Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

JOHANNES SCHMIDT-EHRENBERG
HANS-CHRISTIAN HEGE

Visual Analysis of Molecular Conformations by Means of a Dynamic Density Mixture Model

ZIB-Report 05-02 (July 2005)

Visual Analysis of Molecular Conformations by Means of a Dynamic Density Mixture Model

Johannes Schmidt-Ehrenberg Hans-Christian Hege

We propose an approach for transforming the sampling of a molecular conformation distribution into an analytical model based on Hidden Markov Models. The model describes the sampled shape density as a mixture of multivariate unimodal densities. Thus, it delivers an interpretation of the sampled density as a set of typical shapes that appear with different probabilities and are characterized by their geometry, their variability and transition probabilities between the shapes. The gained model is used to identify atom groups of constant shape that are connected by metastable torsion angles. Based on this description an alignment for the original sampling is computed. As it takes into account the different shapes contained in the sampled set, this alignment allows to compute reasonable average shapes and meaningful shape density plots. Furthermore, it enables us to visualize typical conformations.

1 Introduction

Molecules are flexible structures. They move, vibrate, and interact with other molecules and their environment. Understanding these movements and interactions is essential for the complete comprehension of structure-function relationships, including many aspects of drug design and intermolecular interactions. Information about possible shapes of a molecule is carried by a density in the molecule's state space. Given some reasonably sized molecule, Hybrid Monte Carlo (HMC) simulation techniques allow to compute a set of molecular configurations that approximates this probability density of molecular shapes in thermodynamic equilibrium (see, e.g., [10]). Thanks to the molecular dynamics step in the HMC algorithm, the resulting sequence of states (trajectory) also contains dynamic information. Thus, the molecule's space of shapes can be analyzed [15] for *metastable subsets,* i.e. for regions in shape space that will be left by the molecule with very low probability.

These metastable subsets of shape space can be understood as the different rough shapes the molecule typically can take. However, the subsets can still contain more than one mode of the shape density. These modes can again be understood as typical shapes between which the molecule can change easier than between metastable sets. Overall, this constitutes a hierarchy of metastable sets.

For the visualization of metastable conformations, a partitioning of the shape space is of interest that also separates the modes inside a metastable subset. A further challenge is to find a suitable alignment of the single configurations to each other. In particular, this is important for reasonable averaging and visualization of shapes.

In this paper we will present a method for analyzing molecular dynamics trajectories with respect to typical molecular shapes. After shortly surveying related work (Sect. 2), we will give some background on the treatment of circular coordinates (Sect. 3), in particular dihedral angles, on Hidden Markov Models (Sect. 4), and on the concept of Perron cluster cluster analysis (Sect. 5). Based on this we present

- a technique for partitioning the shape variations of a molecular dynamics trajectory into long-time changes and thermal fluctuations using Hidden Markov Models (Sect. 6)
- a method to determine a hierarchical decomposition of the molecule into rigid sets of interconnected atoms that are connected by metastable degrees of freedom (Sect. 7)
- a new alignment strategy that clearly separates the long-time shapes in cartesian coordinates while minimizing the variance induced by thermal fluctuations (Sect. 8).

Application of the approach is demonstated in Sect. 9. Conclusions and future work are presented in Sect. 10.

2 Related Work

Visual molecular analysis is well established and widely used in research and industry. The multifarious demands in chemical, biochemical and pharmaceutical applications have been addressed by commercial and academic software packages, offering a variety of visual representations of molecules as well as editing functions. However, specific tools for visual shape analysis based on molecular trajectories seem not to exist.

Identification of molecular conformations is a current research topic, see e.g. [16, 4, 3]. Recently, in [6] a method has been proposed - similar to our analysis step - for identification of the most important conformations of a biomolecular system from Metropolis Monte Carlo time series. The authors, however, do not aim at alignment and visualization of long-time shapes.

Alignment is a classical task in molecular science. When two molecules are to be compared in 3D space, alignment is necessary in order to eliminate differences caused by global rigid transformations. Kabsch [8, 9] gives a straightforward method for computing an alignment between two point sets. Pennec [11] develops an approach to align multiple point sets iteratively. Both methods are characterized in more detail in Sect. 8. Huitema and van Liere [7] describe techniques for interactive visualization of protein dynamics, utilizing the concept of essential dynamics [1]. They interpret the dynamics of a protein as a trajectory in a high dimensional space and employ covariance analysis to filter out large concerted motions. Results on visual analysis of metastable molecular conformations on base of time series have been presented in [14].

