II International Conference on Particle-based Methods - Fundamentals and Applications
PARTICLES 2011
E. Oñate and D.R.J. Owen (Eds)

SOFT VERSUS HARD METASTABLE CONFORMATIONS IN MOLECULAR SIMULATIONS

Konstantin Fackeldey*, Susanna Röblitz*, Olga Scharkoi* AND Marcus Weber*

*Zuse Institute Berlin (ZIB)
Takustrasse 7
14195 Berlin, Germany
e-mail: {fackeldey, susanna.roeblitz, scharkoi, weber}@zib.de, www.zib.de

Key words: Proteins, Conformation Space, Meshfree Methods

Abstract. Particle methods have become indispensible in conformation dynamics to compute transition rates in protein folding, binding processes and molecular design, to mention a few. Conformation dynamics requires at a decomposition of a molecule's position space into metastable conformations. In this paper, we show how this decomposition can be obtained via the design of either "soft" or "hard" molecular conformations. We show, that the soft approach results in a larger metastabilitiy of the decomposition and is thus more advantegous. This is illustrated by a simulation of Alanine Dipeptide.

1 Introduction

In practice, molecular simulations are carried out by solving the equations of motion of molecular dynamics. The solution of the ordinary differential equation results in a trajectory in state space (position and momentum space) and is a model for a closed system behavior of the molecule, i.e. a simulation at constant energy. The trajectory is analyzed in position space in order to derive statistical information about the molecular system. In this article, we focus on simulations at constant temperature (canonical ensemble) instead of constant energy (microcanonical ensemble). The dynamical system under consideration is a Markov chain in position space which will be derived in the next section.

Molecular simulations are not as easy as it seems at a first glance. If we observe certain internal coordinates of a simple molecule (like the C-C-C-C torsion angle of butane in Figure 1), we see that the mentioned Markov chain comprises metastabilities. The system jumps between so-called *metastable conformations*. The aim of conformation dynamics is to analyze this jump process, i.e. to identify the metastable conformations in position space, their statistical weights, and to compute the transition probabilities between them [1, 2, 9].

Although it seems to be a good idea to derive this information from a long-term Markov chain, the metastabilities are problematic from a statistical point of view. Since jumps between metastable conformations are rare events, even a long-term simulation (carried out by the largest parallel computing machines) does not provide enough statistical information for the derivation of the transition patterns. Furthermore, counting the number of Markov states per metastable conformation is not a good idea for deriving statistical weights of the conformations, because rapid (global) equilibration is avoided by the rare events.

A rather old idea to circumvent the problem of simulation in the presence of high-energy barriers is given by transition state theory. In this context, many researchers think of a one dimensional reaction coordinate (plotting the reaction coordinate against the free energy level). Two minima representing the educt as well as the product, respectively. An energy barrier between these two states has a local maximum (the transition state). Furthermore, the energy difference between the transition state and the minimum determines the reaction rate.

This simple picture does not hold for all kinds of molecular simulations. Briefly, the higher the barrier the better are the results from transition state theory. In conformation dynamics this insufficient picture is corrected. The metastable conformations are not defined as local minima of a free energy landscape. In contrast, the whole high-dimensional position space is *decomposed* into metastable conformations. More precisely, in Figure 1 the y-axis (representing the position space) may be decomposed into three intervals which are called metastable conformations in this context. Transition state theory searches for a certain point at which the molecular system switches from conformation A to conformation B, whereas a conventional, set-based decomposition approach of conformation dynamics aims at finding high-dimensional transition "hyper planes". In this article, we will replace these transition hyper planes by soft barriers. That means we will replace a set-based decomposition of the position space (hard clustering) by a partition of unity decomposition of the position space using membership functions (soft clustering). Although it seems that this soft clustering leads to rather "unstable" metastable conformations, this is not true. We will give an illustrative example in section 5. In fact, our approach provides conformations with the highest "metastability" and (assuming a perfect discretization) the systematic error of the set-based transition rates described by [8] vanishes. Simply speaking, our more complex picture does hold for all molecular systems and provides an effective analysis of the transition pattern between molecular conformations.

