Cluster Analysis of Molecular Conformations

PETER S. SHENKIN* and D. QUENTIN McDONALD
Department of Chemistry, Columbia University, New York, New York 10027

Received 29 July 1993; accepted 1 March 1994

ABSTRACT

We describe a method for locating clusters of geometrically similar conformers in ensembles of chemical conformations. We first calculate the pairwise interconformational distance matrix in either torsional or Cartesian space and then use an agglomerative, single-link clustering method to define a hierarchy of clusterings in the same space. Especially good clusterings are distinguished by high values of the separation ratio: the ratio of the shortest intercluster distance to the characteristic threshold distance defining the clustering. We also discuss other statistics. The method has been embodied in a program called XCluster, which can display the distance matrix, the hierarchy of clusterings, and the clustering statistics in a variety of formats. XCluster can also write out the clustered conformations for subsequent or simultaneous viewing with a molecular visualization program. We demonstrate the sorts of insight that this approach affords with examples obtained from conformational search and molecular dynamics procedures.

Introduction

Several computational techniques produce large numbers of conformations of a given chemical structure. Examples are molecular dynamics, Monte Carlo sampling, and conformational search methods. Usually one uses such a technique to find a particular quantity of interest (e.g., the lowest-energy conformation or the free energy of the system). Occasions arise, however, when one would like to understand better how the many conformations obtained are related to one another.

Following a conformational search, a user might wish to ask whether the conformations found from natural groupings—whether they occur in structurally related clusters or whether, on the other hand, they are distributed seemingly randomly throughout some region of conformational space. Following a dynamics or Monte Carlo simulation, one might wish to investigate how many structurally distinct classes were visited during the simulation, as well as the time sequence of their having been visited. This can lead to insight into whether

*Author to whom all correspondence should be addressed.

CCC 0192-8651/94/080899-18

the conformational space has been explored fully. We have adapted techniques derived from statistical cluster analysis to investigate questions of this nature and have embodied the resulting method in a program called XCluster. In its most usual mode of operation, XCluster takes as its input a data file containing a list of molecular conformations. The program searches for structurally related clusters based on an analysis of the matrix of pairwise conformational similarities. This matrix is constructed by the program based on a choice by the user from among several pairwise proximity measures. Statistics are calculated for the clusterings found, and the results are presented in a variety of graphical formats.

XCluster can write out the conformational clusters it finds, optimally superimposed and colored, in a molecular format readable by the MacroModel modeling and visualization program.[1] XCluster can also share information with a simultaneously running MacroModel process when the two programs are run simultaneously (e.g., the list of atoms on which XCluster is to operate may be specified by graphically "picking" in a MacroModel window). The program can be run either from an X-windows interface or as a batch program. If a batch run is performed, the results of the analysis may later be visualized using the X interface. All visual displays available from the X interface may be converted to PostScript files for later printing. Finally, XCluster can read arbitrary user-supplied pairwise distance data; thus it can also perform cluster analysis on nonmolecular data.

CLUSTER ANALYSIS

There is no universally agreed on definition of the terms *cluster* or *clustering*. Colloquially, if data points exhibit clustering, they fall into groups such that the similarities within the groups are significantly greater than the similarities between the groups. Algorithms to identify clusters within a set of data points vary tremendously and depend strongly on the nature of the data, which may, for example, be categorical rather than numerical. A general introduction to the subject may be found in the book by Jain and Dubes.[2] Other useful introductions are given by Sokol[3] and, in the context of chemistry, by Zupan.[4] In the language of Jain and Dubes, a clustering, the way we use the term, is first a partitioning of the set of points (molecular conformations) into subsets called clusters. The term *partitioning* implies that the subsets are mutually exclusive and that they cover the set, so that in any clustering, every point (conformation) is a member of exactly one cluster. Our method of assembling clusters is agglomerative, meaning that we start with the separated data points and, in a stepwise manner, unite them into clusters. It is also hierarchical, meaning that clusters forming at later stages (higher clustering levels) are unions of clusters formed earlier; clusters already formed never break apart during this process. Our method may further be described as a single-link clustering, meaning that for two clusters to unite, only a single pair of conformations, one from each of the two, must meet prescribed conditions.

Nearly all clustering methods are based on some notion of proximity, or distance, between pairs of data points. XCluster can calculate the matrix of interconformational distances based on the user's selection from among three choices of proximity measure, to be described in the Methodology section. All the measures we supply are Minkowski metrics, meaning, among other things, that they satisfy the triangle inequality; however, this is not true of all useful proximity measures. In particular, it need not be true of user-defined pairwise distance data supplied directly to XCluster for analysis.

The clustering algorithm used by XCluster begins by calculating the full proximity matrix and then sorting the entries using the quicksort[5] algorithm. Based on the known behavior of quicksort, the runtime of the program should go as $N^2 \log N^2$ in average performance as N (the number of conformations) tends to infinity; however, we find that with N up to about 1000 the calculation of the distance matrix elements dominates the computation, particularly using the Arms measure of proximity (see the section titled Distance Measures for an explanation of this term). Therefore, the overall runtime tends to go as N^2 in this range. Approximate methods that do not employ the full distance matrix may become necessary for data sets containing greater than about 10^4 points.

Several characteristics of the conformational clustering problem make it more amenable to treatment by a special-purpose program than by preexisting statistical or cluster analysis software packages. For each of our distance measures, all the coordinate axes in the corresponding space are commensurate; that is, they have the same units. If this were not true, it would be necessary to perform a preliminary rescaling of the data to make them commensurate. Such rescaling is necessary,

for example, prior to cluster analysis on a set of data points represented as vectors of diverse physical properties. Rescaling can produce artifacts, and because neither our method nor our data require it, we do not rescale. This simplifies our task.

In addition, XCluster incorporates algorithms that treat several aspects of the molecular clustering problem that are not usually encountered in clustering problems of a more general nature. These include the toroidal, as opposed to Cartesian, nature of torsional space; the possible existence of molecular symmetry; the utility of a particular distance measure (Arms; see the Distance Measures section), which provides the proximity matrix much more readily than it does the coordinates of the underlying data points themselves; and the desirability of performing three-dimensional superposition for subsequent display of the clustered conformations in their optimal mutual orientations. XCluster also computes several statistics that we believe to be novel.

PREVIOUS APPLICATIONS OF CLUSTER ANALYSIS AND RELATED METHODS IN CHEMISTRY

Cluster analysis is sometimes viewed as a method of pattern recognition, and pattern recognition methods have been used in analytical chemistry for tasks ranging from automated recognition of pure substances from their complex spectra to inference of the provenance of complex materials (e.g., archaeological artifacts) from trace-element analysis; a review is provided by Kowalski and Wold.[6]

A lot has been published on the application of cluster analysis to chemical database searching. Zupan[4] has given didactic examples of the use of cluster analysis in the hierarchical classification of chemical compounds, based on their molecular graphs. Okada and Wipke[7] have presented a fully worked out method and program that accomplishes this, and Willet[8] has reviewed the field. These authors are most concerned with clusterings defined in the discrete space of chemical graphs rather than a continuous conformational space. These sort of data naturally form a rigorous hierarchy, whereas we, in contrast, seek to determine whether the conformations under study exhibit a meaningful hierarchy.

Several authors have applied cluster analysis techniques to the detection of similarities among protein or nucleic-acid structures or substructures. Gordon and Somorjai[9] applied a method based on fuzzy sets, applied in Cartesian space, to the analysis of a molecular dynamics trajectory of a protein fragment. Karpen et al.[10] applied a nonhierarchical method in torsional space to an analogous trajectory. Gautheret et al.[11] applied a complete-link hierarchical method to the analysis of a dynamics trajectory for a nucleic acid sequence, also in torsion space. Holm and Sander[12] classified structures from the Brookhaven Data Bank using a Monte Carlo clustering algorithm. Their measure of conformational distance is based on the similarity of the internal C_α—C_α distance matrices for a pair of proteins; the comparison is done in a manner that takes possible shifts in sequence alignment into account.

Murray-Rust and Raftery[13] have described the use of single-link clustering in conjunction with principal component analysis for classifying β turns in peptides. More recently, Perkins and Barlow[14] employed a single-link clustering mechanism with user-specified cutoff distance to examine clustering of looped and cyclic structures in Cartesian space; Perkins and Dean[15] have described the use of Ward's method,[16] a different algorithm, for the culling of a 3D database of conformations for a group of flexible molecules. In somewhat similar work, Allen et al.[17] experimented with the use of several clustering algorithms, applied in torsion space, to search for similarities among fragments taken from the Cambridge Structural Database.

Finally, cluster analysis facilities have begun to appear in commercial and public domain software packages designed for chemical use. For example, the MD Toolchest from Wesleyan University[18] contains such tools.

Methodology

DISTANCE MEASURES

If the data supplied to XCluster are molecular conformations, the program can construct the proximity matrix using any of the following measures of pairwise distance between conformations:

1. Root mean square (rms) displacement between pairs of corresponding atoms following optimal rigid-body superposition. (Arms)
2. rms displacement between pairs of corresponding atoms in place (Nrms)

3. rms difference between corresponding torsion angles in pairs of structures (Trms).

