Computer Physics Communications 184 (2013) 2446–2453

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems

Nicolas Blöchliger, Andreas Vitalis *, Amedeo Caflisch

Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

ARTICLE INFO

Article history:
Received 30 October 2012
Received in revised form
7 June 2013
Accepted 12 June 2013
Available online 25 June 2013

Keywords:
Complex system
Trajectory analysis
Scalable algorithm
Minimum spanning tree
Free energy basins

ABSTRACT

Advances in IT infrastructure have enabled the generation and storage of very large data sets describing complex systems continuously in time. These can derive from both simulations and measurements. Analysis of such data requires the availability of scalable algorithms. In this contribution, we propose a scalable algorithm that partitions instantaneous observations (snapshots) of a complex system into kinetically distinct sets (termed basins). To do so, we use a combination of ordering snapshots employing the method's only essential parameter, *i.e.*, a definition of pairwise distance, and annotating the resultant sequence, the so-called progress index, in different ways. Specifically, we propose a combination of cut-based and structural annotations with the former responsible for the kinetic grouping and the latter for diagnostics and interpretation. The method is applied to an illustrative test case, and the scaling of an approximate version is demonstrated to be $\mathcal{O}(N \log N)$ with N being the number of snapshots. Two real-world data sets from river hydrology measurements and protein folding simulations are then used to highlight the utility of the method in finding basins for complex systems. Both limitations and benefits of the approach are discussed along with routes for future research.

1. Introduction

With present day computing resources, large-scale temporal simulations of complex systems can be performed routinely, and time-resolved, experimental data in many dimensions are collected and stored. In both cases, the resultant, very large amounts of data require dedicated, scalable protocols to handle access and analysis [1–3]. Examples can be found in fields such as protein science [4,5], astronomy [6], cell biology [7], or climatology [8] to name just a few.

For a complex system evolving in time, data are present in the form of sequences of instantaneous snapshots (microstates in the language of statistical mechanics) of this complex system, and such a sequence will be referred to as a trajectory throughout. Depending on whether data are synthetic or real, the implied projection of the system to obtain a snapshot may differ, and this may limit spatial resolution. Temporal resolution is limited directly by the instruments or numerical schemes if storage space is not a concern. Even though continuous evolution need not be observed explicitly as a function of time, we will restrict our terminology to this case. Routine analyses of trajectory data may involve computing average properties and their estimated distribution functions in $\mathcal{O}(N)$ time, where N is the number of snapshots. Distribution functions offer hints toward the diversity of states visited by the complex system and their relative weights. Time-resolved analyses provide insight regarding state connectivity and transition rates. Projection onto low-dimensional properties is necessary to render such analyses statistically meaningful and visualizable by conventional means.

If we assume that snapshots follow a well-defined distribution (such as the Boltzmann distribution for particles in the classical limit), these analyses look for spatial domains that are highly populated under the given conditions, *i.e.*, those for which a finite sample yields higher-than-average densities of microstates, preferably through recurrence [9]. Here, recurrence refers to the trajectory's property of entering and exiting subdomains within high density regions several times. The motivation behind this is twofold: (1) characterization of the complex system and communication of results in terms fit for human consumption [10]; (2) derivation of simplified models that provide a meaningful representation of the complex system [11,12]. Such models can preserve coarse-grained dynamical and static properties of the system and enable predictions to be made over vastly extended temporal or spatial domains.

When analyzing trajectories in projected spaces, high density regions are prone to overlap, and plots rarely resolve all of them [13]. This overlap phenomenon can lead to incorrect conclusions regarding the diversity and connectivity of coarse states. Consequently, affordable protocols that require little knowledge of the system *a priori* and that decrease the likelihood of such overlap are of interest. Techniques such as principal component analysis,

* Corresponding author. Tel.: +41 446355597; fax: +41 446356862.
E-mail addresses: n.bloechliger@bioc.uzh.ch (N. Blöchliger), a.vitalis@bioc.uzh.ch (A. Vitalis), caflisch@bioc.uzh.ch (A. Caflisch).

0010-4655/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cpc.2013.06.009

spectral clustering [14] and the related diffusion maps [15], locally linear embeddings [16], cut-based free energy profiles [17], kinetic groupings based on networks [18–21], which are specific cases of community detection algorithms in graphs [22], *etc.* are all in use, but many of them scale superlinearly with N.

Data clustering [23] offers a simple route to the identification of high density domains. Clusters are defined as groups of mutually similar snapshots. Similarity is assessed by a criterion of distance generally requiring an *ad hoc* selection of both a subset of features [24] and a functional form. However, a grouping meant to describe an evolving system should also encode dynamic proximity [25], *i.e.*, given a time resolution, which snapshots constitute a kinetically distinct state? If the system is of atomic scale and at equilibrium, this question aims to identify free energy basins and barriers in a generally high-dimensional phase space [26,27]. Positional coordinates of atoms are often used exclusively given that momenta can likely be ignored out on account of their much shorter autocorrelation times. We note that the language and concepts of statistical physics have proven useful in the analysis of nonphysical systems as well [28], *i.e.*, our adaptation of this language does not imply a restricted domain of application.

In this contribution, we present an algorithm that operates directly on a trajectory. With just the definition of a pairwise distance between snapshots, we are able to generate a one-dimensional plot that allows the identification of states in a joint geometric and kinetic sense, which we will refer to as basins. With standard metrics derived from microstate representations (such as interatomic distances in a flexible molecule), the method relies on the continuity of geometric representations in time. This implies that it may fail for certain classes of discrete systems. The main benefits of our algorithm are that it does not rely on any parameters *per se*, that it is very likely to resolve all basins, and that with the help of reasonable approximations to the exact procedure, the total running time approaches $\mathcal{O}(N \log N)$. The combination of these points is worth emphasizing, since we believe that they constitute a desirable and unique fingerprint of our approach.

The rest of this manuscript is structured as follows. First, we present the key ideas behind the procedure (Section 2.1) and illustrate its utility with a suitable model system (2.2). Next, we describe a computationally efficient and robust approximation to the exact procedure. The scaling of computational cost with data set size and dimensionality is tested explicitly (2.3). This is followed by applying the method to two complex real-world data sets, the first from hydrology (3.1) and the second from protein folding (3.2). We conclude by discussing the advantages and possible problems in comparison with related approaches (4).

2. Methods and proof of concept

2.1. The exact algorithm

Let $T = \{t_1, \ldots, t_N\}$ be a set (trajectory) of N unique snapshots, which usually are representations of the system in $\mathbb{R}^D$, which is the chosen subspace of the original system representation with $D \le D_{system}$. We use any pairwise distance $d : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}_{\ge 0}$ to measure the similarity between two snapshots. This need not be a purely coordinate-dependent function. Below it will prove beneficial for d to be a metric yielding a continuous number space with all $\mathcal{O}(N^2)$ values of d being unique.

