A strategy for analysis of (molecular) equilibrium simulations: Configuration space density estimation, clustering, and visualization

Fred A. Hamprecht, Christine Peter, and Xavier Daura
Laboratory of Physical Chemistry, Swiss Federal Institute of Technology, ETH Zentrum, 8092 Zürich, Switzerland

Walter Thiel
Max-Planck-Institut für Kohlenforschung, 45470 Mülheim, Germany

Wilfred F. van Gunsteren
Laboratory of Physical Chemistry, Swiss Federal Institute of Technology, ETH Zentrum, 8092 Zürich, Switzerland

(Received 16 August 2000; accepted 12 October 2000)

We propose an approach for summarizing the output of long simulations of complex systems, affording a rapid overview and interpretation. First, multidimensional scaling techniques are used in conjunction with dimension reduction methods to obtain a low-dimensional representation of the configuration space explored by the system. A nonparametric estimate of the density of states in this subspace is then obtained using kernel methods. The free energy surface is calculated from that density, and the configurations produced in the simulation are then clustered according to the topography of that surface, such that all configurations belonging to one local free energy minimum form one class. This topographical cluster analysis is performed using basin spanning trees which we introduce as subgraphs of Delaunay triangulations. Free energy surfaces obtained in dimensions lower than four can be visualized directly using iso-contours and -surfaces. Basin spanning trees also afford a glimpse of higher-dimensional topographies. The procedure is illustrated using molecular dynamics simulations on the reversible folding of peptide analoga. Finally, we emphasize the intimate relation of density estimation techniques to modern enhanced sampling algorithms.
 [DOI: 10.1063/1.1330216]

I. INTRODUCTION

With the rapid increase in computational power, simulations of complex systems have reached unprecedented lengths, and the time-honored approach to a first analysis of the resultant configurations, namely visual inspection, has become tedious if not infeasible. In this paper, we wish to sketch a strategy for exploratory data analysis of simulations where populations are Boltzmann distributed according to some energy, loss, or cost function. The aim is to aid the investigator in quickly gaining an overview of the contents of the simulation and turning data into information.

A. Organization of this paper

Even though the strategy is more general, we will make use of concepts, terminology, and data from molecular simulations as an example. For instance, we will refer to the set of all configurations resulting from the simulation as a trajectory, irrespective of whether the underlying sampling algorithm is of the stochastic type (such as Monte Carlo) or the dynamic type (i.e., integrating some equation of motion). One of our aims is to introduce terminology from and build a bridge to the statistical and data mining disciplines, so the corresponding references will be to practical text books rather than the original sources. Also, established technical terms from those fields are typeset in boldface to emphasize that they are not our wording.

The remainder of this section describes the proposed analysis in rough and conceptual terms. Details are provided in Sec. II. A generic algorithm implementation of the analytic strategy is sketched, giving a range of alternative choices at many stages of the procedure. In our view, where choices are given the optimum approach is not yet clear and merits further investigation. Our special concern is to make clear where arbitrary decisions are required, and we have labeled these points as ''choosing'' or ''choice'' throughout.

To make things more specific, we trace one of the many possible paths through this combinational tree of choices in Sec. III and illustrate the analysis methodology using previously published raw data in Sec. IV.

Most building blocks of the propposed analytic strategy can already be found scattered in the literature. This previous work is more aptly summarized after some terminology has been introduced and so we defer citations and comparison to Sec. V before concluding with a discussion in Sec. VI.

B. Conceptual part

In an equilibrated thermodynamic system, the Gibbs free energy difference between two states, ΔG, is given, up to a factor, by the natural logarithm of the ratio of their occupancies, K:

$$\Delta G = -RT \ln K,$$

with R the ideal gas constant and T the temperature of the system. These occupancies are, in turn, given by an integral

over their densities in configuration space. That is, from the densities, the Gibbs free energy surface is known which determines all thermodynamic observables. Using additional models, more chemical properties of the system can be derived. For instance, the different preferred conformations correspond to free energy minima, and transition rates between these can be calculated from a kinetic model employing the free energy of transition states.[1,2]

A trajectory is a discretized path through configuration space with every frame or configuration corresponding to a single point. A choice has to be made as to which degrees of freedom of the system are considered relevant (step 1; numbers refer to the enumeration given in the following section) and how the system should be represented (step 2).

After proper downsampling of the trajectory (step 3), thus removing redundancies, the dissimilarity between all remaining configurations can be evaluated. The resultant dissimilarity matrix (which depends, of course, on the choice of the dissimilarity measure, step 4) can then be used to embed every configuration as a single point in configuration space such that the distances between the points match the dissimilarities previously calculated.

The interactions in a complex system typically confine it to a tiny fraction of the full high-dimensional configuration space.[3] In the case of a molecular system, the nominal dimensionality is three times the number of atoms minus some degrees of freedom for rotation, translation, and possible constraints. The effective dimensionality, however, is much lower due to the specific physics and chemistry of a molecule; for instance, two atoms involved in an unconstrained chemical bond cannot move independently but will always stay at a distance roughly corresponding to the equilibrium bond length. As a consequence, the points can be embedded in a low-dimensional subspace with little loss of information (step 5).

Based on the distribution of points in this low-dimensional subspace, a continuous estimate of the density can be obtained (step 6). A density estimation with no assumptions as to the underlying distribution is called nonparametric. An approach which is well-defined and mathematically well understood consists of centering a **kernel** (typically, these are unimodal symmetric smooth positive functions with unit area) on each data point and summing over all kernels to obtain a **kernel estimate** of the density. The choice of a functional form of the kernel turns out to be of minor importance compared to its **bandwidth** (step 6).

Of course, the resultant density and free energy surface can only be as good as the sampling throughout the simulation, both in terms of the regions that are explored at all and their relative populations. Good statistics concerning the latter can only be expected if multiple transitions over barriers have been observed.

A clustering (also called **unsupervised classification** or **numerical taxonomy**) may be performed on the configurations. A traditional approach is to cluster by maximizing intra-cluster similarity and inter-cluster dissimilarity; we propose instead to base membership on the topography of the estimated free energy surface (step 7).

If a significant part of the information lies in just two or three dimensions, direct visualization (step 8) of the free energy surface becomes meaningful. Otherwise, a glimpse of the density estimate in dimensions greater than three can still be conveyed by showing one graph per cluster with its topology characterizing the relative position of cluster members as well as the overall cluster shape.

More schematic diagrams, finally, can illustrate the observed transition probabilities between clusters and help interpret the dynamics of the system.

C. Overview: Concise description

We propose the following strategy or course of analysis (see next section for details):

(1) Choice of relevant degrees of freedom
(2) Choice of representation
(3) Choice of resampling rate
(4) Choice of measure of dissimilarity between two configurations
(5) Choice of method for embedding configurations in a low-dimensional representation of configuration space
(6) Kernel density estimation in that configuration space with choice of kernel bandwidth
(7) Choice of cluster analysis
(8) Visualization of:
  (a) Free energy surface
  (b) Clusters, cluster members
  (c) Basin spanning trees
  (d) Schematic diagrams
(9) Interpretation

II. TECHNICAL CONSIDERATIONS

A. Step 1: Choice of relevant degrees of freedom

Only the degrees of freedom directly related to the properties of interest should be included in the analysis to prevent an obfuscation of the relevant information.