3 Statistics of Molecular Shapes

For the analysis of molecular shapes some coordinate system is needed that represents the essential aspects of a molecular shape. Since we are not interested in a molecule's absolute position or orientation in 3D space, Cartesian coordinates of the atom positions are not appropriate. Neglecting the need to distinguish between mirror symmetric molecular shapes, it would be sufficient to consider intra-molecular distances between atoms. However, it turns out that the triple of bond lengths, bond angles and dihedral angles is more suitable. Bond lengths and angles can be regarded as nearly constant with respect to the shape variations of interest here. The interesting changes thus can be expressed via *dihedral angles*, which are defined by a sequence of

four atoms where the respective angle is the angle between the two planes spanned by the first three and the last three atoms. Regarding the three bonds that sequentially connect the four atoms, the dihedral describes the rotation of the third bond relative to the first around the axis defined by the middle bond.

As bond lengths and angles are not of interest, we can describe molecular shapes in a coordinate space build by dihedral angles, which have a bounded and periodic value range $[0, 2\pi)$. For statistical analyses this periodicity has to be taken into account. To get statistical informations about our data we cannot apply standard techniques. Naive averaging of angular values may lead to invalid results, because periodicity is ignored. To overcome this problem, we can interpret every angular value α as a point $z(\alpha) = e^{i\alpha}$ on the unit circle in the complex plane. This representation intrinsically reflects the periodicity of angular values and is independent of the choice of an interval of periodicity. Averaging this set of points in the complex plane, we can define a reasonable mean angle $\bar{\alpha}$ by

$$\bar{R}\, e^{i\bar{\alpha}} = \frac{1}{N} \sum_{j=1}^{N} e^{i\alpha_j} \,. \tag{1}$$

As we want to set up a Hidden Markov Model in the space of dihedral angles, we need a probability density function for circular variables. The circular analogon to the normal distribution is the *von-Mises-* or *circular normal* distribution:

$$\mathrm{f_{1D}}(\alpha; \phi, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\left\{\kappa \cos(\alpha - \phi)\right\} \quad \text{with } \phi \in [0, 2\pi) \quad \text{and } \kappa \geq 0 \tag{2}$$

where I_n is the modified Bessel function of order n. f_{1D} is a unimodal distribution with maximum at $\alpha = \phi$ and is symmetric on the interval $[\phi - \pi, \phi + \pi]$. The mean angle $\bar{\alpha}$ from (1) turns out to be a maximum likelihood estimator for the *mean direction* ϕ. The respective maximum likelihood estimator for the *concentration parameter* κ is based on the amplitude $\bar{R}$ of the complex mean in Eq. (1): $I_1(\kappa)/I_0(\kappa) = \bar{R}$. High values of κ correspond to narrow distributions, while the minimal value ($\kappa = 0$) makes the von-Mises distribution uniform. For description of multidimensional distributions of angles we use a tensor product of von Mises distributions.

4 Hidden Markov Models

A *Markov chain* is a sequence of random variables $S_1, S_2, S_3, \ldots$ with state space I that fulfills the so called *Markov property*:

$$P(S_{n+1} = i_{n+1} | S_1 = i_1, \ldots, S_n = i_n) = P(S_{n+1} = i_{n+1} | S_n = i_n) \tag{3}$$

with $i_1, \ldots, i_{n+1} \in I$. If this conditional probability is independent of n, it can be described by a stochastic matrix $T = \{t_{ij}\}$ with $t_{ij} := P(S_{n+1} = j | S_n = i)$ where $i, j \in I$ and $\sum_{j \in I} t_{ij} = 1$. Such a Markov chain is called *homogeneous*.

Hidden Markov Models (HMM), see e.g. [13], are a two-stage probabilistic concept for explaining the course of a time series. The primary assumption of a HMM is that a given time series is based on the realization of a homogeneous Markov Chain with finite state space I. This realization is not directly observable, but only by its influence on the second stage of the model. The HMM associates every state of its Markov chain with a probability density function defined on the value space of the time series to be explained. Depending on the state of the Markov chain realization at a given instant in time, a sample of the respective probability density is

drawn as the observable value of the time series. A HMM is completely specified by the following parameters:

1. Markov chain start distribution $\pi_i = P(S_1 = i), \quad (i \in I)$,
2. Markov chain transition matrix $T = \{t_{ij}\}, \quad (i, j \in I)$, and
3. probability density functions associated to the states of the Markov chain.

