A Meshless Discretization Method for Markov State Models Applied to Explicit Water Peptide Folding Simulations

K. Fackeldey, A. Bujotzek, and M. Weber

Zuse Institute Berlin (ZIB), Takustraße 7, 14195 Berlin, Germany,
fackeldey@zib.de

Summary. Markov State Models (MSMs) are widely used to represent molecular conformational changes as jump-like transitions between subsets of the conformational state space. However, the simulation of peptide folding in explicit water is usually said to be unsuitable for the MSM framework. In this article, we summarize the theoretical background of MSMs and indicate that explicit water simulations do not contradict these principles. The algorithmic framework of a meshless conformational space discretization is applied to an explicit water system and the sampling results are compared to a long-term molecular dynamics trajectory. The meshless discretization approach is based on spectral clustering of stochastic matrices (MSMs) and allows for a parallelization of MD simulations. In our example of Trialanine we were able to compute the same distribution of a long term simulation in less computing time.

Key words: Molecular Dynamics, Conformation Dynamics, Robust Perron Cluster Analysis, Markovianity, Transfer Operator, Partition of Unity

1 Introduction

The dynamics of biomolecules is inherently multiscale since it involves time scales ranging from femtoseconds (bond-bond oscillations) to microseconds (protein folding, binding process). In the literature many methods can be found, which seek to capture this multiscale behavior. In molecular dynamics (MD) [1, 2] for instance, the integration step for the simulation is bounded to femtoseconds, which complicates the simulation of the above mentioned folding (e.g.[3]). On the other hand, coarse graining models [4, 5], which allow for larger or even unphysical time steps, suffer from the fact that the description may be too coarse and thus information is lost. In the last few years, much effort has been invested into combining the advantages of the coarse graining models with the ones of molecular dynamics. One approach is based on so-called Markov State Models (MSM) e.g.[36]. In particular, we mention

the works in the context of Folding at Home [35, 37], the works in the context of conformation dynamics [6, 7, 8] or approaches by other groups also using MSM e.g. [39, 38]. We remark that this list is far from being exhausive. In such models, the different conformational states of a molecular system and its transition probabilities between the states are incorporated, so that the "future" of the system over an arbitrary time span can be predicted. Thereby, it is assumed, that the system has a Markov property, which means that the probability of the system to switch to the next state depends on the current state only. In this context a state (or conformation) is an almost invariant subset in the phase space, where the system is not stable but almost stable (metastable). The performance of this method has been shown in various examples for molecular systems in vacuum. However, to find a MSM which represents the dynamics of a molecule simulated in (explicit) solvent is more difficult [40]. Explicit solvent can spoil the Markov property, such that the long-term behavior of the system is not reproduced correctly by the MSM [16, 14]. A classical MSM is a projection of the continuous dynamics to a low number of discrete (Markov) states, i.e. to a low number of subsets of the conformational space. In this article we want to fill this gap and present simulations of Trialanine in water. In the classical theory of Markov State Models [9, 10], the clusters, corresponding to the molecular conformations of the system, are represented by a collection of characteristic basis functions that yield 1 if the state belongs to the one conformation and zero otherwise. Here, we relax the condition, by softening the hard clustering [11]. More precisely, we allow a state to belong to more than one conformation and assign a degree of membership to each state. As a consequence, this soft clustering allows for a faithful representation of intermediate states, i.e. states which lie in transition regions between metastable conformations. These intermediate states are prevalent in simulation of biomolecules in water, where hydrogen bonds of the surrounding liquid can influence the stability of the molecule [12]. The existence of intermediate states allows for a correct representation of the long-term behavior of molecular systems including explicit solvent. This will be exemplified with Trialanine in water.

2 Basics of Conformation Dynamcis

Let us assume, that the dynamics of a molecular system is given by the Hamiltonian function:

$$H(q,p) = V(q) + K(p) \tag{1}$$

where $K(p)$ is the kinetic energy, which depends on the N generalized momenta $p = (p_1, ..., p_N) \in \mathbb{R}^{3N}$ and $V(q)$ is the potential energy depending on the N generalized positions $q = (q_1, ..., q_N) \in \mathbb{R}^{3N}$. The term $H(q,p)$ is denoted as the internal energy of the system $x = (q,p) \in \Gamma = \Omega \times \mathbb{R}^{3N}$, where Γ is the state space and $\Omega \subset \mathbb{R}^{3N}$ is the position space. We consider a canonical

ensemble, where the number of particles, the volume and the temperature is kept constant. According to the Boltzmann distribution, the positions q and momenta p of each atom in a molecule are then given by:

$$\mu(x) = \mu(q,p) \propto \exp(-\beta H(q,p)). \tag{2}$$

Here, $\beta = 1/k_B T$ is the inverse of temperature T times Boltzmann constant k_B. This canonical density can be split into a distribution of positions $\pi(q)$ and momenta $\eta(p)$, where

$$\pi(q) \propto \exp(-\beta V(q)) \text{ and } \eta(p) \propto \exp(-\beta K(p)). \tag{3}$$