We can now define the following iterative procedure. Choose a starting snapshot $s_1 \in T$ and create the set $S_1 = \{s_1\}$. Initialize the cut function, $c : \{1, \ldots, N\} \to \mathbb{N}$, to 2. Then, for $i = 1, \ldots, N-1$ do the following:

1. Define s_{i+1} as the snapshot in $T \backslash S_i$ realizing the minimum of $d(\cdot, S_i) = \min_{j=1,\ldots,i} d(\cdot, s_j)$.
2. Let $S_{i+1} = S_i \cup \{s_{i+1}\}$.
3. Define $c(i+1) = \sum_{j=1}^{N-1} \zeta_{S_{i+1}}(t_j, t_{j+1})$.

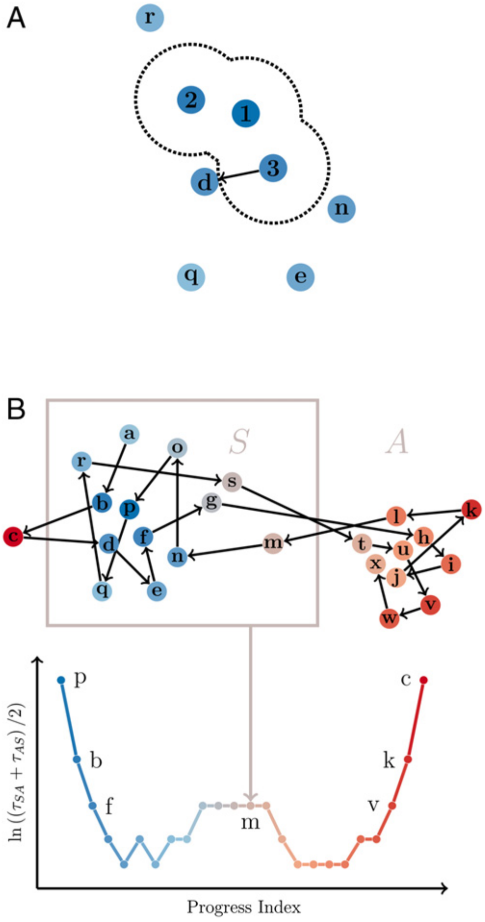

Fig. 1. Schematic highlighting the fundamental components of the algorithm. **A**. A set of points in two dimensions is shown as circles. See 2.1 for details. **B**. The points in **A** are shown as a subset of a larger data set. Arrows and letters indicate progression in time. The color scheme follows the order in which points are added when starting with point **p**, *i.e.*, colors trace the progress index itself. The schematic on the bottom shows values for the inverse logarithm of c at each value of the progress index. An example point and the cut to obtain the respective partitions S_i and A_i are highlighted. Point **c** illustrates an outlier, which are prone to be added last to S. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Here, the function ζ is defined as

$$\zeta_X(t, u) = \begin{cases} 0 & \text{if neither or both } t \text{ and } u \text{ are part of set } X \\ 1 & \text{otherwise.} \end{cases} \qquad (1)$$

The exact progress index of T starting with s_1 is defined as the sequence $S(T, s_1) = (s_1, \ldots, s_N)$. Each entry i is associated with a value for the cut function, $c(i)$. In words, given a starting snapshot, the algorithm finds a unique ordering of the snapshots, and annotates it with the number of transitions between the two partitions defined by all the snapshots that are currently part of the set (S_i) and those that are not yet part of the set ($A_i = T \backslash S_i$). The cut function c is related to the mean first passage time in the implied two-state Markov model via

$$\tau_{\text{MFP}}(A_i \to S_i) + \tau_{\text{MFP}}(S_i \to A_i) = 2N/c(i). \qquad (2)$$

We use τ_{AS} as shorthand notation for $\tau_{\text{MFP}}(A_i \to S_i)$ throughout. In Fig. 1(A), we show an illustration of a trajectory in 2D space with the current set of snapshots 1–3. The order of adding further snapshots would then be **d**, **n**, **r**, **e**, and **q** based on the mutual distance relations. There are no free parameters beyond having to

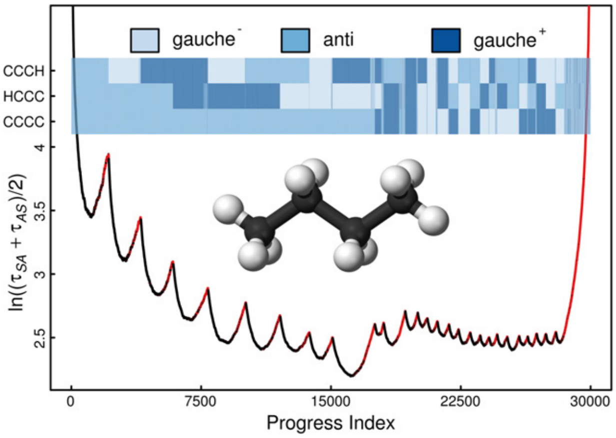

Fig. 2. Illustration of the approach using *n*-butane. The 27 basins of the system are all clearly resolved. Amongst those basins with the CCCC dihedral angle in *anti*, adjacent basins involve the rotation of only one of the methyl groups. This is fortuitous but signifies that the following basin in terms of the progress index is chosen on account of the sampling density in transition regions to any of the preceding ones. This density is higher for transitions involving only a single rotation. Points plotted in red correspond to snapshots that are classified as eclipsed according to the binning strategy described in 2.2 and are found preferentially toward the right half of basins and at the largest values of the progress index in general. The color annotation uses a simplified binning into 120° bins and does not display eclipsed microstates. The implied unit of time on the *y*-axis is a single snapshot, *i.e.*, 250 fs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

define distance relations, and for this purpose we have chosen the canonical tool, *i.e.*, a metric. In Fig. 1(B), the same set of points is shown as part of a longer trajectory. Here, letters indicate temporal order (**a**–**x**), whereas coloring tracks the progress index (blue–red) when using **p** as the starting snapshot. The cut function, *i.e.*, the number of transitions between S_i and $T \backslash S_i$, is illustrated in the lower half of the plot. The logarithm of the inverse of $c(i)$ produces small values if there are many transitions and peaks if there are few transitions. The latter is highlighted in Fig. 1(B) for set S with **m** being the snapshot having been added last. Fig. 1(B) illustrates the hypothesis that maxima in the logarithm of Eq. (2) will correspond to kinetic barriers separating basins to the left from those to the right. Consequently, the cut function should qualitatively encode dynamic properties of the system.

We note that the algorithm has two distinct parts: the progress index generation and the annotation function, here the cut function c. Both components can be treated and modified independently. A determination of the exact progress index is related to finding the minimum spanning tree (MST) of a complete graph with N vertices corresponding to all the t_i and edges with weights given via $d(t_i, t_j)$. The implementation we use scales with an overall complexity near $\mathcal{O}(N^2)$ and is described briefly in the Supplementary Information (SI), S.1.1 (see the Appendix). The exact progress index of T is unique if all possible $d(t_i, t_j)$ are distinct, and a unique progress index does not depend on the order the snapshots appear in T, *i.e.*, it does not contain any kinetic information. By construction, it is not possible for geometrically distinct basins to overlap provided that the sampling is good enough. Moreover, it is worth noting that the progress index does not imply that a given basin is closest kinetically to the one immediately to the left, but rather to any basin to the left.