B. Step 2: Choice of representation

The representation is most straightforward in the space in which the simulation has been performed, e.g., Cartesian or dihedral angle space. In cases where particles can exchange positions without altering the properties of interest, one should choose a representation taking into account this indistinguishability. An example is the ordered list of eigenvalues of an interparticle distance matrix, which is invariant under rotation or permutation.[4]

C. Step 3: Choice of resampling rate

The density estimation that follows assumes the data points it is based on to be **independently identically distributed**. For simulations of finite length, this means that the redundancy introduced by the small time integration step size in dynamic algorithms (required to keep the integration error low) or the small step size in stochastic algorithms (required to keep the acceptance ratio high) should be eliminated. A hint to a necessary, but not sufficient, condition is given by the following consideration: if K independently

identically distributed points $\{x_i\}$; $i=1...K$ are drawn one after the other from a probability distribution with the index specifying the temporal order, then the probability for the spatially nearest neighbor of a point x_k; $k=2...K-1$ to be either x_{k+1} or x_{k-1} is only $2/(K-1)$. In an actual trajectory, the probability for the spatially nearest neighbor to also be a temporal nearest neighbor is much higher, leading to a density estimate that is essentially a ''tube'' through configuration space. As a consequence, the resampling rate from the trajectory should be so low as to eliminate these redundancies. Ideally, the resampling rate should be a function of the density, i.e., lower in low-density regions. In practice, this would require an iterative procedure and a constant resampling rate is chosen.

D. Step 4: Choice of measure of dissimilarity between two configurations

We would like a single scalar δ_{ij} summing up the dissimilarity between two configurations i and j. The set of dissimilarities between all configurations obtained from the trajectory holds information about their relationship which we mean to exploit. The dissimilarity between a configuration and itself should be zero and all dissimilarities should be greater than or equal to zero. The procedure is also simplified if the dissimilarity is symmetric, i.e., $\delta_{ij}=\delta_{ji}$. If the dissimilarity measure obeys, in addition, the triangle inequality, it meets the formal requirements for a **metric**, but this represents a much more stringent criterion.

Metrics for molecular conformations include:

(1) the atom-positional root mean square distance (RMSD) (for proof of metric properties, cf. Ref. 5). A difficulty is that global dissimilarity can completely obscure high local similarity;[6] also, atom-positional RMSD is not very sensitive to greater changes in geometry.

(2) the dihedral angle difference (proof: in Ramachandran type plots with suitably chosen phases taking care of the periodicity, the dihedral angle distance becomes a Euclidean distance, which is a metric; cf. Ref. 7). The problem is that a change of a single dihedral in the middle of an elongated system can cause drastic changes in overall shape,[6,8] whereas the effects on molecular shape of two changes in two dihedrals can approximately cancel, even though a greater dissimilarity is predicted in the latter case.[9]

(3) the distance matrix error (for proof of metric properties, cf. Ref. 10) which measures the dissimilarity between two intramolecular distance matrices. This measure leads to low-dimensional configuration space representations,[11] but becomes problematic when the system can change its handedness[12] because a distance matrix cannot convey chirality.

E. Step 5: Choice of method for embedding configurations in a low-dimensional representation of configuration space

The most straightforward way of embedding a point in configuration space is by simply concatenating all coordinates characterizing the system in the selected representation into a single vector. Redundant degrees of freedom (such as rotation and translation if the properties of interest are invariant thereunder and the representation is in the laboratory frame) may be eliminated.

A more general approach, allowing for the use of different dissimilarity measures, makes use of the set of all dissimilarities between all configurations to embed each one as a point in configuration space. We suggest using **metric multidimensional scaling**,[13] which involves the diagonalization of a centered squared dissimilarity matrix, B.[14] If B is positive semi-definite, the points can be embedded in a Euclidean space so as to satisfy perfectly the supplied distances (note that a mere satisfaction of the triangle inequality is only a necessary, not a sufficient condition[15]). The normalized eigenvectors of B are the principal components of the system and the corresponding eigenvalues indicate the variance along these principal components. The coordinates obtained for the points described by the dissimilarity matrix are unique up to translation and inversion. If B is not positive semi-definite, the parts pertaining to the negative eigenvalues can be discarded or a constant can be added to all off-diagonal elements of B.[13]

The maximum dimensionality of the resulting cloud of points is equal to the number of points described by the dissimilarity matrix minus one. If the eigenvectors of B are ordered by the magnitude of their eigenvalues and the effective dimensionality of the configuration space is low, then the bulk of the information or variance is contained in the first few dimensions. The choice of the precise number of dimensions for further analysis is arbitrary, but should be modest, as we argue in the following and in step 7: while restriction to a lower-dimensional subspace will deteriorate the conservation of the provided dissimilarities, it also circumvents problems associated with the **curse of dimensionality**. This is a term covering many of the features to which our low-dimensional geometric intuition and spatial perception are unaccustomed. Noteworthy in the context of density estimation (step 6) is the vast volume of high-dimensional spaces, quickly making a cloud of points highly sparse (the following examples are taken from Ref. 16, chapter 4.5 and Ref. 3, chapter 7): the number of sample points required to achieve a constant bias and variance rises dramatically (at least exponentially) with the dimensionality. Furthermore, most of the probability mass quickly becomes concentrated in the tails, even of distributions with very light tails. As an example, in one dimension almost 90% of the probability mass of the normal density is confined to $|x|\leqslant 1.6$, whereas in ten dimensions, 99% of the probability mass lies at $|x|>1.6$!

Whatever the particular choice of dimension, one may either simply project all points onto the first few dimensions or project onto a low-dimensional linear subspace maximizing some **projection index**, i.e., some measure of the ''interestingness'' of a certain linear projection (this technique goes under the name of **projection pursuit**[17]).

Alternatively, and if the subspace sampled predominantly by the system is not approximately linear, the points may be mapped to a low-dimensional nonlinear subspace that allows for a more faithful rendering of the supplied dissimilarities. In this case, a **nonlinear mapping** or **nonmetric**

multidimensional scaling[13] may be performed (using either the original dissimilarities or the distances resulting from multidimensional scaling), optimizing criteria such as Kruskal's STRESS[18] (also called Sammon's mapping error[19]). These preserve the supplied dissimilarities much more faithfully, albeit at the cost of the axes losing their simple interpretability: in a linear mapping, a small displacement in the subspace corresponds to an atomic displacement vector of the entire system, and that correspondence is the same throughout the subspace; in a nonlinear mapping, this correspondence is only local.

F. Step 6: Kernel density estimation

The histogram used to be the only widespread nonparametric density estimator before the 1950's when kernel estimators were first introduced. Owing to its practical importance in many technical and scientific areas and spurred by the computational revolution, in the last decades the field of density estimation has become a scientific industry, developed mostly at the interface of statistics and computer science. Note that this step has not been entitled ''Choice of...'': On the one hand, asymptotically all nonparametric methods are kernel methods (Ref. 3, p. 125); on the other hand, kernel estimators (also called Rosenblatt or Parzen estimators) even in the narrow sense are the most appropriate tool for the proposed analysis in our view.