We can now introduce the Hamiltonian equations of motions by

$$\dot{q} = \frac{\partial H}{\partial p} \quad , \qquad \dot{p} = -\frac{\partial H}{\partial q}. \tag{4}$$

For an initial state $x(0) = x_0$, the Hamiltonian flow ϕ^t over time span t is given by

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

where ϕ^t conserves the energy and is reversible. Moreover, the Hamiltonian flow conserves the energy, which implies that the Lebesgue measure on Γ is invariant under ϕ^t.

Let Π_q be the projection of the state (q,p) onto the position q, let Φ^τ be the discrete Hamiltonian flow for a time span τ. It is realized by a velocity verlet integrator with a constant time step and let further p be chosen randomly according to the distribution $\eta(p)$, then

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

describes a Markov process where the $i+1$th state depends on the preceding ith state only. Since $\mu(q,p)$ is symmetric in the sense $\mu(q,p) = \mu(q,-p)$ it can be shown, that the reduced density

$$\pi(q) = \int_{\mathbb{R}^{dN}} \mu(p,q)\, dp$$

is smooth, positive, finite and $\pi(q) = 1$, which allows us to define the transition operator [6]

$$T^\tau u(q) = \int_{\mathbb{R}^{dN}} u(\Pi \Phi^{-\tau}(q,p)) \eta(p) dp, \tag{5}$$

where $u(q)$ is a function on Ω. The operator in (5) can be explained as follows: We apply the Hamiltonian dynamcis backwards to a lag time τ and obtain $\Phi^{-\tau}(q,p)$, which is then projected by Π_q onto the state space. The integral then, averages over all possible initial momentum variables with given Boltzmann distribution η.

This operator on the weighted Hilbert space $L^2_\pi(\Omega)$, which is equipped with the inner product

$$\langle u, v\rangle_\pi := \int_\Omega u(q)v(q)\pi(q)dq\,,$$

is bounded, linear and self-adjoint. This allows us to compute the transition probability between two sets A and B by

$$p(A,B,t) = \frac{\int_A \chi_B(\Phi^\tau(x))\eta(p)dp}{\int_A \eta(p)dp} = \frac{\langle T\chi_A, \chi_B\rangle_\pi}{\langle \chi_A, \chi_B\rangle_\pi},$$

where χ_A ist the characteristic function being 1 in A and 0 otherwise. Based on this, we define a subset $B \subset \Omega$ as almost invariant, if

$$p(B,B,t) \approx 1.$$

and a metastable conformation as a function $C : \Omega \to [0,1]$ which is nearly invariant under the transfer operator T^τ, i.e.

$$T^\tau C(q) \approx C(q). \tag{6}$$

The fundamental idea behind this formulation is, that the transfer operator T^τ in (5) is a *linear* operator, although the ordinary differential equation (4) is (extremely) *non-linear*.

2.1 Discretization of the State Space

The linearization by the above introduced transfer operator 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 thus decompose the state space into metastable sets. More precisely, with respect to (6), we aim at a set $\{C_1, ..., C_{n_c}\}$ such that they form a partition of unity, i.e.

$$\sum_{J=1}^{n_c} C_J(q) = \mathbf{1}_\Omega \quad \forall q \in \Omega\,,$$

where $C_J : \Omega \to [0,1]$ is a function and $C_J(q) \geq 0 \quad \forall q \in \Omega$.

Of course, this decomposition is not given in advance, i.e. the metastable sets have to be detected. To do so, we discretize the state space by meshfree basis functions.