2 Statistical Mechanics

In a canonical ensemble the state of a biomolecule is not described by a single global minimum energy structure, but by a statistical ensemble in a phase space Γ. For $x = (q, p) \in \Gamma = \Omega \times \mathbb{R}^d$ the positions q and momenta p of each atom in the molecule are given

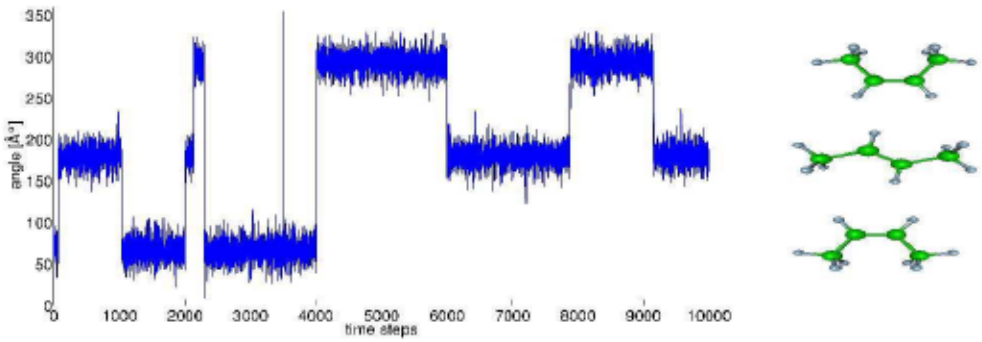

Figure 1: The longterm simulation of butene clearly shows three metastable conformations of the molecule.

according to the Boltzmann distribution:

$$\pi(q,p) \propto \exp(-\beta H(q,p)). \tag{1}$$

Here $\beta = 1/k_B T$ is the inverse temperature T multiplied with the Boltzmann constant k_B, and H denotes the Hamiltonian function which is given by $H(q,p) = V(q) + K(p)$, where $V(q)$ is the potential and $K(p)$ the kinetic energy. This canonical density can be split into a distribution of momenta $\eta(p)$ and positions $\pi(q)$ where

$$\pi(q) \propto \exp(-\beta V(q)) \text{ and } \eta(p) \propto \exp(-\beta K(p)).$$

Let us consider the Hamiltonian dynamics which is given by

$$\dot{q} = p\,,\ \dot{p} = -\nabla V(q), \tag{2}$$

where $\nabla V(q)$ is the gradient of an energy function (the potential) $V : \Omega \to \mathbb{R}$. It is well known, that (2) can be the starting point for a trajectory based description of this system in a microcanonical ensemble. In contrast, we now consider a system which is embedded in a heat bath with constant temperature T in a canonical ensemble. According to (2) the corresponding flow Φ^τ for a time span $\tau > 0$ is given by

$$(q(t), p(t)) = \Phi^\tau(q_n, p_n), \qquad n \in \mathbb{N}.$$

Let Π_q be the projection of the state (q,p) onto the position q and let further p be chosen randomly according to the distribution $\eta(p)$, then

$$q_{i+1} = \Pi_q \Phi^\tau(q_i, p_i)$$

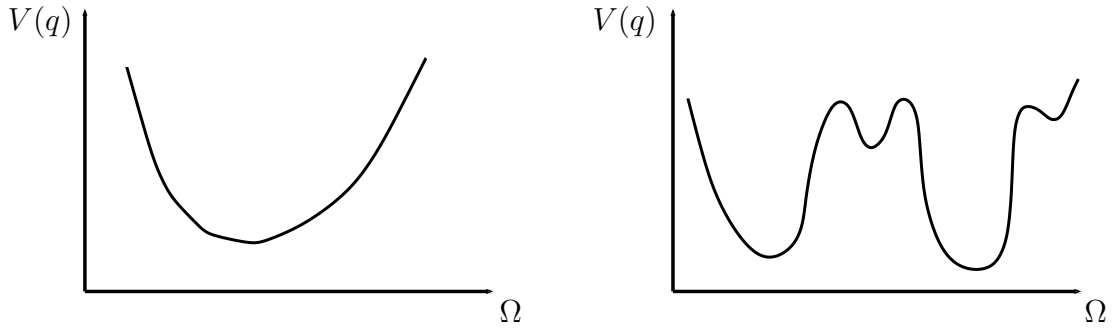

Figure 2: Left: Energy landscape of a rigid molecule with a unique minimum. Right: The energy landscape of biomolecules is in general rough, with multiple local minima.

describes a Markov process. The ith state depends on the preceding state only.

