J Math Chem (2012) 50:2623–2635
DOI 10.1007/s10910-012-0051-5

ORIGINAL PAPER

A coarse graining method for the dimension reduction of the state space of biomolecules

Konstantin Fackeldey · Martina Klimm · Marcus Weber

Received: 11 May 2012 / Accepted: 9 July 2012 / Published online: 22 July 2012

Abstract The simulation of biomolecules requires a vast amount of computer power. One reason can be found in the high dimensionality of the state space. Coarse graining methods attempt to improve the computational performance by reducing the representation of a molecule's dynamics without losing relevant details. Here, we show two methods coarsening the full model by still retaining the details which are necessary for the comprehension of the protein's conformational dynamics. We review the first method, which clusters motions of the particles according to a certain criterion. This approach is a coarse graining strategy in time based on Markov State Models. The second method, the Hierarchical Relevant Descriptor Detector, is a novel technique for coarse graining in space revealing a hierarchy in the descriptors of a protein. Thus, it allows us to describe the relevant motions of a molecule by employing only a minor number of descriptors. The performance of this method is shown in two examples.

Keywords Coarse graining · Markov State Models (MSM) · Dimension reduction · Biomolecules

K. Fackeldey (✉) · M. Klimm · M. Weber
Zuse Institut Berlin (ZIB), Takustr. 7, 14129 Berlin, Germany
e-mail: fackeldey@zib.de

M. Klimm
e-mail: klimm@zib.de

M. Weber
e-mail: weber@zib.de

1 Introduction

The function of proteins in the human body is versatile, they can act as structural proteins for cells, as enzymes or ion channels, to mention a few. Before a protein is in its native state it passes through a process (protein folding) starting from the primary structure (amino acid sequence) and ending in the tertiary structure (folded state). The behavior of a protein in the tertiary structure can be described by the Boltzmann distribution, which relates the maximum entropy state of a protein (equilibrium state) to its energy and temperature in a statistical way. As a consequence, the native state has the lowest free energy. Since the structure of a protein is very flexible, the resulting potential energy surface is rough and has a large number of local minima. On its way from initial to final position the molecule also visits intermediate states or metastable states. In a metastable state, the system stays for a certain time in a subset of the state space before a random force is large enough causing a transition into another metastable state.

From the trajectories of a molecular dynamics (MD) simulation the position and momenta of a biomolecule on the atomistic level are given. Since the first article about MD simulations [1] much effort has been put into the development of fast and reliable algorithms [2,4,18,38,39] and their parallelization [17,21,28,31]. However, in trajectory based MD, the maximum size of the integration step is bounded to femtoseconds whereas protein folding ranges in microseconds which implies the simulation of long trajectories [27,34] or multiple trajectories [3,29,40]. Moreover, having the data of a long trajectory, it is still unclear how to interpret these datas in the high dimensional state space.

Summing up we have two different kinds of challenges, namely a challenge in the *time scale* and a challenge in the *high dimensionality* of the state space. These two aspects lead us to two computational coarse graining strategies, namely a coarse graining strategy in time and a coarse graining strategy in space.

1.1 Coarse graining in time

The first aspect offers us not to use the confining small time step in a trajectory but to consider the conformational changes only, which take place on a coarser time scale. This leads us to Markov State Models (MSM), e. g. [6,7,9–11,15,33] taking advantage of the fact, that for suitable chosen time steps, the transitions appear stochastic and memory less. Thus one seeks in MSM to describe the dynamics of a molecule in terms of transition probabilities between metastable sets. The metastable states can be identified as subsets in the high dimensional conformation space. We thus seek for a metastable decomposition, i.e. a decomposition of the state space into disjoint subsets. We therefore follow the approach of [32] by applying a transfer operator T, describing a momentum weighted fluctuation in a canonical ensemble. More precisely, we employ a Galerkin discretization of T in order to represent the metastable decomposition by linear combinations of characteristic or meshfree basis functions [13,33].

1.2 Coarse graining in space

Here, we consider methods, which allow us to reduce the high dimensionality of the state space by still keeping relevant informations. In general descriptors describe the *structure–activity* relationship in molecules, and they can be defined as a collection of data characterizing the molecule [37]. In contrast to QSAR methods (e.g. [5,12]), we do not want to use descriptors in order to characterize or compare structures of molecules. Here, we answer the question:

How to describe the dynamics of a molecule by using less degrees of freedom?

Of course, this question is closely related to dimension reduction of the state space. In contrast to the above mentioned QSAR methods, we only use dynamical informations of a trajectory. Having these main activators (most relevant descriptors) it allows us, to describe conformational changes in a lower dimensional space than the primal space. For instance, the relevant dynamics (conformational changes) of the well known alanine dipeptide can be described by using only two degrees of freedom, namely the dihedral angles ϕ and ψ. Our scheme reveals from all possible descriptors these two parameters by only using information form a trajectory.

