On metastable conformational analysis of non-equilibrium biomolecular time series *

Illia Horenko[**1] and **Christof Schütte**[***1]

[1] Institut für Mathematik II, Freie Universität Berlin
Arnimallee 2-6, 14195 Berlin, Germany

We present a recently developed clustering method and specify it for the problem of identification of metastable conformations in non-equilibrium biomolecular time series. The approach is based on variational minimization of some novel regularized clustering functional. In context of conformational analysis, it allows to combine the features of standard *geometrical clustering techniques* (like the K-Means algorithm), *dimension reduction methods* (like principle component analysis (PCA)) and *dynamical machine learning approaches* like Hidden Markov Models (HMMs). In contrast to the HMM-based approaches, no a priori assumptions about Markovianity of the underlying process and regarding probability distribution of the observed data are needed. The application of the computational framework is exemplified by means of conformational analysis of some penta-peptide torsion angle time series from a molecular dynamics simulation. Comparison of different versions of the presented algorithm is performed wrt. the *metastability* and *geometrical resolution* of the resulting conformations.

This is a preliminary version. Do not circulate!

Introduction

The field of simulation of large molecular systems has attracted enormous attention with applications ranging from materials science to modelling of highly complex biomolecules like proteins and DNA. Huge amounts of simulation data have been produced and the complexity and thus dimensionality of molecular dynamics simulations is simultaneuosly growing. However, the development of tools for the post-processing of such simulations is still in its infancies. The need for coarse-graining simulation results that allow understanding and appropriate visualization is increasing.

The macroscopic dynamics of typical biomolecular systems is mainly characterized by the existence of biomolecular conformations which can be understood as geometries or structures that are persistent for long periods of time. A typical biomolecular systems possess only few dominant conformations that can be understood as metastable states in state or configuration space [1, 2, 3]. In other words, the effective or macroscopic dynamics is given by a process that hops between the metastable states while within these states some geometric, statistical or dynamical patterns or features are persistent thus being characteristic for the state.

There are a manifold of approaches to the identification of patterns, clusters or features in complex data. In the context of molecular dynamics data the most prominent are geometrical clustering methods like K-Means and fuzzy K-Means (F-K-Means) [4, 5], dimension reduction methods like principal component analysis (PCA) or its variants [6, 7], and Markov- and hidden Markov approaches like HMM-SDE, or HMM-PCA [8, 9, 10]. Unfortunately all of these methods have certain shortcomings: (i) K-Means does not incorporate the dynamical information and is bad for overlapping data; the same is valid for fuzzy-K-Means, (ii) Markov approaches scale unfavorably with dimension and sometimes suffer from long-term memory in the data, (iii) HMMs rely on assumptions about the Markovianity of the hidden process and need an explicit functional form of the probability distribution or likelihood.

* Supported by the DFG research center MATHEON "Mathematics for key technologies" in Berlin.
** E-mail: horenko@math.fu-berlin.de
*** E-mail: schuette@math.fu-berlin.de

We will present an application of the newly developed adaptive FEM-clustering technique [11, 12, 13] to the conformational analysis of biomolecular time series data. The approach is based on finite element method (FEM) discretization of the regularized clustering functional. The clustering functional measures the quality of describing the time series in terms of a fixed number of *local models*. The main structural advantage of the method in biomolecular context is the following: it allows to combine standard *geometrical clustering* or *dimension reduction techniques* (like K-Means or PCA) with *dynamical machine learning approaches* like HMMs. In contrast to HMM-based approaches [14, 8, 10, 9], no a priory assumptions about the Markovianity of the underlying process and the probability distribution of the observed data are needed.

We will demonstrate how to apply the FEM-Clustering framework to identification of the conformational states for the realistic penta-peptide simulation data. In particular, we will compare different forms of the *model distance functional*, investigate their influence on the conformational resolution and compare the resulting mean metastable geometrical structures.

The paper is organized as follows: In Sec. 1 we will follow the modelling steps that relate the concept of locally optimal representation of the data to a specific constrained minimization problem. We then will discuss why and how the problem is regularized and then discretized by FEM. The resulting FEM clustering algorithm contains some free parameters; in Sec. 2 we will consider how to choose these parameters, and what may be the possible pitfalls. Finally, in Sec. 3 we will perform some numerical experiments on realistic molecular dynamics simulation data for a penta-peptide.

1 Finite Element Clustering Method

In the following, we will briefly described the *FEM-Clustering* framework first in general introduced in [11, 12, 13]. Special emphasis will be paid to the numerical and interpretational aspects of the framework in context of high dimensional biomolecular applications.

1.1 Model distance functional

Let $x(t) : [0,T] \to \Psi \subset \mathbf{R}^n$ be the analyzed molecular dynamics (MD) time series describing some molecular degrees of freedom (like atomic coordinates, intramolecular distances, or some torsion angles as functions of time). In the concluding section on numerical experiments we will look at the time series of the essential torsion angles describing the geometrical form of the molecular backbone (n is thus defined by the number of the torsion angles considered). In order to identify the $\mathbf{K}$ *conformational states* characteristic for the analyzed molecular system, we look for $\mathbf{K}$ different *local models* characterized by $\mathbf{K}$ distinct sets of a priory unknown *model parameters*

$$\theta_1, \ldots, \theta_{\mathbf{K}} \in \Omega \subset \mathbf{R}^d, \tag{1}$$

(where d is the dimension of a model parameter space) for the description of the observed time series. That is, the conformational states are *implicitly* characterized by certain patterns related to specific values of the associated parameters. Let

$$g\left(x_t, \theta_i\right) \quad : \quad \Psi \times \Omega \to [0, \infty)\,, \tag{2}$$

be a functional describing the *distance* from the observed molecular configuration $x_t = x(t)$ to the *model* i. In such a case $g(x_t, \theta)$ has to be chosen so that it measures the deviation of x_t from the pattern identified by θ. For a given *model distance functional* g, under *data clustering* we will understand the problem of finding for each t a vector $\Gamma(t) = (\gamma_1(t), \ldots, \gamma_{\mathbf{K}}(t))$ called the *affiliation vector* (or vector of the *cluster weights*) together with model parameters $\Theta = (\theta_1, \ldots, \theta_{\mathbf{K}})$ which minimize the functional

$$\mathbf{L}(\Theta, \Gamma) \quad = \quad \int_0^T \sum_{i=1}^{\mathbf{K}} \gamma_i\left(t\right) g\left(x_t, \theta_i\right) dt \to \min_{\Gamma, \Theta}, \tag{3}$$

subject to the constraints on $\Gamma(t)$:

$$\sum_{i=1}^{\mathbf{K}} \gamma_i(t) = 1, \quad \forall t \in [0,T] \tag{4}$$

$$\gamma_i(t) \geq 0, \quad \forall t \in [0,T], \quad i = 1, \ldots, \mathbf{K}. \tag{5}$$

That is, for each time t the affiliations $\gamma_i(t), i = 1, \ldots, K$ tell us whether the observation x_t belongs to a certain pattern/cluster (i.e., all $\gamma_i(t)$ except one are small), or whether x_t cannot be assigned clearly (i.e., several affiliations are significantly different from 0).