We will refer to these criteria by the abbreviations given in parentheses, which correspond to the command names used in the batch command file of the XCluster program to specify them. These measures can all be used to specify how different two structures are, and they can be thought of as alternative definitions of the conformational distance, d_{ij}, between a pair of structures, i and j. When the Trms method is used, the measure corresponds to the use of the term *conformational distance* by Saunders.[19] XCluster provides for the specification of number-reflectional, number-rotational, and enantiomeric symmetry in its distance calculations to accommodate molecular symmetry, as described by Saunders.[19] When symmetry operations are active, the distance used for clustering is, for each pair, the minimum of the values obtained using all applicable symmetry combinations.

If the Arms distance criterion is specified, least-squares superposition of each pair of structures is carried out by the method of Kabsch,[20,21] and symmetry operations may be specified. This definition of distance does satisfy the triangle inequality† but does not immediately give rise to transformed sets of coordinates for the conformations which are simultaneously consistent with all the distances. Such coordinates may be calculated[22] and will generally lie in a space of greater than three dimensions, but these coordinates are not needed for the analysis described here, and the XCluster program does not calculate them.

In Nrms mode, the root mean square interatomic distance between corresponding atoms in a pair of structures is calculated without any attempt at moving one conformation to best superimpose upon the other, and symmetry operations do not apply. This mode is applicable to the study of a part or parts of a molecule which are allowed freedom to vary in the generation stage while the rest is held fixed. An example might be the conformational search of a loop or loops in a protein.[23]

CLUSTERING

Let us assume that we have a list of N items and have calculated the distances between every pair chosen from the list. These distances, d_{ij}, form a matrix **D**, which is symmetric because $d_{ij} = d_{ji}$ and which has zeroes along the main diagonal because $d_{ii} = 0$. **D** contains $N(N-1)/2$ nonredundant off-diagonal elements.

Let d^* be some distance, which we call a threshold distance. Given a value of d^*, we define two items, i and j, to be in the same cluster if $d_{ij} \leq d^*$. Thus if items a and b are closer together than d^* and items b and c are closer together than d^*, then a, b, and c will be in the same cluster even if $d_{ac} > d^*$. Clearly, if $d^* < \text{Min}(d_{ij})$, then there are N clusters, each consisting of a single item. If $d^* \geq \text{Max}(d_{ij})$, then there is a single cluster that contains all N items. In practice, a single cluster is formed long before d^* approaches $\text{Max}(d_{ij})$ because of the chaining together of items into the same cluster described just above. As d^* is increased through the range of the d_{ij}, clusters agglomerate. Suppose, for example, that clusters E and F (and possibly others) are present at a particular value of d^* and that e and f are specific items belonging to E and F. Because E and F are distinct clusters, we know that $d_{ef} > d^*$. Suppose that upon increasing d^* by a small amount we find that now $d_{ef} \leq d^*$. Under the new value of d^*, E and F are no longer distinct clusters; all their members agglomerate into a single cluster.

†We are indebted to Dr. Mathis Thoma of the CIBA-Geigy corporation for the following demonstration of this. First, let d_{ij} be the rms interatomic separation between conformations I and J following optimal rigid-body superposition of J upon I and let d'_{ij} be the rms interatomic separation between specific coordinate sets i and j for these two conformations. If the molecule in question has n atoms, then $d'_{ij} = d^e_{ij}/n^{1/2}$, where d^e_{ij} is the Euclidean distance between the two coordinate sets, regarding each as a vector in $3n$-dimensional space. Because the d^e measure satisfies the triangle inequality, so does the d' measure.

Now let A, B, and C represent conformations of a molecule, and let a, b, and c represent coordinate sets of the three conformations in arbitrary spatial orientation and position. Suppose we hold A fixed and apply the rigid-body transform to B that minimizes the residual mean-square displacement of corresponding atoms between it and coordinate set a. Call the transformed set of coordinates b'. We have, by definition, $d_{AB} = d'_{ab'}$. Likewise, let us transform C to superimpose optimally upon coordinate set b' and call the transformed coordinates c'. Then we have $d_{BC} = d'_{b'c'}$. From the previous paragraph, we know that $d'_{ac'} \leq d'_{ab'} + d'_{b'c'} = d_{AB} + d_{BC}$. Now suppose we determine the best rigid-body transform to superimpose C upon coordinate set a, and call the new coordinates c''. We then have $d_{AC} = d'_{ac''}$. But because this transform is guaranteed to give the minimum of the d' values between all possible pairs of coordinate sets for A and C, we have $d'_{ac''} \leq d'_{ac'}$, and therefore $d_{AC} \leq d_{AB} + d_{BC}$. This is the desired result, which holds for any conformations A, B, and C, in any order.

Imagine now that we have a sorted list of the d_{ij}. As we set d^* successively to each value on this list, agglomerations will occur only at certain values. If d^* is increased to a new value of d_{ij} such that i and j are already in the same cluster, no agglomeration will occur. An example is provided by the three items a, b, and c of the last paragraph. These items were in the same cluster even though d_{ac} exceeded d^*. If d^* were now increased to d_{ac}, the clustering would be unaffected. Because we start with N isolated data points, after $(N - 1)$ pairwise agglomerations all the points will be in a single cluster. Each agglomeration is considered a new clustering level, L. The procedure just described gives rise to N clustering levels. We consider $L = 1$ to index the state in which each data point constitutes a distinct cluster. Each level is associated with a critical threshold distance, d^*_L. For $L = 1$, we have $d^*_1 = 0$. For the other levels, $d^*_L = d_{ij}$ for some pair of data points i and j. The first agglomeration takes place at $d^*_2 = \text{Min}(d_{ij})$. Each value of L defines a clustering, which is the list of the clusters and their members present at that clustering level.

GENERIC ORDERING

Using these definitions of a clustering and of a cluster, it is possible, given the pairwise distances, to reorder the list of data points (conformations) into a new sequence such that at any clustering level, all points belonging to the same cluster will lie in a contiguous block. We call such a reordering, which is not unique, a generic ordering of the points. The process of building a generic ordering is carried out by a successive reordering of the input list of data items. This is performed as we ascend the sorted distance list. At the start, each data item is in a cluster of its own. At any level, we give each cluster a label, which is the number of the lowest-numbered conformation in it; thus, at the start ($L = 1$), each structure, i, in a cluster labeled i.

The shortest distance on the sorted list—d_{ij}, for some i and j, with $i < j$—is guaranteed to create a new clustering ($L = 2$) because it brings structures i and j, which were formerly in separate clusters, into a single cluster. Thus, at level 1, point i was in cluster i and point j was in cluster j. At level 2, we perform a partial reordering of the list by removing item j from its original position in the list and inserting it immediately after item i. Thus, data points i and j are now contiguous in the reordered list. At the same time, we maintain a record of the fact that at level 2, items i and j both belong to cluster i.

For each succeeding value on the sorted distance list, we first check to see whether the two data items separated by this distance are already in the same cluster; if so, this distance is not a critical threshold distance, and we move on. If the two data items are not in the same cluster, then the current distance is d^*_L for the next value of L (i.e., it is the critical threshold distance for the next clustering level) and we do another partial reordering of the list. Let us again call the two data items i and j. Let these elements be members of clusters I and J, where I and J are the lowest-numbered elements in their clusters. Let us suppose for the sake of discussion that $I < J$. We remove from their place in the list the contiguous items in cluster J and insert them as a group immediately after the last member of cluster I. We also assign all the items previously belonging to cluster J to cluster I.

Thus at each clustering level we take an entire contiguous cluster of data points and move it to a new position immediately following another such contiguous cluster and then join the two groups into a single, larger cluster. At no point do we break up a contiguous cluster previously formed, and at no point do we reorder the data points within a cluster previously formed. Thus, at the end of the process, at clustering level N, where only a single cluster remains, we have created an ordering which preserves the contiguity of clusters at all previous levels.

Figures 1a and 1b illustrate this process. Figure 1a represents an array of points in two-dimensional space; Figure 1b represents the clustering process for this set of points. The parentheses in the figure group those clusters with membership greater than one at each clustering level. At each new clustering level, a single cluster is moved to some location to the left of it in the diagram, if necessary, and then joined to cluster on its left. The order of the data items at the end of this process is the generic ordering.