Dealing with HMMs two questions are typically of interest: First, given a sequence of observations, what are the optimal parameters of the HMM to explain this sequence? And second, given a sequence of observations and the parameters of the HMM, what is the underlying sequence of Markov states? Both question are answered by maximum likelihood estimation, i.e. by choosing the unknown parameters such that the likelihood of the observation sequence gets maximal.

For the first question this leads to the Baum-Welch-algorithm, which is a special case of the iterative Expectation-Maximization algorithm [2]. The iteration assumes a given set of model parameters. In a first step probabilities for the hidden parts of the model, in our case the states of the Markov chain realization, are computed. In a second step new model parameters are computed based on these probabilities. In every iteration cycle the likelihood of the observation given the model parameters increases. The iteration is terminated when the amount of increase drops under a threshold.

Maximum likelihood estimation for the second question is done via the direct Viterbi algorithm. The estimated sequence of states is called *Viterbi path*.

The construction of a HMM for a given series of observations requires the choice of the form of the probability density functions. We use tensor products of von Mises distributions, Eq. (2).

Further, the number of states of the hidden Markov chain has to be determined. In general, there is no way to measure whether a used number of states is appropriate. The achievable likelihood, which is the optimization criterion of the Baum-Welch-algorithm, monotonously increases with the number of states. Thus, it does not have a local maximum that would define an optimal number. In Sect. 6 we will use the concept of metastability to find a suitable number of states.

5 Perron Cluster Analysis

The phrase *metastable conformation* indicates a dynamic aspect of molecular behavior: it denotes approximate molecular geometries that survive the fast oscillations of molecular dynamics. In mathematical terms a metastable conformation is an *almost invariant set of the ensemble*, i.e. a subset of the molecular state space, that a molecular trajectory will only leave after a long time.

To find these metastable subsets of the state space, molecular dynamics is described using a transfer operator approach [5]. The state space is decomposed into subsets and a transfer operator is constructed, that specifies transition probabilities between these sets. Due to the reversibility of the dynamics, spectral analysis of the transfer operator leads to a real valued spectrum with maximal eigenvalue $\lambda_{max} = 1$, while the corresponding eigenvector is constant. If the state space contains l metastable subsets, the $l - 1$ next largest eigenvalues are very close to 1. This so called *Perron Cluster* of eigenvalues can be identified by a spectral gap that separates it from the remaining smaller eigenvalues.

If l has been determined, the metastable subsets can be constructed using the l corresponding eigenvectors, which define a mapping of the states to an approximate simplex in l-dimensional Euclidean space (cf. [3] for details). We can associate the l simplex vertices with the l metastable subsets we are looking for. Applying a linear transformation in the l-dimensional space mapping

the simplex vertices onto the l vectors of an orthonormal basis, we get components for all the mapped sample points with respect to the orthonormal basis that approximately lie between 0 and 1 [18]. These can be interpreted as measures of membership to the respective metastable set. To turn this fuzzy and thereby ambiguous assignment into a definite one, we define a state space element to belong to the metastable set with the maximal membership value.

6 Adapting Hidden Markov Models

As the effort of fitting a HMM to a time series depends quadratically on the number of hidden states, we are interested in models with small numbers of states. Therefore we try to find as small as possible groups of dihedral angles that can be treated as independent from the rest of the molecule. In the first step we combine all dihedral angles into one HMM that rotate around the same bond, as these typically have a strong coupling.

To estimate the number of hidden states we start with a definite overestimate. After the Baum-Welch-iteration, we have a probabilistic decomposition of the molecule's shape space that is defined by the probability densities of the HMM. The transition matrix of the HMM defines a transfer operator on this decomposition and we can apply Perron cluster analysis (cf. Sect. 5). This groups the Markov states into metastable sets. In [6], these sets were reduced to single states with mixture densities. In contrast to that, we replace the mixtures by single von Mises distributions. The resulting HMM is again optimized using the Baum-Welch-algorithm.