When considering this constrained minimization of $\mathbf{L}$ we obviously have to specify the regularity class of Γ, i.e., the function space from which the affiliations Γ may be taken. We will turn unto this question next but first we will give an example of two basic forms of the *model distance functional* (2) relevant for two important classes of cluster models: (I) *geometrical clustering* and (II) *dynamical clustering based on the principle components (dominant PCA modes.*

Example (I): Geometrical Clustering One of the most popular techniques of biomolecular conformational analysis is the so-called *K-means algorithm.* It is based on the iterative minimization of the distance of molecular configurations to a set of K *cluster centers* $\theta_i, i = 1, \ldots, K$. The affiliation to a certain cluster i is defined by the proximity of the observed molecular configuration $x_t \in \Psi$ to the cluster center $\theta_i \in \Psi$. In this case, the *model distance functional* (2) takes the form of the square of the simple Euclidean distance between the points in n dimensions:

$$g(x_t, \theta_i) = \| x_t - \theta_i \|^2, \tag{6}$$

or some weighted variant of it.

Example (II): PCA clustering In many cases the dimensionality of the data x_t can be reduced to few *essential degrees of freedom* without significant loss of information. Dimension reduction becomes crucial when analyzing the data from realistic biomolecular systems (since the dimension n of the analyzed data becomes a limiting factor in the numerical computation). One of the most popular *dimension reduction* approaches used in applications is the method of *essential orthogonal functions (EOFs)* also well-known under the name of *principal component analysis (PCA)* [15]. As was demonstrated recently, it is possible to construct clustering methods based on the decomposition of data sets according to differences in their *essential degrees of freedom* allowing to analyze data of very high dimensionality [10, 16, 17, 18]. If the cluster i is characterized by a linear m-dimensional manifold ($m \ll n$) of *essential degrees of freedom*, the respective model parameter is defined by the corresponding orthogonal projector $\theta_i = \mathcal{T}_i \in \mathbf{R}^{n \times m}$ onto this manifold and the *model distance functional* (2) is given by the Euclidean distance between the original data $\{x_t\}$ and its orthogonal projection on the manifold:

$$g(x_t, \theta_i) = g(x_t, \mathbf{T}_i) = \| x_t - \mathcal{T}_i \mathcal{T}_i^{\mathbf{T}} x_t \|^2 . \tag{7}$$

In context of molecular dynamics, this manifolds describe the *directions of maximal flexibility* of the molecule and can help to understand the differences in its basic dynamical behavior.

1.2 Regularization

We have to minimize $\mathbf{L}$ subject to the constraints (1) and (4-5). The expression in (3) is similar to one that is typically used in the context of *finite mixture models* [19, 20] but is more general, since neither the function $g(\cdot, \cdot)$ nor $\Gamma(\cdot)$ have to be connected to some probabilistic models of the data (which is the case for *finite mixture models*).

However, direct treatment of the problem (3) is hampered by the three following facts: (i) the optimization problem is *infinitely-dimensional* (since $\Gamma(t)$ belongs to some not yet specified function class), (ii) the problem is *ill-posed* since the number of unknowns can be higher then the number of known parameters, and (iii) because of the non-linearity of g the problem is in general *non-convex* and

the numerical solution gained with some sort of *local minimization algorithm* depends on the initial parameter values [21].

It perhaps is even more important that the solution Γ of the above constrained minimization task may be unregular function: To see this let us assume that we already know the minimizer values Θ_* for Θ. Then, the minimizer Γ_* for the affiliation vector Γ has the following form:

$$\gamma_{*,i}(t) = \begin{cases} 1 & \text{if } i = \text{argmin}_j \; g(x_t, \theta_{*,j}) \\ 0 & \text{otherwise} \end{cases}, \tag{8}$$

that is, the datum x_t has perfect affiliation with state i if the model distance functional for x_t is minimal in state i. That is, if the process exhibits strong variability then the affiliations are rather non-smooth functions. Whenever the affiliation functions just take values 0 or 1 we will call them *deterministic* in the following, which is meant in the sense that then for every datum in the time series it is certain to which cluster it belongs.

One of the possibilities to approach the last problem together with the problems (i)-(ii) simultaneously is first to incorporate some *additional information* about the regularity of the observed process (e.g., in the form of *smoothness assumptions* in space of functions $\Gamma(\cdot)$) and then apply a finite Galerkin-discretization of this infinite-dimensional Hilbert space. For example, we can assume the *weak differentiability* of functions γ_i, i. e.:

$$|\gamma_i|_{\mathcal{H}^1(0,T)} = \| \partial_t \gamma_i(\cdot) \|_{\mathcal{L}_2(0,T)} = \int_0^T \left(\partial_t \gamma_i(t)\right)^2 dt \leq C_\epsilon^i < +\infty, \quad i = 1, \ldots, \mathbf{K}. \tag{9}$$

As was demonstrated in [11], the above constraint limits the total number of transitions between the clusters and is connected to the *metastability* of the *hidden process* $\Gamma(t)$.

Another possibility to incorporate *a priori information* from (9) into the optimization is to modify the functional (3) and to write it in the *regularized* form

$$\mathbf{L}^\epsilon(\Theta, \Gamma, \epsilon^2) \;=\; \mathbf{L}(\Theta, \Gamma) + \epsilon^2 \sum_{i=1}^{\mathbf{K}} \int_0^T \left(\partial_t \gamma_i(t)\right)^2 dt \to \min_{\Gamma \in \mathcal{H}^1(0,T), \Theta}. \tag{10}$$

This form of penalized regularization was first introduced by A. Tikhonov for solution of ill-posed linear least-squares problems [22] and was frequently used for non-linear regression analysis in context of statistics [23] and multivariate spline interpolation [24]. In contrast to the aforementioned applications of Tikhonov-type regularization (where the regularization is controlling the smoothness of some non-linear functional approximation of the given data), the regularization of the *averaged clustering functional* (10) allows to control the *metastability* of the assignment $\Gamma(t)$ of the given data to $\mathbf{K}$ distinct a priory unknown clusters, cf. [11].