The kernels are usually smooth, symmetric, and unimodal and are density functions themselves (i.e., they are positive everywhere and integrate to unity), although even this restriction can be relaxed in attempts of reducing bias resulting from oversmoothing. Their support can be finite or infinite. We reiterate that the choice of a functional form of the kernel turns out to be of minor importance (e.g., Ref. 16, Chap. 3.2.2 or Ref. 20, Chap. 2.7) compared to its bandwidth, also denoted window width or smoothing parameter. The bandwidth can be the same for all data points or it can be greater in the tails of the distribution (using **variable** or **adaptive kernel estimates**), i.e., in badly sampled low-density regions.

Different strategies have been proposed for data driven bandwidth selection:

- subjective (manual, interactive) choice based on the appearance of the density or its derivatives[21]
- various cross-validation and bootstrap methods maximizing likelihood or minimizing the integrated square error (see Refs. 16, 20)
- so-called plug-in methods based on formulas that are asymptotically exact for infinitely large samples (see Refs. 16, 20).

The subjective methods yield, by definition, nonreproducible choices. The cross-validation methods usually have difficulty with distributions featuring heavy tails, resulting in oversmoothed density estimates; the plug-in methods require an estimate of the true underlying density or its derivatives and may have problems with multimodality (the presence of many local maxima) or nonnormality. An intermediate strategy may be to parametrize a functional of the dissimilarity matrix predicting the optimal bandwidth for a restricted class of underlying distributions, e.g., from peptidic systems.

There is no universally valid method of determining the optimal bandwidth, just as there is no universally valid cutoff in cluster analysis. Extrinsic knowledge or **metadata**,[22] i.e., information not contained in the supplied numerical data itself, has its rightful place in a meaningful and result-oriented analysis of real-world data. Two entirely unrelated experiments may accidentally give rise to the same data set, but the context of the investigations may require entirely different density estimates based on the same data.

G. Step 7: Choice of cluster analysis

While most clustering criteria optimize some geometric measure of intracluster compactness and inter-cluster separation (for a concise overview of methods, cf. Ref. 23), we propose clustering by membership of local density maxima or, equivalently, free energy minima. This is because we wish to identify ''islands of stability'' in configuration space without *a priori* assumptions about their shape, for instance, biasing in favor of spherical distributions (as in centroid clustering, e.g., Ref. 24) or dense chains (as in single-linkage clustering, e.g., Ref. 5).

For the sake of argument, consider for a moment the topography of a mountainous region and admit that we wish to cluster by membership to local minima. We define the boundaries of membership as the union of all watersheds or ridges, with the catchment regions or basins corresponding to the set of all members of a particular local minimum.

When based on previously embedded data points, the kernel density estimate provides us with a continuous free energy surface; however, the description of the watersheds or cluster boundaries rapidly becomes untractable in higher dimensions. Moreover, a continuous description is not necessary because we only wish to cluster the data that is actually available, i.e., discrete points. Accordingly, a discrete description of the basins is sufficient. We propose constructing disjoint directed graphs assuming the shape of trees, one for each basin or catchment region, encompassing all data points within that basin.

Each tree root should be centered on the data point at which the free energy is lowest. If all vertices were directly connected to the root, the resultant graph would exhibit a primitive topology (comparable to a sea urchin) and not convey much information regarding the arrangement of the data points, let alone the topography. With regard to visualization, we desire something more similar to a **minimal spanning tree** (reminiscent, maybe, of ivy). Indeed, we want the topology to satisfy the graph theoretical definition of a **tree** and we will in the following call it ''basin spanning tree.'' We construct these trees as follows (arguing now in terms of the density rather than the free energy, thus clustering by membership to local maxima rather than minima):

- FOR EACH data point i, find all Delaunay[25] neighbors i_j
  - FOR EACH Delaunay edge e_{i_j}
    check whether the density estimate along that edge attains values below the density estimate at point i; if so, discard that edge;

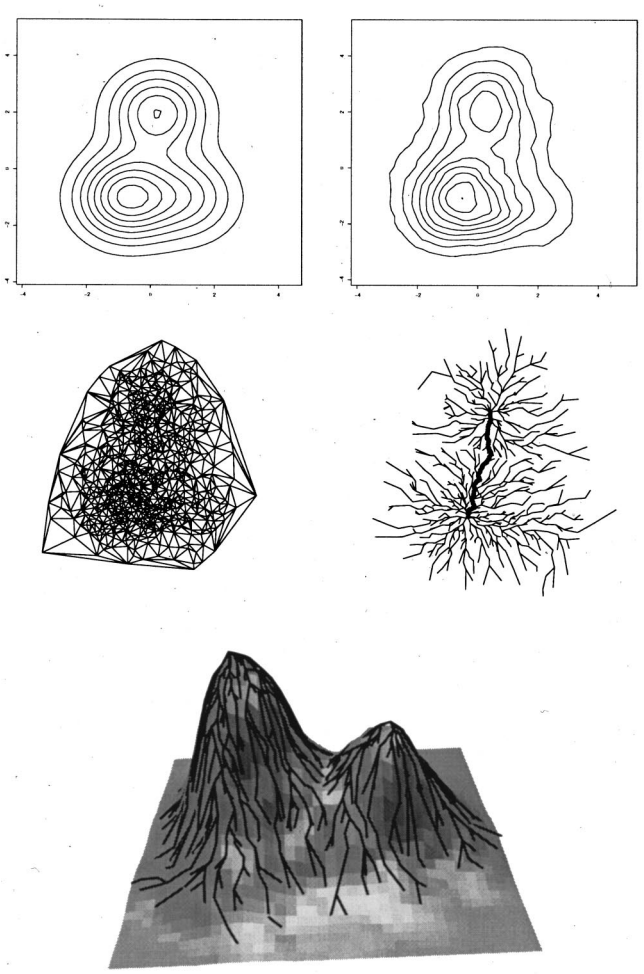

FIG. 1. Illustration of the construction of basin spanning trees. The top left shows an analytical probability density function, the middle left displays 900 points which have been sampled from it and their Delaunay triangulation; the top right shows a density estimate generated from these points and the middle right gives the adjoint basin spanning trees, along with the maximum density path (bold line; see text for definitions). The bottom part renders the density estimate and its basin spanning trees in three dimensions. This example also illustrates that already in two dimensions, a large number of samples is required to accurately estimate a rather simple density. Given a constant number of sample points, the quality of the density estimate that can be obtained from it deteriorates dramatically with growing dimensionality.

-IF there are any remaining edges, calculate the density gradient by dividing the density difference by the distance between the embedded points; let the edge offering the maximum gradient become a new edge of the basin spanning tree;
ELSE point i becomes the root of a basin spanning tree.

The principle is illustrated graphically in Fig. 1: the top left part shows an arbitrary bimodal probability density function which is the sum of three spherical Guassians with unit standard deviation but different weights. The middle left part displays 900 points which have been sampled from the probability density as vertices of their Delaunay triangulation. The top right part shows a density estimate from these 900 points and the middle right illustrates the two basin spanning trees which are disjoint subgraphs of the Delaunay triangulation. The bold line indicates the maximum density path between the two maxima, cf. step 8. In this demonstration, the basin spanning trees only serve to cluster the points and indicate membership. When derived in higher dimensions, they can offer an impression of the higher-dimensional density estimate, cf. step 8. The bottom part gives an alternative rendering of the continuous density estimate along with its basin spanning trees.