We mentioned earlier that a generic ordering is not unique. The aforementioned process always keeps the first data item in the input sequence first in the generic ordering. If the input sequence were shuffled, the first member of the reordered sequence would differ, but it would still be generic. In addition, at any clustering level, the pair of clus-

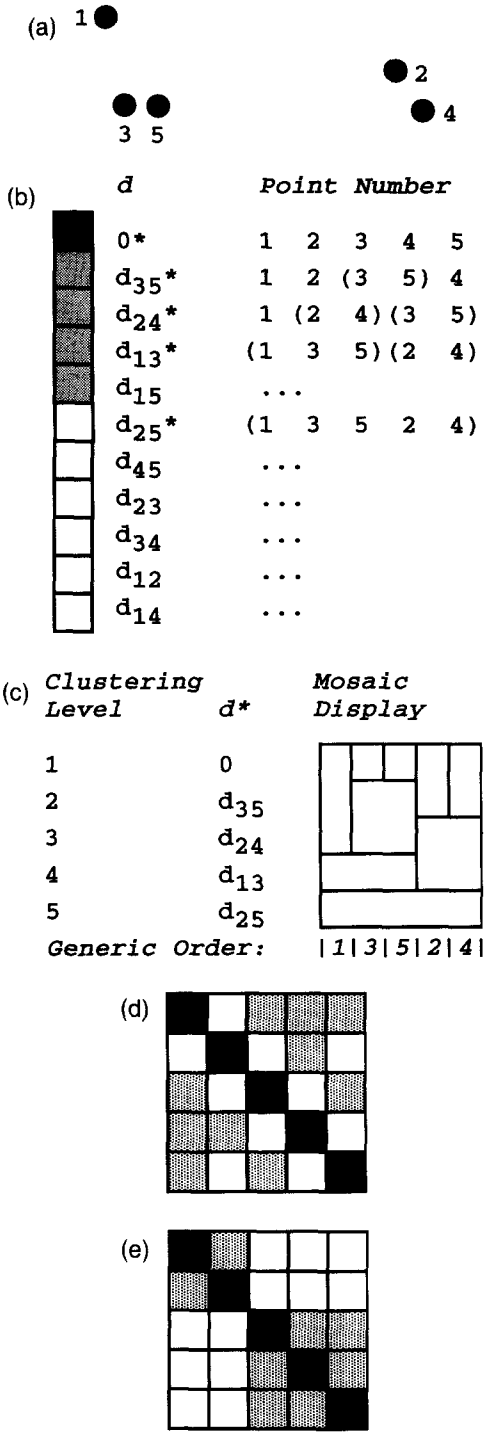

FIGURE 1. (a) Array of points in 2-space. (b) Hierarchical clustering procedure for these points; short distances are encoded by shading; asterisks denote critical threshold distances. (c) Mosaic display for this set of points. (d) Distance map display for these points in input ordering. (e) Distance map in generic ordering; note the on-diagonal low-distance blocks. See text for further details.

ters brought together could be exchanged on the list without destroying the generic property.‡ Because the clustering method used by the XCluster program proceeds by moving entire clusters without altering their internal ordering, the particular generic ordering produced by XCluster would appear to be the one which best preserves the rank ordering of the input items.

The property of the generic ordering that at any clustering level, L, items in the same cluster lie in a contiguous block leads to a display of the entire clustering hierarchy which we call a mosaic. First, at clustering level 1, where each item is in a cluster by itself, the list of data points may be visualized as a series of N bins separated by $N - 1$ dividers. This is shown in the first row of Figure 1c. When the clustering level goes to 2, a single pair of points joins into a cluster, and this corresponds to removing a single divider. At each increase of clustering level, another divider is removed. When all $N - 1$ dividers have been removed, we are at clustering level N, and all points are in a single cluster. In the mosaic display, the dividers are shown as vertical lines, and each clustering level is represented by a horizontal row of bins separated by the dividers. At each level, a horizontal line is placed across the tops of the bins that have just agglomerated. An example is shown in Figure 1c. The mosaic is essentially a form of clustering dendritogram[2]; geometrically, a dendritogram and a mosaic are duals. The advantage of the mosaic representation is that in this format it is easy to see, at any clustering level, which cluster a particular data item (vertical column) belongs to.

In the XCluster implementation, a mouse click on the mosaic display causes the value of the conformation number, clustering level, and critical threshold distance corresponding to the position of the mouse cursor to appear in a small window. This allows the display to be queried and conformations which might be of interest to be identified for later viewing.

In the generic ordering, items which are close to each other in distance tend to be close to each other in the list. Because in building the list pairs separated by short distances are brought together first, the tendency for closeness in distance to be

‡How many generic orderings are there? There appear to be 2^{N-1} of them. At each clustering level beyond the first, the pair of clusters brought together could have their locations exchanged, giving two possibilities for each of $N - 1$ levels. This process creates generic orderings exhibiting all data items in the first position.

correlated with closeness on the list is strongest for the shortest spatial distances. When the generic ordering of the data points is used to assemble the distance matrix, the smallest matrix elements therefore lie closest to the main diagonal. The clusters then appear to grow out of the main diagonal. In XCluster, a visualization of the distance matrix, in either the input or the generic ordering, is called a map display. Figure 1d is the input-ordered map and Figure 1e is the generic-ordered map for the set of points shown in Figure 1a. The symmetry of Figure 1d about the minor diagonal is an accident of the labeling of the points; such symmetry (in contrast to the symmetry about the major diagonal) is not, in general, observed. The appearance of blocks of short distance (high similarity) along the main diagonal in Figure 1e is a consequence of the generic ordering. When the generically ordered distance map for a data set has this apperance, the data items exhibit a significant degree of clustering.

In the XCluster implementation, the distance map may be viewed in either the input or the generic ordering. A color scale represents the values of the matrix elements. In either ordering, the map can be queried: A mouse click on the display gives a readout of the input and generic conformation numbers of the two items whose distance element lies beneath the mouse cursor, with the value of the pairwise distance between these two elements. This allows molecular conformations corresponding to interesting features in the distance map to be identified and later visualized. The user may also specify in an input field a particular threshold distance or clustering level. If this is done, all distance elements with values less than the input distance (or less than the corresponding critical threshold distance value if a clustering level is entered) are blackened on the display. This allows the matrix to be visualized at a finer level of distance resolution than is conveyed by the color coding. When the user specifies a distance or clustering level on the generically ordered display, a set of dividing lines appears at the bottom and side of the display, illustrating where the divisions between the clusters occur at the specified level.

THREE-DIMENSIONAL SUPERPOSITION

The rigid-body transforms that best superimpose the structures are calculated as clustering is performed. Unless the Nrms distance criterion is specified, the transforms are calculated to minimize the rms deviations of the distances between sets of corresponding atoms in a pair of conformations. When Arms distances are specified, the comparison atoms used in the calculation of the transforms are those used in the calculation of the pairwise distances. When Trms distances are specified, the set of atoms used for superposition is the union of the atom sets that define the torsions used in the distance calculation.

Suppose d_{ij} $(i < j)$ is a critical threshold distance, d^*_L, for some value of L, and let I and J be the labels of the clusters to which i and j belong. Assume for the sake of this discussion that $I < J$. If it is, we transform j to best superimpose upon i. (Otherwise, we transform i to best superimpose upon j.) If symmetry operations have been specified, the transform will include a symmetry part as well as a rigid-body translation and rotation.

Once the transform is calculated, it is applied to all elements of cluster J (or cluster I, if we had $J < I$). Thus, an entire cluster is spatially transformed whenever it is joined to another cluster. Once the clustering process is complete, each conformation has been assigned a generic transform, appropriate to all the clusterings. The first conformation in the input order remains untransformed because it is the lowest-numbered input structure. This is analogous to its always being first in the generic ordering which XCluster produces.

When a file of clustered structures is written out in MacroModel format, each structure is first transformed by its generic transform. The coordinates of a given conformation on output are therefore the same regardless of the clustering level specified for the output file. The color of the structures, however, varies with clustering level; at any level, structures in the same cluster are given the same color.

CLUSTERING STATISTICS

Because there are many clusterings to choose from, it is natural to ask, "Which of them—if any—are interesting or significant?" There are two ways to approach this question. The first is to use past experience. If experience dictates that conformations within, say, some particular rms atomic displacement value of each other tend to have properties that are similar in some important way, then there is every reason to examine the clusters that appear at this (or the next lower) value of the critical threshold distance.

On the other hand, it is natural to ask whether the data naturally clump into especially good clus-

ters at some point. Jain and Dubes[2] discuss several statistics that have been proposed for this purpose, but it is clear that no single figure of merit will serve all purposes. We have found that a statistic we have devised, which we call the separation ratio, R_L, of the clustering, correctly identifies situations in which all clusters are clearly defined and well separated. Another statistic, the effective number of clusters, k^*, though not a figure of merit, describes essentially how heterogeneous in size the clusters are at a given level, and tracking this variable as a function of L often gives useful insight into the details of the aggregation process. We have experimented with two additional figures of merit: one, of our devising, which we call the reordering entropy, S_{re}, and one based on analysis of variance (ANOVA). We describe our experiments with these measures as well, although we have not found them useful in the problems we have tried to address thus far.

All of the measures we discuss, as well as straightforward statistics (such as the number of clusters and the average size of a cluster) can be plotted against either clustering level, L, or critical threshold distance, d^*_L, using the facilities of the XCluster program.

