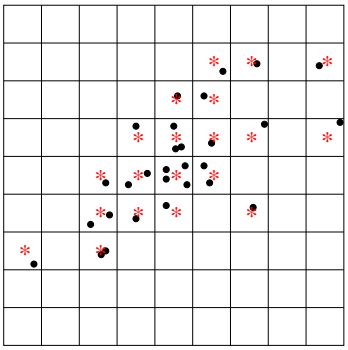

Figure 4: Grid for interpolation and approximation of observables. The points illustrate the given ones, the stars are the mid points of the squares.

4.1 Interpolation Techniques

In a first approach, we covered the domain Ω with patches A_i with midpoints m_i and side length h_i. The covering can be a regular grid as in figure 4. The patches could also have different sizes but must not overlap. Then, the integral is written as

$$< f >= \sum_i |A_i| f(m_i),$$

where $|A_i| = h_i^{\dim}$ denotes the size of patch A_i. There are several possibilities to approximate $f(m_i)$.

Local Approximation First, we take into account the points p_j which are close to m_i, and approximate

$$f(m_i) = \frac{1}{N_i} \sum_{\{j|p_j \in A_i\}} f(p_j) \exp(-\beta \hat{V}(p_j)) \tag{6}$$

where N_i is the number of points p_j in patch A_i.

Global Approximation The pointset $\{p_j\}_{j=1}^N$ can be used to define a set of radial basis functions.

$$\varphi_j(p) = \frac{\exp(-\alpha \operatorname{dist}(p, p_j))}{\sum_j \exp(-\alpha \operatorname{dist}(p, p_j))}.$$

These functions form a partition of unity and can be interpreted as weights of the points p_j in an arbitrary point p. Hence, $f(m_i)$ can be approximated by

$$f(m_i) = \sum_j f(p_j) \varphi_j(m_i).$$

This interpolation technique is frequently used in mesh free methods [16].

5 Numerical Illustrations

We applied our algorithm to epigallocatechine, see figure 5. Its conformational space is described by 7 dihedral angles. Three dihedrals determine the form of the ring R0, three dihedrals at the oxygen atom (denoted by "oxy" in figure 7) determine the position of the ring R3, and one dihedral determines the orientation of the ring R2.

We generated a molecule trajectory of 20000 time steps at 300K by using a hybrid Monte Carlo method combined with replica exchange. From these molecules we chose 1500 molecules evenly spaced in the important part of the dihedral space. We applied a local minimization routine as described in section 2 with penalty parameter $\mu = 100$. Furthermore, we chose 500 equally spaced position states to which we applied an unconstrained minimization of energy. Afterwards, the 2000 elements were reduced to 1500 removing those molecules which had a high potential energy but were close to molecules with low energy. Thereafter we computed the transition rates between points which were closer than a distance $\varepsilon = 0.3$ in the dihedral space.

The eigenvectors corresponding to eigenvalues close to zero are computed by the Jacobi-Davidson algorithm. Besides the gap in the spectrum, we computed the minChi-values [10]. We decided to take the largest possible number of clusters for which the minChi indicator as well as the eigenvalue is acceptable. This results in 18 clusters. The behavior of minChi as well as the resulting membership vectors are shown in figure 6.

By plotting histograms of the dihedral angles for the different clusters, we identified those angles which are significant for the partition:

$$D1:\ 7-9-44-43, \quad D3:\ 9-44-51-6, \quad D4:\ 10-25-19-17.$$

The numbers correspond to the labeling of the atoms in figure 7. The dihedral angles $D3$ and $D4$ indicate the orientation of the two rings $R2$ and $R3$ with respect to rotations while the angle $D1$ determines the form of the ring $R0$ and consequently the positions of the rings $R2$ and $R3$ relatively to each other. The distribution of these angles within single clusters are similar to normal distributions. Therefore, we considered their mean values as significant feature for the identification of a cluster. They are reported in the table. Note that the angles are in the periodic interval $[-180, 180]$. Since single points belong to several clusters, the number of points in the table is not an integer. The last cluster can be considered as numerical artefact because it consists of a single point. The weights were determined by (6) with $h_i = 2\pi/10$ for all i.

The last column displays the lowest energy of all points which would uniquely belong to the corresponding cluster if the soft clustering was transformed into a hard clustering. The table is divided into different blocks. Within the first three blocks, there is a 4-clustering induced by a rotation of the two rings $R2$ and $R3$. The first block is divided from the other ones by the angle $D1$ which corresponds to a change in the positions of rings $R2$ and $R3$. The fourth block in the table represents a third state of angle $D1$ which also forces the angles $D3$ and $D4$ to take positions different to the ones in the previous blocks. However,