A **fuzzy clustering** does not assign objects unequivocally to one or another cluster, but assigns fractional memberships. This often yields a more natural description; in our case, if there are two nearby density maxima connected by a relatively high saddle point, it is not obvious why a point on one side of the saddle should belong entirely to one cluster, while another point lying at a very short distance on the other side of the saddle should belong entirely to the other cluster. Such points should, to some extent, belong to both density maxima. In the present analysis, fuzzy membership of every point to some cluster can be obtained by first approximating the density estimate as a **mixture density** (i.e., a sum of parametric component densities weighted by mixing parameters[26]) and then examining which component contributes how much to the total density at each data point. The parameters can be estimated by means of **maximum likelihood** or extensions such as **expectation maximization**.

In our strategy, obtaining density estimates from one data set with different kernel bandwidths is related to hierarchical cluster analysis.[27,28] Furthermore, there are intimate links between the density estimation and clustering proposed here and **blind source identification** and **regularization**.[29]

H. Steps 8, 9: Visualization and interpretation

If the first two or three dimensions of configuration space carry much of the overall variance, a meaningful representation of the free energy surface can be obtained by plotting iso-contours or iso-surfaces of the free energy, obtained from the negative logarithm of the density estimate. Of course, these representations suffer from the distortions caused by the projection and will generally not be quantitative. We stress again that the free energy surface can only be as good as the sampling in the simulation.

Properties of interest, such as the presence of H-bonds or amount of solvent-accessible surface, can be mapped onto the density.

Also, if conformational states can be defined using continuous boundaries, then their entropies can be calculated through $\int p \log p \, d\tau$, where p is the continuous density estimate with the integration going over the configuration space of a state.

Convergence of a simulation can be monitored by reestimating free energy differences using ever longer portions of the trajectory.

If the effective dimensionality is substantially greater than three, the outcome of the cluster analysis can be used to enhance the amount of information conveyed by a visualization; for instance, the basin spanning trees can be plotted with the locations of the vertices mapped down to three dimensions. Overlapping basin spanning trees can still visually convey a distinctness of clusters that a simple projection

does not reveal. Conceptually, this is equivalent to contouring of the density arising from different clusters separately and plotting all contours together, giving rise to intersecting contour lines. Additionally and by construction, the basin spanning trees allow gaining, by means of their topology, some appreciation of the topography of the higher-dimensional density estimate. This provides, along with the quest for a good density estimate, another motivation for a careful and modest choice of dimensionality for analysis: Unless the number of data points is vast, every point becomes a Delaunay neighbor of most others in higher dimensions. This is not only computationally unfavorable, but also makes for a primitive topology of the resulting basin spanning trees, such that they cannot reveal much about the relative arrangement of cluster members and about the topography of the high-dimensional free energy surface.

Summing up, acute angles in the basin spanning tree can indicate a local (intra-cluster) effective dimensionality higher than the one visualized (but not higher than the one used in the density estimate). Overlapping basin spanning trees can indicate a global (intercluster) effective dimensionality higher than the one visualized.

As a more schematic representation, the cluster centers (i.e., the basin spanning tree roots) can be drawn as spheres with their volume representing the populations (the number of members of one cluster); cylinders connecting these spheres can then represent the transition probability (as in Ref. 8), obtained either directly from the trajectory or from the overlap of clusters, which can be calculated as an integral over all space of the product of mixture components in a mixture density approximating the density estimate.

A subjective choice of the kernel bandwidth may be guided by the resultant clustering; a good choice should lead to a robust estimate such that minor variations in the bandwidth do not lead to splitting or coalescence of significant clusters. A graphical technique based on this idea has been illustrated in Ref. 28.

Given all basin spanning trees, a minimum energy path between two adjacent clusters can be found by identifying those two leaves between the two corresponding basin spanning trees that are Delaunay neighbors and feature the highest minimum density along their shared Delaunay edge. From that edge, representing the transition between the two clusters, the path can be followed to the respective cluster centers. This strategy will not find favorable paths that involve passages through other clusters. This discretized minimum energy path will encompass only configurations that were actually visited during the simulation. These may, in turn, be used to visualize the nature of the transition. The minimum free energy path can deviate strongly from a straight line if the two clusters involved are significantly nonspherical.

III. CHOICE OF METHODS

This section describes our selections in the combinatorial tree of choices. We have usually implemented and applied the most straightforward choice.

All analyses were performed with *R*.[30] I/O and molecular superposition was handled by Fortran and Perl programs.

We used trajectories from previously published studies on a β-heptapeptide[31] (200 ns at 298 K and 340 K, 50 ns at 350 K and 360 K, all in methanol) and an aminoxy acid trimer[32] (70 ns at 293 K and 340 K in chloroform and 25 ns at 300 K and 340 K in water).

We eliminated the degrees of freedom of all atoms

- from the solvent
- from the first and last residues in the β-heptapeptide
- from the protective groups in the aminoxy acid trimer
- that are not directly covalently connected to the molecular backbone.

A Cartesian representation was chosen with a resampling rate of 10 ps.

We used atom-positional RMSD between 500 to 1000 configurations sampled evenly from one (or more, in the case of the joint embedding of different simulations as in Fig. 4) molecular simulation trajectory to perform a metric multidimensional scaling. Distances from all configurations relative to the ones used in multidimensional scaling were employed to embed these additional points according to Ref. 33 into a space of a dimensionality accounting for 95% of the total variance. The points were then projected down to a low dimension (two or three), without attempting to relax the distortions thus introduced. The resulting distance matrix was too large to store in memory, so its elements were recomputed when required.

Delaunay triangulation was performed using qhull.[34] Delaunay neighbors were stored in a list structure. The Epanechnikov kernel (the tip of a paraboloid, thus with finite support)[16] was used for density estimation, and the bandwidth was chosen as three times the median of the nearest-neighbor distances [this dependence on the number of configurations N is not proportional to the correct asymptote $D^{1/(D+4)}N^{-1/(D+4)}$ (Ref. 16, p. 85) for dimensionality D]. The Delaunay graph was used for a range search, i.e., to find for every point all others that lie within a distance given by the kernel bandwidth.

For visualization purposes, the estimated density was calculated at all mesh points of a two- or three-dimensional grid. (To accelerate repeated calculations with different bandwidths, a list structure was used once again.)

Three-dimensional iso-surfaces were computed using polyr.[35] Visualizations were performed with *R* or Geomview.[36]

For clustering, basin spanning trees as defined previously were approximated by discarding all Delaunay edges longer than the kernel bandwidth and assuming that all remaining edges are valid candidates, with the winner maximizing the density gradient. We have also implemented the strict and much more computing-intensive definition (results not shown), but have found the approximation to yield the desired topology.

IV. RESULTS

Figure 2 gives the cumulative sum of the eigenvalues from metric multidimensional scaling for two different systems: the β-heptapeptide represents a benign case where the

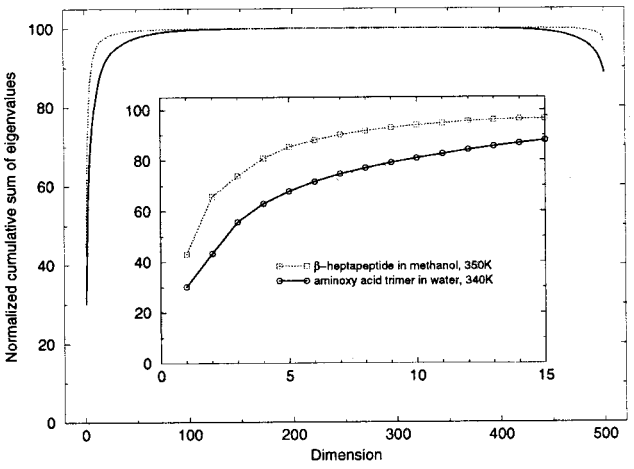

FIG. 2. Eigenvalues from metric multidimensional scaling based on backbone atom-positional RMSD dissimilarity matrices for two oligoamide simulations, illustrating a benign and an unfavorable case.

two (three) principal components carry 66% (74%) of the overall variance; the negative eigenvalues, corresponding to imaginary coordinates, amount to 4% of the sum of all positive eigenvalues.