2 Metastability and coarse graining in time

A molecule passes through intermediate states on its way from the initial to the final state. Thereby, a protein in an intermediate state is not stable but almost stable (metastable), which means that the molecule stays for a certain time span in this configuration before it switches to another state. In the following, we give the above mentioned terms *metastability* and *conformational change* a mathematical foundation. Thereby we employ MSM by identifying metastable states as subsets in the phase space Γ and the dynamics by transition probabilities of a transition matrix. More precisely, the entries of the transition matrix give the probability that a system being in metastable state A switches to a metastable state B (cp. Fig. 1). We remark, that the different configurations belonging to a certain metastable state are only kinetically closely related, but not necessarily "geometrically" or "homotopically" similar. However, Hammond's postulate [20] states that energetically closely related configurations are also geometric similar.

Mathematically, we understand a conformation as a part of the conformation space Ω, which comprises the collection of structurally related configurations. We now consider n atoms of a molecular system with the spatial coordinates $q_j \in \mathbb{R}^3$, $j = 1, \ldots, n$, and their n generalized momenta $p_j \in \mathbb{R}^3$. Then $(q, p) \in \Gamma$ where $\Gamma = \Omega \times \mathbb{R}^{3n}$ is the phase space and Ω the position space. These states are distributed according to the Boltzmann distribution

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

Here $\beta = 1/(k_B T)$ is the inverse of the temperature T multiplied with the Boltzmann constant k_B, and H denotes the Hamiltonian function which is given by $H(q, p) =$

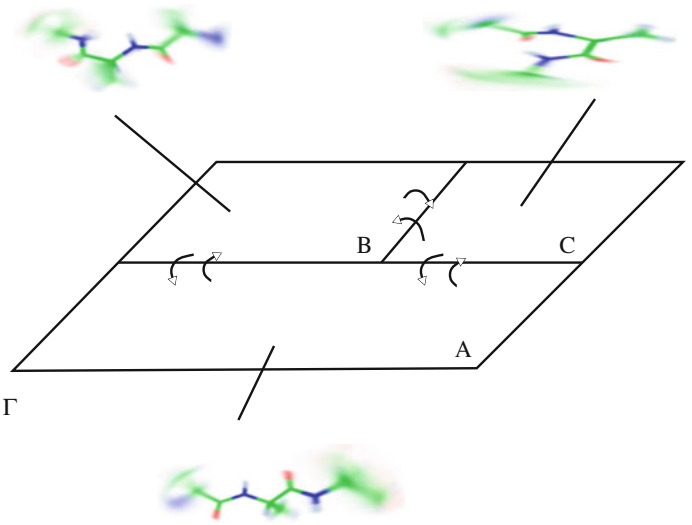

Fig. 1 Sketch of a phase space Γ and its metastable subsets A, B and C

$K(p) + V(q)$, where $K(p)$ is the kinetic energy and $V(q)$ is the potential energy. The canonical density can be split into a distribution of momenta $\eta(p)$ and positions $\pi(q)$ that is

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

A direct computation of the distribution is often hardly feasible, since for its evaluation a high dimensional integral has to be solved. For the dynamics of the system we can employ a corresponding flow Φ^τ for a time span $\tau > 0$. This Hamiltonian flow is given by

$$(q_i, p_i) = \Phi^{i\tau}(q_0, p_0), \quad i = 1, \ldots, n.$$

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

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

describes a Markov process. Note, that the ith state depends on the preceding state only. It can be shown, that this assumption of Markovianity implies the time independence of the corresponding transfer operator (e.g. [32]). Hence, we will define the above described metastability in a mathematical framework. Let us introduce the Boltzmann weighted scalar product $\langle f, g \rangle := \int f(q) g(q) \pi(q) dq$ and the characteristic function $\chi_A(q)$, being 1 if $q \in A$ and 0 otherwise. Then we can define the

conditional probability of the system to move during time span τ from subset A_1 to subset A_2 by

$$p(A_1, A_2, \tau) = \frac{\langle \chi_{A_1}, T^\tau \chi_{A_2} \rangle}{\langle \chi_{A_1}, \chi_{A_2} \rangle},$$

with the transfer operator

$$T^\tau f(q) = \int_{\mathbb{R}^{3n}} f(\Pi_q \Phi^{-\tau}(q, p)) \eta(p)\, dp \tag{2}$$

which has been introduced by Schütte [32]. The operator in (2) can be explained as follows: We apply the Hamiltonian dynamcis backwards to a lag time τ and obtain $\Phi^{-\tau}(q, p)$, which is projected by Π_q onto the state space. The integral then, averages over all possible initial momentum variables with given Boltzmann distribution η. These tools enable us to characterize a subset $A \subset \Omega$ metastable if it is almost invariant under the transfer operator T^τ, i.e.

