Adaptive spectral clustering for conformation analysis

Fiete Haack*, Susanna Röblitz†, Olga Scharkoi†, Burkhard Schmidt** and Marcus Weber†

*Institut für Informatik, Universität Rostock Albert-Einstein-Str. 21, D-18059 Rostock, Germany
†Zuse Institut Berlin,Takustraße 7, D-14195 Berlin, Germany
**Institut für Mathematik, Freie Universität Berlin, Arnimallee 6, D-14195 Berlin, Germany

Abstract. Markov state models have become very popular for the description of conformation dynamics of molecules over long timescales. The construction of such models requires a partitioning of the configuration space such that the discretization can serve as an approximation of metastable conformations. Since the computational complexity for the construction of a Markov state model increases quadratically with the number of sets, it is desirable to obtain as few sets as necessary. In this paper we propose an algorithm for the adaptive refinement of an initial coarse partitioning. A spectral clustering method is applied to the final partitioning to detect the metastable conformations. We apply this method to the conformation analysis of a model tri-peptide molecule, where metastable β- and γ-turn conformations can be identified.

Keywords: molecular dynamics simulations, peptides, metastable conformation, spectral clustering
PACS: 36.20.Ey, 87.14.ef, 87.15.ap

INTRODUCTION

Molecular dynamical systems are often characterized by the existence of metastable conformations – subsets of molecular configurations where the dynamical system spends a very long time before it rapidly switches to another conformation. This set based point of view of metastable conformations differs from the classical definition of conformations as minima of the energy landscape because it also takes into account entropic barriers. Usually, there exist considerably more energy minima than metastable conformations. Multiple minima can well belong to one metastable conformation if there are frequent transitions between these minima. The identification of metastable conformations together with their life times and transition patterns is essential for the analysis of the system's long term behavior. Initiated by the pioneering work of Dellnitz, Deuflhard, and Schütte [1, 2, 3, 4], a multi-scale method, called *conformation dynamics*, has been developed. Its main objective is the identification of metastable conformations together with their life times and transition patterns. In this mixed deterministic/stochastic approach, the dynamics is modeled as a Markov process in a discretized finite state space, which results in a nearly decomposable transition probability matrix. By considering the transition probabilities as similarities between the states, the application of a spectral clustering algorithm, the Robust Perron Cluster Analysis (PCCA+), reveals the metastable conformations.

The discretization is defined in terms of basis functions that form a partition of unity, e.g., characteristic functions of Voronoi cells. Since the computational complexity increases quadratically with the number of basis functions, we aim at a partitioning of the state space with as few basis functions as possible, but as many basis functions as necessary to resolve the metastable conformations. Therefore, we have developed an algorithm for an adaptive decomposition of the state space starting from a very coarse decomposition [5]. The refinement in our scheme is not based on geometric clustering but takes into account the intrastate dynamical behavior.

SPECTRAL CLUSTERING BY PCCA+

In molecular simulations, for example molecular dynamics (MD) or hybrid Monte Carlo (HMC), we deal with successive configurations $o_1 \to o_2 \to \ldots \to o_n$ of a molecular trajectory. Via decomposition of the position space Ω into sets $\Omega = \bigcup_{i=1}^{N} A_i$, $A_i \cap A_j = \emptyset$, we can define a transition probability matrix P. The (unknown) transition probabilities can be estimated by simply counting transitions within a certain lag-time τ between the states A_i,

$$P_{ij}^{\tau} \approx \frac{\#[o_k \in A_i, o_{k+1} \in A_j]}{\#[o_k \in A_i]}, \quad i,j = 1,\ldots N. \tag{1}$$

The stationary distribution $\pi = [\pi_1, \ldots, \pi_n]^\top$ of P can be estimated by $\pi_i = \#[o_k \in A_i] / \sum_j \#[o_k \in A_j]$. Thus, the discretization of the configuration space will be limited by the statistical information that must be available to compute the matrix entries of P with sufficient accuracy.

In case of a decomposable Markov chain an appropriate permutation of objects according to their connectedness results in a block-diagonal matrix P with k blocks. This matrix has a k-fold eigenvalue $\lambda = 1$. The corresponding eigenvectors $X = [x_1, \ldots, x_k]$ are piecewise constant on the blocks and can thus be used to identify the clusters. In fact, the rows of X can be considered as vertices of a $(k-1)$-dimensional simplex. Every object can be assigned to one of the k vertices and thus to one of the k clusters.

