7 Application to Large Systems

In this section we want to demonstrate that the algorithmic strategy presented in Section 3.2 can be applied to identify biomolecular conformations even for large systems as, for instance, small biomolecules with hundreds of atoms. For large systems, we have to face two particular problems:

1. How to approximate the stationary distribution in a high–dimensional space?

2. How to decompose the high–dimensional state space in order to discretize the propagator?

We will address these problems in the following.

7.1 Monte Carlo methods

The typical approach to sample the canonical distribution in a high–dimensional space is via Monte Carlo techniques. There is an extremely rich and varied literature on Monte Carlo methods (see, e.g., [44, 75]) and every converging method would allow to realize the problem of sampling the invariant distribution. In addition, we may also apply molecular dynamics based techniques, e.g., constant temperature sampling of the canonical distribution [54, 2]. It is widely known, that Monte Carlo simulations may suffer from possible "trapping problems" [48]. This kind of problem occurs when the Monte Carlo Markov process gets trapped near a local potential energy minimum due to high energy barriers so that a proper sampling of the entire state space within reasonable computing times is impossible. As illustrated in [68], this phenomenon is related to the existence of metastable subsets for the Monte Carlo Markov process.

There exists various strategies addressing the trapping problem. Especially the so–called extended ensemble methods, which are based on reweighting techniques, are gaining significant popularity. Recently, Fischer presented a promissing alternative approach, the uncoupling–coupling Monte Carlo (UCMC) technique [20]. It links Monte Carlo sampling methods with the algorithmic strategy to the identification of metastable subsets, as described in this thesis. Loosely speaking, it exploits a clustering of the state space, when metastability starts to become "visible" but is far from causing trapping problems. As outline in [22] this may allow to sample the canonical distribution with reasonable computational effort.

7.2 Adaptive Discretization Techniques

There are different ways of facing the second problem, the decomposition of the state space. We assume that the canonical distribution has properly

been sampled by some Monte Carlo method. Then the following possibilities arise.

Essential Degrees of Freedom. Typical biomolecular systems contain hundreds or thousands of atoms. If we would generate a decomposition of the state space by simply decomposing every degree of freedom, the number of subsets within the decomposition and thus the dimension of the stochastic transition matrix would grow exponentially with the size of the molecular system. Chemical insight into biomolecular systems allows to circumvent this "curse of dimensionality". Conformations of biomolecules are mostly described in terms of a few *essential degrees of freedom*. In the subspace of essential degrees of freedom most of the positional fluctuations occur, while in the remaining degrees of freedom the motion can be considered as "physically constrained". *Based on the sampling of the canonical distribution*, we may determine essential degrees of freedom either in the position space according to Amadei et al. [1] or in the space of internal degrees of freedom, e.g., dihedral angles [35], by statistical analysis of circular data. Either case is based on a principal component analysis of the sampling via analyzing a covariance matrix. As shown in [35], this procedure may results in a tremendous reduction of the number of degrees of freedom and, consequently, in a moderate number of subsets within the decomposition when discretizing the essential variables only. The principal component analysis is a linear approach to essential degrees of freedom. A characterization and identification of more general nonlinear essential degrees of freedom is subject to investigations within a current research project [70] and part of a current diploma thesis [78].

Self–Organizing Maps. An alternative approach is to discretize the propagator by means of self–organizing maps, a special kind of neural networks. Self–organizing maps allow to cluster the Monte Carlo sampling data by assigning each sampling point to the nearest neurons, each of them representing a subset of the decomposition. We have demonstrated its successful application to sampling data of biomolecular systems in [28]. More advanced extensions, such as "box–neurons" and a hierarchical embedding, have recently be designed [26, 27].

Clustering Algorithms. A third approach of decomposing the state space is based on clustering the sampling data by means of clustering algorithms (see, e.g., [37] and cited references). These methods cluster according to structural similarity: The set of sampling points is partitioned into disjoint subsets with the property that two states belonging to the same subset are in some sense structural closer to each other than two states belonging to different subsets. A crucial question is the design of appropriate measures

of structural similarity. In the biomolecular application context, these measures can either be based on the Cartesian coordinates of the molecules or on the internal degrees of freedom. In contrast to the former the latter approach is invariant under rotations and translations of the entire molecule. For an application to biomolecular systems see [35].

Solving the Eigenvalue Problem. Finally, we want to remark that although the stochastic transition matrix resulting from the discretization may be quite large, it turns out to be sparse in our application context. Furthermore, since the algorithmic strategy is based solely on the dominant eigenvalues and corresponding eigenfunctions, we can apply subspace oriented iterative techniques (see, e.g., [64]) to solve the eigenvalue problem. It is important to notice that the convergence rate of those methods depends only on the spectral gap between the cluster of dominant eigenvalues and the remaining part of the spectrum and is *independent of the size of the stochastic transition matrix* and hence of the number of discretization subsets.

However, it should be clear that any refinement process of the discretization is limited by the quality of the underlying sampling data, since the approximation quality of the stochastic transition matrix is based on the interplay between sampling data and fineness of the discretization (see Sec. 5.3).