Separation Ratio

Recall that there are N values of critical threshold distance, d^*_L. For any value of L between 2 and $N - 1$, we define R_L, the separation ratio of clustering L, as d^*_L/d^*_{L+1}. Because d^*_{L+1} is the next value of d_{ij} on the sorted distance list after d^*_L that creates a new clustering, d^*_{L+1} must be the shortest distance between any two data points which, at level L, are not in the same cluster. Similarly, d^*_L is the greatest nearest-neighbor distance between any two items that, at level L, are in the same cluster. If R_L is high, it means that the minimum distance between two items not in the same cluster is high compared to the critical threshold distance defining the clustering; in this situation, all clusters are well separated with respect to the greatest nearest-neighbor distance within any cluster.

R_L is a pessimistic measure; for it to be high, all clusters must be well separated. It is possible for some clusters to be well separated but for R_L to be close to unity because other clusters are close together. As an aid to identifying situations of this nature, we also define the separation ratio of an individual cluster, which is the ratio of the shortest distance between any item in that cluster and an item in any other cluster to the critical threshold distance. Individual clusters may have high separation ratios even when R_L is low.

Although R_L is a pessimistic measure at high clustering levels, misleadingly high values of this statistic may occur at low levels. We sometimes find pairs of structures that are nearly superimposable in our data sets. For example, we might be looking for clusters of ring conformations in the output of a conformational search of a system which also has sidechains. Two structures which vary significantly only in the sidechain degrees of freedom will exhibit ring conformations that aggregate at very low values of d^*_L, giving rise to a large value of R_L at some low L value. Another thing that can cause this is imperfect convergence of a minimizer; this can lead to structure pairs which lie within the same energetic basin and are just different enough to be regarded as distinct by the generating program. In our experience, R_L values greater than about two occurring at high values of L (between, say, $N - 5$ and $N - 1$), indicate interesting clusterings.

For the data displayed in Figure 1a, a maximum in the separation ratio would appear at clustering level 4, where two clusters are present. This maximum would have a value of about 3.0, the ratio of d_{25} to d_{13}. This tells us that the best clustering is the one in which two clusters are present—one containing input items 1, 3, and 5, and the other containing items 2 and 4. At level 3, where three clusters are present, the separation ratio would be about 1.6, the ratio of d_{13} to d_{24}. This clustering is also reasonable to the eye; it is the one in which 3 and 5 form one cluster, 2 and 4 form another cluster, and 1 is in a third cluster by itself. But the separation ratio statistic provides a quantitative criterion: one that tells us that the clustering at level 4 is in some sense better than that at level 3.

Effective Number of Clusters

Suppose that at some clustering level we observe k clusters with populations $n_1, n_2, \ldots, n_k$. We use the symbol x_i for n_i/N, the fraction of items in cluster i. We then define the clustering entropy by

$$S_{cl} = -\sum_{i=1}^{k} x_i \ln x_i$$

and the effective number of clusters by $k^* = \exp(S_{cl})$. The properties of this measure have been discussed elsewhere,[24,25] but to summarize in the

present context, when the x_i are all equal, k^* will be equal to k, the actual number of clusters present. If k' of the clusters are large and of approximately equal magnitude and the remainder are much smaller, then k^* will be approximately equal to k'.

At clustering levels 1 and N we have, respectively, N clusters of one item each and one cluster of N items. At these levels, k^* is equal to the actual number of clusters present. Between these limits, k^* generally lies below k because it is rare for the clusters to all be of equal size. Thus a plot of k versus clustering level is generally concave when viewed from above. [In any event, k^* is guaranteed to decrease monotonically as cluster level increases. Suppose two clusters, I and J, with fractional populations x_I and x_J, agglomerate with an increase in clustering level; recall that this is the only type of event that can occur with such an increase. Prior to the agglomeration, the contribution of I and J to S_{cl} is $S_{before} = -x_I \ln x_I - x_J \ln x_J$; afterward, the contribution of the combined cluster is $S_{after} = -(x_I + x_J) \ln(x_I + x_J) = -x_I \ln(x_I - x_J) - x_J \ln(x_I + x_J)$. x_I and x_J are positive, by definition; therefore, $\ln(x_I + x_J) > \ln(x_I), \ln(x_J)$, because the logarithm increases monotonically with its argument. It follows that $S_{after} < S_{before}$, and, therefore, that k^* decreases as agglomeration proceeds.]

Sometimes a few large clusters form early, and then grow either slowly by accretion of individual items as clustering index increases or else remain approximately constant in size as outlying items agglomerate to form new clusters. In these situations a plot of k^* versus clustering level tends to exhibit a broad, nearly flat region. In contrast, when, at a given clustering level, k^* becomes smaller by about one, it means that two clusters of nearly equal size have just agglomerated. The agglomeration process can be observed in detail in the mosaic display; the plot of k^* versus L may be viewed as a stage-by-stage summary of this behavior.

Reordering Entropy

At a given clustering level, in how many ways can the generically ordered conformation list be reordered without destroying the contiguity of elements belonging to the same cluster? We do not require that the ordering remain generic (i.e., that it preserve contiguity at all clustering levels). We require merely that it preserve contiguity at the given level. At $L = 1$, where each item is in a cluster by itself, there are $N!$ ways of reordering the list because the first item can be in any of N positions, the second can be in any of $(N - 1)$ positions, etc. At the Nth level, where all items are present in a single cluster, the ordering is again irrelevant, and we again have $N!$ possibilities. But between these bounding levels some reorderings break up clusters and are forbidden. If some clustering exhibits k clusters having $n_1, n_2, n_3, \ldots, n_k$ members, then within each cluster the members can be freely reordered, giving $\prod_{i=1}^{k} n_i!$ possibilities. In addition, the k clusters themselves can be reordered on the list, giving $k!$ additional possibilities. The total number of allowed reorderings is then given by

$$W = k! \prod_{i=1}^{k} n_i!$$

In analogy to Boltzmann's $S = k_B \ln W$, we define the reordering entropy as $S_{re} = \ln W$.

S_{re} goes through a minimum as clustering level or threshold increases. It can be shown that there is a tendency for S_{re} to be minimal when one has a large number of small clusters, and in fact we have observed this in chemical examples. The minimum value of S_{re} defines a unique clustering level, but, contrary to our *a priori* intuition, in our work thus far we have not found this level to correspond to clusterings that are intuitively good.

ANOVA

We have also experimented with a figure of merit based on the statistical method of analysis of variance.[26] In this method the variance of some parameter within subpopulations is compared to the variance of the same parameter within either the entire data set or, more usually, with the variance of the set of subpopulation means for this parameter. If the subpopulations exhibit significantly smaller variances, internally, than the comparison group, then the parameter differs significantly among the subgroups. The Fisher F test is usually used as the criterion of significance.

In most applications of ANOVA, the subgroups are distinguished by having been subjected to different treatments (e.g., the application of fertilizer or not to two otherwise comparable fields of a crop), but in our application the subgroups are defined by the clusters at a given level, and the variance we calculate is the square of the radius of gyration of a conformational cluster, considered as a point clouds in conformation space. The calculation is greatly simplified by the use of Lagrange's

theorem,[27] which states that for an ensemble of N points, the population variance is equal to $(1/N^2)$ $\Sigma_{i<j}\, d_{ij}^2$. This allows us to calculate the variances from the distance matrix without calculating the means (centroids) of the clusters in conformation space.

For each clustering we calculate F as the ratio of the sample variance between clusters to that within clusters, and we calculate the one-way probability, P, of the F value given the number of degrees of freedom of the numerator and denominator, which are $k - 1$ and $N - k - 1$, respectively. P represents the probability that a value of F or greater would have been obtained by means of random sampling from two distributions with the given degrees of freedom, subject to the assumption that the distributions are normal (an assumption which may not hold for our examples). Thus smaller values of P correspond to greater significance. The usual assertion is that F was significant at the P level, with P usually expressed in percent.

We have found that the F values we obtain in all of our examples are highly significant but that the examples that exhibit stronger clustering exhibit more significant values of P. For example, the cycloheptadecane example discussed in Results gives $-\log_{10} P$ values of about 10, whereas the roseotoxin B example gives values of about 300. However, when the best clustering is intuitively clear, as in the latter example, where it occurs at the two-cluster level, the ANOVA result fails to reflect intuition. Instead, for all data sets we have examined, the ANOVA $\log_{10} P$ parallels closely the L dependence of S_{re}, going through a minimum at a low clustering level. We find this observation intriguing but do not have an explanation.

Results

CONFORMATIONAL SEARCH

Roseotoxin B

Roseotoxin B, a toxic metabolite of peanut fungal mold, is essentially a cyclic hexapeptide with some unusual residue types. Its structure is shown in Figure 2. An MM2 study of the conformational preferences of this molecule concluded[28] that the cross-ring hydrogen bonds which are exhibited in a pair of beta turns are a consequence of the inherent rigidity of the 19-membered ring and are not in themselves responsible for the backbone structure observed in X-ray studies.[29]

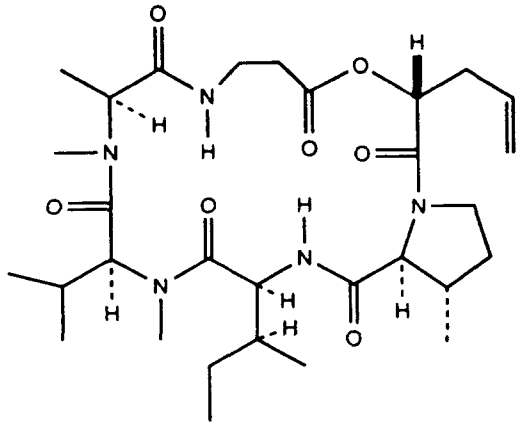

FIGURE 2. Structure of roseotoxin B.