$$p(A, A, \tau) = \frac{\langle \chi_A, T^\tau \chi_A \rangle}{\langle \chi_A, \chi_A \rangle} \approx 1 \tag{3}$$

which can be restated as

$$p(A, A, \tau) \approx 1 \iff \langle \chi_A, T^\tau \chi_A \rangle \approx \langle \chi_A, \chi_A \rangle.$$

Here we can attain to two conclusions: For the first, metastable sets can be identified by computing the eigenfunctions of the propagator T^τ. For the second, the more the fraction in (3) is close to one, the more metastable is the set in A. Let us now assume, that we can partition the space into N metastable sets, i.e. $C_1, \ldots, C_N$, which approximate metastable eigenfunctions of the operator T^τ. Their characteristic basis functions are given by $\chi_{C_1}, \ldots, \chi_{C_N}$. Thus, the dynamics of the system can be approximated by the transition probabilities between the metastabilities. The resulting linear operator $P : \text{span}(\chi_{C_1}, \ldots, \chi_{C_N}) \to \text{span}(\chi_{C_1}, \ldots, \chi_{C_N})$ is given by the stochastic matrix

$$(P_{ij})_{i,j=1,\ldots,N} \quad \text{with} \quad P_{ij} = \frac{\langle T^\tau \chi_{C_i}, \chi_{C_j} \rangle}{\langle \chi_{C_i}, \chi_{C_i} \rangle}. \tag{4}$$

According to the celebrated Perron Frobenius theorem (e. g. [23]) the row stochastic matrix P has the eigenvalues $\lambda_1, \ldots, \lambda_N$ which can be arranged such that $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_N$. Moreover, let v_i be the N dimensional right eigenvector of eigenvalue λ_i and u_i the left eigenvector, then $\lambda_1 = 1, v_1 = e = (1, \ldots, 1)^T, u_1 = \pi_N$ and $\lambda_1 > \lambda_i \quad i = 2, \ldots, N$. Here, π_N is a N-dimensional vector, whose elements are the stationary probabilities of the N characteristic functions. Hence, the right eigenvectors $v_1, \ldots, v_N$ form an eigenbasis which can be used to express any vector x as:

$$x = \sum_{i=1}^{N} \alpha_i v_i = \sum_{i=1}^{N} \langle x, v_i \rangle v_i .$$

Since $Pv_j = \lambda_j v_j$ we have

$$Px = P \sum_{j=1}^{N} \langle x, v_j \rangle v_j = \sum_{j=1}^{N} \langle x, v_j \rangle P v_j = \sum_{j=1}^{N} \lambda_j \langle x, v_j \rangle v_j .$$

We thus can explain the dynamics of a molecule in terms of transition probabilities (given by Matrix P) between subsets in the conformation space.

3 Coarse graining in space: HRDD

The introduction of the conformation concept allowed us for a coarse graining in time, which we established in the foregoing section. Here, we exploit, that in most molecules a conformational change can be detected by very few descriptors. One prominent example of this fact is the Ramachandran plot, where even larger molecules can be described admissible by using only two descriptors per amino acid. In the following, we introduce a novel method, which reveals a hierarchy in the descriptors of a molecule and thus allows to map the relevant motions (e.g. conformational change) by using only a few degrees of freedom (descriptors). Having such a hierarchy, we can start with a one dimensional model, by taking only the first (and most relevant) descriptor. Using the second most relevant descriptor we can approximate the motion of the integral by a two dimensional model and so on. This method does not use any deeper chemical details, in fact, it only relies on the long-term trajectory. The spatial structure of each molecule can be characterized by its descriptors $D^i, i = 1, \ldots, n_d$, which are a collection of all internal degrees of freedom of the molecule such as distances between two arbitrary atoms, bond or dihedral angles between three or four bonded atoms. However the full number of descriptors of a molecule n_d to specify it, would lead to an over-determined system, which means that we have a redundancy in the descriptors. For instance, pentane consists of 17 atoms, thus it has $17 \times 3 - 6 = 45$ degrees of freedom. Even if we only consider all possible distances between two atoms, we obtain $17 \times 16/2 = 136$ possible descriptors. This also leads to the assumption, that the descriptors of a molecule adhere a hierarchy, according to their influence on the spatial structure of the molecule. As a consequence, only a fraction of all possible descriptors, the *relevant descriptors* suffice to characterize the spatial structure and, thus, we do not need the high dimensional full phase space, but a low dimensional space, the so called *conformational space*.

The detection of the relevant descriptors can either be done by employing detailed chemical expert knowledge or by analyzing the time series of the molecule's dynamics. In the following, we concentrate ourselves to the latter by introducing the Hierarchical Relevant Descriptor Detector (HRDD). Let us, therefore, assume that we have a sampling covering the whole conformational space of a molecule with at least two conformations.