It can be shown, that this assumption of Markovianity implies that the corresponding Liouville operator is time independent e.g. [9]. By projecting this Liouville operator onto the position space the behavior of the system can be described by a transition function [9], which is given by

$$p(\tau, f, h) = \int_\Omega T^\tau f(q)h(q)\pi(q)\,dq, \tag{3}$$

where

$$T^\tau f(q) = \int_{\mathbb{R}^d} f(\Pi_q \Phi^\tau(q,p))\eta(p)\,dp. \tag{4}$$

This construction offers many advantages for the analysis of molecular processes. The fundamental idea behind this formulation is, that the transfer operator T^τ in (4) is a *linear* operator although the ordinary differential equation (2) is (extremely) *non-linear*. This linearization allows for a Galerkin discretization of T^τ and thus for a numerical approximation of eigenfunctions and eigenvalues of the discrete spectrum of T^τ.

We take advantage of the fact, that the behavior of molecules can be well described by its structurally related configurations (metastable conformations). Mathematically speaking, a metastable conformation is a function $C : \Omega \to [0,1]$ which is nearly invariant under the transfer operator T^τ, i.e.

$$T^\tau C(q) \approx C(q). \tag{5}$$

In the following we show how the metastable conformations can be computed via a discretization of the position space.

3 Discretization

In order to "resolve" or to identify the metastabilities we need a discretization of the position space. At this stage, we define a decomposition of Ω, which we need in order to employ our techniques. Let therefore $\{\theta_i\}_{i=1}^N$ be a set of basis function and $\Omega_i = \text{supp}(\theta_i)\,\forall i = 1, ..., N$. We say that the basis functions $\{\theta_i\}_{i=1}^N$ are a *hard decomposition* of Ω, if

i) Ω_i is measurable and $|\Omega_i| > 0$ for $i = 1, ..., N$

ii) $|\Omega_i \cap \Omega_j| = 0$ if $i \neq j$

iii) $\sum_{i=1}^{N} \theta_i(q) = \mathbf{1}_\Omega$, $q \in \Omega$.

As we have already mentioned, the position space is high dimensional which prohibits any usage of meshbased methods like finite elements. Thus we take advantage of meshfree methods, more precisely we consider a Voronoi tessellation. We choose characteristic basis functions $\{\chi_i\}_{i=1}^N$ with $\chi_i : \Omega \to \{0, 1\}$ defined by

$$\chi_i(q) = \mathbf{1}_{\Omega_i}(q) := \begin{cases} 1 & \text{if } q \in \Omega_i \\ 0 & \text{otherwise} \end{cases} .$$

Obviously, these basis functions suit the requirements of a hard decomposition. Moreover it is easy to see that the $\chi_i, ..., \chi_N$ form a partition of unity, i.e.

$$\sum_{i=1}^{N} \chi_i(q) = 1 \text{ and } \chi_i \geq 0 \qquad \forall i.$$

In terms of the characteristic basis functions $p(\tau, \chi_i(q), \chi_j(q))$ describes the transition probability between the two sets Ω_i and Ω_j. In other words, it describes the ratio of trajectories starting in $q \in \Omega_i$ with Boltzmann distributed momenta $p \in \mathbb{R}^d$ and ending in Ω_j after timespan $\tau > 0$.

Having now a discretization of Ω we can give the metastabilities a more precise meaning. To do so, we aim at a set $\{C_1, ..., C_{n_c}\}$ such that $\sum_{J=1}^{n_c} C_J(q) = \mathbf{1}_\Omega \quad \forall q \in \Omega$ where $C_J : \Omega \to [0, 1]$ is a function. Then we can define each C_J as a linear combination of the basis functions $\{\chi_i\}_{i=1}^N$. More precisely

$$C_J(q) = \sum_{i=1}^{N} G_{iJ}\chi_i(q), \qquad J = 1, ..., n_c. \tag{6}$$

Here and in the forthcoming, we use capital $I, J, ...$ indices for numbering of the metastable conformations. In order to employ a Galerkin discretization, we define
$\langle g, f\rangle_\pi = \int_\Omega f(q)g(q)\pi(q)\, dq$ and insert (6) into (5), s.th.