The data on the aminoxy acid trimer show the worst case we have encountered: the two (three) principal components carry only 43% (56%) of the overall variance and the negative eigenvalues amount to 11% of the sum of all positive eigenvalues, thus making metric multidimensional scaling less suitable. For the same worst case, the distribution along the first 15 dimensions is illustrated in Fig. 3. The increasing normality with higher dimension forms the theoretical basis of the essential dynamics method.[37]

Due to space limitations, we show results only on the aminoxy acid trimer in the following, thus selecting the more difficult case to show the robustness of the strategy for analysis.

The top of Fig. 4 shows iso-surfaces of the estimated configuration space densities for the aminoxy acid trimer in chloroform at 293 K (dark) and 340 K (light). Much of the density lies close to part of the surface of a cylinder. From this figure, it is also apparent that two simulations in the same solvent and at different temperatures explore largely the same parts of configuration space with the exception of the space around cluster numbers 15, 19 (cf. Fig. 5) which is apparently disfavored entropically at higher temperatures. Change of solvent, on the other hand, drastically changes the ensemble generated in the simulation, illustrated by the almost orthogonal density estimates in the lower part of Fig. 4. The viewpoint was chosen arbitrarily to enhance the spatial impression and has been retained throughout.

This figure also illustrates the indifferentiation[38] of the atom-positional RMSD, leading to an artificial sphericity of the points produced by the embedding.

Figure 5 displays the location of all cluster centers or basin spanning tree roots relative to the density estimate for the aminoxy acid trimer in chloroform at 293 K (normal type) and for selected cluster centers in water at 340 K (boldface). The former are all inside the isosurface which has been

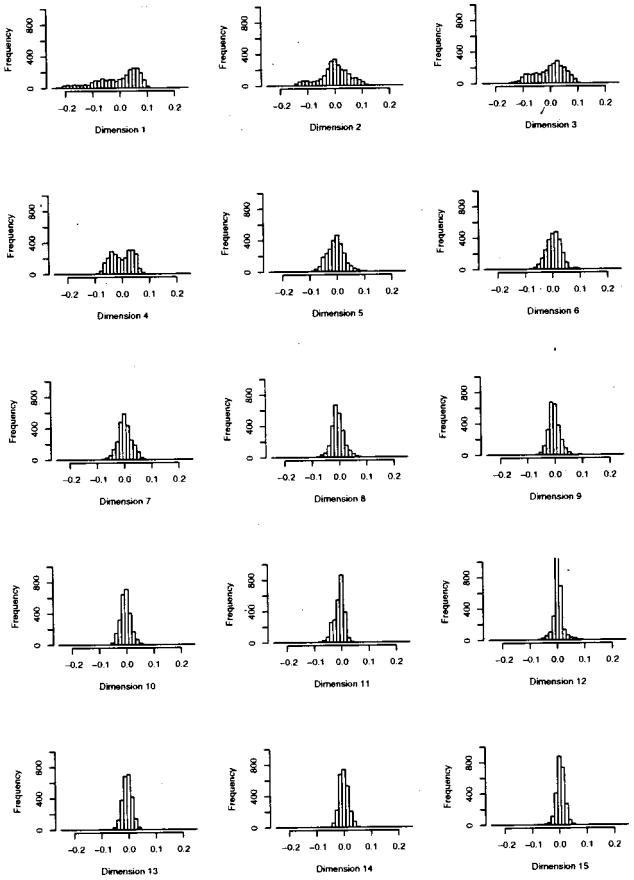

FIG. 3. Distribution of points along the first principal components for the unfavorable case of the aminoxy acid trimer in water at 340 K.

rendered transparently. The sizes of all these basin spanning trees are given in Fig. 6.

Figure 7 shows which clusters are visited over time by the aminoxy acid in chloroform at 293 K, and Fig. 8 illustrates the structures corresponding to different cluster centers, along with some members.

Figure 9 shows the density estimate along the two first principal components. The viewpoint is chosen differently from the one before, but the three principal mountain ranges correspond, in descending altitude, to clusters no. 1, 8, and 15 and their surroundings. Floating above the density estimate, the basin spanning trees derived in three dimensions are shown, with the colors coding the three-dimensional density estimate. The trees are seen to overlap heavily when projected down to two dimensions. The projection to low dimensionality can only produce interpoint distances shorter than or equal to those in the higher dimension. The clustering radius used on the same data in Ref. 32 was 0.07 nm.

In our implementation making use of Gower's embedding[33] and of neighborhood and range lists, the computational complexity is bounded, for dimensionality $D>2$, by the Delaunay triangulation; an optimal algorithm for it is $\mathcal{O}(N^{D/2})$ in the worst case. In practice, this estimate is too pessimistic[39] and construction of the range lists is the most time-consuming part. The computational cost for the examples supplied was of the order of several hours on a PC as compared to weeks for the simulations on similar machines.

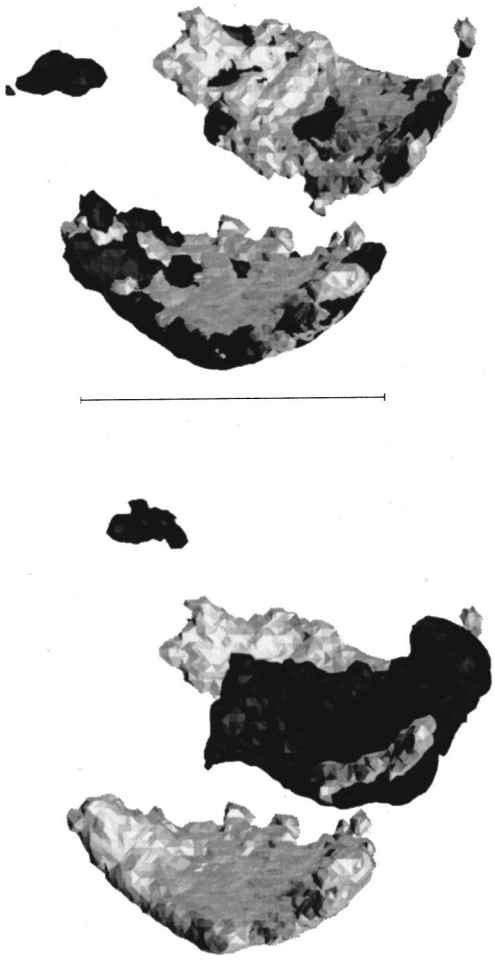

FIG. 4. Top: three-dimensional iso-contours for density estimate of configurations embedded in configuration space by metric multidimensional scaling for an aminoxy acid trimer in chloroform at 293 K (dark) and at 340 K (light). The two simulations explore similar regions of configuration space. Bottom: iso-contours for the aminoxy acid trimer in chloroform at 340 K (light) and in water at 340 K (dark). The different solvent leads to a completely different distribution in configuration space.