For all possible and given descriptors $D^i, i = 1, \ldots, n_d$, we employ a fine uniform discretization of (the subset of) the conformational space represented by the sampled data. For the ease of notation we assume that the number of discretization intervals (bins) N is equal for all descriptors, though, we remark that our method allows for different dimensions in each descriptor.

For a given trajectory, we calculate the transition matrix $P^i \in \mathbb{R}^{N\times N}$, for each descriptor D^i, i. e. increase the entries of P^i_{ab} and P^i_{ba} by one if the trajectory switches within two consecutive time-steps from bin a to bin b. In order to gain a row-stochastic matrix, that is $\sum_{b=1}^{N} P^i_{ab} = 1\ \forall a = 1, \ldots, N$, we normalize each row by dividing each entry by its row-sum. Following the discussion of the foregoing section we compute for each matrix P^i the corresponding second largest eigenvalue λ^i_2 and select the respective descriptor D^k with

$$\lambda^k_2 = \max\{\lambda^{(i)}_2 \,|\, i = 1, \ldots, n_d\}, \tag{5}$$

i.e. largest of the second largest eigenvalues.

If this λ^k_2 is close to 1, then the descriptor D^k characterizes at least two metastable conformations within the sampling and we can continue to calculate the actual number of metastabilities in this descriptor D^k. Otherwise the algorithm stops in this level (but might continue in another branch, cp. Fig. 2b). The number of clusters n_c is gained by estimating all eigenvalues $\lambda^k_l, l = 1, \ldots, N$, which fulfill

$$|\lambda^k_l - 1| < tol, \quad l = 1, \ldots, n_c. \tag{6}$$

The tolerance value must satisfy Eq. (6) for $l = 1, 2$, since we already required that λ^k_2 is close to 1 and, thus, at least two metastabilities exist. We remark, that this tolerance has to be chosen with care: If the tolerance is to large, the number of clusters is large and thus we might have a poor reduction of the dimension. However if the tolerance is to small relevant dynamical information of the molecule might get lost.

For the decomposition of the space into clusters $C_l, l = 1, \ldots, n_c$, we compute the corresponding eigenvectors $v^k_l, l = 1, \ldots, n_c$, building a basis according to the respective eigenvalues λ^k_l. Finally the algorithm is recalled for each cluster found in the last iteration that covers sufficient data. For illustration purpose Fig. 2a shows a flow chart of the recursive algorithm.

Each non decomposable cluster is related to a metastable conformation. Starting from the top, the number of branches is given by all second largest eigenvalues satisfying tol. Going down one arm corresponds to "freezing" the first metastable state described by the most relevant descriptor. In the next branch the metastabilities except for the frozen parts are computed. For the illustration a rooted tree shall be drawn, see Fig. 2b. The number of branches on each level ("width") is given by the number of eigenvalues according to criterion (6). The height of the tree is determined by (5).

Each leaf is a non decomposable cluster and, thus, a metastable conformation M_l. Each parent corresponds to a descriptor. The higher the tree-order (height of a tree from root to leaf) the more descriptors are necessary to characterize a metastable conformation. In Fig. 2b the algorithm finds the descriptor D^{k_1} in the first instance, which