Let $\{x_i\}_{i=1}^N$ be a set of points from which we construct the meshfree basis functions by Shepard's approach [32] :

$$\varphi_i(q) = \frac{\exp(-\alpha\|q - x_i\|)}{\sum_{i=1}^N \exp(-\alpha\|q - x_i\|)}$$

such that

$$\forall q \in \Omega : \quad \sum_{i=1}^{N} \varphi_i(q) = 1 \text{ and } \bigcup_{i=1}^{N} \text{supp}(\varphi_i) = \Omega \,.$$

Finally, this allows us to represent each C_J as a linear combination of the basis functions $\{\varphi_i\}_{i=1}^N$, so that

$$C_J(q) = \sum_{i=1}^{N} G_{iJ}\varphi_i(q), \qquad J = 1, ..., n_c, \tag{7}$$

where each entry G_{iJ} of the matrix $G \in \mathbb{R}^{N\times n_c}$ represents the contribution of the ith basis function to the conformation C_J. By employing a Galerkin discretization of (5), we obtain the stochastic matrix

$$P_{ji}^{\tau} = \frac{\langle \varphi_j, T^{\tau}\varphi_i\rangle_{\pi}}{\langle \varphi_j, \varphi_j\rangle_{\pi}} = \int_{\Omega} T^{\tau}\varphi_i(q) \frac{\varphi_j(q)}{\int_{\Omega} \varphi_j(q)\pi(q)dq} dq \tag{8}$$

Taking advantage of the partition of unity property of the set φ_i, we can then localize the global quantities. More precisely, associated with each φ_i, a partial density is given by

$$\pi_i(q) = \frac{\varphi_i(q)\pi(q)}{\int_{\Omega} \varphi_i(q)\pi(q)}. \tag{9}$$

Note that the term in the denominator of (9) represents the localized thermodynamical weight. Analogous to (6), we then obtain for the coefficients G_{iJ}

$$P^{\tau}\mathbf{g}_J \approx \mathbf{g}_J, \tag{10}$$

where $\mathbf{g}_J = [G_{1J}, G_{2J}, \ldots, G_{NJ}]^T$.

From the knowledge of the transition matrix P^{τ} we can identify the metastable states as a linear combination of the meshfree basis functions. This fact is based on the Frobenius Perron theory, which we first introduce for decoupled Markov states in the following section.

2.2 Decoupled Markov states

In the case of a completely decoupled Markov State Model, we do not have metastable states, but n_c stable states. The transition matrix P can be rearranged by permutations such that it has a block diagonal structure, i.e.

$$P = \begin{pmatrix} P_1 & & & \\ & P_2 & & \epsilon \\ \epsilon & & P_3 & \\ & & & \ddots \end{pmatrix} \tag{11}$$

where $\varepsilon = 0$. Each submatrix P_I, $I = 1, ..., n_c$ is associated with a right eigenvector u_I and an eigenvalue λ_I, such that $P_I u_I = \lambda_I u_I$. This vector can be trivially embedded such that it is also an eigenvector of the matrix P. For a row stochastic matrix, i.e. $\sum_j P_{ij} = 1$, we can employ the Frobenius - Perron theory, which states that its largest eigenvalue is 1, the corresponding left eigenvector gives the stationary distribution and the elements of the right eigenvector are all identical. Consequently, the matrix P has n_c eigenvalues equal to 1. Thus, the dominant right eigenvectors u_I have the form

$$u_I = (c_1, ..., c_1, c_2, ..., c_2, c_3, ..., c_3, ..., c_{n_c}, ..., c_{n_c}).$$

Summing up, in a decoupled Markov process, the subspace of the dominant eigenvalue 1 is spanned by n_c piecewise constant eigenvectors. Thus, the metastable states can be identified by the identical entries in the right eigenvectors. Or, alternatively, a linear transformation of the eigenvector basis provides a set of n_c characteristic vectors, such that for the vector u_I corresponding to the I-th cluster all elements are zero except for $c_I = 1$.

2.3 Nearly Decoupled Markov states & Water

In the following, the Markov process is only nearly decoupled, i.e., $\varepsilon > 0$. At the beginning, let us restate the results of the Frobenius-Perron theory from the foregoing section by rewriting G as

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

where X represents the eigenvectors of P^τ corresponding to eigenvalues close to one, and the non-singular, still unknown transformation matrix $\mathcal{A} \in \mathbb{R}^{n_c \times n_c}$.

Due to the definition of the conformations and the basis functions, we can state the following constraints:

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)

The transformation matrix $\mathcal{A}$ is not uniquely defined by these constraints. Given the eigenvectors of P^τ and a transformation matrix $\mathcal{A}$, the matrix elements G_{iJ} are computed according to (iii). The conformations C_I are then constructed according to (7). In order to yield an "optimal" transformation matrix, a corresponding optimization problem is formulated. One possibility is to maximize the "crispness" of the conformations (i.e. $\frac{\langle C_I, C_I \rangle_\pi}{\langle C_I, \mathbf{e} \rangle_\pi}$ close to one), which can be achieved by maximizing

$$I(\mathcal{A}; X, \pi) = \frac{1}{n_c} \sum_{I=1}^{n_c} \frac{\langle C_I, C_I \rangle_\pi}{\langle C_I, \mathbf{e} \rangle_\pi} \leq 1, \tag{12}$$

where $\mathbf{e}$ denotes the constant function equal to 1, cf. [13]. PCCA+ solves the maximization problem (12) subject to the linear constraints (i) through (iii). Given a good starting guess, this convex global maximization problem can be solved by local maximization methods as described in [8, 11].