1.3 FEM-discretization

Let $\{0 = t_1, t_2, \ldots, t_{N-1}, t_N = T\}$ be a finite subdivision of the time interval $[0, T]$ with uniform timestep Δ_t. We can define a set of continuous functions $\{v_1(t), v_2(t), \ldots, v_N(t)\}$ called *hat functions* or *linear finite elements* [25]

$$v_k(t) \;=\; \begin{cases} \frac{t - t_k}{\Delta_t} & 2 \leq k \leq N-1, t \in [t_{k-1}, t_k], \\ \frac{t_{k+1} - t}{\Delta_t} & 2 \leq k \leq N-1, t \in [t_k, t_{k+1}], \\ \frac{t_2 - t}{\Delta_t} & k = 1, t \in [t_1, t_2] \\ \frac{t - t_{N-1}}{\Delta_t} & k = N, t \in [t_{N-1}, t_N]. \end{cases} \tag{11}$$

Assuming that $\gamma_i \in \mathcal{H}^1(0, T)$, standard FEM theory tells us we can approximate the continuous solution by means of a Galerkin ansatz with ansatz space

$$V_N = \text{span}\{v_k,\, k = 1, \ldots, N\}.$$

The discretized constrained minimization task then reads

$$\tilde{\mathbf{L}}^{\epsilon} = \sum_{i=1}^{\mathbf{K}} \left[a(\theta_i)^{\mathbf{T}} \tilde{\gamma}_i + \epsilon^2 \tilde{\gamma}_i^{\mathbf{T}} \mathbf{H} \tilde{\gamma}_i \right] \rightarrow \min_{\tilde{\gamma}_i, \Theta}, \tag{12}$$

$$\sum_{i=1}^{\mathbf{K}} \tilde{\gamma}_{ik} = 1, \quad \forall k = 1, \ldots, N, \tag{13}$$

$$\tilde{\gamma}_{ik} \geq 0, \quad \forall k = 1, \ldots, N, \quad i = 1, \ldots, \mathbf{K}, \tag{14}$$

where $\tilde{\gamma}_i = (\tilde{\gamma}_{11}, \ldots, \tilde{\gamma}_{iN})$ is the vector of *discretized affiliations* to cluster i, yielding the approximate affiliations

$$\tilde{\gamma}_i^N(t) = \sum_{k=1}^{N} \tilde{\gamma}_{ik} v_k(t),$$

while

$$a(\theta_i) = \left(\int_{t_1}^{t_2} v_1(t) g(x_t, \theta_i) dt, \ldots, \int_{t_{N-1}}^{t_N} v_N(t) g(x_t, \theta_i) dt \right), \tag{15}$$

is a vector of *discretized model distances* and $\mathbf{H}$ is the symmetric tridiagonal *stiffness-matrix* of the linear finite element set with $2/\Delta_t$ on the main diagonal, $-1/\Delta_t$ on both secondary diagonals and zero elsewhere.

Standard FEM theory tells us that the solution $\tilde{\gamma}^N$ of the discretized problem converges to the continuous one γ for $N \rightarrow \infty$. However for finite N we will have to face a discretization error. Whenever we solve the above minimization problem (12-14) for fixed cluster model parameters, we will denote the discretization error by $\delta_N = \|\gamma - \tilde{\gamma}^N\|_{\mathcal{L}^2(0,T)}$. Standard techniques from numerical mathematics are available using reliable estimators of the discretization error for controlling N adaptively such that some pre-defined tolerance is undercut [25].

Whenever the process under investigation is not observed in continuous time but in form of a time series with discrete time lags, then the discretized model distances a from (15) have to be computed based on these discrete information. That is, the integrals in (15) have to be replaced by appropriate quadrature formula. This poses no problem as long as the time lags in the time series are sufficiently small compared to the uniform timestep Δ_t of the FEM discretization.

If $\epsilon^2 = 0$, then the above minimization problem (12-14), can be solved analytically wrt. $\tilde{\gamma}_{il}$ for a fixed set of *cluster model parameters* Θ resulting in

$$\tilde{\gamma}_{il} = \begin{cases} 1 & i = \operatorname{argmin}_j \int_{t_l}^{t_{l+1}} v_l(s) g(x_s, \theta_j) ds, \\ 0 & \text{otherwise}, \end{cases} \tag{16}$$

in analogy to the continuous solution (8) for $\epsilon^2 = 0$.

If $\epsilon^2 > 0$, for some *fixed set of cluster model parameters* $\Theta^{(l)}$, the minimization problem (12-14) reduces to a sparse quadratic optimization problem with linear constraints which can be solved by standard tools of sparse quadratic programming (sQP) with computational cost scaling as $\mathcal{O}(N \log(N))$ [26].

In addition, the minimization problem (12 -14) wrt. the parameters Θ for a *fixed* set of *discretized cluster affiliations* $\tilde{\gamma}_i$ is equivalent to the *unconstrained minimization problem*

$$\sum_{i=1}^{\mathbf{K}} a(\theta_i)^{\mathbf{T}} \tilde{\gamma}_i^{(l)} \rightarrow \min_{\Theta}. \tag{17}$$

For both of the K-Means and K-EOFs *model distance functionals* (6,7), the above problem (17) can be solved *analytically* and *explicit* estimators for the cluster parameters can be given.

Therefore, the resulting *FEM-clustering algorithm* can be implemented as the following iterative numerical scheme:

FEM-clustering Algorithm.