To find further correlations, we determine the Viterbi paths of all HMMs and compute for every two paths x and y the following entropy based measure of association [12]:

$$U(x,y) = 2\,\frac{H(x) + H(y) - H(x,y)}{H(x) + H(y)} \tag{4}$$

where $H(x)$ and $H(y)$ are the state distribution entropies of the single paths and $H(x,y)$ denotes the entropy of the combined state distribution. The value of $U(x,y)$ will range between 0 and 1, with 0 representing complete independence; $U(x,y) = 1$ on the other hand indicates complete dependence. Thus, pairs of HMMs with high values of U are merged into one common HMM. An intial value for this merged HMM can be generated by building all possible combinations of states from both original HMMs. After the optimization, this HMM is again reduced by Perron Cluster Analysis.

7 Rigid Substructures

In the following, we will propose a policy to divide the shape variations of a molecular dynamics trajectory into long-time changes that lead to substantially different shapes, and thermal fluctuations around those shapes. We will use information from the HMMs that describe the various groups of dihedral angles in the molecule.

We distinguish between trivial HMMs with only one state and HMMs with multiple states. In case of a single state HMM, no hidden Markov chain exists and any shape variability is expressed by the variance of the corresponding probability density of this single state. For our purposes, we consider dihedral angles that are described by a single state HMM to be of constant shape, i.e. we interpret their complete shape variation as thermal noise.

If a dihedral angle is described by a HMM with multiple states, it changes between different shapes that are characterized by the corresponding probability densities. Therefore, the four atoms of the dihedral cannot be altogether in one rigid structure. Nevertheless, if we consider

parts of the trajectory where the Viterbi path remains in the same state, the dihedral can be treated as constant in these subtrajectories.

In order to perform an alignment that takes these insights about rigid substructures into account, we build up a tree that specifies for every step of the trajectory to which other steps it has to be aligned and with respect to which atoms. Every node specifies a set of atoms, every edge corresponds to a HMM. The atoms of a node are considered to build a substructure of constant shape. The atoms of a child node can be added to this structure, if the trajectory is resolved into subtrajectories whose Viterbi paths with respect to the connecting node's HMM stay constant.

After the preliminary determination of all maximal rigid sets of interconnected atoms, we perform the following steps:

1. Choose a rigid structure containing central atoms of the molecule to be the root node.
2. For all leaf nodes of the tree:
   - Follow the root path of the current leaf node and collect all atoms contained in the nodes along the path. We will call this the *current rigid structure*.
   - Check all unused HMMs for a dihedral angle that overlaps in three atoms with the current rigid structure. If a HMM meets this criterion, collect all atoms of the dihedrals described by this HMM and remove those atoms, that are already contained in the tree. From these atoms build a child node of the current leaf node and associate the connecting edge with the HMM.
   - Check all unused rigid substructures for an overlap of at least 3 atoms with the union of the current rigid structure and one of the newly created nodes. If you find such a rigid substructure add it to the respective newly created node.
3. While unused rigid structures exist, repeat step 2

8 Alignment

Since visualization of conformations takes place in Cartesian coordinates, it is necessary to assign global positions and orientations to the geometries and thereby to define a relative alignment between them. As has been detailed for example in [14], this can be done by superimposing the atomic positions via rigid-body transformations. On the one hand, a reference shape can be chosen to which all other shapes are pairwise aligned. On the other hand, an iterative algorithm can be used that in every step aligns a shape to the current mean of all other shapes. Although requiring some higher computational effort, this approach is superior to the first one, as it does not depend on the arbitrary choice of a reference.

In the following, we will introduce an extension of the second approach that uses the hierarchy of rigid structures constructed in Sect. 7. The tree of rigid structures specifies a hierarchy of atom sets together with sets of time steps in which the respective structure is considered to be constant. Therefore, we keep one mean for the atoms connected to the root node. If the HMM that connects a child node with the root has three possible states, we compute three means for the atoms corresponding to the child node. Only those time steps contribute to one of these means that have the same state in the Viterbi path of the HMM. If a grandchild node of this child node is connected by another HMM that again has three states, we have to deal with nine different mean structures for the atoms of the grandchild, because now two Viterbi paths with

three states each have to be considered and, as the HMMs are independent, all nine combinations can arise.