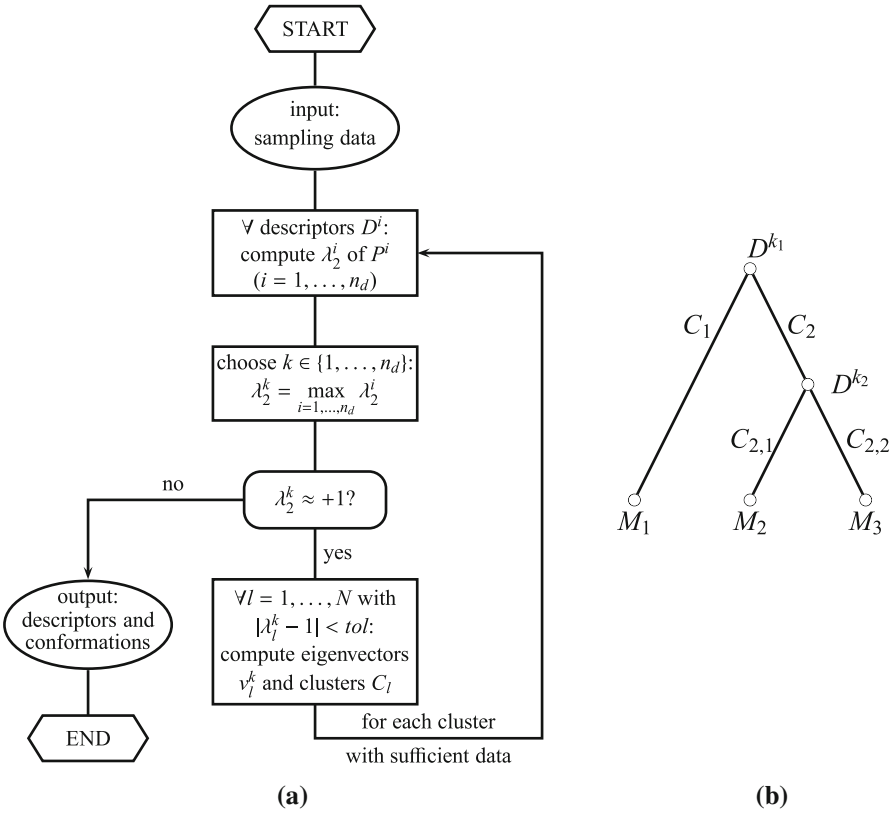

Fig. 2 The recursive algorithm, starting with sufficient sampling data of a molecule with at least two conformations supposed. **a** Flow chart. **b** Rooted tree

decomposes the whole conformational space into two clusters C_1 and C_2. The first one is not decomposable and, thus, a metastable conformation M_1 characterized by the single descriptor D^{k_1}. The second cluster is most metastable in descriptor D^{k_2} and comprises two clusters $C_{2,1}$ and $C_{2,2}$, which prove to be metastable conformations M_2 and M_3 characterized by both D^{k_1} and D^{k_2}.

We remark, that such a hierarchical scheme has also been applied in the context of temperature and decomposition of the conformation space [15].

4 Applications

For testing the applicability of the method we analyzed the conformations and descriptors of two small molecules, pentane and alanine dipeptide. Both molecules have been investigated years before, see [32] and [35] respectively.

To obtain the data for our algorithm, we ran different trajectories with Hybrid Monte Carlo (HMC) [14] and MD in vacuum at 300 K. In advance, we parametrized the molecules according to different force fields. In case of HMC simulation we used the Merck

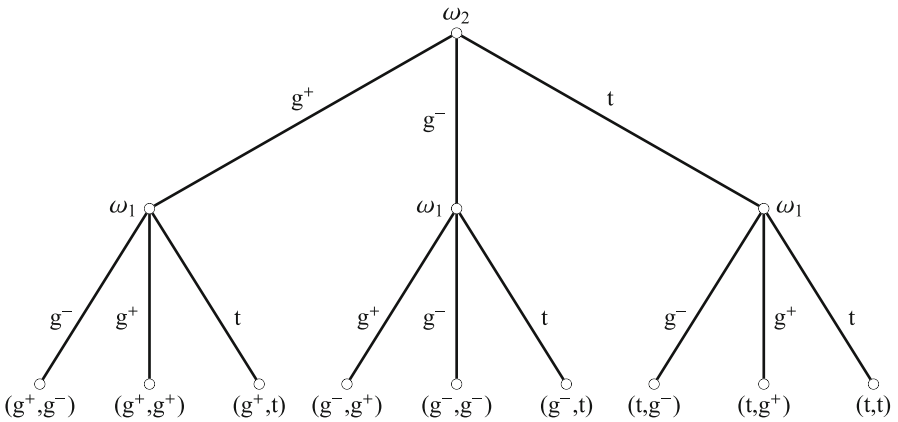

Fig. 3 The resulting tree for pentane with nine conformations described by (ω_2, ω_1). On the leaves the types of conformations are shown

Molecular Force Field (MMFF) [19] for both. For the MD simulation with Gromacs we chose the Amber Force Field ffAmber99sb [22] for pentane and the Optimized Potential for Liquid Simulations All-Atoms (OPLS-AA) [24] for alanine dipeptide.

For the HMC samplings five chains were started in parallel, for each HMC step a 60 fs MD trajectory was calculated to generate a trial state. Altogether the samplings covered 6 ns for pentane and 12 ns for alanine dipeptide. We reached an average acceptance probability of more than 95 %. The convergence was monitored by the Gelman-Rubin acceptance criterion [16] with a threshold value of 1.2.

In order to generate MD trajectories we used Gromacs [4,21,26,36] with a modified Berendsen thermostat [8], where the correct distribution of the energy is enforced by the a constructed force. Furthermore we took an integration step of 1 fs and a reference temperature of 300 K, and simulated over the same time as with HMC, i. e. 6 ns for pentane and 12 ns for alanine dipeptide. To have a comparable amount of data we wrote out the coordinates every 60 fs.

4.1 Pentane

Pentane is able to adopt nine metastable conformations which can be described by the two dihedral angles ω_1 between the carbon atoms C_1, C_2, C_3, C_4 and ω_2 between C_2, C_3, C_4, C_5 (e.g. [25]). Out of a number of 138 descriptors of the pentane molecule (both the distances between any two atoms and the dihedral angles between the bonded carbon atoms), HRDD has selected the two relevant ones ω_1 and ω_2. The dihedral angles can take the values $\pm 180°$ (trans), $60°$ (gauche$^-$) and $-60°$ (gauche$^+$) which we want to abbreviate to t, g^- and g^+, respectively. As shown in Fig. 3 our algorithm found all nine conformations and the two descriptors ω_1 and ω_2 both for the sampling data generated by HMC and by MD, the chosen tolerances are shown in Table 1. In Fig. 4 the absolute number of states for the respective conformations are given.