Because cyclic peptides are important models for understanding secondary structure in proteins and because roseotoxin B has been well characterized by three separate X-ray structure determinations,[29] we have performed an extensive conformational search of all torsions in the molecule in order to assess the effectiveness of methodology developed in this laboratory for modeling bio-organic molecules in solution. The potential function chosen for this study was the AMBER* force field, which includes recently determined parameters for peptide backbone substructures.[30] The search was performed in the context of the GB/SA continuum model for chloroform solvation[31]; thus we expect that the ensemble of conformations obtained reflects that exhibited in the low-polarity organic solvents from which the crystals analyzed in the X-ray work were precipitated.

A systematic conformational search of a molecule this large is not practical, and therefore a stochastic search procedure was used. This procedure has been shown to be efficient for cyclic systems.[32] Even so, a large number of structures needed to be examined, and the search consisted of 50,000 structure-generation and minimization cycles performed on a network of eight UNIX workstations. The initial search, for which we kept structures within 50 kJ/mol of the lowest energetic minimum found, located 323 unique conformations. We then minimized these structures exhaustively, discarding those which were more than 3.0 kcal/mol above the lowest minimum found. A total of 192 structures remained, and these are all likely to be exhibited significantly in solution.

The chief feature which we wished to examine was the conformation of the 19-membered ring

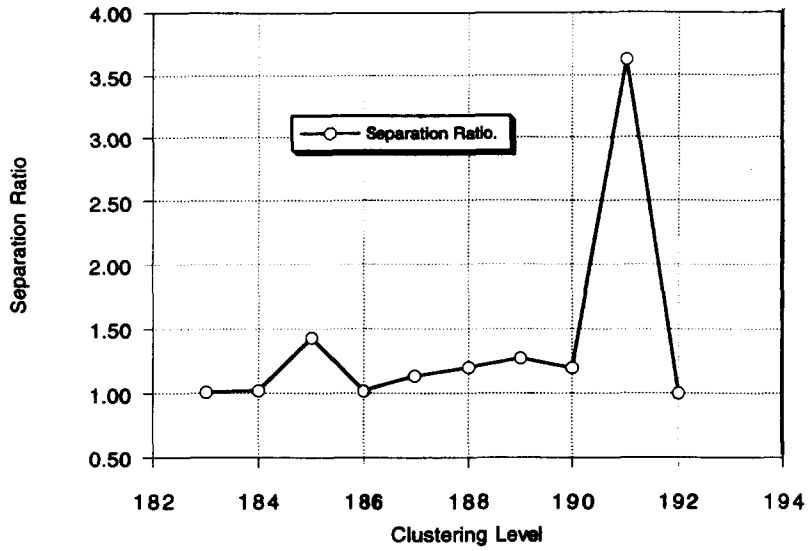

FIGURE 3. Separation ratio versus clustering level for the top 10 clustering levels, roseotoxin B conformational search, Trms on all ring torsions.

formed by the peptide backbone. We therefore applied the Trms mode of XCluster to the ring torsions of the $N = 192$ structures. Figure 3 shows a plot of separation ratio versus clustering level over the last 10 levels for this run. The peak at level 191 indicates that the clustering at this level is likely to be significant. This conclusion is reinforced by examination of the corresponding distance matrix in the generic ordering (Fig. 4a). Here two large on-diagonal blocks can be identified, each containing a number of subblocks. Recall that at clustering level N all structures are united into a single cluster. At level 191 $= N - 1$ two clusters are present. By writing out the two clusters in the optimal superposition and visualizing them using MacroModel, we could see (Fig. 5, bottom center) that the two clusters differ in the value of the torsion about the C_α—C_β bond of the beta-alanine residue: This torsion exhibits a gauche-plus value in one cluster and a gauche-minus value in the other.

In a second run we applied the Trms method to this torsion alone. The separation-ratio plot exhibited a much sharper peak than the one we observed in the run in which the analysis was applied to all ring torsions. As in the previous run, the peak occurred at clustering level 191, but here the peak separation ratio was 34, instead of the value of 3.6 observed in the all-ring-torsion run. The membership of the two clusters obtained at this level is identical in both analyses. It is clear from this that the C_α—C_β bond of the beta-alanine residue is in some sense a fiducial torsion, defining two major ring conformations for this molecule; as this torsion goes, so goes the ring. Further confirmation is to be found from the results of yet a third run, an Arms run on the ring atoms. This run also produced a peak in the separation ratio (value 2.8) at level 191, and again the members of the two clusters were identical to those found at this level in the two Trms runs. Thus, when clustering is especially strong and clear cut, as it is here, the same results seem to be obtained whether the proximity matrix is built using distances defined in torsional space or in Cartesian space.

Further insight into the ensemble of roseotoxin B conformations can be obtained by comparing the appearance of the distance map in the initial input order (Fig. 4b) with its appearance in the generic order (Fig. 4a). The conformations were supplied to the XCluster program in order of increasing energy; thus Figure 4b reflects this ordering. At a first glance, this figure appears chaotic; the strong block-diagonal structure exhibited in Figure 4a is absent. This implies that structures which are close to each other energetically are not generally similar to each other in ring geometry. Closer examination of Figure 4b shows, however, that there are indeed small on-diagonal blocks containing two to four structures each. On examination of the structures forming these small groups, we found them to differ significantly in the conformations of the sidechains only. From this it can be inferred that some sidechain conformational preferences in solution are only weakly coupled to the ring conformational

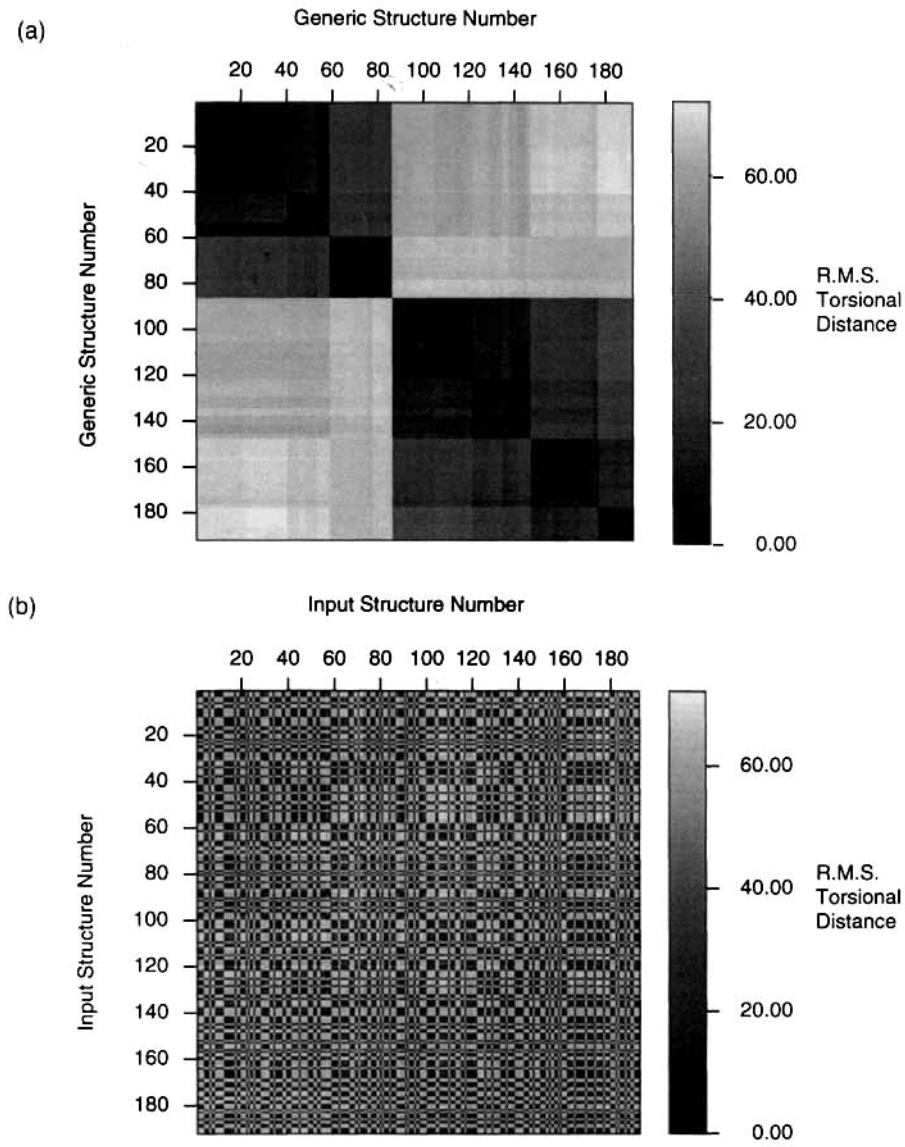

FIGURE 4. Distance maps, roseotoxin B conformational search, Trms on all ring torsions. (a) Generic structure ordering. (b) Input structure ordering.

preferences and that the molecule exhibits small groups of structures with similar energies and similar ring geometries but varying sidechain geometries.