In the following, M will denote the number of atoms in the molecule and N the number of time steps in the trajectory. Further, we write $\mathcal{T}$ for the tree constructed in Sect. 7 and $C_1, \ldots, C_L$ for the *rigid groups of atoms* that correspond to the nodes of $\mathcal{T}$. It holds: $C_k \cap C_l = \emptyset,\ (k \neq l)$ and $\bigcup_{k=1}^{L} C_k = \{1, \ldots, M\}$.

If C_k is a rigid group of atoms, then C_k^* is the union of all groups of atoms corresponding to the nodes of $\mathcal{T}$ that build the connecting path from the root of $\mathcal{T}$ to the node corresponding to C_k. G_k is the number of subtrajectories for which C_k^* is considered to be of constant shape. We denote with S_{kg} ($g \in \{1, \ldots, G_k\}$, $k \in \{1, \ldots, L\}$) the set of time steps that are in the respective subtrajectories. It holds $S_{kg} \cap S_{kh} = \emptyset,\ (g \neq h;\ g, h \in \{1, \ldots, G_k\})$ and $\bigcup_{g=1}^{G_k} S_{kg} = \{1, \ldots, N\}$. With $S_k(t)$ we denote the set of time steps that contains a step t with respect to a C_k^*.

Let the original cartesian coordinates of all time steps be $\boldsymbol{x}_{\boldsymbol{a}}^{(\boldsymbol{t})}$, where $t \in \{1, \ldots, N\}$ indicates the time step and $a \in \{1, \ldots, M\}$ the atom. Associated with every time step t we assume a weight factor w_t and define $W_{kg} = \sum_{t \in S_{kg}} w_t$ and $W_k(t) = \sum_{t \in S_k(t)} w_t$. We also define the weighted barycentric coordinates

$$\hat{\boldsymbol{x}}_{\boldsymbol{a}}^{(\boldsymbol{t})} = \boldsymbol{x}_{\boldsymbol{a}}^{(\boldsymbol{t})} - \frac{\sum_{k=1}^{L} (1 - \frac{w_t}{W_k(t)}) \sum_{b \in C_k} \boldsymbol{x}_b^{(t)}}{\sum_{k=1}^{L} (1 - \frac{w_t}{W_k(t)}) \cdot |C_k|}, \tag{5}$$

where $|C_k|$ is the number of atoms in C_k.

The aligned coordinates are $\tilde{\boldsymbol{x}}_{\boldsymbol{a}}^{(\boldsymbol{t})} = R^{(t)} \hat{\boldsymbol{x}}_{\boldsymbol{a}}^{(\boldsymbol{t})} + \boldsymbol{q}^{(t)}$, with $R^{(t)}$ a rotation matrix and $\boldsymbol{q}^{(t)}$ a translation vector. The determination of $R^{(t)}$ and $\boldsymbol{q}^{(t)}$ for $t \in \{1, \ldots, N\}$ will be described in the following. We seek an alignment that minimizes

$$V = \sum_{k=1}^{L} \sum_{a \in C_k} \sum_{g \in G_k} W_{kg} \sum_{\substack{t,s \in S_{kg} \\ t \neq s}} \frac{w_t w_s}{W_{kg}^2} (\tilde{\boldsymbol{x}}_{\boldsymbol{a}}^{(\boldsymbol{t})} - \tilde{\boldsymbol{x}}_{\boldsymbol{a}}^{(\boldsymbol{s})})^2. \tag{6}$$

This is the sum of variances of all atoms of the molecule, but, for rigid groups of atoms where the trajectory decomposes into different sets of time steps ($G_k > 1$), the variances are computed per set (S_{kg}) and then summed up using the set weights W_{kg}. Solving $\nabla_{\boldsymbol{q}^{(r)}} V = 0$ for $r \in \{1, \ldots, N\}$ we get

$$\boldsymbol{q}^{(r)} = \frac{\sum_{k=1}^{L} \sum_{a \in C_k} \sum_{\substack{t \in S_k(r) \\ t \neq r}} \frac{w_t}{W_k(r)} \tilde{\boldsymbol{x}}_{\boldsymbol{a}}^{(\boldsymbol{t})}}{\sum_{k=1}^{L} \sum_{a \in C_k} \left(1 - \frac{w_r}{W_k(r)}\right)} \qquad (r \in \{1, \ldots, N\}) \tag{7}$$