2.2. Illustration with labeled *n*-butane

Let the linear alkane *n*-butane be described by three dihedral angles specifying rotations around all three carbon–carbon bonds (see Fig. 2). We assume atoms to be labeled such that the degeneracy of states can be resolved. In our chosen description, each dihedral angle has three distinct potential energy minima at 180°, 60°, and −60° corresponding to *anti*, *gauche*$^+$, and *gauche*$^-$ conformations. The potential has threefold symmetry for the methyl groups but favors *anti* for the central dihedral angle. It is expected that the system has access to $3^3 = 27$ coarse, metastable states. This is a good example for the algorithm presented in 2.1 since the low dimensionality and good knowledge of the system allow us to characterize basins and transitions independently.

Using stochastic dynamics simulations (see SI, S.1.4.1), we generated a classical trajectory of 30 000 snapshots under conditions such that recurrent sampling of all 27 basins is observed. Fig. 2 shows a plot generated by the algorithm described in 2.1 based on a trajectory with a time resolution of 250 fs and using a distance function defined on the three dihedral angles [29]. Clearly, we can resolve all basins, which is in contrast to cut-based free energy profiles used in prior work [29]. To confirm that the indicated basins do indeed correspond to the 27 expected ones, a color map representing an independent annotation based on binning the three degrees of freedom separately is shown. This correspondence is also established in Fig. S.1 with the help of box plots. Both figures reveal an asymmetry for snapshots within basins: points in highest density regions appear toward the left, and points in lower density ("fringe") regions appear toward the right. The latter correspond to eclipsed states, which have maximal enthalpy for this system. The asymmetry within each basin is a natural consequence of the way the progress index is constructed and annotated.

Further exploration of this system is meant to analyze two critical issues. First, what is the impact of the trajectory's time resolution? Second, can a connection between the results in Fig. 2 and an independent analysis of the thermodynamics and kinetics of this system be established?

We expect the progress index annotated with c as in Fig. 2 to successively lose its pertinent features if time resolution becomes so coarse that the various basin-to-basin transitions can no longer be resolved. We note that such a trajectory will eventually look random, which implies that the cut function just reports on the relative sizes of the two partitions, and not on (time-)local groupings of snapshots. This is indeed the case as shown in Fig. 3. For a resolution beyond 6 ps, the profile relaxes to a smooth, parabolic shape, which can be rationalized based on combinatorial arguments. We plot as a dashed line in Fig. 3 the analytically derived prior expected for a completely random trajectory (see SI, S.1.2). The result in Fig. 3 is obtained despite the fact that the progress index still orders the snapshots in fundamentally the same way as at finer time resolution. To make this point, a color map analogous to Fig. 2 is shown in Fig. 3 for the progress index derived from the 31.25 ps case. Therefore, the lack of features in Fig. 3 is not a result of overlap in the way one would encounter it in histogram- [17,30] or cut-based methods [29]. This is a significant advantage of our approach.

To perform an independent analysis of thermodynamics and kinetics, we constructed a set of macrostates by creating a 3D histogram with cubic bins of side length 60°. Bins are called eclipsed unless all three dimensions are centered at one of the three potential energy minima. Thus, 3^3 out of 6^3 macrostates are not eclipsed, and those correspond to the 27 basins. The resultant sequence of macrostates can be used to infer the transition matrix of an underlying Markov state model (MSM). From the MSM, pairwise τ_{MFP} values can be computed. If we now consider the progress index, at each point, we have a given MSM state annotation of the points immediately to the left (smaller values of the progress index) and to the right (larger values of the progress index). We may then infer the dominantly populated macrostate to either side via maximum likelihood guesses. With the knowledge of those two guesses, L and R, for each point of the progress index, we can plot the sum $\tau_{\mathrm{MFP}}(L \rightarrow R) + \tau_{\mathrm{MFP}}(R \rightarrow L)$. If $L \equiv R$, the result is directly proportional to the inverse of the probability of $L \equiv R$. Conversely,

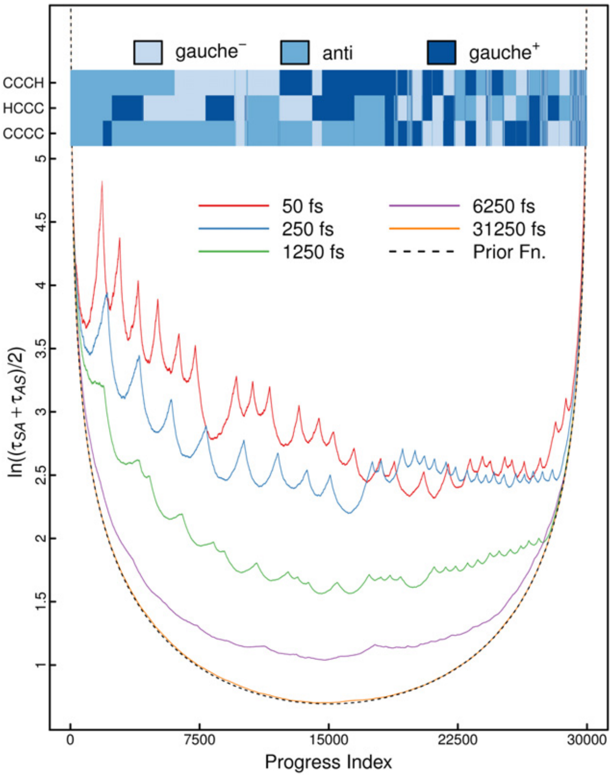

Fig. 3. Dependence of progress index and annotation function on temporal resolution. Data comparable to Fig. 2 are shown for decreasing temporal resolution. Features are successively lost, and at 31.25 ps the annotation becomes indistinguishable from that expected for a completely random trajectory (prior function). For the cases of 1.25 and 6.25 ps, it is apparent that the strong inherent curvature of function c impedes the identification of basins if they are small and/or temporal resolution is poor. For each curve the implied unit of time on the y-axis is a single snapshot of the respective trajectory, *i.e.*, the saving frequency or temporal resolution itself. As in Fig. 2, a color annotation is shown, here for the 31.25 ps case. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

if the point is in a barrier region, $L \neq R$ and the result measures the kinetic proximity of two neighboring macrostates. These data are shown as the green curve in Fig. 4. Comparison with the original profile shows that there is no quantitative relationship between the two plots. It is therefore impossible to obtain quantitative thermodynamic or kinetic information from c. This is expected because the cut function measures kinetics in a crude two-state assumption (A and S above) and not between individual basins.