Setting of optimization parameters and generation of initial values:

· Set the number of clusters $\mathbf{K}$, regularization factor ϵ^2, discretization error tolerance δ, additional internal parameters of the distance model g, and the optimization tolerance TOL

· Set the iteration counter $l = 1$

· Choose random initial $\tilde{\gamma}_i^{(1)}, i = 1, \ldots, \mathbf{K}$ satisfying (13-14)

· Calculate $\Theta^{(1)} = \underset{\Theta}{\operatorname{argmin}}\, \tilde{\mathbf{L}}^{\epsilon}\left(\Theta, \tilde{\gamma}_i^{(1)}\right)$ solving the problem (17)

Optimization loop:

do

· Compute $\tilde{\gamma}^{(l+1)} = \operatorname{argmin}_{\tilde{\gamma}}\, \tilde{\mathbf{L}}^{\epsilon}\left(\Theta^{(l)}, \tilde{\gamma}\right)$ satisfying (13-14) by applying sQP and adapting N until the discretization error is less than δ (for $\epsilon^2 > 0$), or by applying (16) (if $\epsilon^2 = 0$)

· Calculate $\Theta^{(l+1)} = \underset{\Theta}{\operatorname{argmin}}\, \tilde{\mathbf{L}}^{\epsilon}\left(\Theta, \tilde{\gamma}_i^{(l+1)}\right)$ solving the problem (17)

· $l := l + 1$

while $\left|\tilde{\mathbf{L}}^{\epsilon}\left(\Theta^{(l)}, \tilde{\gamma}_i^{(l)}\right) - \tilde{\mathbf{L}}^{\epsilon}\left(\Theta^{(l-1)}, \tilde{\gamma}_i^{(l-1)}\right)\right| \geq \mathrm{TOL}$.

Major advantage of the presented algorithm compared to HMM-based strategies [14, 9, 17, 18] and to finite mixture models [19, 20] is that no a priori assumptions about the probability model for hidden and observed processes are necessary in the context of the *FEM-clustering algorithm*.

2 Selection of parameters

The quality of the clustering very much depends on the original data, especially on the length of the available time series. The shorter the observation sequence is, the bigger the uncertainty of the resulting estimates. The same is true, if the number $\mathbf{K}$ of the hidden states is increasing for the fixed length of the observed time series: the bigger $\mathbf{K}$, the higher will be the uncertainty for each of the resulting clusters. Therefore, in order to be able to statistically distinguish between different hidden states, we need to get some notion of the *model robustness*, and how it is influenced by the selection of the cluster number $\mathbf{K}$, the regularization factor ϵ^2, and the details of the model distance functional g. This can be achieved through the postprocessing of the clustering results and analysis of the transition process and model parameters estimated for each of the clusters.

Identification of optimal K. Let us assume for a moment, that ϵ^2 and the details of g are fixed, and reflect on the question of how to select $\mathbf{K}$. If there exist two states with overlapping confidence intervals for each of the respective model parameters, then those are statistically indistinguishable, $\mathbf{K}$ should be reduced and the optimization repeated. In other words, confidence intervals implicitly give a natural upper bound $\mathbf{K}_{max}$ for the number of possible clusters. Algorithmically, one starts with some large $\mathbf{K}$, performs clustering, compares the confidence intervals of the resulting cluster model parameters and sets $\mathbf{K} = \mathbf{K} - 1$ if the confidence intervals are overlapping. If at certain step of this procedure all of the confidence intervals are non-overlapping, the correspondent value $\mathbf{K}$ is equal to the *maximal number of statistically distinguishable robust clusters* $\mathbf{K}_{max}$ for given data and some chosen *model distance functional*.

Another possibility to estimate the optimal number of clusters can be used, if the identified transition process $\Gamma(t)$ is shown to be Markovian for given $\mathbf{K}, \epsilon^2$. Markovianity can be verified applying some standard tests, e. g., one can check the *generator structure* of the hidden process, see [27]. In such a case the hidden transition matrix can be calculated and its spectrum can be examined for a presence of the *spectral gap*. If the *spectral gap* is present, then the number of the dominant eigenvalues (i.e., eigenvalues between the *spectral gap* and 1.0) gives the number of the metastable clusters in the system [28]. Positive verification of the hidden process' Markovianity has an additional advantage: it allows to construct a *reduced dynamical model* of the analyzed process and to estimate some dynamical characteristics of

the analyzed process, e.g., one can calculate *relative statistical weights*, *mean exit times* and *mean first passage times* for the identified clusters [18].

Choosing the model distance functional g. It should not be necessary to emphasize that the results of the minimization will strongly depend on the selection of the functional g. In Sec. 3 we will demonstrate that, for example, there are significant differences between PCA-based and geometrical clustering even if applied to the same time series. Furthermore the selection of internal parameters of g will also be decisive. For example, when considering PCA-based clustering we can choose different values for the dimension m of the low-dimensional manifold to be identified; we will comment on this in Sec. 3 also. However, in general there is no algorithmic scheme for choosing g or internal parameters in g; these choices should not be made without careful analysis of the system under consideration and of the properties of interests, perhaps in combination with appropriate model discrimination approaches.

Selection of regularization factor ϵ^2. As was demonstrated in [11], there is a connection between the *regularization factor* ϵ^2 and *metastability* of the resulting data decomposition. This means that respective mean exit times for the identified clusters get longer and the corresponding cluster decompositions become more and more *metastable*. Careful inspection of the transition process $\Gamma(t)$ identified for different values of ϵ^2 is essential for choosing the appropriate value of ϵ^2. In order to illustrate this let us consider the one-dimensional time series shown in Fig. 1. The data obviously exhibits two different states both with significant metastability. For $\epsilon = 0$ and $K = 2$, however, we find two cluster centers $\theta_1 = -1$ and $\theta_2 = 1$ and the affiliation functions get the form

$$\gamma_i(t) = \begin{cases} 1 & \text{if } i = 1, x_t <= 0 \text{ or } i = 2, x_t > 0 \\ 0 & \text{otherwise} \end{cases} ,$$

such that they are very rough functions, i.e., both exhibit almost permanent jumps between 0 and 1. The right hand side panel of Fig. 1 shows the FEM-Kmeans affiliation functions for several $\epsilon > 0$. We observe that for $\epsilon^2 = 0.1$ we get almost step functions with jumps in the right places while for larger ϵ^2 the functions are mollified and do no longer clearly separate the two obvious clusters/states. When increasing ϵ^2 further the affiliations tend to become constant functions since the the regularization part of the functional dominates the data-based part.