In order to determine $R^{(r)}$, we isolate those parts of V that depend on $R^{(r)}$:

$$-2 w_r \sum_{k=1}^{L} \sum_{a \in C_k} \left[\sum_{\substack{t \in S_k(r) \\ t \neq r}} \tfrac{w_t}{W_k(r)} (\tilde{\boldsymbol{x}}_{\boldsymbol{a}}^{(\boldsymbol{t})} - \boldsymbol{q}^{(r)}) \right] R^{(r)} \hat{\boldsymbol{x}}_{\boldsymbol{a}}^{(\boldsymbol{r})}. \tag{8}$$

To minimize this term we have to perform a pairwise alignment of the barycentric coordinates of time step r to the barycentric coordinates of the mean of all other time steps that belong to the same alignment groups $S_k(r)$. On this basis, we perform the following algorithm:

1. Set $R^{(1)} = \mathbb{1}$ and $\boldsymbol{q}^{(1)} = 0$.
2. Initialize all $R^{(r)}$ and $\boldsymbol{q}^{(r)}$ for $r \in \{2, \ldots, N\}$ by aligning time step r to time step 1 with respect to C_1.
3. Loop over all time steps $r \in \{1, \ldots, N\}$ and recompute $\boldsymbol{q}^{(r)}$ using (7) and $R^{(r)}$ by minimizing Eq. (8), but assuming $L = 1$ in Eq. (7) and Eq. (8).
4. Repeat step 3, thereby continuously increasing the influence of atoms group C_k with $k > 1$, but keeping the maximal change of an atom position under a threshold by increasing the influence slow enough.
5. Stop the iteration when all atoms have full influence and the maximal change of an atom position drops under another, lower threshold.

The computation time of a single iteration depends linearly on the number of time steps and the number of atoms.

9 Results

To demonstrate the described algorithm we use a Hybrid Monte Carlo sampling of a pentane molecule with 15,000 samples and another sampling of trialanine with about 500,000 samples. The density of pentane describes nine shape clusters of which seven are metastable conformations. The relevant shape variations of pentane are described by two dihedral angles, each of which consists of four of the five carbon atoms. Both have three typical values 0.4π, π, and 1.4π which are found by fitting an HMM as described in Sect. 4. The analysis by HMMs took about 15 minutes. Hence, the first three atoms of the first dihedral angle can be considered to be a rigid structure. Fig. 1 shows a comparison between (1) an alignment that only superposes these three atoms and (2) our new approach that takes the three atoms as root structure and aligns the other two carbons with respect to their shape groups.

The first alignment looks only on three atoms and shifts most of the trajectory's shape variance to the disregarded rest of the molecule. Thus, it gives a clearly defined geometry for the aligned carbons, but blurs the rest of the molecule. Our new method results in more variance for the positions of the three base atoms, but delivers a much clearer image of the rest of the molecule. The structure of all the nine shape clusters is clearly visible. The computation of the new alignment took about 15 seconds on a Pentium4 1.8 GHz Notebook. To visualize the aligned trajectories we accumulate the density over all geometries [14], where a geometry is the wireframe representation of an aligned time step. In accumulating the density, we count for every node of a uniform grid, how many geometries overlap with it. The densitiy is then visualized using isosurfaces or direct volume rendering (Fig. 1). The respective visualization for the bigger molecule trialanine is depicted in Fig. 2.

10 Conclusion and Future Work

We have presented an algorithm that divides the shape variations of a molecular dynamics trajectory into long-time changes which lead to substantially different shapes and thermal fluctuations