$$\begin{aligned} \langle T^\tau(\sum_{i=1}^{N} G_{iJ}\chi_i), \chi_j\rangle_\pi &\approx \langle \sum_{i=1}^{N} G_{iJ}\chi_i, \chi_j\rangle_\pi \\ \sum_{i=1}^{N} \langle T^\tau G_{iJ}\chi_i, \chi_j\rangle_\pi &\approx \sum_{i=1}^{N} \langle G_{iJ}\chi_i, \chi_j\rangle_\pi. \end{aligned} \tag{7}$$

Since $\langle \chi_i, \chi_j \rangle_\pi = \delta_{ij} \langle \chi_j, \chi_j \rangle_\pi$ we obtain

$$\sum_{i=1}^{N} G_{iJ} \langle \chi_i, \chi_j \rangle_\pi = G_{jJ} \langle \chi_j, \chi_j \rangle_\pi.$$

Thus (7) is equivalent to

$$\sum_{i=1}^{N} G_{iJ} \langle T^\tau \chi_i, \chi_j \rangle_\pi \approx G_{iJ} \langle \chi_j, \chi_j \rangle_\pi. \tag{8}$$

Dividing both sides of (8) by $\langle \chi_j, \chi_j \rangle_\pi$ we can define

$$P^\tau_{ji} = \frac{\langle \chi_j, T^\tau \chi_i \rangle_\pi}{\langle \chi_j, \chi_j \rangle_\pi} = \int_\Omega T^\tau \chi_i(q) \frac{\chi_j(q)}{\int_\Omega \chi_j(q)\pi(q)dq} dq$$

and we obtain for the coefficients G_{iJ}

$$P^\tau \mathbf{g}_J \approx \mathbf{g}_J, \tag{9}$$

where $\mathbf{g}_J = [G_{1J}, G_{2J}, \ldots, G_{NJ}]^T$. The stochastic matrix P^τ describes the transition probabilities between the basis functions.

We remark that the computation of the above integral is a challenging task, since the underlying space is high dimensional. To overcome this, we employ strategies from particle methods. In detail, we apply Markov chain Monte Carlo methods [4] in each Voronoi cell to generate a local Boltzmann distribution $\pi_i(q)$ for each of the basis functions $\{\chi_i\}_{i=1}^N$, i.e.

$$\pi_i(q) = \frac{\chi_i(q)}{\int_\Omega \chi_i(q)\pi(q)dq}.$$

The sampled positions q are propagated by molecular dynamics according to Φ^τ with randomized initial momenta. With these data we compute the entries of P^τ. So far we have not given any details for the matrix G in (6). Let us therefore point out the two following aspects:

- The coefficients G_{iJ} can then be computed as a linear combination

  $$G = X\mathcal{A}$$

  of the eigenvectors X of P^τ corresponding to eigenvalues close to one. For the n_c metastable conformations, T^τ has a cluster of eigenvalues λ_i close to one, i.e. $1 = \lambda_1 > \lambda_2 > \ldots > \lambda_{n_c} = 1 - \epsilon \gg \lambda_{n_c+1} \ldots$ [9]. Therefore, single eigenvectors are ill-conditioned, whereas the invariant subspace $\mathcal{X} = \mathrm{span}(g_1, \ldots, g_{n_c})$ is well conditioned. The matrix $\mathcal{A} \in \mathrm{R}^{n_c \times n_c}$ is some unknown non-singular transformation matrix. Every matrix G obtained by such a transformation of eigenvectors satisfies the *invariance condition* (9).

  Among all possible transformation matrices $\mathcal{A}$, we would like to find one that results in vectors $\mathbf{g}_J$ with special properties.