By PCCA+, the concept of metastable sets is replaced by the concept of metastable membership functions. A conformation of the molecular system is given by a membership function, which may also attain values inside the interval $[0, 1]$. Although this approach seems to be of technical interest only, it is the key to formulate Markov State Models for molecular systems including explicit water molecules. Usually, Markov State Models are said to be invalid in the case of explicit water simulations, because

1) transitions between molecular conformations can not be modeled as jump-processes, due to the fact that there is a diffusive part in the system. The system is assumed to be non-Markovian.
2) the dynamics of the system is highly non-linear and can not be captured by a linear matrix P.
3) a spectral gap of T^τ for identifying the number of metastable conformations does not exist.

Ad 1): A jump inside a Markov State Model, which is based on membership functions, is not a jump between two subsets of the conformational space. Diffusive processes can be modeled correctly by allowing for a "soft" decomposition of the conformational space, cf. [14].
Ad 2): Non-linear (stochastic) dynamics corresponds to a linear Fokker-Planck-operator. The main problem is given by the projection error of this linear operator to a low-dimensional transition matrix P. In principle, a correct projection is possible. However, this projection has to include all relevant degrees of freedom which may also include the "conformational state" of the water molecules. Note that only degrees of freedom which are necessary to *distinguish* between different conformations of the system are relevant.
Ad 3): In explicit water simulations, the timescales of conformational transitions are not well-separated from the time scales of "internal fluctuations", such that T^τ does not provide a spectral gap. In this situation, the projection error of set-based Markov State Models might be high, because the spectral gap plays an important role in its error estimation, cf. [15]. However, if continuous membership functions are allowed, it is only mandatory to identify an invariant subspace of T^τ, the spectral gap does not enter in the error estimator directly, cf. [16, 14]. Additionally, the spectral gap is not the only possibility to identify the number of clusters in PCCA+, cf. [17].

3 Simulation of Trialanine in an Explicit Solvent

The structure of Trialanine at a pH value of 1 was created by using the visualization software Amira [18]. All molecular dynamics simulations were

performed using the molecular dynamics software package GROMACS 4.5 [30]. Trialanine was parameterized according to the Amber-99SB force field [19]. As protonated C-terminal alanine is not part of the standard force field, we created and parameterized a novel residue in compliance with the Amber-99SB force field by using the software Antechamber from AmberTools 1.5 [20, 21, 22], with charges calculated by the AM1-BCC method. [23, 24]. For the explicit solvent system, we used the TIP4P-Ew water model [25]. Trialanine was placed in a rhombic dodecahedron solvent box of about 4.2 nm side length. In order to neutralize the overall charge, we placed a single negative counter ion in the box. The energy of the system was minimized by using the steepest descent algorithm. Afterwards, a 200 ps simulation was performed during which the position of all heavy atoms of Trialanine was restrained in order to settle the solvent molecules. The output of this run was used as starting configuration for a molecular dynamics run of 100 ns. In order to maintain a constant temperature of 300 K and a pressure of 1 bar, velocity rescaling [26] and Berendsen weak coupling [27] were applied. A twin range cut-off of 1.0/1.4 nm for van-der-Waals interactions was applied and the smooth particle mesh Ewald algorithm [28] was used for Coulomb interactions, with a switching distance of 1.0 nm. Bond length oscillations of bonds involving hydrogen atoms were contrained using the LINCS algorithm [29], allowing for an integration step of 1 fs.

For the ZIBgridfree sampling, the same parameter settings were adopted, save molecular dynamics run time, which was set to 500 ps per node. The ZIBgridfree algorithm uses short local sampling in order to evaluate the partial densities associated with the n basis functions φ_i. In order to ascertain thorough local sampling, each local sampling is confined to the essential support of the corresponding basis function. In this process, the short time trajectories can be parallelized on the force field level (such as provided by state-of-the-art molecular dynamics code). A second level of parallelization is given by the possibility to perform multiple local samplings at the same time, as the basis functions can be evaluated independent of each other. The ZIBgridfree sampling algorithm was implemented using a Python framework built around the GROMACS molecular dynamics code (publication in preparation). In ZIBgridfree, the conformational space discretization is based on internal degrees of freedom. Only degrees of freedom which indicate conformational changes are relevant. Here we follow the work of Mu et al. [31] by taking the two central dihedral angles Φ and Ψ indicated in Figure 1. In Figure 2, we plotted the histograms of the two dihedrals in explicit water simulations. The ZIBgridfree simulation shows good agreement with the 100 ns long term molecular dynamics trajectory. It can also be found in literature [33, 31, 34], that the P_{II} (polyglycine II) conformation is predominant in explicit water. The mean dihedral angles and the statistical weights of the four conformations calculated by ZIBgridfree are given in Table 1. However, as observed by [31], the choice of the force field has a dramatic influence on the Boltzmann distribution of the states and thus, a comparison with results from simulations with differ-