Generally, the matrix P constructed from practical data is not decomposable. However, if there are k hidden clusters, P has a cluster of eigenvalues $1 = \lambda_1 > \lambda_2 > \ldots > \lambda_k > 1 - \varepsilon$ near the Perron eigenvalue $\lambda_1 = 1$. The rows y_i of the corresponding eigenvectors still nearly form a simplex. The goal of PCCA+ is to identify the vertices of a simplex σ_{k-1} such that all points y_i are located within the simplex (simplex condition). Then every point y_i can be assigned to one of the k vertices and thus to one of the k clusters by a certain *membership vector* $\chi(i,:) = [\chi(i,1), \ldots, \chi(i,k)]$. The identification of such a simplex is equivalent to finding a non-singular transformation matrix $\mathscr{A}$ such that $\chi = X\mathscr{A}$, $\chi \geq 0$ (positivity), and $\sum_{j=1}^{k} \chi(:,j) = 1$ (partition of unity). Among the feasible transformation matrices we search for a matrix $\mathscr{A}$ such that the resulting membership vectors $\chi(i,:)$ are as characteristic as possible. This can be achieved by maximizing the objective function

$$I(\mathscr{A}; X, \pi) = \frac{1}{k} \sum_{i=1}^{k} \frac{\langle \chi_i, \chi_i \rangle_\pi}{\langle \chi_i, e \rangle_\pi} \leq 1, \tag{2}$$

where e denotes the vector with all entries equal to 1. One has to maximize a convex function with linear constraints, which is not a trivial task. The optimization problem can be solved by the Nelder-Mead algorithm provided that a good initial guess for $\mathscr{A}$ is available. This starting guess is obtained by the *inner simplex algorithm* as described in [6].

Once the membership functions χ_i have been computed, one can compute a coarse grained transition probability matrix P_c by projecting the original matrix P onto the metastable conformations, $P_c = (\chi^\top \pi_D \chi)^{-1} \chi^\top \pi_D P \chi$, where π_D denotes a diagonal matrix with the vector π on the diagonal. The matrix P_c is not necessarily a stochastic matrix because it can have negative entries if the membership functions χ_i are far from being characteristic. However, P_c has row sum one and is the correct propagator for densities restricted to the metastable conformations [7].

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

ADAPTIVE DECOMPOSITION OF THE STATE SPACE

Instead of a (Voronoi) discretization of the object space Ω by means of characteristic basis functions, we use *soft characteristic* global functions to decompose Ω in order to avoid statistically unreliable transition probabilities during the refinement procedure. However, once the final decomposition has been obtained, we reconvert the soft partitioning back into a Voronoi tessellation in order to compute transition probabilities between the Voronoi cells. This reconversion is necessary to make the transition probabilities independent from any shape parameter which will be introduced with the global functions. To be precise, we use n radial basis functions with nodes $\{q_1, \ldots, q_n\}$ with a Gaussian similarity measure following the partition of unity method of Shepard [8]:

$$\phi_i(o_k) = \frac{\exp(-\alpha\, d(o_k, q_i)^2)}{\sum_{j=1}^{n} \exp(-\alpha\, d(o_k, q_j)^2)}, \quad i = 1, \ldots n, \tag{3}$$

where $d(o_k, q_i)$ denotes an appropriate distance measure, for example the Euclidian distance. The basis functions can be interpreted as membership functions since they are non-negative and form a partition of unity. In analogy to (1) we define the kinetic similarity $K^{(L)}(i,\, j)$ between two basis functions ϕ_i and ϕ_j for lag-time $L\tau$ as

$$K^{(L)}(i,\, j) := \frac{\sum_{k=1}^{N} \phi_i(o_k) \phi_j(o_{k+L})}{\sum_{k=1}^{N} \phi_i(o_k)}. \tag{4}$$

This formula will be used in the following to detect metastabilities within basis functions.

Note that it is not possible to separate two different metastable sets in the process of clustering if they are covered by only one basis function. With the following locally adaptive partitioning algorithm we aim to improve the initial

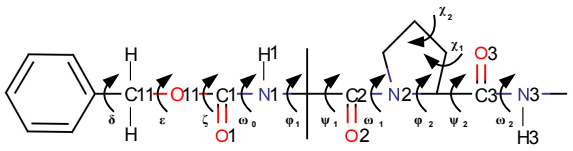

FIGURE 1. Primary structure of ZAibProNHMe model peptide defining 12 dihedral angles, see also Ref. [10].