Are there alternative annotation functions to consider? Here, we define a 'localized' cut function as follows:

$$l(i) = \sum_{j=1}^{N-1} \zeta_{B_i(n_l(i))}(t_j, t_{j+1}) \zeta_{C_i(n_l(i))}(t_j, t_{j+1}). \tag{3}$$

In Eq. (3), partition $B_i(n_l(i))$ is defined as $S_{i-1} \backslash S_{i-1-n_l(i)}$, and partition $C_i(n_l(i))$ is defined as $S_{i-1+n_l(i)} \backslash S_{i-1}$. This corresponds to a restriction of the cut function to contributions from points in the trajectory that are near in the progress index, and function l is expected to offer better resolution than c for reasonable choices of $n_l(i)$. A progress index annotated with l is shown in Fig. 4 as well. Due to the peculiar nature of the system, the parameter $n_l(i)$ in Eq. (3) is chosen in accordance with average basin sizes (see the caption to Fig. 4). There is very good correspondence between these results and the thermodynamic information inferred from the MSM. However, Fig. 4 shows that peak heights are not correlated beyond both sets appearing to populate two dominant ranges of values. Quantitative correspondence is unlikely because the cut function defined by Eq. (3) is not equivalent to well-defined kinetic information within the underlying three-state MSM. However, function l does appear to be able to provide higher resolution when it comes to identifying basins. This is highlighted by comparison of Figs. 3 and S.2 for the 1.25 ps case, which reveals that the inherent curvature of function c may limit basin delineation before the time resolution approaches characteristic transition times of the system. If meaningful values for $n_l(i)$ can be found, annotation with l is likely to provide more information.

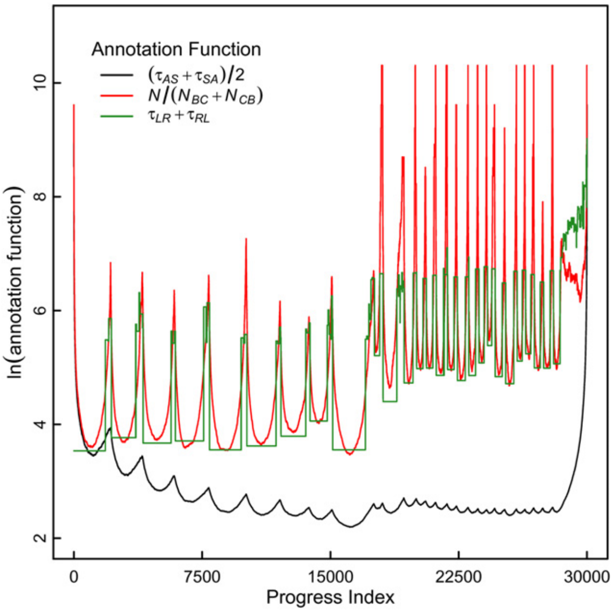

Fig. 4. Quantitative kinetic and thermodynamic interpretation of two annotation functions. The standard annotation function via Eq. (1) is reproduced identically to Fig. 2 (black). The localized annotation function defined in Eq. (3) is shown in red. Because basins have two standard sizes (assumed to be 1600 snapshots if the central torsion is in *anti* and 400 snapshots otherwise), we generated data with $n_l(i)$ set to fixed values of either 1600 or 400 snapshots. For the curve shown in the plot, values were simply interpolated to convert from the case with $n_l(i) = 1600$ to the case with $n_l(i) = 400$ over values of the progress index of 17 300–17 700. Lastly, the green curve shows results from an underlying MSM as described in 2.2. The width for constructing the maximum likelihood guess of assigning basins L and R was 100 snapshots throughout. The implied unit of time on the y-axis is a single snapshot, *i.e.*, 250 fs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.3. An approximate algorithm operating in near-linear time

Because the exact algorithm as described in 2.1 and expanded upon in the SI, S.1.1, requires approximately $\mathcal{O}(N^2)$ time, it is impractical for large data sets. In this section, we outline conceptually the implementation of an approximate algorithm that operates in $\mathcal{O}(N \log N)$ time. A detailed description is found in the SI, S.1.3.

Briefly, a spanning tree is constructed with Borůvka's algorithm [31], which works by successively merging subtrees using nearest neighbors. However, instead of considering rigorous nearest neighbors for each subtree, we instead consider a set of nearby snapshots, which is extracted from preorganizing the data via hierarchical clustering [29]. A hierarchical grouping means that snapshots are partitioned into groups of similar objects (clusters) for a range of resolutions. The set of nearby snapshots is then constituted from the union of all clusters that the subtree spans, and which are not yet part of the subtree. This is done for the finest

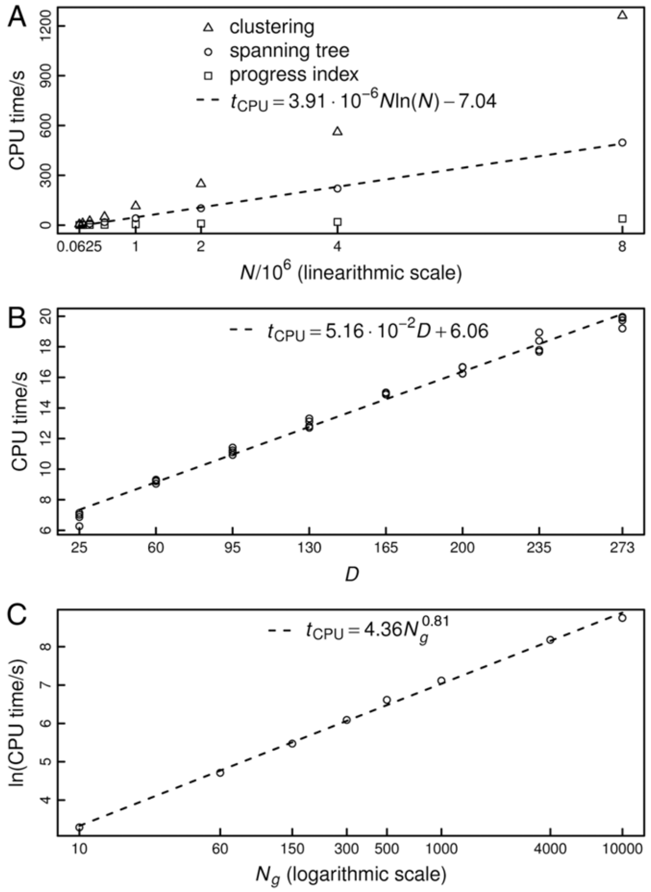

Fig. 5. Runtime analysis for the approximate version of the proposed algorithm. **A**. The cost for computing the SST is shown as a function of $N \log N$, *i.e.*, the expected complexity. We also show a linear fit and the costs for the tree-based clustering and generation of the progress index from the SST. An apparent scaling exponent from a double logarithmic plot of cost *vs.* N (not shown) is 1.15, close to the expected value of 1.08 for this range of values for N. Data are for the case where the number of clusters at the leaf level of the hierarchical clustering grows linearly with N, *i.e.*, the average cluster size is roughly constant (see S.1.3). **B**. Computational cost of SST construction as a function of dimensionality. D was adjusted as described in SI, S.1.4.2. As expected, cost is linear in D, and four independent trials yield similar answers. **C**. Computational cost of SST construction as a function of the number of guesses, N_g. Because N_g will eventually exceed the size of the restricted search space, it is expected that cost scales sublinearly with N_g. This is confirmed by the double logarithmic plot.

resolution level still yielding a nonempty set. If the set is larger than a parameter, N_g, N_g guesses are taken instead of searching the entire set. These two approximations limiting the search space mean that for a given subtree the number of candidate edges is within a constant upper bound, which gives the desired complexity of $\mathcal{O}(N \log N)$. The output is a "short" spanning tree (SST) used for the generation of progress index and annotation functions analogously to the MST in the exact case. Qualitative neighbor relations are expected to be preserved in the SST with the approximations primarily leading to randomization within basins.