7.3 Analyzing a Small Biomolecule

This section illustrates the performance of the algorithmic approach to the triribonucleotide adenylyl(3'-5')cytidylyl(3'-5')cytidin (r(ACC)) model system in vacuum, see Figure 11. Its physical representation is based on the GROMOS96 extended atom force field [77], resulting in $N = 70$ atoms, hence $\Omega = \mathbf{R}^{210}$ and $\Gamma = \mathbf{R}^{420}$. The internal fluctuations are modeled w.r.t. the Hamiltonian system with randomized momenta. For details see [35].

The sampling of the canonical distribution was generated using an adaptive temperature hybrid Monte Carlo[15] (ATHMC) method [21] at $T = 300$K resulting in the sampling sequence $q_1, \ldots, q_{32000} \in \Omega$. The dynamical fluctuations within the canonical ensemble were approximated by integrating $M = 4$ short trajectories of length $\tau = 80$fs starting from each sampling point. To facilitate transitions, analogous to the ATHMC sampling, the momenta were chosen according to the momenta distribution $\mathcal{P}(p)$ corresponding to four different temperatures between $300K - 400$K and reweighted afterwards. This resulted in a total of $4 \times 32.000 = 128.000$ transitions.

The configurational space was discretized using all four essential degrees of freedom, which were identified by means of a statistical analysis of the sampling data (see Sec. 7.2), resulting in $d = 36$ discretization subsets.

[15] ATHMC is part of the earlier mentioned UCMC method (see Section 7.1).

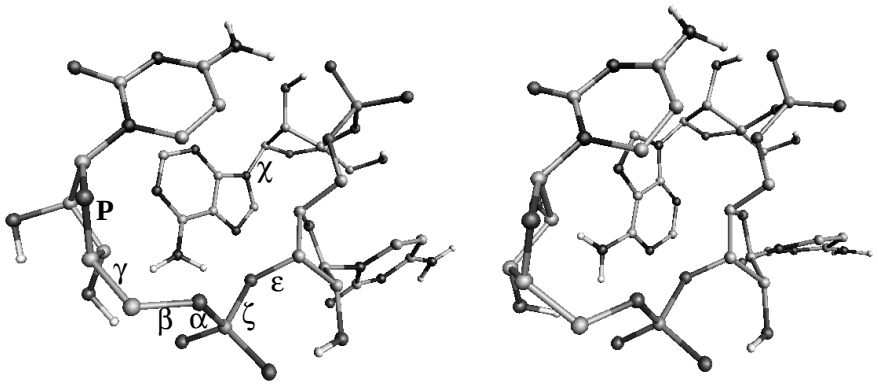

Figure 11: Two representatives of different conformations of r(ACC). Left: The χ angle around the first glycosidic bond is in *anti* position (-175 degrees) and the terminal ribose pucker P is in C(3')endo C(2')exo conformation. Right: The χ angle is in *syn* position (19 degrees) and the terminal ribose in C(2')endo C(3')exo conformation. Visualization by amira [42].

Then the 36×36 stochastic transition matrix S was computed based on the 128.000 transitions taking the different weighting factors into account. The computation of the eigenvalues of S close to 1 yielded a cluster of eight eigenvalues with a significant gap to the remaining part of the spectrum, as shown in the following table:

k	1	2	3	4	5	6	7	8	9	...
λ_k	1.00	0.99	0.98	0.97	0.96	0.95	0.93	0.90	0.81	...

Finally, we computed conformations based on the corresponding eight eigenvectors of S via the identification algorithm presented in Section 5.4. We identified eight conformations; their statistical weights and metastabilities are shown in the following table:

conformations	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
statistical weight	0.11	0.01	0.12	0.03	0.32	0.04	0.29	0.10
metastability	0.99	0.94	0.96	0.89	0.99	0.95	0.98	0.96

The transition probabilities between the different conformations are visualized schematically in Figure 12. The matrix allows to define a hierarchy between the clusters, which is inherent to the algorithm. On the top level, there are two clusters, one consisting of the conformations $C_1, \ldots, C_4$ and the other consisting of the conformations $C_5, \ldots, C_8$. This structure corresponds to the two 4×4 blocks on the diagonal. On the next level, each

of these clusters splits up into two subclusters yielding four conformations $\{C_1, C_2\}$, $\{C_3, C_4\}$, $\{C_5, C_6\}$, $\{C_7, C_8\}$. On the bottom level, each cluster is further divided resulting in eight conformations.

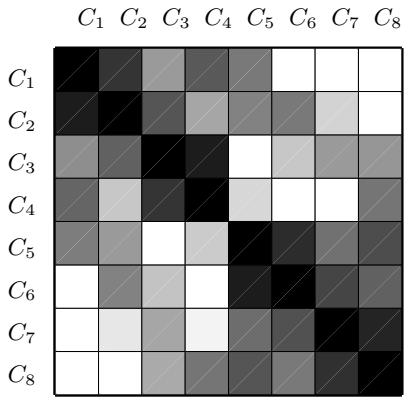

Figure 12: Schematical visualization of the transition probabilities $p(\tau, C_i, C_j)$ between the conformation C_i (row) and C_j (column). The colors are chosen according to the logarithm of the corresponding entries: from $p \approx 0$ (light) to $p \approx 1$ (dark).