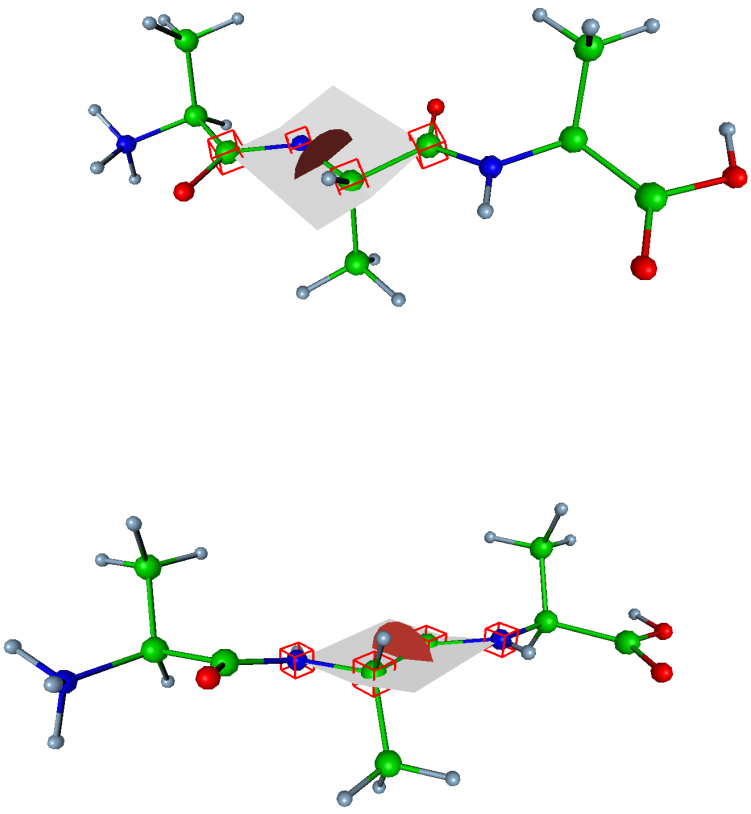

Figure 1. Top: dihedral angle Φ. Bottom: dihedral angle Ψ.

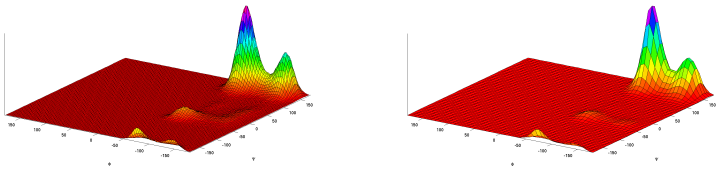

Figure 2. Left: Trialanine in explicit water, computed by molecular dynamics (sampling time 100 ns), Right: Trialanine in explicit water, computed by ZIBgridfree (joint sampling time 10 ns).

Table 1. Mean dihedral angles and statistical weights of the conformations.

conformation	(Φ, Ψ)	weight
helical	($\approx$ -143, $\approx$ -18)	0.0183
α_R	($\approx$ -77, $\approx$ -40)	0.0583
β	($\approx$ -143, $\approx$ 158)	0.3199
P_{II}	($\approx$ -77, $\approx$ 151)	0.6034

ent force fields is hardly possible. The four identified conformations are given in Figure 3 and 4. Since we have applied Markov State Models and continuous membership functions in order to identify the conformations, they are given by a family of microstates. Hence, conformations are best represented by density clouds, and not by single representatives. What is the difference

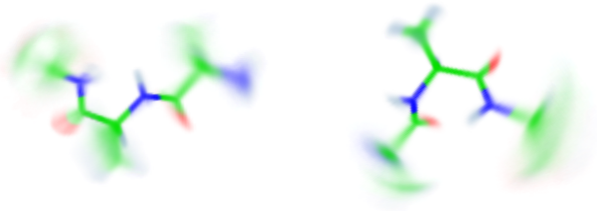

Figure 3. Left: Helical structure of Trialanine, cluster with the lowest weight. Right: α_R structure of Trialanine, cp. with Table 1. The smeared-out microstates suggest that the conformations are represented by density clouds.