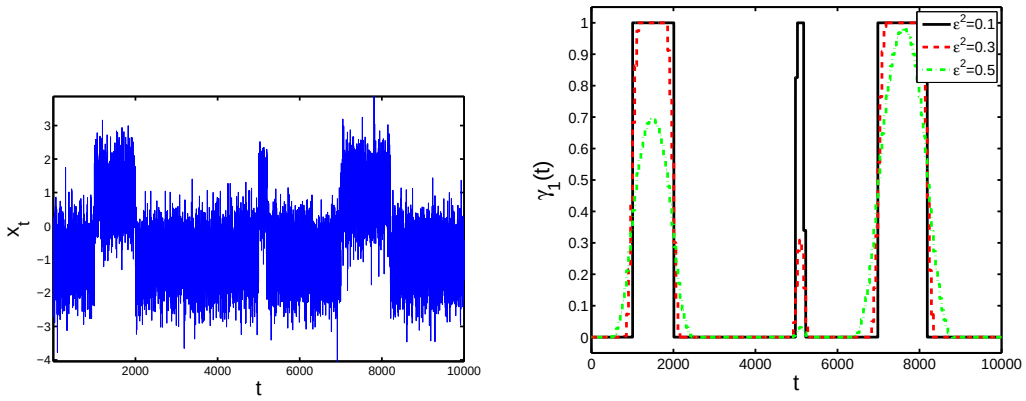

Fig. 1 One-dimensional time series (left) and FEM-Kmeans affiliation function γ_1 against time for $K = 2$ and different values of ϵ (right).

3 Analysis of a penta-peptide molecular dynamics trajectory

We will use simulation data of an artificial penta-peptide, consisting of a capped chain of five amino-acids: Glutamine-Alanine-Phenylalanine-Alanine-Argenine, shown in Fig. 2. The peptide is itself an

interesting object to study, as it is a small molecule which is able to form salt bridges, an important and still not well understood matter. We will not concern with this subject but rather use a trajectory of the peptide for demonstration purpose of our algorithm only. The trajectory was obtained from an MD-simulation in vacuum using the NWChem software package [**?**, 29]. The integration time step was set to 1 femtosecond, while the coordinates were written out every 200 femtoseconds. The trajectory we use consists of 100000 points thus covers a length of 20 nanoseconds. What can be seen in the trajectory is the folding of the peptide from a spread out structure where only the two long side chains interact (the salt bridge) to a more compact structure and very stable structure, see Fig. 2.

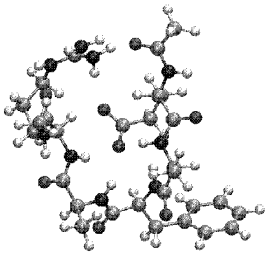

Fig. 2 Snapshot of the analyzed MD-trajectory of the penta-peptide, consisting of a capped chain of five amino-acids (Glutamine-Alanine-Phenylalanine-Alanine-Argenine).

In the subsequent section we will look at the time series of the essential torsion angles describing the geometrical form of the molecular backbone ($n = 10$ is thus the number of the torsion angles considered; periodicities were removed). For an illustration of the resulting time series see Fig. **??** below. The reduction of the original time series (atomic Euklidean coordinates) to torsion angles has been done mainly for the sake of illustration.

3.1 Clustering via FEM-Kmeans

First we apply the proposed algorithm based on the K-means distance functional (6).

As before, we chose the discretization error tolerance $\delta = 0.0001$. Again, for different, non-extreme values of the regularization factor ϵ the optimal number of clusters $\mathbf{K}_{max}$ was determined according to the procedure described in Sec. 2, see Fig. 8 below.

When applying FEM-K-means with the regularization factor $\epsilon^2 = 0.1$ and the associated optimal cluster number $\mathbf{K}_{max} = 4$, the resulting affiliation functions exhibit clear jumps such that we can again assign every time in the time series to exactly one cluster. This assignment results in the cluster assignment shown in Fig. 3. Fig. 3 also exhibits the coarse-grained description of the molecular dynamics time series in terms of the K-means distance functional (6): on the long time scale, the effective dynamics can be described as a *folding* process, i.e., starting with the unfolded conformation 3, the molecule finally *folds* into the β-sheet conformation 1 while passing through unfolded conformation 4 and partially-folded conformation 1. Respective mean configurations of the identified conformational states are shown in the left part of the Fig. 3 in form of 3D probability density plots of the backbone (generated by all states of the time seires in the respective conformation).

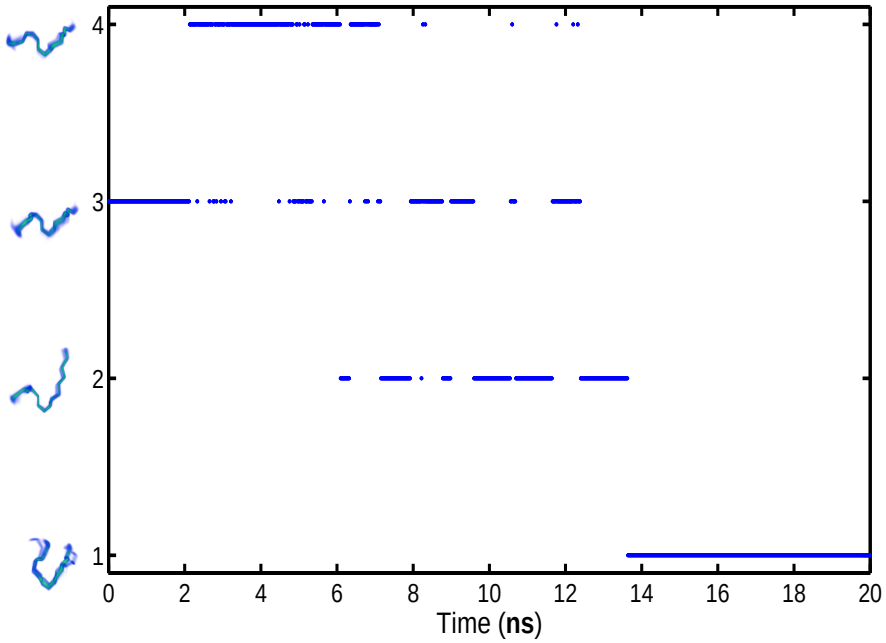

Fig. 3 Cluster assignment identified for $\epsilon = 0.1$, $\mathbf{K}_{max} = 4$, and $\delta = 0.0001$ obtained via FEM-K-means. 3D probability density plots of the respective cluster states are shown in the left side of the plot. See text for further explanation.