Close examination of Figure 4a reveals subblocks within the two main on-diagonal blocks. Some of these subblocks exhibit further structure. This reflects a hierarchy of subclasses of ring geometry within each of the two large conformational families. Critical threshold distances at clustering levels corresponding to the formation of the subblocks show that the subblocks differ from each other far less than do the two large families. A description of the full hierarchy of roseotoxin B ring conformations is reserved for a future publication.

Cycloheptadecane

Conformational search of the 17-membered alkyl ring (C_{17}) has been used as a test of search methodology. A previously reported exhaustive study using the MM2 force field located 262 conformations within 3.0 kcal/mol of the global minimum. We have performed a similar search with the MM3 force field using the same search strategy used for roseotoxin B. With this force field only 132 confor-

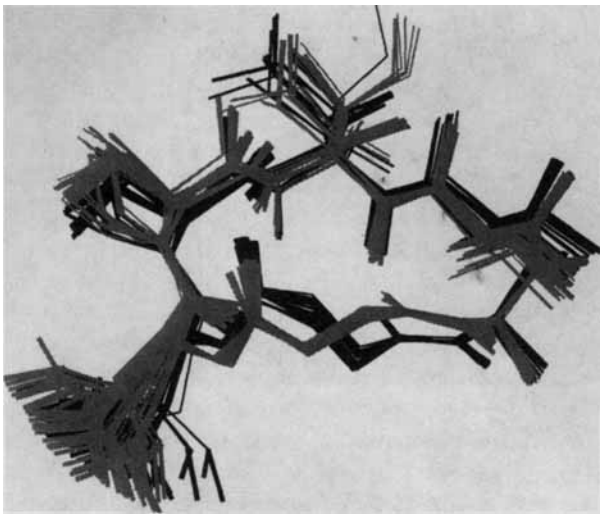

FIGURE 5. Superimposed roseotoxin B conformations, gray-scale-coded according to cluster membership; Trms on all ring torsions, clustering level 191 (two clusters present).

mations are found within 3.0 kcal/mol of the minimum energy conformation located. This is consistent with the work of Saunders,[19] who, in studies of medium-size rings, reported fewer conformations when MM3, rather than MM2, was used.

We were interested in whether these 132 conformations exhibited any structural features which would allow for classification of the ensemble members into clusters. An XCluster analysis was performed on all ring torsions (Trms), with full account taken of the ring symmetry when determining the distances between conformations, as described by Saunders.[19] For each pair, this leads to $4 \times 17 = 68$ comparisons; the shortest distance among the 68 comparisons is taken as the corresponding distance matrix element.

We observed little or no clustering of these 132 conformations. None of our clustering statistics indicated significant clustering at any level, and the generically ordered distance map (Fig. 6) exhibits no significant block-diagonal structure. Our superposition algorithm performs the appropriate symmetry operations when writing out clusters of conformations, and the superimposed conformations reveal no clustering visible to the eye. This result confirms the traditional view that a ring as large as C_{17} is floppy rather than strained. The strong clustering observed for roseotoxin B is presumably a consequence of ring strain. We also draw the important conclusion that the analysis and visualization methods provided by XCluster can reveal the absence of clustering, as well as its presence.

MOLECULAR DYNAMICS

Pentane

We have found cluster analysis to be a useful tool for gaining insight into the progress and possible convergence of molecular dynamics runs. To illustrate this, we have chosen a simple system—

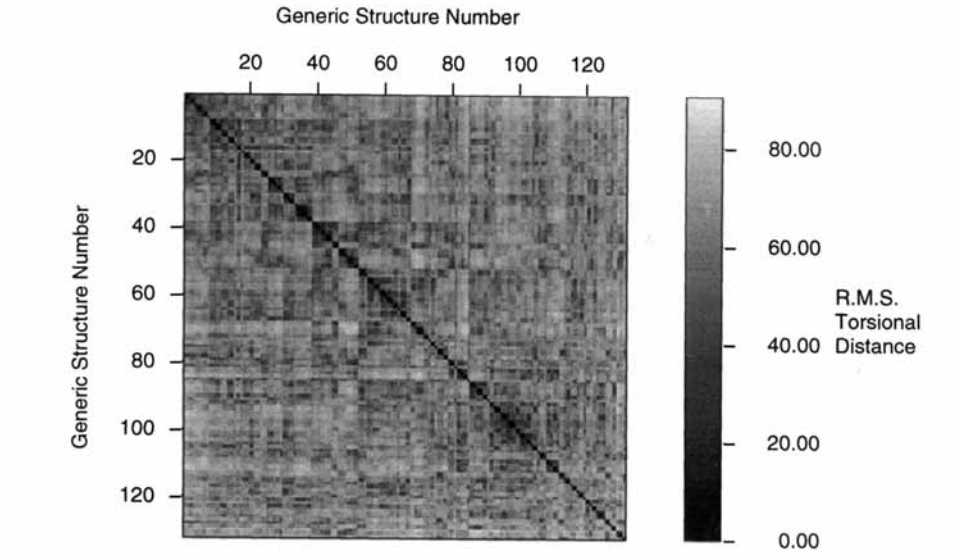

FIGURE 6. Distance map, cycloheptadecane conformational search, Trms on all ring torsions, generic structure ordering, full symmetry.

pentane in the united atom approximation—and performed two simulations, using the stochastic dynamics[33] method and the AMBER* force field,[30] as implemented in MacroModel.[1]

The conformational space of the pentane system is described conveniently by considering the 2—3 and 3—4 bonds to each have three minima, termed trans (t, $\Theta = 180°$), gauche-plus ($+$, $\Theta = +60°$), and gauche-minus ($-$, $\Theta = -60°$). The combination of these gives nine minima. Cluster analyses were performed on these two torsion angles starting with a time-ordered sequence of structures obtained from the runs, with and without allowance made for molecular symmetry. With number-order-reflective and enantiomeric symmetry, the nine minima coalesce into four groups: ($++$, $--$) ($-+$, $+-$), (tt), ($t+$, $t-$, $+t$, $-t$). A fully convergent dynamics simulation should explore all four basins.

The first simulation was performed at 300 K with a timestep of 1.5 fs for 200 ps. During the simulation 200 structures were saved, giving a rate of one structure per picosecond. A second simulation was performed under the same conditions but for 20 ns, again saving 200 structures, this time at a rate of one structure per 100 ps. For each dynamics simulation we performed two XCluster analyses: one incorporating full symmetry and one without specifying symmetry. We will see that these two cluster analyses provide different sorts of information regarding the results for the simulations.

Figure 7a shows the distance matrix in the generic ordering for the 200-ps run, incorporating the full symmetry of the molecule. Three on-diagonal blocks are apparent, which seems to indicate that three clusters have been found. This is confirmed by the plot of separation ratio against clustering level, which displays a maximum at level 198, where three clusters appear. Thus only three of the four conformational groups expected were sampled during the run. Further analysis shows that group ($++$, $--$) is absent. Interestingly, this structure does not correspond to the highest-energy minimum of the four groups, and thus our sampling of the space is highly non-Boltzmann. We failed to sample the ($++$, $--$) basin only because we did not allow the simulation to proceed long enough.

Figure 7b shows the distance map for the same dynamics run in the input ordering, without specifying a symmetry. The input ordering, of course, reflects the time sequence of structure collection. This figure exhibits a number of large on-diagonal blocks. This means that for long periods of time (2 to 60 ps), the simulation explores sets of structures that are geometrically similar; in fact, each such block corresponds to the exploration of a basin surrounding one of the minima in the conformation space. This run was initiated in the (tt) conformation; the structure flipped into and explored the regions about other minima in the course of the simulation.

Note the presence of a number of off-diagonal blocks, which in every case connect two on-diagonal blocks—one in the vertical and one in the horizontal direction. For example, the first and the fifth on-diagonal blocks are connected by an off-diagonal block. Because the blocks represent structure pairs which are geometrically similar, this tells us that the conformations in the first and fifth on-diagonal blocks are similar; in fact, investigation shows that structural pairs spanning the two on-diagonal blocks are as similar as pairs within either one. We conclude from this that after exploring three other local minima in the conformational space, the molecule returns in block 5 to the (tt) basin of attraction whence it started.

Consider now the second on-diagonal block in Figure 7b. There are no off-diagonal low-distance blocks associated with it, as one can see by examining the display in a vertical and a horizontal direction starting within the block. This means that structures similar to those within the block are never sampled again during the trajectory, once this block is exited. In other words, the basin of attraction within which these structures reside is never revisited in the course of the simulation. The revisiting of every basin of attraction, preferably many times, is a necessary, but not sufficient, condition for convergence of a dynamics run. Thus, examination of this distance map demonstrates that this run has not converged.