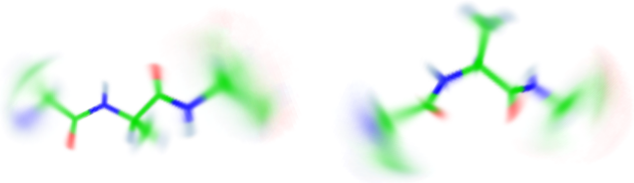

Figure 4. Left: β structure of trialanine. Right: P_{II} structure of trialanine, cluster with the largest weight.

between vacuum and explicit water simulations for the given molecular system? In contrast to vacuum simulations, we observe that the conformations β and P_{II} overlap. Seemingly, the transition between these two conformations is not jump-like. This statement is true if we try to define subsets of the conformational space identified as β or as P_{II}. A Markov State Model based on

subsets would lead to a systematic error (and probably to a wrong equilibrium density). However, in the presented approach, subsets are replaced by membership functions. The "Markovian jump" does not take place between two subsets, it takes place between two *overlapping* densities, see also Figure 5.

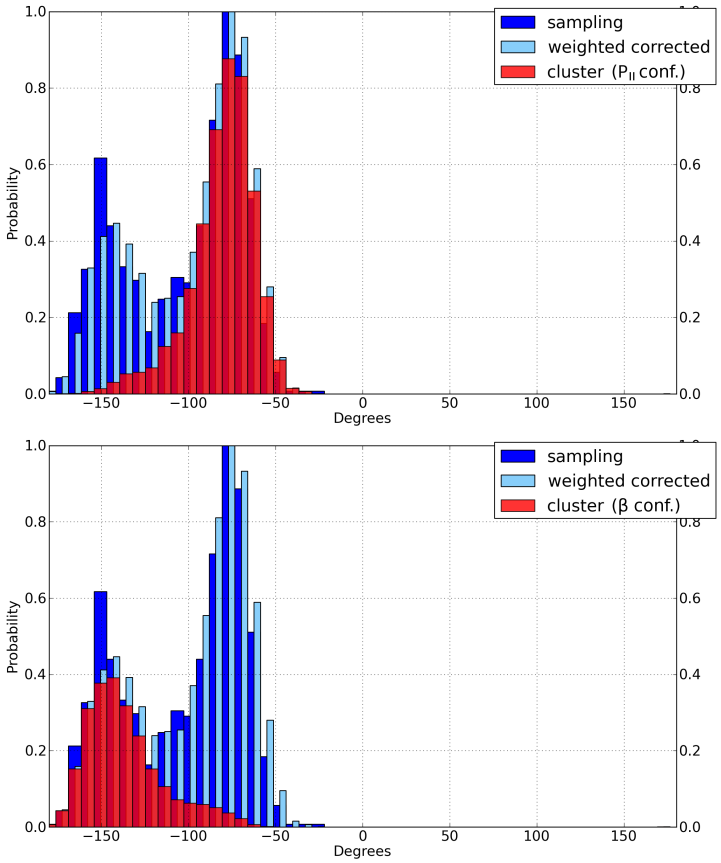

Figure 5. Distribution of torsion angle Φ. Dark blue histogram: Distribution obtained by long term MD (sampling time 100 ns). Light blue histogram: Joint reweighted distribution obtained by ZIBgridfree (joint sampling time 10 ns). Red histogram top: Density corresponding to Φ-part of the P_{II} conformation. Red histogram bottom: Density corresponding to Φ-part of the β conformation..

For the given example, the spectral gap ($\lambda_4 = 0.97$ and $\lambda_5 = 0.94$) for the calculation of C_I via PCCA+ is not significant as it has been expected for the analysis of an explicit water simulation. Since the memebership functions allow for each state a certain "degree of memebership", we call these kind

of clusters *soft*. The MSM for those membership functions (conformations according to Table 1) is given by

$$P_{\text{soft}}(1ps) = \begin{pmatrix} 0.981 & 0.015 & 0.002 & 0.002 \\ 0.005 & 0.995 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.982 & 0.018 \\ 0.000 & 0.000 & 0.009 & 0.991 \end{pmatrix},$$

its spectrum exactly reproduces the eigenvalues of the Galerkin discretization of the transfer operator. The construction of a set-based MSM (*hard cluster*) leads to

$$P_{\text{hard}}(1ps) = \begin{pmatrix} 0.981 & 0.017 & 0.001 & 0.001 \\ 0.007 & 0.993 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.964 & 0.036 \\ 0.000 & 0.000 & 0.017 & 0.983 \end{pmatrix}.$$