The scaling with data set size (see SI, S.1.4.2) is demonstrated in Fig. 5(A) for a fixed value for N_g of 20. Clearly, a plot of computational cost *vs.* $N \log N$ is roughly linear. As can be seen, the cost for the construction of the SST along with the generation of data pairs for progress index and annotation function is less than that of the tree-based clustering. Fig. 5(A) implies that we can identify basins in a data set composed of 8×10^6 snapshots with a dimensionality of $D = 273$ in less than an hour on a single core of a modern desktop machine. Fig. 5(B) shows the dependence of computational cost on D. This is expected to be linear, since the dimensionality of representation only matters for computations of distance, the total number of which is roughly constant. This expectation is confirmed by Fig. 5(B).

3. Results

3.1. Hydrology data for rivers near Portland, Oregon

While *n*-butane is a perfect example for the algorithm, real world data sets may not be, especially if they describe the evolution of a complex system that is not fundamentally stochastic in nature. We constructed an example from hydrology parameters measured at various river sites near Portland, Oregon, USA, over a period of about 5 years. Measured quantities include pH, conductance, discharge (volume flux), temperature, and oxygen content. River parameters are expected to be influenced by ambient weather, specific climatic events such as snowmelt, and local geography. Seasonal patterns generate data sets that are likely to show recurrence (similar seasons in subsequent years give rise to similar river conditions), but that are not random. These data are challenging for the following reasons:

1. Measurements are performed with low accuracy and may contain outliers caused by malfunctioning devices or short-term, local contaminations.
2. Subtle trends observed over multiple years may render conditions locally more similar than compared to analogous times in other years, and recurrence of conditions is weak overall due to the (small) number of years in the data set. This challenges the annotation function that relies on good mixing within a basin.
3. Most measured parameters produce uninformative histograms on their own. In conjunction with the first point, this challenges the geometrical separability of these data, *i.e.*, the pairwise distance spectrum is expected to be relatively featureless (see Fig. S.3).

We note that the data set is small enough ($N = 87\,840$ and $D = 15$) that we can use the exact algorithm. Fig. 6 plots the progress index annotated with both c and l, and the kinetic annotation confirms the challenging nature of these data. Profiles are sparse in well-resolved features and allow the identification of two larger basins with unclear size along with a number of smaller basins, *e.g.*, for values of the progress index around $1.6 \cdot 10^4$ or $8 \cdot 10^4$. The color annotation of the input data supports this interpretation. These data were normalized, centered, and subjected to noise before being fed into the algorithm (see S.1.4.3). Red colors indicate high values, and hence the first major basin is a putative warm season with high water temperatures, high conductivity (σ), low river levels, and low oxygen concentrations. The second major basin (up to $4.5 \cdot 10^4$) corresponds to a putative cold season with generally inverted parameters. We can confirm these assignments by using the time annotation of the progress index shown in the bottom part of Fig. 6. These highlight that the data in the first basin indeed come from the warmest and driest months (July–September) and that the data in the second basin come from the extended winter months (November–April).

The rest of the plot reveals a few well-defined regions of homogeneous conditions that often come from specific years. These are not always well-resolved in terms of functions c or l, and one important problem contributing to this lack of resolution is lack of recurrence. This is seen most clearly for winter and spring of 2008 found at progress index values of 5 to $6 \cdot 10^4$ and indicated by linear correlation of progress index and real time. Cut values become nearly invariant, which limits the use of these annotation functions for nonrecurrent, but kinetically partitioned data. As a counterexample, the mid-summer months of 2008 found

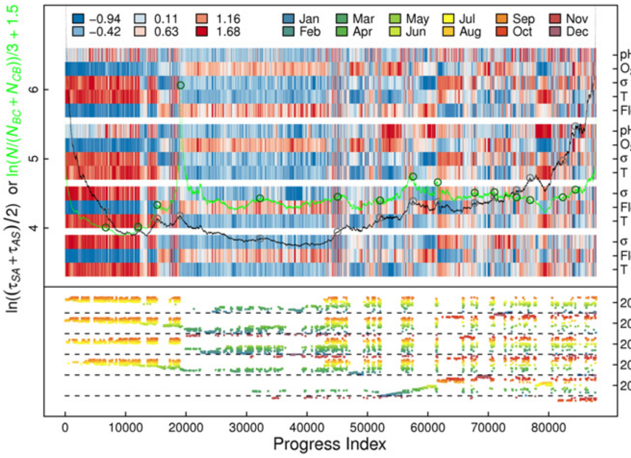

Fig. 6. Application of the exact algorithm to hydrology data. The annotation functions, c and l with a fixed $n_l = 10\,000$, derived from the progress index are plotted against the progress index as black and green curves, respectively. The data for function l were scaled and shifted as indicated in the axis label. The implied unit of time on the y-axis is a single snapshot, *i.e.*, 30 minutes. A color annotation similar to the one in Figs. 2 and 3 is shown along with these plots. Data are centered and normalized as described in the SI, S.1.4.3, and a uniform color scheme is used (legend in the upper left-hand side). Data come from four stations (that are offset visually) and encompass different measurements as indicated on the right-hand side. The lower half of the figure annotates the progress index temporally with an additional monthly color code meant to highlight the yearly patterns (legend in the upper right-hand side). Finally, circles highlight barriers identified via a measure of the locality of the progress index as described in Fig. S.4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

at progress index values near $8 \cdot 10^4$ yield water conditions with recurrence according to both c and l. If c and l fail, it is also possible to identify barrier regions via the locality of the progress index that is known from the MST. Essentially, each snapshot is added to the set S on account of a specific edge to a specific "parent" vertex, whose position in the progress index is known. If this position is not nearby (not local), we can speculate that we have encountered a barrier region (see Fig. S.4). Putative barriers derived this way are plotted as circles in Fig. 6 and seem to offer potential in delineating basins for nonrecurrent data. Finally, a more detailed analysis is given in S.2.1. In particular, Figs. S.5 and S.6 explore differences between the exact and approximate algorithms. The latter is used exclusively for the final data set.

3.2. Reversible folding of a β-sheet miniprotein

As a final test, we apply the approximate algorithm to a complex system analyzed extensively in previous works [32,33,29]. Beta3S is a 20-residue polypeptide that undergoes reversible folding transitions at 330 K on the high ns time scale if a suitable computational model is utilized [34]. The native basin is a three-stranded β-sheet, but various other enthalpic basins are known and populated (for further details, see SI, S.1.4.2).