Table 1 The eigenvalue tolerances of the pentane simulation

	$\lambda_2^k \approx +1$	$\lvert\lambda_l^k - 1\rvert < tol$
HMC	0.05	0.1
MD	0.01	0.05

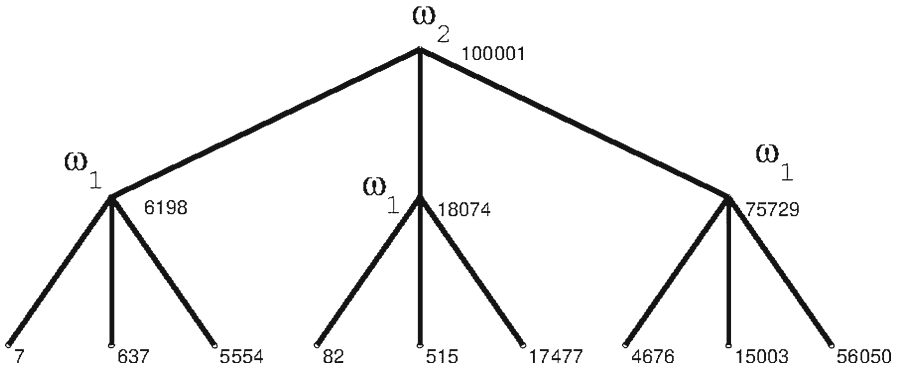

Fig. 4 The resulting tree for pentane with nine conformations described by (ω_2, ω_1), the numbers on the leave show the number of states according to the respective conformation

4.2 Alanine dipeptide

The alanine dipeptide which is a, by acetyl and methyl, terminally-blocked alanine amino acid, has a more complex structure and the metastable conformations are not that well-defined which reflects in the inconsistent results of different articles, see Table 4 in [35] for an overview of some of them. Additionally the conformational space in vacuum is much smaller than in solution. Thus, we considered 175 possible descriptors of alanine dipeptide, thereof 171 distances and four dihedral angles. For the methylene group at C_β we chose the united atom presentation since rotation in this side group leads to structurally identical molecules. The algorithm should automatically identify the descriptors which include the most relevant information about conformational transitions. In vacuum the method found the two dihedral angles ϕ (C–N–Cα–C) and ψ (N–Cα–C–N) describing three metastable conformations, namely C_7^{eq} for $\phi \approx -80°$ and $\psi \approx 70°$, C_5 for $\phi \approx -150°$ and $\psi \approx 155°$, and C_7^{ax} for $\phi \approx 80°$ and $\psi \approx -50°$, see Fig. 5a (for both cases: HMC and MD). In Fig. 5b the corresponding 2-dimensional plot clearly shows the three conformations of alanine dipeptide

This Fig. 5b shall be used as well to illustrate the division of the conformational space for the two dihedral angles in case of the MD simulation. The color typifies the absolute number of states visited by the MD simulation with the coordinates rounded to two decimal places. In the first instance the conformational space is divided along the ϕ-axis into two clusters. The first cluster cannot be divided furthermore, this subset matches the conformation C_7^{ax}, while the second cluster can be split along the ψ-axis into two clusters, which are finally the two other conformations C_7^{eq} and C_5.

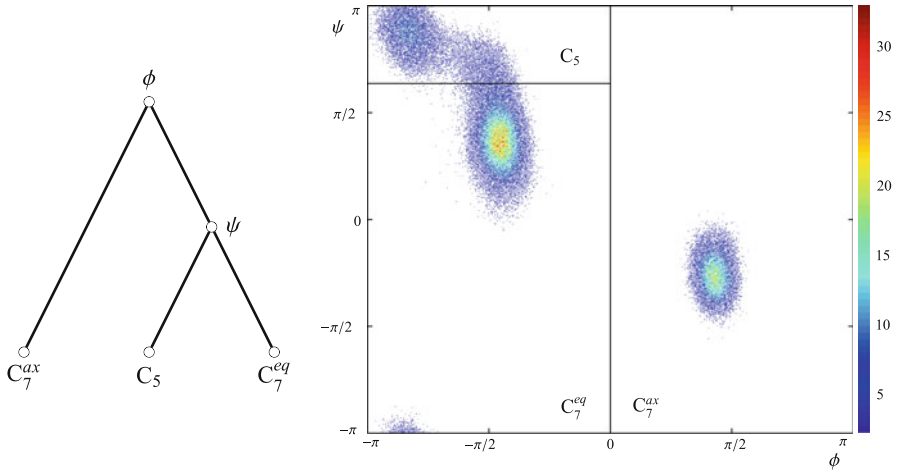

Fig. 5 The results for alanine dipeptide: three conformations described by (ϕ, ψ). **a** Rooted tree. **b** The ϕ-ψ-frequency-plot (the *color* describes the absolute number of states visited by the MD sampling) (Color figure online)

5 Conclusion and outlook

This paper is geared on two aspects: The first one is the conformational change and the second one is the observation of hierarchies in the descriptors. These two aspects lead us to coarse graining strategies making the computation of molecule's dynamics more tractable by requiring less computational power.