Although the difference is small, this matrix does not reproduce the correct spectrum. Especially the transition probabilities between the overlapping conformations are overestimated. This can be seen in the following way. If we compute the diagonal elements of a transition matrix P_{ij} by using $\langle \cdot, \cdot \rangle_\pi$-normalized eigenfunctions ϕ_i in equation (8) instead of characteristic functions, then (according to the Rayleigh-Ritz principle) these elements are maximized (see also [Huisinga-Diss]). In fact, PCCA+ provides a linear combination of approximations of eigenfunctions of T^τ. Thus, the trace of $P(1ps)$ is lower, if we restrict the corresponding approximation space to characteristic functions. This means, the outer diagonal elements are higher in the set-based approach. Therefore, the transition probabilities are overestimated if we do not account correctly for the overlapping conformations.

4 Conclusions

The discretization of the phase space by meshfree basis functions allows us to localize the densities and thus to compute them by short MD runs which are confined to the support of the meshfree basis functions. This enables us to run several MD trajectories in parallel, allowing for parallelization on both force field and basis function level. After these short localized MD simulations, the independent localized densities are reweighted according to the Boltzmann distribution. In our example we observed that this sampling scheme can be applied to small molecular systems in explicit water. Additionally, the theory of MSMs is not restricted to vacuum simulations and allows for the analysis of spatially overlapping conformations. In our future work, we plan to investigate MSMs applied to large molecules (such as proteins) in explicit water.

References

1. M. Griebel, S. Knapek and G. Zumbusch, *Numerical Simulation in Molecular Dynamics*, Springer, 2007
2. A. R. Leach, *Molecular Modelling: Principles and Applications* , Prentice Hall, Harlow, 2001
3. W. F. van Gunsteren and H. J. C. Berendsen, *Algorithms for macromolecular dynmaics and constraint dynamics*, Mol. Phys. **34** (1977), 1311–1327.
4. H. Gohlke and M. F. Thorpe, *A Natural Coarse Graining for Simulating Large Biomolecular Motion*, Biophys. J. **91**, (6),2115–2120, 2006
5. C. Chennubhotla and I. Bahar *Markov methods for hierarchical coarse-graining of large protein dynamics* J. Comp. Bio, **14**(6),765–776, 2007
6. C. Schütte, *Conformational Dynamics: Modelling, Theory, Algorithm, and Application to Biomolecules*. Habilitation Thesis, Fachbereich Mathematik und Informatik, Freie Universität Berlin, 1999.
7. 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 Comp. Biophysics, Academic Press, **151**,146–169, 1999.
8. 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, 2005.
9. J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Model. Simul.*, 5(4):1214–1226, 2006.
10. W. C. Swope and J. W. Pitera. Describing protein folding kinetics by molecular dynamics simulation. 1. Theory. *J. Phys. Chem. B*, 108:6571–6581, 2004.
11. M. Weber. *Meshless Methods in Conformation Dynamics*. Doctoral thesis, Department of Mathematics and Computer Science, Freie Universität Berlin, 2006. Published by Verlag Dr. Hut, München.
12. G. Rose, P. Fleming, J. Banavar, and A. Maritan. *A backbone-based theory of protein folding*. Proc. Natl. Acad. Sci. U.S.A., 103(45): 16623-33, 2006.
13. S. Röblitz. *Statistical Error Estimation and Grid-free Hierarchical Refinement in Conformation Dynamics*. Doctoral thesis, Department of Mathematics and Computer Science, Freie Universität Berlin, 2008.
14. M. Weber, 2011: *A Subspace Approach to Molecular Markov State Models via a New Infinitesimal Generator*. Habilitation Thesis, Fachbereich Mathematik und Informatik, Freie Universität Berlin.
15. W. Huisinga. *Metastability of Markovian systems: A transfer operator based approach in application to molecular dynamics*, Freie Universität Berlin, Doctoral Thesis, 2001
16. M. Sarich, F. Noe and C. Schütte, *On the Approximation Quality of Markov State Models*. Mult. Mod. Sim. **8**, 1154 – 1177, 2010
17. M. Weber, W. Rungsarityotin und A. Schliep, *An Indicator for the Number of Clusters Using a Linear Map to Simplex Structure*, 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, M. Spiliopoulou et al (eds.) Studies in Classification, Data Analysis, and Knowledge Organization 103–110, Springer2006,