Fig. 7 shows representative results for the approximate progress index coupled to the simple annotation function, c. The first thing to note is the strong similarity of the plot in Fig. 7 to cut-based free energy profiles based on the same data set [32,33,29] (see also Figs. S.7 and S.8). The native basin, which the starting snapshot is part of, encompasses about 35% of the data. Secondary structure, *i.e.*, DSSP annotations [35] to the progress index are shown as well in a color plot for individual residues. These confirm the correct topology for a three-stranded β-sheet. For large values of the progress index, we find a basin comprised of an ensemble of structures rich in α-helix. In between, there is a mix of smaller, enthalpic basins that usually share part of their topology with the native state, and entropic regions without consistent order formation. Based on the DSSP annotation, it appears that function c resolves all structurally homogeneous sets of microstates suggesting that the system exhibits sufficient recurrence over the aggregate sampling time of 20 μs. This holds even for tiny basins such as the one seen at values of the progress index just past $6 \cdot 10^5$. Fig. S.8 shows the annotation with l and highlights that c provides sufficient resolution for this system.

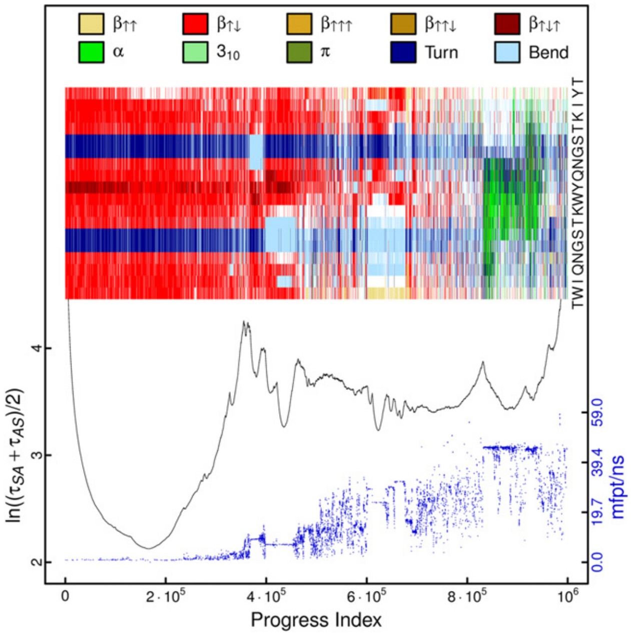

Fig. 7. Application of the approximate algorithm to Beta3S. The distance function in use is the coordinate root mean square deviation computed over backbone oxygen and nitrogen atoms of residues 3–18 after pairwise alignment. The progress index obtained with $N_g = 200$ is plotted against annotation function c. For each trajectory snapshot, we also computed DSSP annotations [35] that are presented as a color annotation (legend on top, one-letter codes for individual amino acids on the right-hand side). Only every 20th snapshot is shown to keep the size of the original vector image manageable. Lastly, we show a further kinetic annotation by plotting independent τ_{MFP} values to the native basin (small values of the progress index) for selected snapshots. The selected snapshots are the centroids of those nodes in the MSM used to generate the τ_{MFP} values, which encompass at least 10 snapshots (about 8000). Tests with values for N_g as small as 20 yielded comparable results (not shown). For plotting details, please refer additionally to the caption of Fig. S7.

There are two questions we want to address. First, are the resolved basins in fact kinetically homogeneous? To this extent, we constructed a network of conformational transitions based on the tree-based clustering and conformational root mean square deviations exactly as described in prior work [29] (this is also the exact same clustering used for data preorganization when generating the SST). Using a target node in the native basin as reference, we proceeded to determine the τ_{MFP} values for all other nodes. If a node contains at least 10 snapshots, the value for τ_{MFP} is plotted in Fig. 7 for all those snapshots at their respective positions in the progress index. This simple annotation confirms that – at least in reference to the native basin – the basins identified by our proposed approach are indeed kinetically homogeneous. To further address this, Fig. S.7 shows a correlation analysis of the cut-based free energy profile based on the same clustering with the results in Fig. 7. The conclusions are the same. As a corollary, a lack of kinetic homogeneity seen for example around values of the progress index of $5 \cdot 10^5$ or $8 \cdot 10^5$ correlates with parts of the profile, for which c does not indicate the presence of a basin.

The second question is with regard to the ordering of the basins by the progress index. The annotation with τ_{MFP} makes the point that there is weak correlation between a distance in the progress index and a distance in kinetics (see also Fig. S.7). This is expected, since the sampling density in transition regions no longer represents a ruler for kinetic distance to a specific basin once multiple basins have been incorporated into set S. In analogy to cut-based free energy profiles, this also means that neighbor relations are not necessarily meaningful for larger values of the progress index as discussed in the context of Fig. 2. In summary, for this more appropriate data set compared to 3.1, the proposed scheme provides exactly the information we expected to obtain with no obvious limitations or errors in annotation.

4. Discussion and conclusions

In this contribution, we have presented a new algorithm for sorting and annotating sets of data that are the result of continuous evolution. The sorting component, *i.e.*, the progress index, is derived in both an exact form with modest computational complexity and in an approximate form that is computationally efficient and scalable to very large data sets (see Fig. 5). Such scalable algorithms are increasingly sought after due to the routine generation and storage of massive trajectories given present day computing resources [6,36,2]. The second component, *i.e.*, the various annotation functions used throughout, generally scale as $\mathcal{O}(N)$ and are of lesser cost than the progress index generation. The two components combine to yield one-dimensional plots that are able to distinguish kinetically grouped sets of microstates in complex systems that exhibit sufficient recurrence (mixing) both within and amongst basins. There are no parameters controlling size, number, or other properties of basins, and the algorithm is agnostic beyond the fact that we have to define a pairwise measure of similarity. We believe that the combination of minimal user input and high computational efficiency makes our proposed scheme a useful one.

The total runtime for generating Fig. 7 was on the order of minutes for a trajectory of 10^6 snapshots. This highlights the utility of the approach in quickly and reliably partitioning a complex system into an annotated set of basins. We are unaware of alternative methods offering comparable amounts of information at this cost. The strengths of the approach rest on the use of all snapshots, *i.e.*, the lack of any binning or other *a priori* grouping (the auxiliary clustering is for efficiency only (see 2.3), and has no direct bearing on the results (see Fig. S.6)). The kinetic annotation functions, c and l (see 2.1 and 2.2), operate relative to the time resolution of the data and will correctly lump all snapshots together if the latter is too coarse (see Figs. 3 and S.2). Actual failure is possible if small basins are placed in regions of high inherent curvature (see Fig. 3). This is an issue of the signal-to-noise ratio, and we expect it to be corrected by increasing the amount of data or using a different starting point. Any lack of recurrence is a potentially more critical issue and is encountered in Fig. 6. However, it need not result from non-stochastic evolution of the system, but can also result from an inappropriately high dimensionality in representation. In the latter case, the point density becomes so low everywhere that basins are no longer identifiable.