3.2 Clustering via FEM-KPCA

We now can repeat the same numerical experiment based on the PCA model distance functional (7). We select $m = 1$, i.e., just the most flexible mode is used for determining distances. We chose the same discretization error tolerance $\delta = 0.0001$ as in the previous example. Based on these parameter settings and for different, non-extreme values of the regularization factor ϵ the optimal number of clusters $\mathbf{K}_{max}$ was determined according to the procedure described in Sec. 2, see Fig. 8 below.

Application of FEM-KPCA then leads to the affiliation functions as shown in Fig. 4 for different values of the regularization factor ϵ (and $\mathbf{K} = 8$). We observe that the functions for $\epsilon = 0$ show many jumps between purely deterministic affiliations, while for $\epsilon^2 = 0.5$ they still exhibit rather sudden (but less frequent transitions between otherwise almost deterministic affiliations, and for $\epsilon^2 = 1$ the jumps become mollified and cluster affiliations are no longer deterministic but "'fuzzy"'.

For the remaining experiments with FEM-KPCA we choose the regularization factor $\epsilon^2 = 0.1$ and the associated optimal cluster number $\mathbf{K}_{max} = 8$. The resulting affiliation functions exhibit clear jumps such that we can assign every time in the time series to exactly one cluster (the one with the largest affiliation value). This assignment results in the coarse-grained description of the MD time series as shown in Fig. 5. Similar to the coarse-grained description in terms of the K-means distance functional (6), see Fig.3, the FEM-KPCA algorithm allows to interpret the overall dynamics as a *folding* process, starting with the unfolded state 1 and finally ending up in the folded β-sheet conformation 5.

Last but not least, one should ask whether the choice $m = 1$ is sensible based on the results of FEM-KPCA. Therefore we computed the covariance matrices of the eight clusters identified before and analysed the spectrum of these matrices. The results are collected in Table 3.2. We observe that in seven of the eight states there is one clearly most flexible mode which justifies the choice $m = 1$ in retrospective. However, the eighth state seems to be a kind of a collective transition state (see Fig. 5) that collects everything not belonging into the first seven states. It also has a lowest statistical weight among other states.

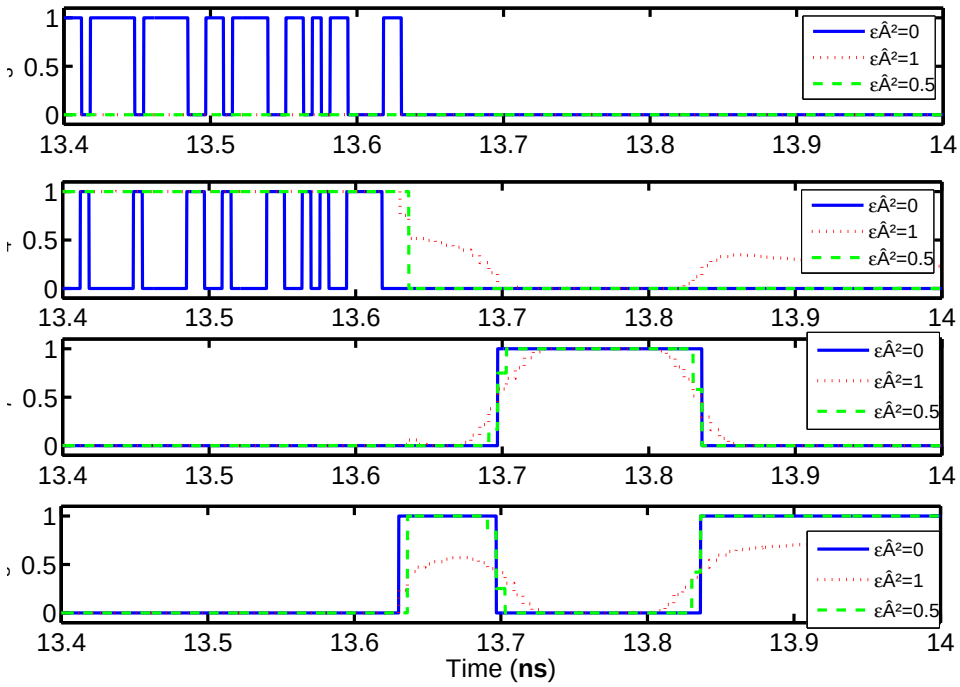

Fig. 4 Comparison of the affiliation functions γ_i as identified by FEM-KPCA for different values of the regularization factor ϵ, and $\delta = 0.0001$, $m = 1$, and cluster number $\mathbf{K} = 8$. For the sake of transparency the plots show only part of the time axis and just four of the eight affiliation functions.

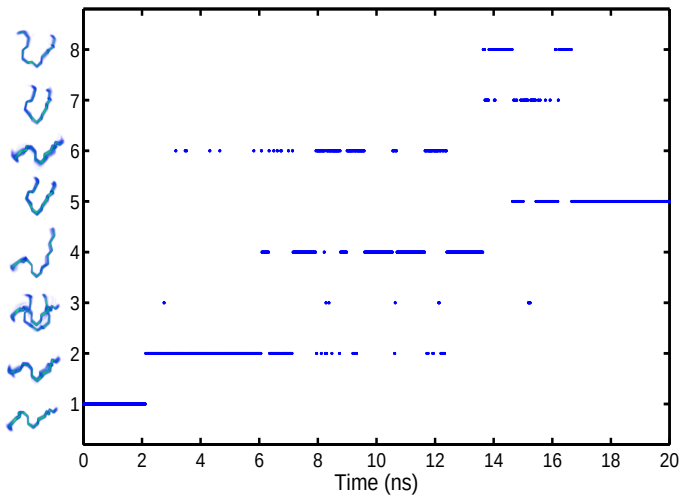

Fig. 5 Cluster assignment identified for $\epsilon = 0.1$, $\mathbf{K}_{max} = 8$, $m = 1$, and $\delta = 0.0001$ obtained via FEM-KPCA. 3D probability density plots of the respective cluster states are shown in the left side of the plot. See text for further explanation.

Cluster No.	1st ev.	2nd ev.	3rd ev.	4th ev.	5th ev.
1	699.9	365.4	335.9	193.9	144.8
2	772.2	413.1	350	318.3	217.7
3	26331	603	398	268	200
4	1157	392.3	322.8	293.1	242.6
5	1930	236.5	210.3	163.3	137.7
6	1029	613.5	560.3	329.7	201.8
7	1322	306.3	227.5	205.2	157.5
8	446.9	325.4	297.6	210.0	203.1

Table 1 Leading five eigenvalues of the covariance matrix for each of the eight clusters discussed in the text.