R is an interpreter and a compiled executable would shorten execution time significantly.

V. RELATION TO PREVIOUS WORK

Multidimensional scaling techniques and/or clustering have been used for the analysis of molecular simulations in a series of investigations.

The earliest attempt at reaching a low-dimensional representation of configuration space that we could trace is by Diamond[40] and Levitt.[41,42] The embedding of additional configurations relative to some reference structures has been demonstrated by Abagyan and Argos.[43]

Hierarchical clustering was used by Shenkin and McDonald[5] and Rooman *et al.*[44] Two clustering algorithms were compared and single linkage found to be inappropriate by Torda and van Gunsteren.[9] Clustering based on convex molecular hulls was shown by Lin *et al.*[45] The data used in the present study have already been analyzed with a nonhierarchical cluster analysis.[31,32]

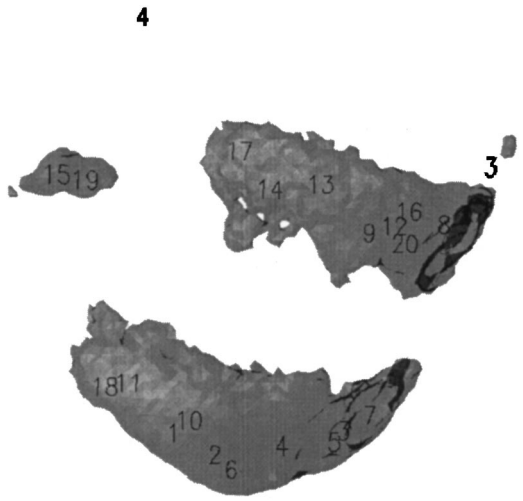

FIG. 5. For the aminoxy acid trimer in chloroform at 293 K, location of the first 20 cluster centers relative to the iso-contour previously shown (rendered transparently). The bold figures show the locations of the cluster centers three and four from the simulation in water at 340 K, indicating regions that are only sampled well in the water simulations.

An early reference to hierarchical configurational clustering and local principal components analysis is due to Murray-Rust and Raftery.[6] Multidimensional scaling and hierarchical clustering were employed by Troyer and Cohen.[46] Hierarchical clustering of conformations and an approximative visualization thereof assuming normality was performed by Caves *et al.*[47] Embedding of clusters in 2D using a nonlinear mapping and schematic representation of transition probabilities has already been demonstrated in a paper by Karpen *et al.*[8] Clustering by basins on the free energy surface has also been envisaged, but not performed, by these authors as well as by Bravi *et al.*[48]

A basin volume estimation using the convex hull of clusters should be more precise than one based on ellipsoids as used in Ref. 49.

Becker and Karplus have indirectly characterized and visualized an energy surface through a disconnectivity graph carrying topographical information;[2] see also Levy and

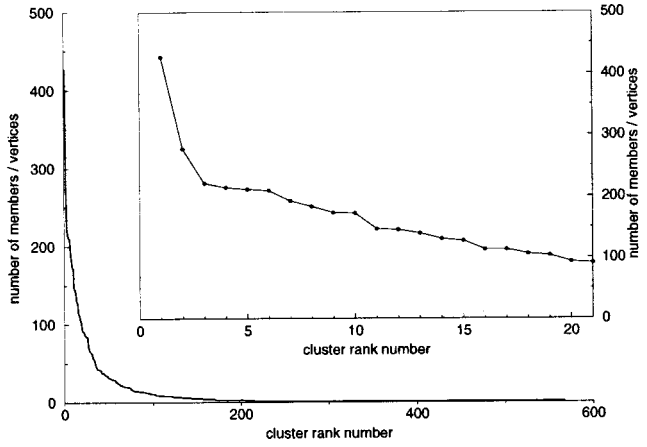

FIG. 6. Sizes of all clusters or basin spanning trees for the aminoxy acid trimer in chloroform at 293 K. Clusters rank number 217 and higher are singletons.

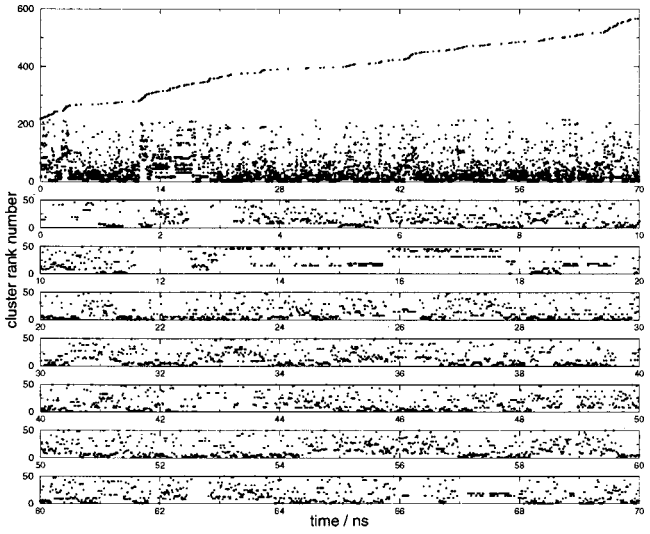

FIG. 7. Top: membership of each configuration to different clusters; clusters are ordered according to size and first visit. Clusters rank number 217 and higher are singletons. Bottom: detailed membership to largest 50 clusters over time.

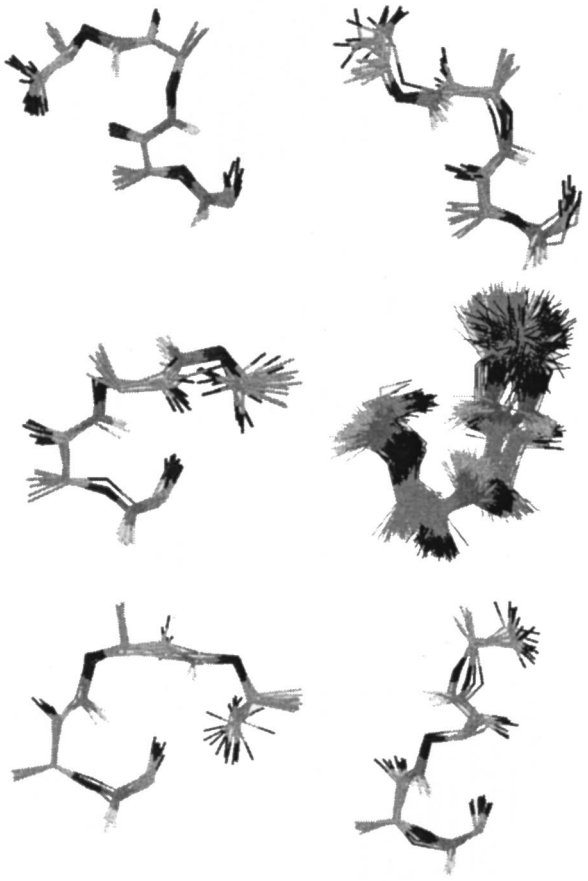

FIG. 8. For the aminoxy acid trimer in chloroform at 293 K, cluster center and the nine configurations mapped closest to it for cluster no. 1 (top left), cluster no. 9 (top right), cluster no. 15 (middle left); entire cluster no. 15 (middle right). For the aminoxy acid trimer in water at 340 K, cluster center and the nine configurations mapped closest to it for cluster no. 3 (bottom left) and cluster no. 4 (bottom right). Only the atoms used in the superposition and atom-positional RMSD calculation are shown; the spread in the clusters shows both their density and the shortcomings of the mapping to a low dimension. The two distinct conformations visible in the middle right figure are not resolved into two separate clusters with the current choice of dissimilarity measure, dimensionality, and kernel bandwidth.