The last comment above implies that the utility of data processing algorithms of this type rests on the appropriateness of the distance function. This is a very fundamental problem, but there is little rigorous work comparing combinations of different classes of distance functions coupled to different representations of a complex system [37]. A more active and closely related area of research is that of finding suitable reaction coordinates for complex systems that preserve correct, coarse-grained kinetics and thermodynamics [38,33,39,40]. Viewed as a simple grouping scheme [23], our approach offers the advantage over the majority of algorithms that there is no parameter controlling the number or size of clusters. Moreover, comparable groupings are normally the result of a two-stage process: efficient, fine-grained clustering is followed by suitable refinement [41]. Our approach shares a strong formal similarity with the OPTICS clustering algorithm that also utilizes a combination of sorting and annotation [42]. We emphasize that few algorithms in this class operate at such low time complexity, *e.g.*, [43,44,29]. The reliance on geometric continuity during system evolution is shared explicitly with methods computing eigenvectors of a kernel-based density estimate given the full $\mathcal{O}(N^2)$ Laplacian matrix, *i.e.*, diffusion maps [15,39]. These methods not only require choosing a kernel function (or at least parameter(s) for it), but the reliance on the Laplacian matrix renders them infeasible for data sets exceeding $\sim 10^5$ snapshots. Lastly, we briefly mention path sampling approaches. With suitably chosen end points, these methods can yield comparable information [45–48], because they directly probe kinetic connectivity of different sets of microstates. Of course, they are conjoined with the sampling protocol itself, *i.e.*, they are not pure analysis schemes applicable to any continuous data set, and require significant human input. This is also true for metadynamics [49] and many related approaches, *e.g.*, a recent approach to sequential basin discovery [50].

The algorithm as described here has been implemented in the CAMPARI software package [51], and the current development version is available from the authors on request (campari.software@gmail.com). Ongoing work is targeting three areas. First, can we automatize feature selection using an appropriate criterion of optimality, *i.e.*, is it possible to eliminate the need to manually define a distance function? Second, for the localized cut function, l, is there an iterative, but efficient procedure that determines a suitable value of $n_l(i)$ for all snapshots? The current restriction to one or a few values of n_l clearly lacks general utility. Third, can we identify additional annotation functions that can be quantitatively related to relevant time scales of the system? We believe that addressing these questions opens up fruitful avenues for future research toward routine analysis of large data sets continuous in time.

Acknowledgments

We thank Dr. Ting Zhou for sharing data on Beta3S used for runtime analysis [52]. This work was supported by a grant of the Swiss National Science Foundation to A. C.

Appendix. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.cpc.2013.06.009.

References

[1] I. Antcheva, M. Ballintijn, B. Bellenot, M. Biskup, R. Brun, N. Buncic, P. Canal, D. Casadei, O. Couet, V. Fine, L. Franco, G. Ganis, A. Gheata, D. Gonzalez Maline, M. Goto, J. Iwaszkiewicz, A. Kreshuk, D. Marcos Segura, R. Maunder, L. Moneta, A. Naumann, E. Offermann, V. Onuchin, S. Panacek, F. Rademakers, P. Russo, M. Tadel, ROOT—A C++ framework for petabyte data storage, statistical analysis and visualization, Comput. Phys. Comm. 180 (2009) 2499–2512.

[2] D. Hasenkamp, A. Sim, M. Wehner, K. Wu, Finding tropical cyclones on a cloud computing cluster: using parallel virtualization for large-scale climate simulation analysis, in: J. Qiu, G. Zhao, C. Rong (Eds.), 2010 IEEE Second International Conference on Cloud Computing Technology and Science, Indianapolis, IN, USA, November 30–December 3, 2010, CloudCom, IEEE Computer Society Conference Publishing Services, Los Alamitos, CA, USA, 2010, pp. 201–208.

[3] E.E. Schadt, M.D. Linderman, J. Sorenson, L. Lee, G.P. Nolan, Computational solutions to large-scale data management and analysis, Nature Rev. Genet. 11 (2010) 647–657.

[4] K. Lindorff-Larsen, S. Piana, R.O. Dror, D.E. Shaw, How fast-folding proteins fold, Science 334 (2011) 517–520.

[5] G. Settanni, F. Rao, A. Caflisch, Φ-value analysis by molecular dynamics simulations of reversible folding, Proc. Natl. Acad. Sci. USA 102 (2005) 628–633.

[6] V. Springel, S.D.M. White, A. Jenkins, C.S. Frenk, N. Yoshida, L. Gao, J. Navarro, R. Thacker, D. Croton, J. Helly, J.A. Peacock, S. Cole, P. Thomas, H. Couchman, A. Evrard, J. Colberg, F. Pearce, Simulations of the formation, evolution and clustering of galaxies and quasars, Nature 435 (2005) 629–636.