3.3 Comparison between FEM-KPCA and FEM-K-means

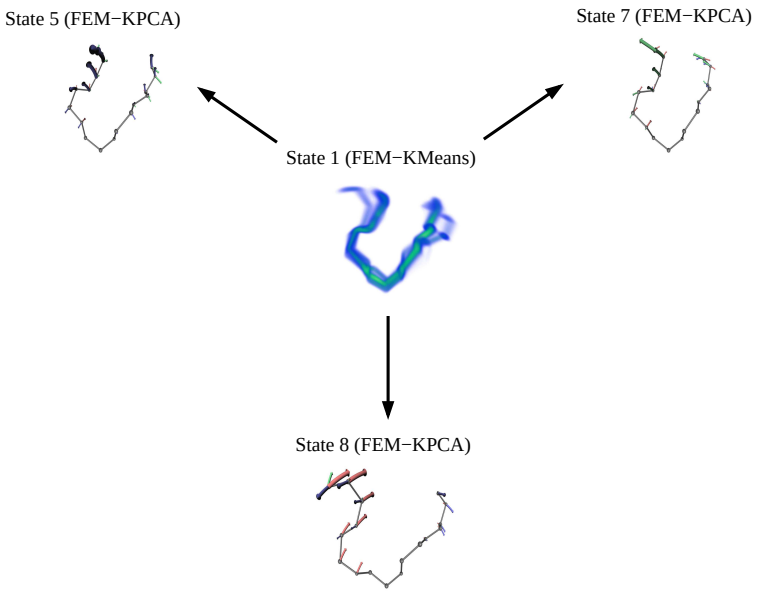

Fig. 6 Comparison of flexibility: The β-hairpin cluster state 1 (identified by the FEM-K-means algorithm, see Fig. 3) can be further subdivided in three cluster substates (identified by the FEM-KPCA algorithm, see Fig. 5). The arrows denote the dominant dynamical modes identified by the FEM-KPCA algorithm

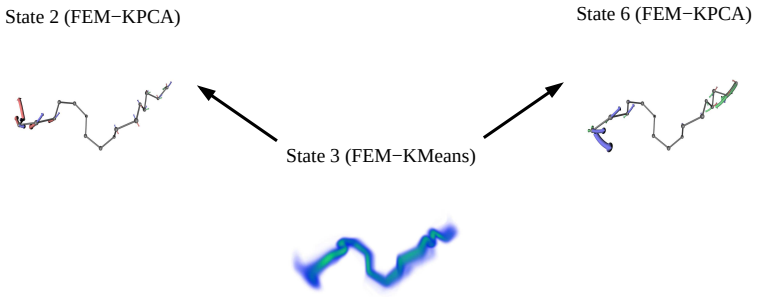

Fig. 7 Comparison of flexibility: the unfolded cluster state 3 (identified by the FEM-K-means algorithm, see Fig. 3) can be further subdivided in two cluster substates (identified by the FEM-KPCA algorithm, see Fig. 5). The arrows denote the dominant dynamical modes identified by the FEM-KPCA algorithm

By comparing Figs. 5 and 3, we see that the resulting cluster assignments describe qualitatively the same folding process. However, it seems that the FEK-KPCA approach reveals several *geometrically redundant* states having the same mean configurations (compare the configurations of states 5,7 and 8 or states 2 and 6, see Fig. 5). To explain this peculiarity, one has to recall the meaning of the PCA model distance functional (7): it allows to distinguish different conformational states due to the differences in m *essential dynamical modes* of the analyzed data. In context of molecular dynamics, these quantities describe the *normal modes* of the molecule in the respective state. It means that the conformations having the same (or very similar) mean configurations can be distinguished by the FEM-KPCA algorithm because of the differences in the local dynamics/flexibility. Figs. 6 and 7 demonstrate that application of the FEM-KPCA procedure allows to decompose the FEM-K-means states (e.g. states 1 and 3) into states with (almost) the same mean configurations but with different *essential flexibility*.

Therefore, when comparing the results of FEM-KPCA and FEM-K-means, the importance of the choice of the model distance functional g comes obvious. Let us now concentrate on the difference

of the optimal number $\mathbf{K}_{max}$ of statistically distinguishable clusters depending on the choice of the regularization parameter ϵ^2. Fig. 8 shows that for the data at hand, the resolution of FEM-KPCA is finer than that of FEM-K-means in the sense that FEM-KPCA allows to observe more statistically distinguishable clusters based on the same data for the same regularization factor. We should be aware, however, that this result depends on the scaling of the data under consideration. In any case we observe that for both approaches the number of clusters decreases with increasing regularization factor while there average metastability increases.

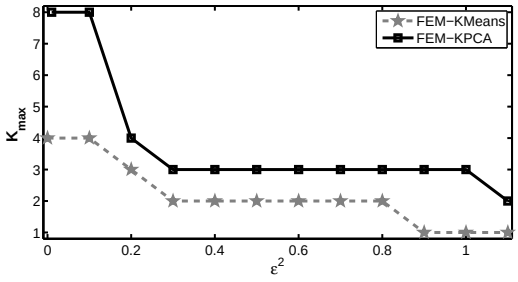

Fig. 8 Comparison of the *maximal number of statistically distinguishable conformational states* $\mathbf{K}_{max}$ obtained for different *regularization factors* ϵ^2. FEM-KPCA: black, solid; FEM-K-means: gray, dashed.

4 Conclusion

We presented an application of the FEM-clustering approach [11, 12, 13] to the analysis of time series from molecular dynamics applications. The main methodological advantage of this scheme is that it allows to combine the features of the standard *geometrical clustering* and *dimension reduction techniques* (like K-Means or PCA) with *dynamical machine learning approaches* like HMMs for analysis of *multidimensional biomolecular time series*. In contrast to HMMs, no explicit assumptions about the Markovianity of the underlying dynamical process and probability distribution of the observables are needed. Another advantage of the proposed framework is that it is flexible wrt. the choice of the form of the *model distance functional* that describes the conformational molecular states.

When working with multidimensional molecular data, it is very important to be able to extract some reduced dynamical description out of it (e.g., in form of *hidden transition pathes* or *reduced dynamical models*). In order to control the reliability of the results, one has to analyze the sensitivity of obtained conformational states wrt. the length of the time series and the number K of the identified clusters. We demonstrated how the *maximal number of statistically distinguishable conformational states* $\mathbf{K}_{max}$ can be identified.