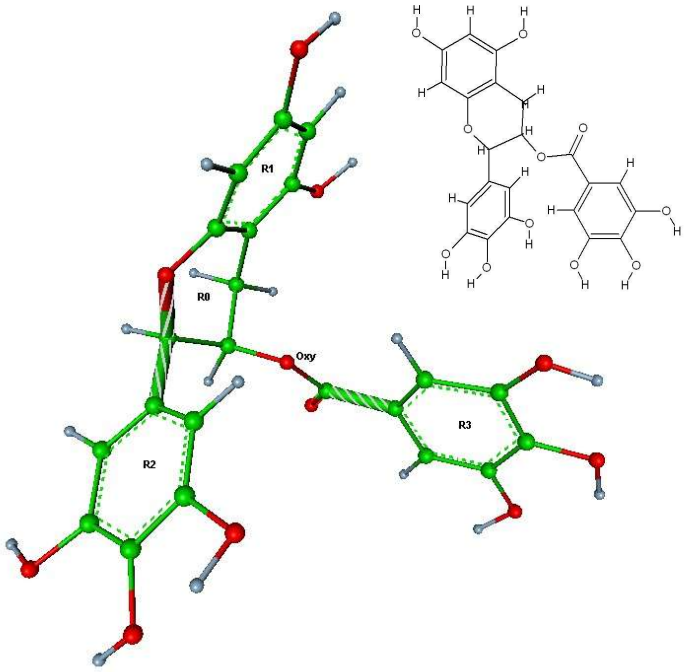

Figure 5: 3d representation and line formula of epigallocatechine. The 3d plot was generated with *amira* [2].

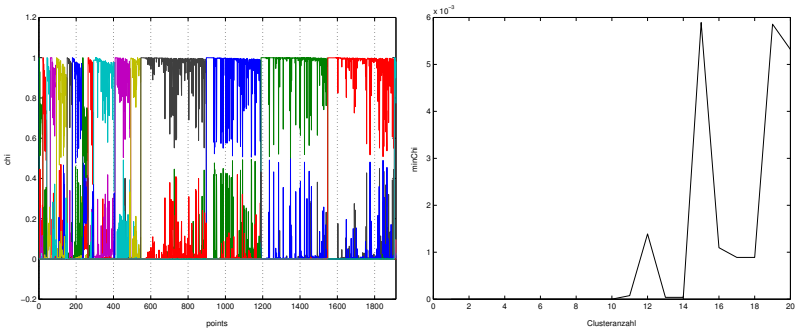

Figure 6: Transformed eigenvectors of the transition rate matrix for epigallocatechine and the minChi-values for different numbers of clusters.

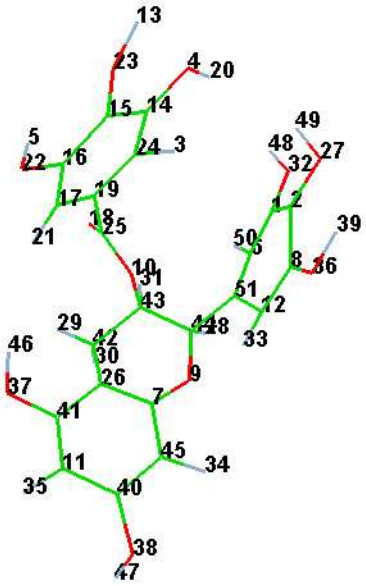

Figure 7: 3d representation of epigallocatechine with labeled atoms.

	D1	D3	D4	points	weight	minimal energy
cluster 1	48.2	-38.2	-3.4	356.0	31.9	115.9
cluster 2	47.8	147.6	-1.9	349.1	13.3	116.4
cluster 3	48.3	144.8	175.4	363.8	33.2	116.4
cluster 4	48.3	-37.7	177.7	290.6	20.5	116.6
cluster 5	-33.8	-33.4	-9.0	24.6	0.058	129.9
cluster 6	-33.0	145.7	0.9	18.4	0.140	130.3
cluster 7	-30.7	139.7	177.0	84.0	0.420	129.9
cluster 8	-32.2	-47.5	-178.4	116.6	0.189	130.2
cluster 9	-29.7	-48.5	11.1	60.7	0.061	130.5
cluster 10	-30.1	137.0	4.5	54.6	0.158	130.7
cluster 11	-34.8	151.6	173.8	29.8	0.016	135.2
cluster 12	-36.0	-30.8	174.9	31.5	0.048	134.9
cluster 13	-47.4	-136.3	12.2	54.5	0.034	134.7
cluster 14	-49.0	-141.1	-178.0	15.8	0.001	141.3
cluster 15	-50.7	-132.7	-13.9	22.3	0.001	141.5
cluster 16	-53.1	39.4	179.7	30.9	0.0004	143.6
cluster 17	-43.5	-132.1	-74.9	9.2	6e-5	150.4
cluster 18	48.4	-	-	1.5	1e-8	320.6

Table 1: Mean values of significant dihedral angles for different clusters.

apart from cluster 13, they have a vanishing weight due to their high potential energy.