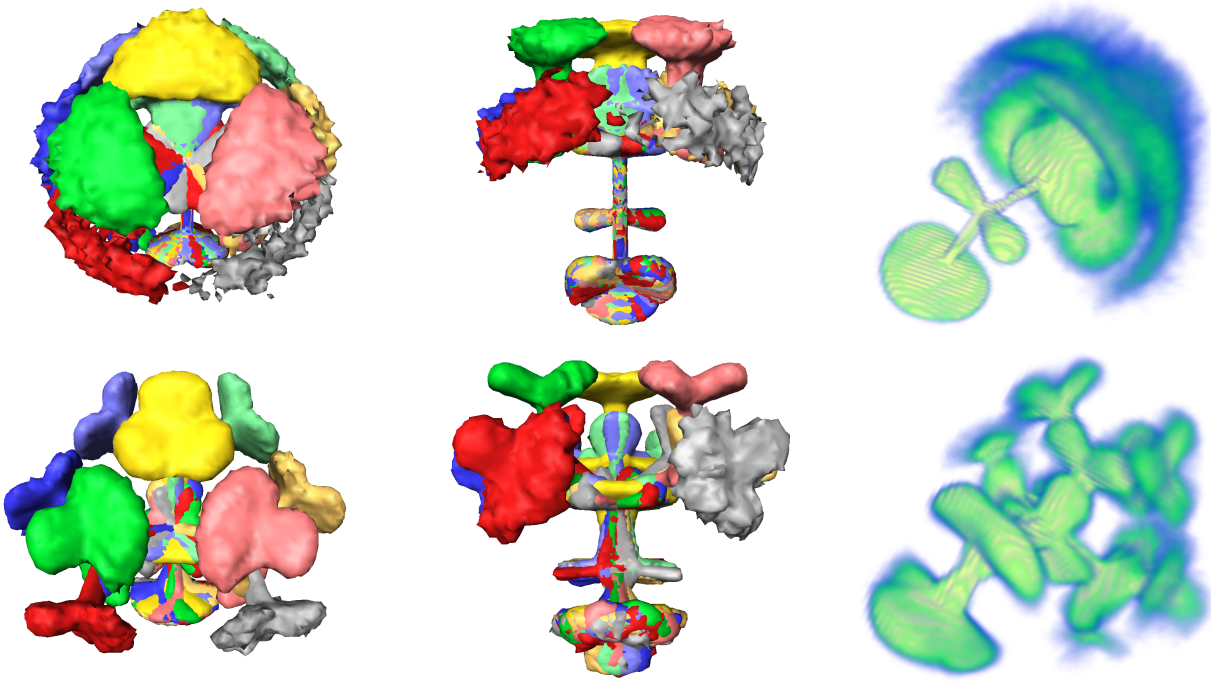

Figure 1: *Comparison of alignment strategies for a trajectory of the pentane molecule with 15,000 time steps that can be divided into 9 shape clusters. All images of a row depict the same 3D geometry from different viewing perspectives, while the two images in a column show approximately the same view on the geometries resulting from the two alignment strategies. In the left and middle column, the clusters are visualized by isosurfaces of their corresponding configuration densities. In the right column, the complete density is visualized by direct volume rendering.* Top row: *alignment by minimizing the positional variance of three carbon atoms.* Bottom row: *the new approach taking the same three carbons as the root structure which has constant shape over the whole trajectory. For the fourth carbon, three possible positions relative to the root structure are assumed and nine for the fifth atom.*

around those shapes. Groups of dihedral angles are analyzed by fitting Hidden Markov Models. In contrast to Perron cluster analysis based on uniform discretizations of dihedral angles, HMMs allow metastable clusters with fuzzy borders. We identified the fluctuation of the HMMs' distribution functions with thermal fluctuations of the molecule, while the state changes of multistate HMMs were interpreted as long-time changes. Thus, the combination of the different states of the multistate HMMs defines classes of different shapes. Using this decomposition we defined an alignment strategy that tries to mimimize the variance induced by thermal fluctuations. Thus, we got a clear depiction of the different shapes that the molecule takes on in its long-time changes. For the entire aligned trajectory, as well as the subtrajectories belonging to the long-time shapes, we accumulated configuration densities. These were visualized using isosurfaces and direct volume rendering. All the described techniques are integrated in the visualization system Amira [17]. Application to larger biochemically relevant molecules will be subject of further investigation. The algorithmic complexity of the analysis by HMMs depends linearly on the trajectory