Two different forms of the FEM-clustering method, *FEM-K-means* and *FEM-KPCA*, were compared and the influence of the chosen model distance functional on the *metastability* and *clustering resolution* was investigated. It was shown that: (i) increasing the regularization parameter ϵ^2 increases the metastability of the identified conformational states, (ii) the PCA-based metric (7) used in the FEM-KPCA-algorithm allows for a higher resolution wrt. the identified conformational states for all levels of regularity, (iii) local PCA modes identified with the FEM-KPCA-algorithm can help to understand the differences between the conformations in terms of *molecular flexibility*.

It has been demonstrated that the impact of implicit method assumptions (like the choice of the model distance metric used in the algorithm or the selection of the regularization factor) on the results of the analysis and its interpretation is enormous, and requires insight in the nature of the data under investigation as in almost all clustering approaches. The algorithm in its present form is still far from allowing black box applications. Further investigations will be needed to come closer to this aim.

Acknowledgements

We are thankful to T. Frigato and E. Meerbach (FU-Berlin) who provided us with the penta-peptide MD-trajectory. We would also like to thank J. Smidt-Ehrenberg for his assistance with AMIRA-visualisation. The work was partially supported by the DFG SPP 1276 MetStroem "Meteorology and Turbulence Mechanics" and DFG Research Center "Matheon".

References

[1] 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.

[2] Christof Schütte and Wilhelm Huisinga. Mathematical analysis and simulation of conformational dynamics of biomolecules. In P. G. Ciaret and J.-L. Lions, editors, *Handbook of Numerical Analysis X*, volume Computational Chemistry, pages 699–744. Elsevier, 2003.

[3] 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.

[4] S. Hayward and H. J.C. Berendsen. Systematic analysis of domain motions in proteins from conformational change: New results on citrate synthase and T4 lysozyme. *Proteins*, 30(2):144–154, 1998.

[5] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy cluster analysis.* John Wiley and Sons, New York, 1999.

[6] J. Evanseck L. Caves and M.Karplus. Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci.*, 7(3):646–666, 1998.

[7] S. Neidle S. Haider, G. Parkinson. Molecular dynamics and principal components analysis of human telomeric quadruplex multimers. *Biophys. J.*, 95(1):296–311, 2008.

[8] I. Horenko, E. Dittmer, and Ch. Schuette. Reduced stochastic models for complex molecular systems. *to appear in SIAM Comp. Vis. Sci.*, 2005. (available via biocomputing.mi.fu-berlin.de).

[9] I. Horenko and C. Schuette. Likelihood-based estimation of langevin models and its application to biomolecular dynamics. *SIAM Mult. Mod. Sim.*, 7(2):802–827, 2008.
[10] I. Horenko, J. Schmidt-Ehrenberg, and Ch. Schütte. Set-oriented dimension reduction: Localizing Principal Component Analysis via Hidden Markov Models. In R. Glen M.R. Berthold and I. Fischer, editors, *CompLife 2006*, volume 4216 of *Lecture Notes in Bioinformatics*, pages 98–115. Springer, Berlin Heidelberg, 2006.
[11] I. Horenko. Finite element approach to clustering of multidimensional time series. *submitted to SIAM J. of Sci. Comp., (available via biocomputing.mi.fu-berlin.de)*, 2008.
[12] I. Horenko. On clustering of non-stationary meteorological time series. *submitted to the Journal of Climate, (available via biocomputing.mi.fu-berlin.de)*, 2008.
[13] I. Horenko. On robust estimation of low-frequency variability trends in discrete markovian sequences of atmospherical circulation patterns. *to appear in the Journal of Atmospherical Sciences, (available via biocomputing.mi.fu-berlin.de)*, 2008.
[14] I. Horenko, E. Dittmer, A. Fischer, and Ch. Schuette. Automated model reduction for complex systems exhibiting metastability. *SIAM Mult. Mod. Sim.*, 5:802–827, 2006.
[15] I.T. Jolliffe. *Principal Component Analysis*. Springer, 2002.
[16] I. Horenko. On simultaneous data-based dimension reduction an hidden phase identification. *J. of Atmos. Sci.*, 65(6), 2008.
[17] I. Horenko, R. Klein, S. Dolaptchiev, and Ch. Schuette. Automated generation of reduced stochastic weather models i: simultaneous dimension and model reduction for time series analysis. *SIAM MMS*, 6(4), 2008.
[18] I. Horenko, S. Dolaptchiev, A. Eliseev, I. Mokhov, and R. Klein. Metastable decomposition of high-dimensional meteorological data with gaps. *J. of Atmos. Sci.*, 65(10), 2008.
[19] G. McLachlan and D. Peel. *Finite mixture models*. Wiley, New–York, 2000.
[20] S. Fruhwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer, 2006.
[21] P. Deuflhard. *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms*, volume 35 of *Computational Mathematics*. Springer, Heidelberg, 2004.
[22] A. Tikhonov. On the stability of inverse problems. *Dokl. Akad. Nauk SSSR*, 39(5):195–198, 1943.
[23] A. Hoerl. Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58:54–59, 1962.
[24] G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.
[25] D. Braess. *Finite Elements: Theory, Fast Solvers and Applications to Solid Mechanics*. Cambridge University Press, 2007.
[26] P. Gill, W. Murray, M. Saunders, and M. Wright. A Schur-complement method for sparse quadratic programming. *Technical report, STANFORD UNIV CA SYSTEMS OPTIMIZATION LAB*, 1987.
[27] Ph. Metzner, I. Horenko, and Ch. Schuette. Generator estimation of Markov jump processes based on incomplete observations nonequidistant in time. *Physical Rewiev E*, 227(1), 2007.
[28] Ch. Schütte and W. Huisinga. Biomolecular conformations can be identified as metastable sets of molecular dynamics. In P. G. Ciaret and J.-L. Lions, editors, *Handbook of Numerical Analysis*, volume X, pages 699–744. Elsevier, 2003.
[29] R.A. Kendall, E. Apra, D.E. Bernholdt, E.J. Bylaska, M. Dupuis, G.I. Fann, R.J. Harrison, J. Ju, J.A. Nichols, J. Nieplocha, T.P. Straatsma, T.L. Windus, and A.T. Wong. High performance computational chemistry: An overview of NWChem a distributed parallel application. *Computer Physics Communications*, 128:260–283, 2000.