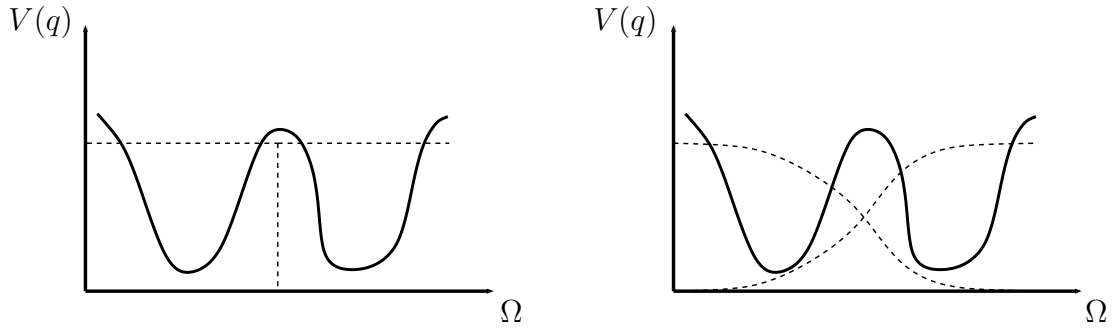

Figure 3: Left: A double welled potential (solid line), and the hard partition of the metastable configurations (dashed line). Right: Same double welled potential with a soft partition of metastabilties.

- The matrix G plays an important role, since each entry g_{iJ} of G relates the basis function χ_i to the metastable conformation C_J. In the original work [2], the conformations $\{C_1, ..., C_{n_c}\}$ are built as hard decomposition, such that the matrix G has only the entries 0 and 1. However correspondingly, to the definition of the hard decomposition, we can also define a decomposition *soft* if it meets the same conditions as a hard decomposition except for the fact, that

  ii*) $|\Omega_i \cap \Omega_j| \geq 0$ if $i \neq j$,

  i.e. we allow an overlap of the Ω_i.
  As a consequence the entries in the matrix G can take all values between 0 and 1 [3]. This allows us to assign for each basis function χ_i a certain degree of membership to each conformation. In other words, the ith row of G shows how much the ith basis function "contributes" to each metastability. In Figure 3 the difference between the soft and hard decomposition is shown.

Since every soft clustering can always been relaxed to a hard one, our goal is to find a soft partitioning of the position space. Thus, we have to find a non singular transformation matrix $\mathcal{A}$ such that G describes a soft partitioning. The computational details will be explained in the following section.

4 Clustering

The conditions on G discussed in the previous section can be summarized as follows:

(i) $G_{iJ} \geq 0 \quad \forall i \in \{1, \ldots, N\},\ J \in \{1, \ldots, n_c\}$ (positivity)

(ii) $\sum_{J=1}^{n_c} G_{iJ} = 1 \quad \forall i \in \{1, \ldots, N\}$ (partition of unity)

(iii) $G = X\mathcal{A}$ where $P^\tau X = X\Lambda$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{n_c})$, $\mathcal{A}$ non-singular (invariance)

Among the feasible transformation matrices we seek for a matrix $\mathcal{A}$ such that the resulting membership vectors $\mathbf{g}_J$ are as characteristic as possible ($|\Omega_i \cap \Omega_j| \approx 0$).

This can be achieved by maximizing the objective function

$$I(\mathcal{A}; X, \pi) = \frac{1}{n_c} \sum_{i=1}^{n_c} \frac{\langle \chi_i, \chi_i \rangle_\pi}{\langle \chi_i, \mathbf{e} \rangle_\pi} \leq 1, \tag{10}$$

where $\mathbf{e}$ denotes the vector with all entries equal to 1. One has to maximize a convex function with linear constraints. The optimization problem can be solved by the Nelder-Mead algorithm provided that a good initial guess for $\mathcal{A}$ is available. This starting guess is obtained by the *inner simplex algorithm* as described in [10]. The maximization of (10) subject to the constraints (i) to (iii) is called *Robust Perron Cluster Analysis* (PCCA+) [3, 7].

Once the membership vectors $\mathbf{g}_J$ have been computed, one can compute a coarse grained transition probability matrix P_c by projecting the original matrix P^τ onto the metastable conformations,

$$P_c = (G^\top w_D G)^{-1} G^\top w_D P^\tau G, \tag{11}$$

where w_D denotes a diagonal matrix with the stationary distribution w of P ($w^T P^\tau = w^T$)on the diagonal. The matrix P_c is not necessarily a stochastic matrix because it can have negative entries if the membership vector $\mathbf{g}_J$ are far from being characteristic. However, P_c has row sum one and is the correct propagator for densities restricted to the metastable conformations [6]. In fact, $\det(P_c)$ is a measurement for the metastability of the decomposition defined by G [11]. It holds

$$P_c = \mathcal{A}^{-1} \Lambda \mathcal{A}, \;\; \text{thus} \;\; \det(P_c) = \prod_{i=1}^{n_c} \lambda_i.$$

It has been shown that for any G satisfying (i) and (ii), $\det(P_c)$ can be bounded from above by $\prod_{i=1}^{n_c} \lambda_i$ [11]. Thus condition (iii) ensures a decomposition with maximal metastability.