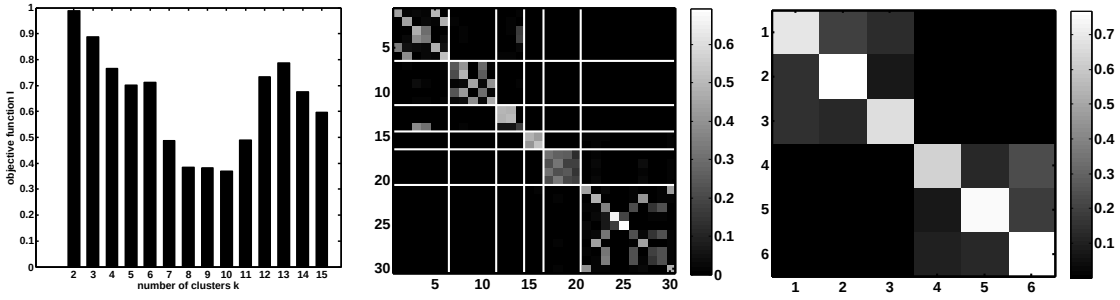

FIGURE 2. Results from the Perron Cluster analysis of the transition matrix generated from trajectory simulation of the model peptide for $T = 600$K. (a) Values of the objective function for different numbers of clusters. (b) Re-ordered transition matrix. (c) Coarse-grained transition matrix.

selection of nodes and thus find an optimal soft partition of the object space. The main idea is to check each local basis function for the existence of further metastabilities and, if found, to refine the basis function by adding a user-defined number of s nodes that represent the metastable sets. For one specific basis function ϕ_i, the algorithm has the following structure:

1. Select all objects o_j with $\phi_i(o_j) > \phi_k(o_j) \forall k \neq i$.
2. Perform the k-means algorithm [9] with s clusters on the selected objects. Choose the objects nearest to the s computed centroids as new temporal nodes $\{\widetilde{q}_{i1}, \ldots, \widetilde{q}_{is}\}$.
3. Compute the kinetic similarity matrix $K^{(L)}$ based on the temporal set of basis functions $\{\widetilde{\phi}_{i1}, \ldots, \widetilde{\phi}_{is}\}$ with

$$\widetilde{\phi}_{il}(x) = \frac{\exp(-\alpha\, d(q_{il}, x)^2)}{\sum_{j=1}^{s} \exp(-\alpha\, d(q_{ij}, x)^2)}, \quad l = 1, \ldots s.$$

4. Select from the temporal nodes the ones for which $K^{(L)}(i,i) > \rho$, $1 > \rho \sim 1 - \varepsilon$, $i = 1, \ldots, s$.
5. Replace the old node q_i with all accepted new nodes.

Steps 3 and 4 aim at checking whether the clusters proposed by the k-means algorithm really separate different metastable sets. Only in this case the basis function will be refined.

After a successful iteration the complete partitioning is recomputed based on the updated list of representing nodes, and the above algorithm is applied again to all newly added basis functions. The iteration continues until no new basis functions are added.

APPLICATION: CONFORMATIONS OF MODEL TRI-PEPTIDE

The various clustering methods introduced in the previous sections shall be illustrated here for a simulation of the ZAibProNHMe model peptide which is of considerable pharmaceutic interest in antibiotic peptides [11]. This molecule has been chosen because the combination of the two methyl groups (in Aib) and the steric restrictions introduced by the pyrrolidine ring (in Pro) cause a strong competition between γ (7-membered ring with H-bond between H3 and O2) and β (10-membered ring with H-bond between H3 and O1) turn structures, see Fig. 1, coexisting at similar energies [12, 10]. Molecular dynamics trajectories with a duration of 39 ns have been generated using the Merck Molecular Force Field [13] where a Nosé-Hoover thermostat ensures sampling of a canonical (NVT) ensemble for