We briefly reviewed the detection of metastable sets by MSM. We revealed that the second largest eigenvalue is equivalent to the maximal autocorrelation coefficient. In other words, the side motions in a conformation have a low variance.

The main focus of this paper is the introduction of HRDD (Hierarchical Relevant Descriptor Detector) as a method for the identification of the relevant internal degrees of freedom of molecular structures. Whereas principal component analysis (PCA) extracts the coordinate directions with maximal variance of data, HRDD selects the degrees of freedom which include the "rarest" transitions, therefore, leading to a good separation of conformations. The performance of this method has been shown by two examples, namely pentane and alanine dipeptide. In Fig. 3, the relevant torsion angles of pentane have been identified. For alanine dipeptide HRDD has selected two descriptors such that the three conformations can clearly be seen in the corresponding two-dimensional plot Fig. 5b. Of course, efforts on larger datasets comprising larger molecules are planned in order to attest the applicability to proteins. Therefor, detailed chemical knowledge could be used to reduce the initial amount of possible descriptors (e. g. regarding the atoms of the backbone only) and, thus, improve the performance of the algorithm.

After having successfully applied the HRDD method to our exemplary molecules we also want to critically examine it. The first point concerns the bounds for the eigenvalues (5) and (6). As already mentioned above these barriers affect the height and the

"width" of the illustrating tree. However, we cannot give a general definition but adjust the values individually for the underlying simulation. The second part refers to the strictness of the division of the conformational space. Since each time step of the simulation will be assigned to one conformation, the resulting conformations comprise transition states as well and might be, therefore, artificially enlarged. Anyway, the algorithm reveals the correct number of conformations and their relevant descriptors.

During the development of this paper MD simulations for trialanine in explicit water were run in our workgroup, which incited us to test HRDD for this molecule in explicit water. As the respective publication is still in preparation, we will forbear from describing the simulation and the exact results in detail. Anyway, the analysis with HRDD revealed three conformations described by the two dihedral angles ϕ and ψ, which seem to almost coincide with the results in the paper of Mu and Stock [30]. After the publication of the original paper referring to the simulation of trialanine and after some more examinations of our method with explicit water we intend to publish this and further results in detail.

References

1. B.J. Alder, T.E. Wainwright, Phase transition for a hard sphere system. J. Chem. Phys. **27**, 1208–1209 (1957)
2. E. Barth, K. Kuczera, B. Leimkuhler, R.D. Skeel, Algorithms for constrained molecular dynamcis. J. Comp. Chem. **16**, 1192–1209 (1995)
3. A. Beberg, D. Ensign, G. Jayachandran, S. Khaliq, V. Pande, *Folding@home: Lessons From Eight Years of Volunteer Distributed Computing* (IEEE Computer Society Press, Los Alamitos, CA, 2009), pp. 1–8
4. H.J.C. Berendsen, D. van der Spoel, R. van Drunen, Gromacs: a message-passing parallel molecular dynamics implementation. Comput. Phys. Commun. **91**, 43–56 (1995)
5. D. Bonchev, D. Rouvray, *Chemical Graph Theory: Introduction and Fundamentals* (Gordon and Breach Science Publisher, London, 1990)
6. G.R. Bowman, K.A. Beauchamp, G. Boxer, V.S. Pande, Progress and challenges in the automated construction of Markov state models for full protein systems. J. Chem. Phys. **131**(12), 124101 (2009)
7. N.V. Buchete, G. Hummer, Coarse master equations for peptide folding dynamics. J. Phys. Chem. B **112**(19), 6057–6069 (2008)
8. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. J. Chem. Phys. **126**(1), 014101 (2007)
9. J.D. Chodera, F. Noé, Probability distributions of molecular observables computed from markov models. II: Uncertainties in observables and their time-evolution. J. Chem. Phys. **133**(10), 105102 (2010)
10. J.D. Chodera, N. Singhal, V.S. Pande, K.A. Dill, W.C. Swope, Automatic discovery of metastable states for the construction of markov models of macromolecular conformational dynamics. J. Chem. Phys. **126**(15), 155101 (2007)
11. J.D. Chodera, W.C. Swope, J.W. Pitera, K.A. Dill, Long-time protein folding dynamics from short-time molecular dynamics simulations. Multiple Model. Simul. **5**(4), 1214–1226 (2006)
12. M. Dehmer, K. Varmuzar, D. Bonchev (eds.), *Statistical Modelling of Molecular Descriptors in QSAR/QSPR* (Wiley-VCH, Weinheim, 2012)
13. P. Deuflhard, M. Weber, Robust Perron cluster analysis in conformation dynamics. Lin. Alg. Appl. **398**, 161–184 (2005). doi:10.1016/j.laa.2004.10.026
14. S. Duane, A.D. Kennedy, B.J. Pendleton, D. Roweth, Hybrid monte carlo. Phys. Lett. B **195**(2), 216–222 (1987)
15. A. Fischer, C. Schuette, P. Deuflhard, F. Cordes, *Hierarchical Uncoupling-Coupling of Metastable Conformations*, vol. 24 (Springer, Berlin, 2002), pp. 235–259
16. A. Gelman, D. Rubin, Inference from iterative simulation using multiple sequences. Stat. Sci. **7**(4), 457–511 (1992)