Since the number of clusters n_c is unknown in advance, it is recommended to run the cluster algorithm several times with different input values for n_c and to choose the "best" solution for which $I(\mathcal{A}; X, \pi)$ is maximal.

5 Example

We demonstrate the application of our algorithm to the model system alanine dipeptide in vacuum with the mmff forcefield [5], Figure 4. For the discretization, we chose $N = 504$ molecular configurations from a high temperature (1000 Kelvin) molecular dynamics trajectory as defining nodes of our Voronoi basis functions $\{\chi_i\}_{i=1}^N$. As distance measure, we use the Euclidean distance in the space spanned by the four backbone torsion angles $\omega_1, \ldots, \omega_4$. We thus ignore variability in other degrees of freedom, which is justified by the fact that the torsion angles are the slow degrees of freedom and that is what we are interested in. Within every basis function, a Markov chain Monte Carlo method

Figure 4: *Left:* Spatial structure of alanin dipeptid. *Right:* Chemical structure of alanin dipeptide.

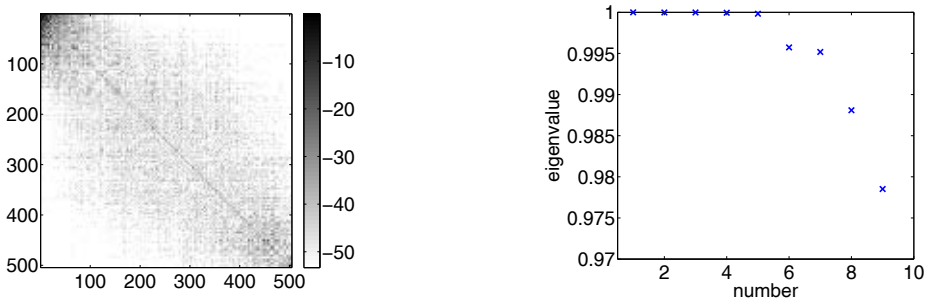

Figure 5: *Left:* Image of the 504×504 transition probability matrix P (for ease of visualization, we plotted the element-wise logarithm $\log(P)$ instead of P). *Right:* The first 10 eigenvalues of P. The first 5 eigenvalues form a cluster that is clearly separated from the rest of the spectrum.

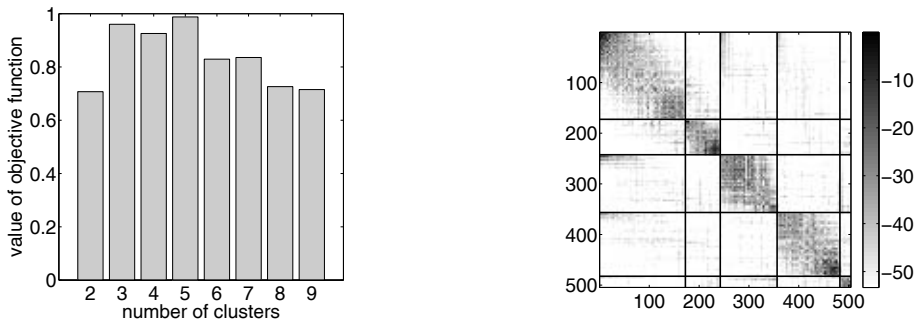

Figure 6: *Left:* The value of the objective function (10) for different numbers of clusters. The maximum value is achieved for $n_c = 5$ clusters. *Right:* Image of $\log(P)$ with rows and columns resorted according to the decomposition into 5 clusters.