We emphasize that this conclusion could have been drawn from observation of the input-ordered conformational distance map even without the prior insight into the specifics of pentane conformation embodied by the statement that the system should exhibit four symmetry-related conformational minima. Please note, too, that although the generically ordered distance map and separation-ratio graph are best examined with full symmetry specified, questions of convergence are best examined using the input-ordered map without symmetry specification. For this simple example, similar conclusions could have been drawn by monitoring the dihedral angles during the simu-

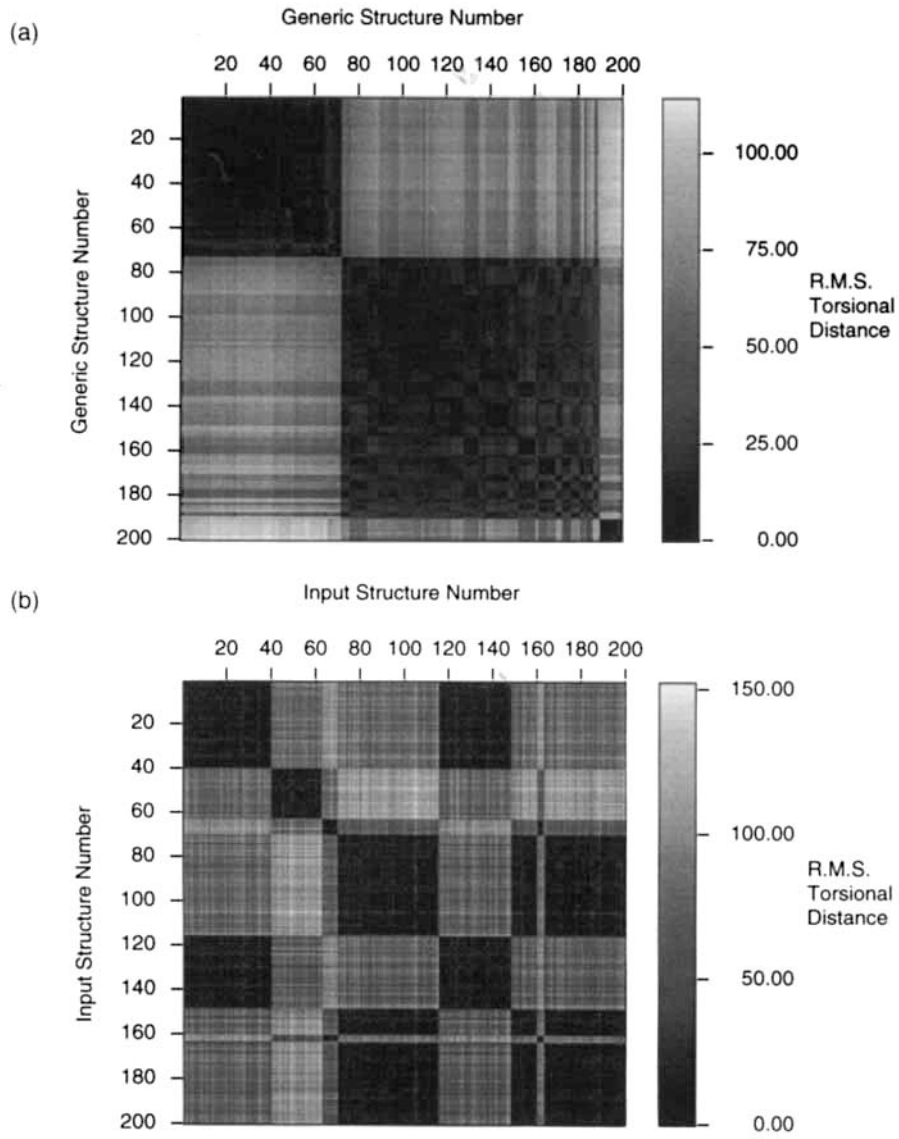

FIGURE 7. Distance maps, united-atom pentane, 200-ps stochastic dynamics, 200 structures. Trms on both dihedral angles. (a) Generic structure ordering, full symmetry. (b) Input structure ordering, no symmetry.

lation; however, for more complex molecular systems the input-ordered distance map provides a concise way to represent the sequence of gross conformational changes that occur during a dynamics simulation.

Figure 8b exhibits the input-ordered distance map from a Trms analysis of the 20-ns simulation without symmetry specification. This map exhibits on-diagonal blocks that are much smaller than those of the corresponding 200-ps map. The reason, of course, is that the structures are sampled 100 times less frequently here than in the former run. Nevertheless, every on-diagonal block appears to have multiple corresponding off-diagonal blocks, which implies that every minimum has been sampled many times. Although this does not prove that the full conformational space has been explored, it at least indicates that it may have been.

Figure 8a exhibits the generically ordered distance map from a Trms analysis of the 20-ns run, specifying full symmetry. Here, four distinct on-diagonal blocks appear. The fourth block has a small population and is barely visible at the lower right of the diagram. This block does not correspond to the $(++, --)$ basin which was absent in the results shown in Figure 7, but rather to the $(+-, -+)$ conformation, whose energetic minimum is the highest of the four conformations:

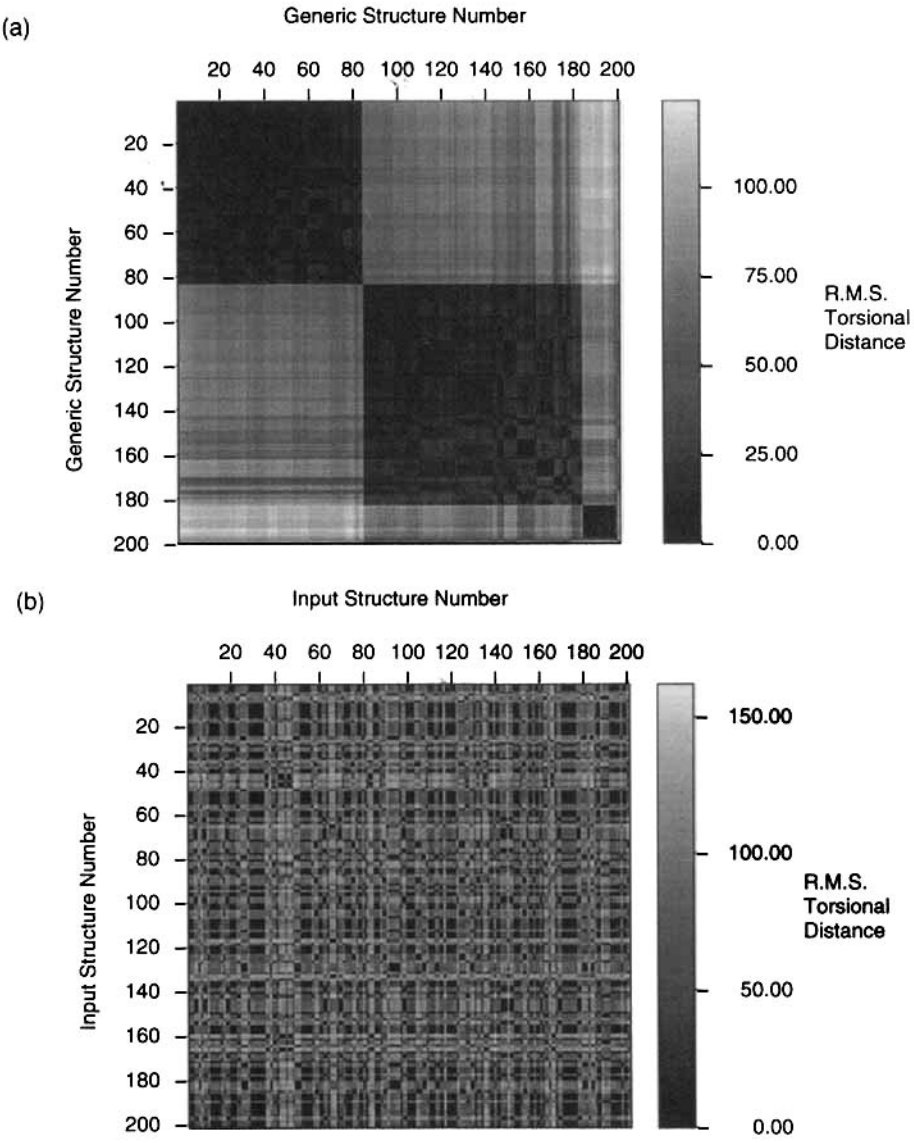

FIGURE 8. Distance maps, united-atom pentane, 20-ns stochastic dynamics, 200 structures. Trms on both dihedral angles. (a) Generic structure ordering, full symmetry. (b) Input structure ordering, no symmetry.

about 11 kJ above that of the global (*tt*) minimum. The appearance of four blocks indicates the appearance of four conformational clusters; this is borne out by a peak in the separation ratio at the four-cluster level. This is in accord with our *a priori* understanding of the pentane conformational system. Furthermore, a calculation taking into account only the energy of each minimum and the configurational entropy of each, as described by the symmetry-related groupings of the nine local minima, reveals that the four symmetry-related minima are sampled with Boltzmann frequency within experimental error.§ This gives us further confidence in the convergence, or at least the self-consistency, of this run, and implies that the differences between

§The energetic minimum of each conformational class was determined by means of energy minimization in MacroModel. A partition function was calculated in the form $Z = \sum_{i=1}^{4} g_i e^{-E_i/RT}$, where the g_i are the symmetry-based degeneracies and the E_i are the molecular mechanics energies. The expected frequency of appearance of each conformational class was calculated as $g_i e^{-E_i/RT}/Z$. This was observed to be equal, in all cases, to the actual frequency of appearance of the class, within experimental error.

the rotational/vibrational entropies of the various structural minima are small.