[7] Y. Wang, J.Y.-J. Shyy, S. Chien, Fluorescence proteins, live-cell imaging, and mechanobiology: seeing is believing, Annu. Rev. Biomed. Eng. 10 (2008) 1–38.
[8] B.P. Kirtman, C. Bitz, F. Bryan, W. Collins, J. Dennis, N. Hearn, J.L. Kinter III, R. Loft, C. Rousset, L. Siqueira, C. Stan, R. Tomas, M. Vertenstein, Impact of ocean model resolution on CCSM climate simulations, Clim. Dyn. 39 (2012) 1303–1328.
[9] N. Marwan, M.C. Romano, M. Thiel, J. Kurths, Recurrence plots for the analysis of complex systems, Phys. Rep. 438 (2007) 237–329.
[10] I.G. Kevrekidis, C.W. Gear, G. Hummer, Equation-free: the computer-aided analysis of complex multiscale systems, AIChE J. 50 (2004) 1346–1355.
[11] S. Itzkovitz, R. Levitt, N. Kashtan, R. Milo, M. Itzkovitz, U. Alon, Coarse-graining and self-dissimilarity of complex networks, Phys. Rev. E 71 (2005) 016127.
[12] J.D. Halley, D.A. Winkler, Classification of emergence and its relation to self-organization, Complexity 13 (2008) 10–15.
[13] M. Sips, B. Neubert, J.P. Lewis, P. Hanrahan, Selecting good views of high-dimensional data using class consistency, Comput. Graph. Forum 28 (2009) 831–838.
[14] M. Filippone, F. Camastra, F. Masulli, S. Rovetta, A survey of kernel and spectral methods for clustering, Pattern Recognit. 41 (2008) 176–190.
[15] B. Nadler, S. Lafon, R.R. Coifman, I.G. Kevrekidis, Diffusion maps, spectral clustering and reaction coordinates of dynamical systems, Appl. Comput. Harmon. Anal. 21 (2006) 113–127.
[16] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[17] S.V. Krivov, M. Karplus, One-dimensional free-energy profiles of complex systems: progress variables that perserve the barriers, J. Phys. Chem. B 110 (2006) 12689–12698.
[18] F. Noé, I. Horenko, C. Schütte, J.C. Smith, Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states, J. Chem. Phys. 126 (2007) 155102.
[19] 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 (2007) 155101.
[20] S. Muff, A. Caflisch, Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein, Proteins: Struct. Funct. Bioinform. 70 (2008) 1185–1195.
[21] J.M. Carr, D.J. Wales, Folding pathways and rates for the three-stranded β-sheet peptide Beta3s using discrete path sampling, J. Phys. Chem. B 112 (2008) 8760–8769.
[22] A. Lancichinetti, S. Fortunato, Community detection algorithms: a comparative analysis, Phys. Rev. E 80 (2009) 056117.
[23] R. Xu, D. Wunsch II, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (2005) 645–678.
[24] A. Jain, D. Zongker, Feature selection: evaluation, application, and small sample performance, IEEE Trans. Pattern Anal. 19 (1997) 153–158.
[25] B. Keller, X. Daura, W.F. van Gunsteren, Comparing geometric and kinetic cluster algorithms for molecular simulation data, J. Chem. Phys. 132 (2010) 074110.
[26] W. Huisinga, C. Best, R. Roitzsch, C. Schütte, F. Cordes, From simulation data to conformational ensembles: structure and dynamics-based methods, J. Comput. Chem. 20 (1999) 1760–1774.
[27] D.J. Wales, Energy landscapes: some new horizons, Curr. Opin. Struct. Biol. 20 (2010) 3–10.
[28] C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics, Rev. Modern Phys. 81 (2009) 591–646.
[29] A. Vitalis, A. Caflisch, Efficient construction of mesostate networks from molecular dynamics trajectories, J. Chem. Theory Comput. 8 (2012) 1108–1120.
[30] M. Daszykowski, B. Walczak, D.L. Massart, Projection methods in chemistry, Chemometr. Intell. Lab. Syst. 65 (2003) 97–112.
[31] J. Nešetřil, E. Milková, H. Nešetřilová, Otakar Borůvka on minimum spanning tree problem, translation of both the 1926 papers, comments, history, Discrete Math. 233 (2001) 3–36.
[32] S.V. Krivov, S. Muff, A. Caflisch, M. Karplus, One-dimensional barrier-preserving free-energy projections of a β-sheet miniprotein: new insights into the folding process, J. Phys. Chem. B 112 (2008) 8701–8714.
[33] B. Qi, S. Muff, A. Caflisch, A.R. Dinner, Extracting physically intuitive reaction coordinates from transition networks of a β-sheet miniprotein, J. Phys. Chem. B 114 (2010) 6979–6989.
[34] P. Ferrara, J. Apostolakis, A. Caflisch, Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations, J. Phys. Chem. B 104 (2002) 5000–5010.
[35] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.
[36] D.E. Shaw, M.M. Deneroff, R.O. Dror, J.S. Kuskin, R.H. Larson, J.K. Salmon, C. Young, B. Batson, K.J. Bowers, J.C. Chao, M.P. Eastwood, J. Gagliardo, J.P. Grossman, C.R. Ho, D.J. Ierardi, I. Kolossváry, J.L. Klepeis, T. Layman, C. McLeavey, M.A. Moraes, R. Mueller, E.C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, S.C. Wang, Anton, a special-purpose machine for molecular dynamics simulation, in: D. Tullsen, B. Calder (Eds.), Proceedings of the 34th Annual International Symposium on Computer Architecture, San Diego, CA, USA, June 9–13, 2007, ISCA'07, ACM, New York, NY, USA, 2007, pp. 1–12.
[37] P. Cossio, A. Laio, F. Pietrucci, Which similarity measure is better for analyzing protein structures in a molecular dynamics trajectory? Phys. Chem. Chem. Phys. 13 (2011) 10421–10425.
[38] R.B. Best, G. Hummer, Reaction coordinates and rates from transition paths, Proc. Natl. Acad. Sci. USA 102 (2005) 6732–6737.
[39] M.A. Rohrdanz, W. Zheng, M. Maggioni, C. Clementi, Determination of reaction coordinates via locally scaled diffusion map, J. Chem. Phys. 134 (2011) 124116.
[40] S.V. Krivov, Is protein folding sub-diffusive? PLoS Comput. Biol. 6 (2010) e1000921.
[41] T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: an efficient data clustering method for very large databases, in: J. Widom (Ed.), SIGMOD'96: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, QC, Canada, June 4–6, 1996, ACM Press, New York, NY, USA, 1996, pp. 103–114.
[42] M. Ankerst, M.M. Breunig, H.-P. Kriegel, J. Sander, OPTICS: ordering points to identify the clustering structure, in: J. Clifford, R. King (Eds.), SIGMOD'99: Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, Philadelphia, PA, USA, May 31–June 03, 1999, ACM Press, New York, NY, USA, 1999, pp. 49–60.
[43] A. Hinneburg, D.A. Keim, Optimal grid-clustering: towards breaking the curse of dimensionality in high-dimensional clustering, in: M.P. Atkinson, M.E. Orlowska, P. Valduriez, S.B. Zdonik, M.L. Brodie (Eds.), Proceedings of the 25th VLDB Conference, Edinburgh, Scotland, September 7–10, 1999, Morgan Kaufmann, San Francisco, CA, USA, 1999, pp. 506–517.
[44] R.L.F. Cordeiro, A.J.M. Traina, C. Faloutsos, C. Traina Jr., Halite: fast and scalable multi-resolution local-correlation clustering, IEEE Trans. Knowl. Data Eng. 25 (2013) 387–401.
[45] P.G. Bolhuis, D. Chandler, C. Dellago, P.L. Geissler, Transition path sampling: throwing ropes over rough mountain passes, in the dark, Annu. Rev. Phys. Chem. 53 (2002) 291–318.
[46] D.J. Wales, Discrete path sampling, Mol. Phys. 100 (2002) 3285–3305.
[47] W. E, W. Ren, E. Vanden-Eijnden, Simplified and improved string method for computing the minimum energy paths in barrier-crossing events, J. Chem. Phys. 126 (2007) 164103.
[48] P. Faccioli, Characterization of protein folding by dominant reaction pathways, J. Phys. Chem. B 112 (2008) 13756–13764.
[49] A. Laio, M. Parrinello, Escaping free-energy minima, Proc. Natl. Acad. Sci. USA 99 (2002) 12562–12566.
[50] Y.V. Sereda, A.B. Singharoy, M.F. Jarrold, P.J. Ortoleva, Discovering free energy basins for macromolecular systems via guided multiscale simulation, J. Phys. Chem. B 116 (2012) 8534–8544.
[51] A. Vitalis, A. Steffen, N. Lyle, A.H. Mao, R.V. Pappu, Campari v1.0, http://sourceforge.net/projects/campari (accessed 30.10.12).
[52] T. Zhou, A. Caflisch, Free energy guided sampling, J. Chem. Theory Comput. 8 (2012) 2134–2140.