17. M. Griebel, S. Knapek, G. Zumbusch, *Numerical Simulation in Molecular Dynamics, Texts in Computational Science and Engineering*, vol. 5 (Springer, Berlin, 2007)
18. W. van Gunsteren, H. Berendsen, Algorithms for brownian dynamics. Mol. Phys. **45**(3), 637–647 (1982)
19. T.A. Halgren, Merck molecular force field. I–V. J. Comput. Chem. **17**(5–6), 490–641 (1996)
20. G.S. Hammond, A correlation of reaction rates. J. Am. Chem. Soc. **77**, 334–338 (1955)
21. B. Hess, C. Kutzner, D. van der Spoel, E. Lindahl, Gromacs 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J. Chem. Theory Comput. **4**(3), 435–447 (2008)
22. 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 **65**, 712–725 (2006)
23. C.R. Johnson, R.A. Horn, *Matrix Analysis,* Chapter 8 (Cambridge University Press, Cambridge, 1990)
24. W.L. Jorgensen, D.S. Maxwell, J. Tirado-rives, Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. J. Am. Chem. Soc. **118**(45), 11225–11236 (1996)
25. A. Leach, *Molecular Modelling: Principles and Applications,* Chapter 5, 2nd edn. (Addison Wesley Longman Limited, Reading, 2003)
26. E. Lindahl, B. Hess, D. van der Spoel, Gromacs 3.0: a package for molecular simulation and trajectory analysis. J. Mol. Model. **7**(8), 306–317 (2001)
27. E. Lyman, D.M. Zuckerman, Ensemble-based convergence analysis of biomolecular trajectories. Biophys J. **91**(1), 164–172 (2006)
28. S. Marsili, G.F. Signorini, R. Chelli, M. Marchi, P. Procacci, Orac: a molecular dynamics simulation program to explore free energy surfaces in biomolecular systems at the atomistic level. J. Comput. Chem. **31**(5), 1106–1116 (2010)
29. L. Monticelli, E. Sorin, D. Tieleman, V. Pande, G. Colombo, Molecular simulation of multistate peptide dynamics: a comparison between microsecond timescale sampling and multiple shorter trajectories. J. Comput. Chem. **29**(11), 1740–1752 (2008)
30. Y. Mu, D.S. Kosov, 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**(21), 5064–5073 (2003)
31. J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale, K. Schulten, Scalable molecular dynamics with namd. J. Comput. Chem. **26**, 1781–1802 (2005)
32. C. Schütte, Conformational Dynamics: Modelling, Theory, Algorithm and Application to Biomolecules. Habilitation thesis, Freie Universität Berlin (1999)
33. C. Schütte, A. Fischer, W. Huisinga, P. Deuflhard, A direct approach to conformational dynamics based on hybrid monte carlo. J. Comput. Phys. **151**(1), 146–168 (1999)
34. L.J. Smith, X. Daura, W.F. van Gunsteren, Assessing equilibration and convergence in biomolecular simulations. Proteins Struct. Funct. Genet. **48**, 487–496 (2002)
35. P.E. Smith, The alanine dipeptide free energy surface in solution. J. Chem. Phys. **111**(12), 5568–5579 (1999)
36. D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, H.J.C. Berendsen, Gromacs: fast, flexible and free. J. Comput. Chem. **26**(16), 1701–1718 (2005)
37. R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics* (Wiley-VCH, Weinheim, 2009)
38. M. Tuckermann, B. Berne, Molecular dynamics in systems with multiple time scales. J. Comput. Chem. **95**, 8362–8364 (1992)
39. M. Tuckermann, B. Berne, A. Rossi, Molecular dynamics in systems with multiple time scales. J. Comput. Chem. **94**, 1465–1469 (1991)
40. G.A. Worth, F. Nardi, R.C. Wade, Use of multiple molecular dynamics trajectories to study biomolecules in solution: the ytgp peptide. J. Phys. Chem. B **102**(32), 6260–6272 (1998)