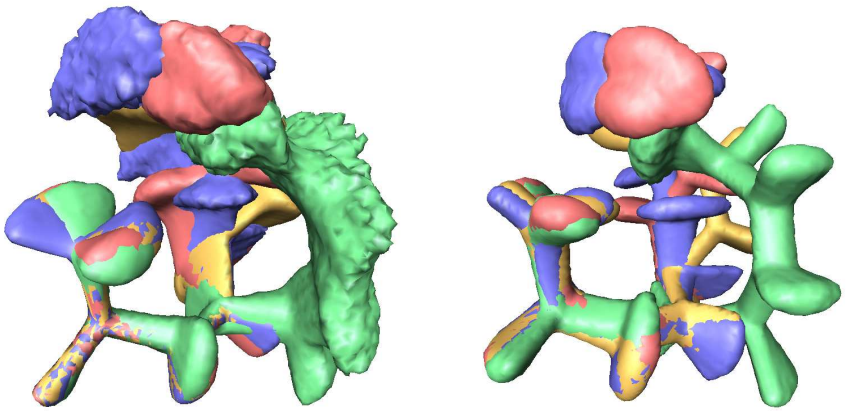

Figure 2: *Comparison of alignment strategies for a trajectory of trialanine with about 500,000 time steps. The 4 most important clusters are visualized by isosurfaces of their corresponding configuration densities.* Left image: *alignment by minimizing the positional variance of 5 selected atoms.* Right column: *the new approach taking the same 5 atoms as the root structure which has constant shape over the whole trajectory.*

length and the number of atoms. The number of necessary Markov states increases with the number of atoms and quadratically increases the computational effort. While we succeeded in lower dimensions by using random start condition we expect this to be problematic in higher dimensions. Regarding the alignment, we are not expecting relevant problems, as it has a linear dependence on the problem size.

11 Acknowledgments

We would like to thank Frank Cordes for kindly providing the data for the example molecules as well as for detailed discussions and Peter Deuflhard for continuous support.

References

[1] Andrea Amadei, A. B. M. Linssen, and Herman J. C. Berendsen. Essential dynamics of proteins. *Proteins: Structure, Function and Genomics*, 17(4):412–425, 1993.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Stat. Soc.*, 39(1):1–38, 1977.

[3] P. Deuflhard and M. Weber. Robust Perron cluster analysis of conformation dynamics. *Lin. Alg. Appl. (Special Issue on Matrices and Mathematical Biology)*, 398 C:161–184, 2005.

[4] Peter Deuflhard, Wilhelm Huisinga, Alexander Fischer, and Christof Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled markov chains. *Lin. Alg. Appl.*, 315:39–59, 2000.

[5] Peter Deuflhard and Christof Schütte. Molecular conformation dynamics and computational

drug design. In J. M. Hill and R. Moore, editors, *Applied Mathematics Entering the 21st Century. Proc. ICIAM 2003*, pages 91–119, Sydney, Australia, 2004.

[6] Alexander Fischer, Sonja Waldhausen, and Christof Schütte. Identification of biomolecular conformations from incomplete torsion angle observations by hidden markov models. Preprint, FU Berlin, Dept. of Mathematics and Computer Science, 2004.

[7] Henk Huitema and Robert van Liere. Interactive visualization of protein dynamics. In *Proceedings of IEEE Vis2000*, pages 465–468, October 2000.

[8] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica A*, 32:922–923, 1976.

[9] Wolfgang Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica A*, 34:827–828, 1978.

[10] B. Mehlig, D. W. Heermann, and B. M. Forrest. Hybrid Monte Carlo method for condensed-matter systems. *Phys. Rev. B*, 45:679–685, 1992.

[11] Xavier Pennec. Multiple registration and mean rigid shape - Application to the 3D case. In K.V. Mardia, C.A. Gill, and Dryden I.L., editors, *Image Fusion and Shape Variability Techniques (16th Leeds Annual Statistical Workshop)*, pages 178–185. University of Leeds, UK, July 1996.

[12] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C - Second Edition*, chapter 14, pages 632–636. Cambridge University Press, 1992.

[13] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of IEEE*, 77(2):257–286, February 1989.

[14] Johannes Schmidt-Ehrenberg, Daniel Baum, and Hans-Christian Hege. Visualizing dynamic molecular conformations. In *Proceedings of IEEE Visualization 2002*, pages 235–242, 2002.

[15] Christof Schütte. Conformational dynamics: Modelling, theory, algorithm, and application fo biomolecules. Habilitation Thesis, Dept. of Mathematics and Computer Science, Free University Berlin, 1998.

[16] Christof Schütte, Alexander Fischer, Wilhelm Huisinga, and Peter Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys.*, 151:146–168, 1999.

[17] Detlev Stalling, Malte Westerhoff, and Hans-Christian Hege. Amira: A highly interactive system for visual data analysis. In Charles D. Hansen and Christopher R. Johnson, editors, *The Visualization Handbook*, chapter 38, pages 749–767. Elsevier, 2005.

[18] Markus Weber. Clustering by using a simplex structure. Report 04-03, Zuse Institute Berlin, February 2004.