18. D. Stalling, M. Westerhoff and H .-Ch. Hege. *Amira: A Highly Interactive System for Visual Data Analysis*, in: Hansen, Ch. D.; Johnson, Ch. R.; (eds.), The Visualization Handbook, Chapter 38, 749- 767, Elsevier 2005.
19. V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling. *Comparison of multiple Amber force fields and development of improved protein backbone parameters*. Proteins 2006, **65**, 712 – 725.
20. D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K.M. Jr. Merz, A. Onufriev, C. Simmerling, B. Wang, R. J. Woods, *The Amber biomolecular simulation programs*. J. Comput. Chem. 2005, **26**, 1668 – 1688
21. J. Wang, W. Wang,J. Caldwell, P. A. Kollman, D. A. Case, *Development and testing of a general Amber force field*. Comput. Chem. , **25**, 1157 – 1174, 2004
22. J. Wang, W. Wang, P. A. Kollman, D. A. Case,*Automatic atom type and bond type perception in molecular mechanical calculations*. J. Mol. Graphics Model. , **25**, 247 – 260, 2006
23. A. Jakalian, B. L. Bush, D. B. Jack, C. I. Bayly, *Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method*. J. Comput. Chem. , **21**, 132 - 146, 2000
24. A. Jakalian, D. B. Jack, C. I. Bayly. *Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation*. J. Comput. Chem. , 23, 1623 – 1641, 2002
25. H. W. Horn, W. C. Swope, J. W. Pitera. *Characterization of the TIP4P-Ew water model: Vapor pressure and boiling point*. J. Chem. Phys., 123, 194504, 2005
26. G. Bussi, D. Donadio, M. Parrinello. *Canonical sampling through velocity rescaling*. J. Chem. Phys. , 126, 014,101, 2007
27. H. J. C. Berendsen, J. P. M. Postma, A. DiNola,J. R. Haak.*Molecular dynamics with coupling to an external bath*. J. Chem. Phys. , 81, 3684 – 3690, 1984
28. U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen. *A smooth particle mesh Ewald method*. J. Chem. Phys. , 103, 8577 – 8592, 1995
29. B. Hess and H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije. *LINCS: A linear constraint solver for molecular simulations*. J. Comput. Chem. , 18, 1463 – 1472, 1997
30. B. Hess, C. Kutzner, D. van der Spoel, E. Indahl,*GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation*. J. Chem. Theory Comput. , 4(3), 435 – 447, 2008
31. Y. Mu, D. Kosov and G. Stock *Conformational Dynamics of Trialanine in Water.2 Comparison of AMBER, CHARMM, GROMOS, and OPLS Force Fields to NMR and Infrared Experiments* J. Phys. Chem. B **107**, 5064– 5073, 2003
32. D. Shepard *A two-dimensional interpolation function for irregularly spaced data* Proceeding of the 1968 23rd ACM national conference, 517 – 524 , 1968
33. Y. Mu and G. Stock *Conformational Dynamics of Trialanine in Water: A Molecular Dynamcis Study* J. Phys. Chem. B **106**, 5294– 5301, 2002
34. O. Engin, M. Sayar and B. Erman *The introduction of hydrogen bond and hydrophobicity effects into the rotational isomeric states model for conformational analysis of unfolded peptides* Phys. Biol. **6** 016001, 2009

35. G. R. Bowman, X. Huang and V. S. Pande *Using generalized ensemble simulations and Markov state models to identify conformational states* Methods **49**, 2 p. 197-201, 2009
36. J. -H. Prinz, H. Wu, M. Sarich, B. Keller, M. Fischbach, M. Held, J. D. Chodera, C. Schütte and F. Noe, *Markov models of molecular kinetics: Generation and validation* J. Chem. Phys., **134** :174105 , 2011
37. A. Beberg and V. S. Pande, *Folding@home: lessons from eight years of distributed computing* IEEE Int. Symp. on Parallel & Distributed Processing, 1-8 (2009)
38. A. Pan and B. Roux *Building Markov state models along pathways to determine free energies and rates of transitions* J Chem Phys. 2008;**129**(6):064107.
39. S. Sriraman, I. Kevrekidis and G. Hummer *Coarse master equation from Bayesian analysis of replica molecular dynamics simulations* J Phys Chem B. 2005; **109**(14):6479 – 84.
40. John D. Chodera, William C. Swope, Jed W. Pitera, and Ken A. Dill *Long-Time Protein Folding Dynamics from Short-Time Molecular Dynamics Simulations* Multiscale Model. Simul. 5, pp. 1214–1226