Discussion

Cluster analysis provides a means of gaining insight into the nature of large ensembles of molecular conformations. When the ensemble is the output of a conformational search, it provides us with an answer to the question, "Do the conformations fall into several geometrically related classes, or do they in some sense sample a unimodal distribution in conformation space?" In certain instances the answer may have chemical implications. For example, it was found recently[34] that a host molecule which strongly discriminates between enantiomers of α-methylbenzylamine in binding studies exhibits a single, undifferentiated cluster when the results of a conformational search are examined with XCluster. A closely related compound which exhibits much weaker discrimination exhibits several clusters. Apparently, the stiffness which forces the strongly discriminating host to adopt only a single solution conformation is largely responsible for its binding behavior, and this effect is seen only in an analysis of the entire ensemble, not in the analysis of any single conformation.

A comparison between distance maps when exhibited in generic and input ordering can lead to additional insights. If the maps are dissimilar, it means that whatever parameter orders the input structures is not highly correlated with geometric similarity. For example, in our roseotoxin B study we found that the energetic ordering of the structures was not highly correlated with geometric similarity of the ring backbone.

When cluster analysis is performed on structures sampled from a dynamics simulation, the clusters tend to correspond to basins of attraction explored during the run. In the generic ordering, the number of blocks along the main diagonal may be counted, and in favorable cases a corresponding peak will appear in the plot of separation ratio versus clustering level. The number of basins visited during the simulation may thus be determined. Insight into the course of the run may be obtained by examining the distance map in the input ordering; here, an on-diagonal block corresponds to the exploration of a single basin of attraction during a contiguous segment of time. Off-diagonal blocks, when they occur, always connect on-diagonal blocks and reveal multiple visits to the same regions of conformational space. Examination of the input-ordered distance map can reveal whether every minimum observed has been visited more than once. Because this is a necessary condition for convergence, the distance matrix provides a means of testing for the feasibility of convergence.

The separation ratio provides an indication of the best clustering. This statistic is not perfect: When it is high, a good clustering is certainly present, but when it is low, it does not rule out the existence of at least some good clusters. More work is needed to determine what the expectation value for the maximum value of this statistic might be for relevant null hypotheses, such as a random spherical Gaussian or uniform point-cloud distribution in high-dimensional spaces. Better statistics than this one may be available. In addition, care must be taken when interpreting this statistic at low clustering levels, particularly when only a substructure (e.g., the ring atoms) of the molecule of interest is subjected to cluster analysis. In this situation, several essentially identical substructure conformations may occur for molecules which differ outside the substructure (e.g., in the sidechains). This leads to spurious peaks of essentially infinite separation ratio at low clustering levels.

The single-link clustering method has a tendency to chain together early in the agglomerative process clusters that intuition might regard as distinct. We have been able to concoct two-dimensional examples in which the distance matrix exhibits a largely block-diagonal form but in which the clusters found by the program do not correspond to intuition and the separation ratio exhibits no strong clustering. This is due to the characteristics of the single-link method. The absence of such problems in the chemical systems we have studied may or may not be a typical of ensembles of chemical conformations in general; application of this method to more examples is the only way to resolve this question.

Nevertheless, we do see the need to experiment with other hierarchical clustering methods, in particular Ward's method.[16] Another alternative, the complete-link method,[2] is known to replace the key deficiency of the single-link method (its tendency to chain clusters together, and thus hide real distinctions) with a complementary deficiency of its own: It often fails to incorporate into a growing cluster data items which intuition dictates should be included; nevertheless, it may prove useful. We expect to supply additional methods as options in future versions of the program.

Acknowledgments

We would like to thank Professor W. Clark Still of Columbia University for helpful discussions and for support of this work. Hany Farid of the University of Pennsylvania, Dr. Barr Bauer of Arris Pharmaceutical Company, and Drs. Mathis Thoma and Greg Paris at the CIBA-Geigy Corporation supplied useful suggestions early in the development of the program, and we thank them as well. Finally, we would like to thank the National Science Foundation (CHE 92-08253, to W. C. Still) and the National Institutes of Health, Division of Research Resources (P41-RR06892, to R. Friesner) for partial support.

References

1. F. Mohamadi, N. G. J. Richards, W. C. Guida, R. Liskamp, C. Caufield, G. Chang, T. Hendrickson, and W. C. Still, *J. Comp. Chem.*, **11,** 440 (1990).
2. A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, NJ, 1988.
3. R. R. Sokol, In *Classification and Clustering*, J. Van Ryzin, Ed., Academic Press, New York, 1977, p. 1.
4. J. Zupan, *Algorithms for Chemists*, John Wiley & Sons, New York, 1989.
5. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipies in FORTRAN*, 2nd ed., Cambridge University Press, Cambridge, UK, 1992, p. 320.
6. B. R. Kowalski and S. Wold, In *Handbook of Statistics*, Vol. 2, P. R. Krishnaiah and L. N. Kanal, Eds., North-Holland, Amsterdam, 1982, p. 673.
7. T. Okada and T. Wipke, *Tetrahedron Comp. Meth.*, **2,** 249 (1989).
8. P. Willett, *Similarity and Clustering in Chemical Information Systems*, Research Studies Press, Letchworth, UK, 1987.
9. H. I. Gordon and R. L. Somorjai, *Proteins: Struct., Func., Genetics*, **14,** 249 (1992).
10. M. E. Karpen, D. J. Tobias, and C. L. Brooks III, *Biochem.*, **32,** 412 (1993).
11. D. Gautheret, F. Major, and R. Cedergren, *J. Mol. Biol.*, **229,** 1049 (1993).
12. L. Holm and C. Sander, *J. Mol. Biol.*, **233,** 123 (1993).
13. P. Murray-Rust and J. Raftery, *J. Mol. Graphics*, **3,** 50 (1985).
14. T. D. J. Perkins and D. J. Barlow, *J. Mol. Graphics*, **8,** 156 (1990).
15. T. D. Perkins and P. M. Dean, *J. Comp.-Aid. Mol. Design*, **7,** 155 (1993).
16. J. H. Ward, *J. Am. Stat. Assoc.*, **58,** 236 (1958).
17. F. H. Allen, M. J. Doyle, and R. Taylor, *Acta Cryst.*, **B47,** 50 (1991).
18. G. Ravishankar and D. L. Beveridge, personal communication.
19. M. Saunders, *J. Comp. Chem.*, **12,** 645 (1991).
20. W. Kabsch, *Acta Cryst.*, **A32,** 922 (1976).
21. W. Kabsch, *Acta Cryst.*, **A34,** 827 (1978).
22. G. Young and A. S. Householder, *Psychometrika*, **3,** 19 (1938).
23. R. M. Fine, H. Wang, P. S. Shenkin, D. L. Yarmush, and C. Levinthal, *Proteins: Struct., Func., Genetics*, **1,** 342 (1986).
24. R. M. Swanson, *J. Chem. Educ.*, **67,** 206 (1990).
25. P. S. Shenkin, B. Erman, and L. D. Mastrandrea, *Proteins: Struct., Func., Genetics*, **11,** 297 (1991).
26. L. L. Havilcek and R. D. Crain, *Practical Statistics for the Physical Sciences*, American Chemical Society, Washington, DC, 1988.
27. P. J. Flory, *Statistical Mechanics of Chain Molecules*, John Wiley & Sons, New York, 1969, p. 5.
28. J. P. Snyder, *J. Am. Chem. Soc.*, **106,** 2393 (1984).
29. J. P. Springer, R. J. Cole, J. W. Dorner, R. H. Cox, J. L. Richard, C. L. Barnes, and D. van der Helm, *J. Am. Chem. Soc.*, **106,** 2388 (1984).
30. D. Q. McDonald and W. C. Still, *Tetrahedron Lett.*, **33,** 7743 (1992).
31. W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, *J. Am. Chem. Soc.*, **112,** 6127 (1990).
32. M. Saunders, K. N. Houk, Y-D. Wu, W. C. Still, M. Lipton, G. Chang, and W. Guida, *J. Am. Chem. Soc.*, **112,** 1419 (1990).
33. W. F. van Gunsteren and H. J. C. Berendsen, *Mol. Simulation*, **1,** 173 (1988).
34. Shawn Erickson, personal communication.