$T = 600$ K [14]. In the first step, the adaptive geometric clustering technique is applied to the time series of the 12 torsional coordinates indicated in Fig. 1. Starting from 10 seed nodes with width parameter $\alpha = 1/(2\sigma^2)$, $\sigma = 20^\circ$, the basis is adaptively refined leading to 30 basis functions after 3 iteration steps. The corresponding transition matrix P for a lag-time $L\tau = 7.8$ ps is nearly-decomposable which is reflected by several eigenvalues near unity, separated by a spectral gap from the remaining ones. In the second step, the PCCA+ spectral clustering technique is applied to transform eigenvectors to membership vectors. The dependence of the objective function I defined in (2) on the number of clusters is shown in Fig. 2 (a). Based on this metastability criterion we choose this number to be 6. The re-ordering of rows and columns of the transition matrix according to the 6 clusters is shown in Fig. 2 (b), where a set A_i is assigned to the cluster j with maximal membership $\chi(i,j) = \max \chi(i,:)$. The final, coarse-grained transition matrix P_c is shown in Fig. 2 (c). There are six metastable sets with near unity probabilities to stay within each of the clusters and rather low probabilities of undergoing a jump between the sets within the time span of 7.8 ps. The resulting clusters are in all cases centered around known minima [10] of the molecular potential energy surface corresponding to different β- and γ-turn structures, see forthcoming paper.

CONCLUSION

We have demonstrated that adaptive spectral clustering can successfully be applied to the analysis of time-series from molecular dynamics simulations. Since the computational complexity does not scale exponentially with the number of degrees of freedom, the method is believed to be applicable to large molecules such as larger peptides or proteins.

REFERENCES

1. M. Dellnitz, and O. Junge, *SIAM J. Numer. Anal.* **36**, 491–515 (1999), URL http://www.jstor.org/stable/2587207.
2. C. Schütte, *Conformational Dynamics: Modelling, Theory, Algorithm, and Application to Biomolecules*, Habilitation thesis, Department of Mathematics and Computer Science, Freie Universität Berlin (1999), URL http://opus.kobv.de/zib/volltexte/1999/407/.
3. P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte, *Lin. Alg. Appl.* **315**, 39–59 (2000), URL http://dx.doi.org/10.1016/S0024-3795(00)00095-1.
4. P. Deuflhard, "From Molecular Dynamics to Conformational Dynamics in Drug Design," in *Trends in Nonlinear Analysis*, edited by M. Kirkilionis, S. Krömker, R. Rannacher, and F. Tomi, Springer, 2003, pp. 269–287.
5. F. Haack, *Representative Spectral Clustering for Large Data Sets applied to Gene Expression Data*, Master's thesis, Freie Universität Berlin (2009), available from fiete.haack@uni-rostock.de.
6. M. Weber, and T. Galliat, Characterization of transition states in conformational dynamics using fuzzy sets, ZIB-Report 02-12, Zuse Institute Berlin (2002), URL http://opus.kobv.de/zib/volltexte/2002/680/.
7. S. Kube, and M. Weber, *J. Chem. Phys.* **126** (2007), URL http://dx.doi.org/10.1063/1.2404953.
8. D. Shepard, "A two-dimensional interpolation for irregularly spaced data," in *Proc. 23rd ACM Nat. Conf.*, 1968, pp. 517 – 524.
9. J. Hartigan, and M. Wong, *J. Royal Stat. Soc.* **28**, 100–108 (1979), URL http://www.jstor.org/stable/2346830.
10. H. Zhu, M. Blom, I. Compagnon, A. M. Rijs, S. Roy, and G. von Helden, *Phys. Chem. Chem. Phys.* **12**, 3415–3425 (2010), URL http://dx.doi.org/10.1039/b926413b.
11. A. Aubry, D. Bayeul, H. Brückner, N. Schiemann, and E. Benedetti, *J. Pept. Sci.* **4**, 502 (1998), URL http://dx.doi.org/10.1002/(SICI)1099-1387(199812)4:8<502::AID-PSC171>3.0.CO;2-N.
12. B. Di Blasio, V. Pavone, M. Saviano, A. Lombardi, F. Nastri, C. Pedone, E. Benedetti, M. Crisma, M. Anzolin, and C. Toniolo, *J. Am. Chem. Soc.* **114**, 6273 (1992), URL http://dx.doi.org/10.1021/ja00042a001.
13. T. A. Halgren, *J. Comput. Chem.* **20**, 730–748 (1999), URL http://dx.doi.org/10.1002/(SICI)1096-987X(199905)20:7<730::AID-JCC8>3.0.CO;2-T.
14. S. Nosé, *J. Chem. Phys.* **81**, 511–519 (1984), URL http://dx.doi.org/10.1063/1.447334.