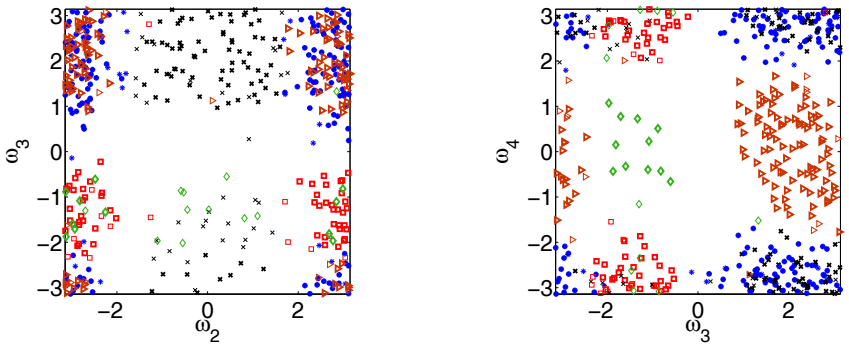

Figure 7: Coloring of the defining nodes of the Voronoi basis functions. Bold markers indicate that the basis function belongs to the cluster with probability larger than 0.8, whereas the other markers indicate memberships smaller than 0.8. *Left:* Torsion angles 2 and 3. *Right:* Torsion angles 3 and 4. It can be seen that the clusters identified by PCCA+ are isolated in at least one slow degree of freedom.

was applied to generate configurations distributed according to the partial densities $\pi_i(q)$. These configurations were propagated according to the flow Φ^τ with $\tau = 39$ femtoseconds. With these data we computed the entries of the transition probability matrix P^τ, which are visualized in Fig. 5. P^τ has a cluster of 5 eigenvalues close to one, and the value of the objective function (10) is also maximal for $n_c = 5$. Thus we computed the membership matrix G for $n_c = 5$ clusters. The metastability of this decomposition amounts to $\det(P_c) = 0.342$. For comparison, we calculated the relaxation of G towards a hard decomposition $\tilde{G}$, i.e.

$$\tilde{G}_{iJ} = \begin{cases} 1, & \text{if } J = \arg\max_j G_{ij} \\ 0, & \text{else} \end{cases} .$$

The metastability of the hard decomposition amounts to only 0.264. This hard decomposition, however, can be used to visualize the metastable conformations. Fig. 7 shows the nodes of the basis functions colored according to the final decomposition.

6 Conclusions

Starting from a discretization of the position space Ω we employed two different ways to describe the metastable configurations as a linear combination of eigenvectors. In the first method we took coefficients 0 or 1, and named these metastable conformations *hard*. In the second method we used *soft* metastable conformations by allowing the coefficients to take values between 0 and 1. We have employed an example molecule and compared the performance of the soft versus hard metastable conformations. In good agreement with our theory, the soft decomposition leads to a larger metastability than the hard decomposition.

REFERENCES

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

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

[3] P. Deuflhard and M. Weber. Robust Perron cluster analysis in conformation dynamics. *Lin. Alg. Appl.*, 398:161–184, 2005.

[4] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation*. Academic Press, 2 edition, 2002.

[5] T. Halgren and B. Nachbar. Merck molecular force field. iv. conformational energies and geometries for mmff94. *J. Comput. Chem.*, 17(5-6):587–615, 1996.

[6] S. Kube and M. Weber. A coarse graining method for the identification of transition rates between molecular conformations. *J. Chem. Phys.*, 126(2):0241203, 2007.

[7] S. Röblitz and M. Weber. Fuzzy spectral clustering by PCCA+. WIAS Report 26, WIAS Berlin, Berlin, 2009.

[8] M. Sarich, F. Noé, and Ch. Schütte. On the approximation quality of Markov state models. *Mult. Mod. Sim.*, 8(4):1154–1177, 2010.

[9] Ch. Schütte. *Conformational Dynamics: Modelling, Theory, Algorithm, and Application to Biomolecules*. Habilitation thesis, Department of Mathematics and Computer Science, Freie Universität Berlin, 1999.

[10] M. Weber and T. Galliat. Characterization of transition states in conformational dynamics using Fuzzy sets. ZIB-Report 02-12, Zuse Institute Berlin, 2002.

[11] Marcus Weber. *Meshless Methods in Conformation Dynamics*. PhD thesis, Freie Universität Berlin, 2006.