6 Conclusion and Outlook

We have demonstrated how pointwise information about the potential energy landscape of a molecule can be used to identify metastable conformations. As an advantage of our algorithm, we know the stationary distribution of the underlying Markov jump process in advance which turned out to be useful for the eigenvector computation. One conformation comprises different basins of attraction of local minima which can be connected by paths over low energy barriers while different conformations are separated by high energy barriers. Points in transition regions are used to obtain information about the shape of these conformations in the dihedral space. They serve for the approximation of the cluster weights.

Of course, we can only extract the information which is contained in the given points. In future, we aim to include more points in regions where they are necessary in order to improve the results, but without a complete re-computation of the eigenvectors. Furthermore, it remains to examine the influence of certain parameters, for example the transition length ε or the penalty parameter μ, on

the results.

Acknowledgment The authors like to thank P. Deuflhard for his suggestions concerning mesh free methods. Furthermore, thanks to J. Schmidt-Ehrenberg for his patience while explaining programming details in *amira*, and thanks to H. Meyer for the discussions concerning point based methods.

References

[1] *Algebraic Multigrid*, chapter 4, pages 73–130. SIAM, 1987.

[2] Amira—advanced visualization, data analysis and geometry reconstruction, user's guide and reference manual. Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB), Indeed–Visual Concepts GmbH and TGS Template Graphics Software Inc., 2000.

[3] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, editors. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. SIAM, Philadelphia, 2000.

[4] R.B. Bapat and T.E.S. Raghavan. *Nonnegative Matrices and Applications*. Cambridge University Press, 1997.

[5] D. Baum. Multiple semi-flexible 3d superposition of drug-sized molecules. Technical Report 04-52, Zuse Institute Berlin (ZIB), 2004.

[6] T. Baumeister and F. Cordes. A new model for the free energy of solvation and its application in protein ligand scoring. Technical Report 04-51, Zuse Institute Berlin (ZIB), 2004.

[7] L. J. Billera and P. Diaconis. A geometric interpretation of the Metropolis-Hastings algorithm. *Statistical Science*, 16(4):335–339, 2001.

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

[9] 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.

[10] 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.

[11] Diederik R. Fokkema, Gerhard L. G. Sleijpen, and Henk A. van der Vorst. Jacobi–Davidson style QR and QZ algorithms for the reduction of matrix pencils. *SIAM J. Sci. Comput.*, 20(1):94–125, 1998.

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

[13] T.A. Halgren. *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] S. Kube and M. Weber. Conformation kinetics as a reduced model for transition pathways. Technical Report 05-43, Zuse Institute Berlin, 2005.

[16] G. R. Liu. *Mesh Free Methods – Moving beyond the Finite Element Method*. CRC Press, 2002.

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

[18] H. Meyer, M. Weber, and A. Riemer. HuMfree. Software package for HMC-simulation and conformation analysis based upon C++ classes of amiraMol [19] using the Merck Molecular Force Field [13, 14] implemented by T. Baumeister and parametrized by F. Cordes. Robust Perron Cluster Analysis implemented by M. Weber and J. Schmidt-Ehrenberg. VERX (extrapolation method based on Verlet) implemented by U. Nowak, Status: January 2005. Software owned by the Zuse Institute Berlin.

[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] 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. Preprint. Available via http://www.math.fu-berlin.de/~biocomp.

[22] G. L. Sleijpen and F. W. Wubs. Exploiting multilevel preconditioning techniques in eigenvalue computations. *SIAM J. Sci. Comput.*, 25(4):1249–1272, 2003.

[23] G. L. G. Sleijpen, A. G. L. Booten, D. R. Fokkema, and H. A. Van der Vorst. Jacobi-Davidson type methods for generalized eigenproblems and polynomial eigenproblems. *BIT*, 36:3:595–633, 1996.

[24] G. L. G. Sleijpen and H. A. Van der Vorst. A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 17(2):401–425, 1996.

[25] G. L. G. Sleijpen, H. A. Van der Vorst, and E.Meijerink. Efficient expansion of subspaces in the Jacobi-Davidson method for standard and generalized eigenproblems. *Electronic Transaction on Numerical Analysis*, 7:75–89, 1998.

[26] W. F. van Gunsteren, S. R. Billeter, A. A. Eising, P. H. Hünenberger, P. Krüger, A. E. Mark, W. R. P. Scott, and I. G. Tironi. *Biomolecular Simulation: The GROMOS96 Manual and User Guide.* vdf Hochschulverlag AG, ETH Zürich, 1996.

[27] Wei Wu W. J. Stewart. Numerical experiments with iteration and aggregation for Markov chains. *ORSA Journal on Computing*, 4:336–350, 1996.

[28] M. Weber and H. Meyer. ZIBgridfree - adaptive conformation analysis with qualified support of transition states and thermodynamic weights. Technical Report ZIB 05-17, Zuse Institute Berlin, 2005.

[29] 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.