Becker.[50] Guo *et al.*[51] have plotted free energy surfaces with respect to the radius of gyration, number of native contacts and number of hydrogen bonds. Other plots of energy surfaces[49,52] were constructed by smoothing over enthalpic energies of individual configurations. This is problematic because single force-field terms correlate badly with free energy; see figures in Refs. 52, 53. The nonequivalence of clustering by molecular similarity or by the topography of an energy surface has been illustrated by Becker.[54]

Alternative approaches to the visualization of distortions arising from a mapping into a low dimension are given by Bienfait and Gasteiger.[55]

VI. DISCUSSION

We have made an attempt at a description of an analytic strategy that is both general and consistent. Our generic description should encompass a large variety of particular approaches and allow for them to be conveniently defined by specifying the choices in Sec. I C. Within the proposed strategy, we have in many steps opted for the simplest choice. This has led to a procedure the robustness of which we have illustrated by showing results on the case with the highest intrinsic dimensionality we have encountered. The merit of alternative techniques, many of which are mentioned in Sec. II, should be evaluated in a systematic fashion.

We hold that multidimensional scaling techniques in conjunction with cluster analysis can help in quickly gaining an overview of the data produced in a long simulation of a complex system in equilibrium and enhance its intuitive understanding. For instance, Fig. 5 suggests that there are three distinct regions of configurational space which are sampled and that looking at only 3 out of the first 15 clusters may suffice to gain a first impression of the system's behavior. The utility of multidimensional scaling then lies in telling the investigator which few clusters out of a large number should be scrutinized, preventing his or her being overwhelmed by the sheer amount of data.

We believe that making the detour of estimating a configuration space density is a more well-defined approach to the visualization of free energy surfaces than those previously pursued.

We also argue that this detour may be an asset in cluster analysis: Clustering by the topography of the free energy surface has (especially in the limit of long simulations, allowing for density estimates with both low bias and low variance) the potential of revealing similar but distinct conformational states that are separated by a high but narrow free energy barrier. These states would, due to their geometrical similarity, be lumped together by centroid clustering and related methods using a fixed cutoff.

Proper choice of dimensionality for the analysis is indispensable for a good density estimate and profitable visualization of density estimates in more than three dimensions. It is, however, the choice of kernel bandwidth that is of paramount importance: The topography of the free energy sur-

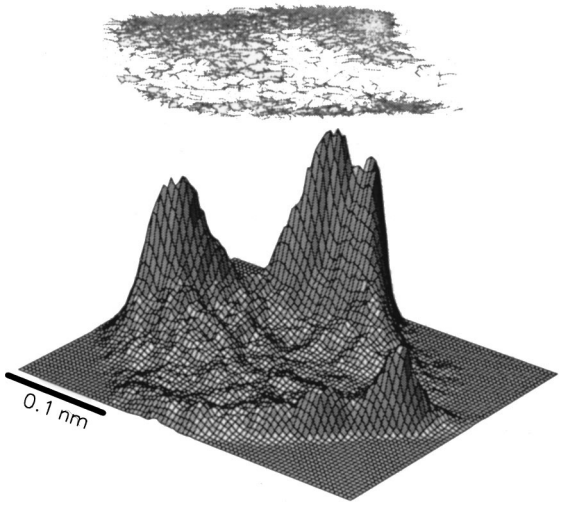

FIG. 9. Two-dimensional projection of the configuration space density estimate for the aminoxy acid trimer in chloroform at 293 K. Floating above are all basin spanning trees derived from a three-dimensional density estimate and clustering with the color coding for the three-dimensional density estimate. The overlap of the basin spanning trees gives an impression of the projection error.

face depends on it, with greater bandwidths provoking coalescence of individual mountains or modes of the density estimate. It is through this parameter that prior (chemical) knowledge can enter the analysis. Ultimately, the bandwidth should always be specified by the human expert, as a function of the class of systems under study and the questions posed to it.

There are two basic types of enhanced sampling algorithms: those driving the system *to* an undersampled region—often requiring manual intervention and detailed prior knowledge—and those driving the system *away* from oversampled regions.

Our perspective on analysis reveals intimate links to the latter approach, where well explored regions of configuration space are typically made less favorable by adding some kind of repulsive potential, thus encouraging transitions over barriers to previously unexplored regions. This idea has been exploited in global optimization methods such as combinatorial optimization (e.g., tabu search[56]), evolutionary algorithms (niching techniques, e.g., Ref. 57) and molecular dynamics (e.g., conformational flooding,[58] local elevation search[59]). We argue that many of these methods implicitly perform a (coarse) configuration space density estimation. In multidimensional adaptive umbrella sampling,[60] for instance, the histogram is chosen as a nonparametric density estimator and a smooth density is obtained by fitting the histogram with a set of basis functions which have to be selected manually. This two-step procedure is mathematically less well-behaved than a kernel density estimate with a kernel that is smooth everywhere.

Furthermore, the preselection of axes that are used to span the low-dimensional configuration space (e.g., radius of gyration and proximity to native fold[61]) may facilitate interpretation, but lead to mapping errors that are greater than if one lets the system decide—as in multidimensional scaling—which its most important dimensions are.

The reader can now anticipate where enhanced sampling methods may go in the future: Instead of (subjectively) deciding when a basin has been sampled sufficiently and then adding a repulsive potential[58] or refining the density estimate after each of a series of runs,[60] we propose using an *online* density estimation based on a data-driven bandwidth selection method that *continuously* updates and refines the density estimate as the simulation proceeds. A configurational linear subspace should be chosen by the system itself, and the kernel density estimate obtained therein should then be taken into account alongside with the potential function in every step of the simulation. To reduce computational cost and memory requirements, the kernel density estimate (the sum of thousands of kernels) can be replaced by a mixture density (the sum of only a few components; see Sec. II G; Step 7). In the further course of the simulation, that mixture density can be used in addition to the increasing number of kernels as a repulsive potential until an updated mixture density replaces the newly accumulated kernels, and so on.

Ultimately, the converged density estimate should completely flatten out those regions of the free energy surface which are accessible to the system at the given temperature. Driving the system away from where it has already spent much time arguably introduces less bias than attracting it to where it has not been before, as is the case with all kinds of reaction path methods.

We believe that enhanced sampling methods should take advantage of the more advanced density estimation techniques that have become available.

ACKNOWLEDGMENTS

We would like to thank M. Mächler from the Seminar for Statistics for valuable advice and the contributors to R, qhull, Geomview and polyr for developing such powerful tools and making them freely available.

[1] J. P. K. Doye and D. J. Wales, J. Chem. Phys. **105**, 8428 (1996).
[2] O. M. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).
[3] D. W. Scott, *Multivariate Density Estimation* (Wiley, New York, 1992).
[4] F. R. Manby, R. L. Johnston, and C. Roberts, MATCDY **38**, 111 (1998). In systems composed of different chemical elements, separate distance matrices are required for each element.
[5] P. S. Shenkin and D. Q. McDonald, J. Comput. Chem. **15**, 899 (1994).
[6] P. Murray-Rust and J. Raftery, J. Mol. Graphics **3**, 50 (1985).
[7] T. F. Havel and K. Wüthrich, J. Mol. Biol. **182**, 281 (1985).
[8] M. E. Karpen, D. J. Tobias, and C. L. Brooks III, Biochemistry **32**, 412 (1993).
[9] A. E. Torda and W. F. van Gunsteren, J. Comput. Chem. **15**, 1331 (1994).
[10] T. F. Havel, Bull. Math. Biol. **45**, 665 (1983).
[11] N. Elmaci and S. Berry, J. Chem. Phys. **110**, 10606 (1999).
[12] F. E. Cohen and M. J. E. Sternberg, J. Mol. Biol. **138**, 321 (1980).
[13] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling. Monographs on Statistics and Applied Probability* (Chapman & Hall, London, 1995).
[14] One element of B is given by $[B]_{rs}=1/2(-\delta_{rs}^2+1/n\Sigma_{s=1}^n\delta_{rs}^2+1/n\Sigma_{r=1}^n\delta_{rs}^2-1/n^2\Sigma_{r=1}^n\Sigma_{s=1}^n\delta_{rs}^2)$ with δ_{rs} as the dissimilarity between configurations r and s and n the number of configurations. The need for diagonalization of B makes metric multidimensional scaling impractical on more than a few thousand configurations on current computers. However, additional points can be embedded in that space, using only dissimilarities to the points previously embedded (Ref. 33). The nature of the embedding will depend on which points were used to span the space; in an iterative procedure, it is conceivable to use either a diverse subset of points, leading to similar distortions for all points, or to use a representa-

tive subset, reducing the distortion in those regions which are most heavily populated.
[15] J. C. Gower and P. Legendre, J. Classif. **3**, 1 (1986).
[16] B. W. Silverman, *Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability* (Chapman & Hall, London, 1986).
[17] M. Jones and R. Sibson, J. R. Stat. Soc. Ser. A **150**, 1 (1987).
[18] J. B. Kruskal, Psychometrika **29**, 1 (1964).
[19] J. W. Sammon, Jr., IEEE Trans. Comput. **C-18**, 401 (1969).
[20] M. P. Wand and M. C. Jones, *Kernel Smoothing. Monographs on Statistics and Applied Probability* (Chapman & Hall, London, 1995).
[21] B. W. Silverman, Biometrika **65**, 1 (1978).
[22] D. J. Hand, *Advances in Intelligent Data Analysis* (Springer, New York, 1997), pp. 1–14.
[23] J. M. Barnard and G. M. Downs, J. Chem. Inf. Comput. Sci. **32**, 644 (1992).
[24] H. L. Gordon and R. L. Somorjai, Proteins: Struct., Funct., Genet. **14**, 249 (1992).
[25] A space with a number of embedded points can be decomposed into cells, the Voronoi polyhedra, each comprising one point such that the space enclosed by a cell is closer to the corresponding point than to any other. A Delaunay diagram is the dual of its Voronoi diagram (see any textbook on Computational Geometry): every two points whose Voronoi polyhedra share a face become Delaunay neighbors. If the points are nondegenerate, their Delaunay graph is a triangulation. Delaunay triangulations are optimal in the sense of avoiding long, skinny triangles.
[26] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York, 1973).
[27] K. Rose, E. Gurewitz, and G. C. Fox, Phys. Rev. Lett. **65**, 945 (1990).
[28] M. C. Minnotte and D. W. Scott, J. Comput. Graph. Stat. **2**, 51 (1993).
[29] B. D. Ripley, *Pattern Recognition and Neural Networks* (Cambridge University Press, Cambridge, 1997).
[30] R. Ihaka and R. Gentleman, J. Comput. Graph. Stat. **5**, 299 (1996). http://www.r-project.org/.
[31] X. Daura, W. F. van Gunsteren, and A. E. Mark, Proteins: Struct., Funct., Genet. **34**, 269 (1999).
[32] C. Peter, X. Daura, and W. F. van Gunsteren, J. Am. Chem. Soc. **122**, 7461 (2000).
[33] J. C. Gower, Biometrika **55**, 582 (1968).
[34] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, ACM Trans. Math. Softw. **22**, 469 (1996). http://www.geom.umn.edu/locate/qhull.
[35] J. J. Jensen, http://hendrix.imm.dtu.dk/software/.
[36] T. Munzner, S. Levy, and M. Phillips, http://www.geom.umn.edu/sotware/download/geomview.html.
[37] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Proteins: Struct., Funct., Genet. **17**, 412 (1993).
[38] J. de Leeuw and I. Stoop, Psychometrika **49**, 391 (1984).
[39] S. Fortune, *In Computing in Euclidean Geometry*, edited by D.-Z. Du and F. Hwang, *Vol. 4 of Lecture notes series on Computing*, 2nd ed. (World Scientific, Singapore, 1995), pp. 225–265.
[40] R. Diamond, J. Mol. Biol. **82**, 371 (1974).
[41] M. Levitt, J. Mol. Biol. **82**, 393 (1974).
[42] M. Levitt, J. Mol. Biol. **168**, 621 (1983).
[43] R. Abagyan and P. Argos, J. Mol. Biol. **225**, 519 (1992).
[44] M. J. Rooman, J. Rodriguez, and S. J. Wodak, J. Mol. Biol. **213**, 327 (1990).
[45] T.-H. Lin, J.-J. Lin, Y.-F. Huang, and J.-H. Liu, J. Chem. Inf. Comput. Sci. **39**, 622 (1999).
[46] J. M. Troyer and F. E. Cohen, Proteins: Struct., Funct., Genet. **23**, 97 (1995).
[47] L. S. D. Caves, J. D. Evanseck, and M. Karplus, Protein Sci. **7**, 649 (1998).
[48] G. Bravi, E. Gancia, A. Zaliani, and M. Pegna, J. Comput. Chem. **18**, 1295 (1997).
[49] O. M. Becker, J. Mol. Struct.: THEOCHEM **398–399**, 507 (1997).
[50] Y. Levy and O. M. Becker, Phys. Rev. Lett. **81**, 1126 (1998).
[51] Z. Guo, C. L. Brooks III, and E. M. Boczko, Proc. Natl. Acad. Sci. U.S.A. **94**, 10161 (1997).
[52] O. M. Becker, J. Comput. Chem. **19**, 1255 (1998).
[53] X. Daura, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark, J. Mol. Biol. **280**, 925 (1998).
[54] O. M. Becker, Proteins: Struct., Funct., Genet. **27**, 213 (1997).
[55] B. Bienfait and J. Gasteiger, J. Mol. Graph. Mod. **15**, 203 (1998).
[56] F. Glover, ORSA J. Comput. **1**, 190 (1989).
[57] D. Beasley, D. R. Bull, and R. R. Martin, Evolut. Comput. **1**, 101 (1993).
[58] H. Grubmüller, Phys. Rev. E **52**, 2893 (1995).
[59] T. Huber, A. E. Torda, and W. F. van Gunsteren, J. Comput.-Aided Mol. Design **8**, 695 (1994).
[60] C. Bartels, M. Schaefer, and M. Karplus, J. Chem. Phys. **111**, 8048 (1999).
[61] Z. Guo and C. L. Brooks III, Biopolymers **42**, 745 (1997).