Metric and Multidimensional Scaling: Efficient Tools for Clustering Molecular Conformations

Miklos Feher* and Jonathan M. Schmidt

Nanodesign Inc., Suite 300, Research Park Centre, 150 Research Lane, Guelph, Ontario N1G 4T2, Canada

Received August 10, 2000

The application of metric and multidimensional scaling to conformer ensembles was demonstrated in this work. An automated process was devised to cluster and assign group memberships and cluster representatives. The method allows rapid clustering, leading to intuitive results that can be visually inspected. Multidimensional scaling was found to be superior to metric scaling for clustering conformers. The performance of different hierarchical clustering algorithms was compared using multidimensional plots, and the group average method was found to perform best.

INTRODUCTION

There are many areas of computational chemistry where considering a small number of groups, families, or clusters of conformations is more favorable than dealing with the original multitude of conformers. Finding clusters is also useful for identifying common features among conformers, as well as reducing the total number considered in modeling or 3D-QSAR applications. It is also often helpful to visualize the complex relationship among these conformer groups.

The word clustering generally implies the identification of groups such that the similarities within the groups are significantly greater than those between the groups.[1,2] A number of well-documented methods exist for clustering conformations.[3−6] Nearly all of these are based on some kind of distance measure between pairs of conformers, using either the distance matrix or a selection of dihedral angles. The most popular clustering methods for conformational selection are based on the hierarchical and Jarvis−Patrick schemes. For most clustering methodologies, determining the optimum number of clusters is a difficult issue.[2,4]

Metric and multidimensional scaling are well-characterized methods in statistics for identifying groups or clusters in data.[1,7,8] Although they appear to be used widely in other areas of science, they have not yet found many applications in molecular modeling. In the latter area, multidimensional scaling was first suggested as a tool to visualize conformer sets[9] and later as a possible tool for conformer selection.[10] It has also been applied as a visualization tool to display conformer distributions in proteins based on their α-carbon[11] or inter-ring[12] distances. However, the use of multidimensional scaling as a general cluster analysis tool for conformations has not been shown, and its performance on a range of examples has not been demonstrated. In addition, its application to identifying cluster membership and cluster representatives has not been studied. The aim of the present work was to develop a method for selecting conformers from groups, to identify figures of merit for this method, and to establish its usefulness in comparison with other clustering tools.

METRIC SCALING AND MULTIDIMENSIONAL SCALING

The basic principles of metric scaling have been described previously.[1,7,8] Metric scaling (also called principal coordinate analysis to distinguish it from principal component analysis) is a technique for reducing the dimensionality of a problem. If we have the distance matrix for a set of data points, the method projects these to lower dimensional space in such a way that each entity is represented by a point and the geometric distance between points is proportional to the dissimilarity of the two entities. The technique involves computing the eigenvalues of the so-called **F** matrix, derived from the original distance matrix by simple transformations.[1] The eigenvectors corresponding to the k largest positive eigenvalues give the best k dimensions in which to represent the object. If there are negative eigenvalues, the distance matrix cannot be represented in Euclidean space.

Metric scaling, which is described in statistics as a "Q-technique",[1] and principal component analysis (PCA), which is classified as an "R-technique",[1] are two different methods of factor analysis. If the distance matrix is composed of Euclidean distances, the scores obtained from the two methods are identical. However, even in this case they are different from each other computationally, since metric scaling is faster on general $n \times p$ matrices ($n < p$) than principal component analysis.[1]

There are many cases in which direct proportionality between the distances of the points after scaling and those of the original objects is not necessary and it is sufficient to preserve the rank order of the original distances. This is achieved in multidimensional scaling (also called nonmetric scaling). The criterion that is used to measure the closeness of fitted distances to the original ones is called STRESS (acronym for standardized residual sum of squares). It provides a measure of fit defined as the sum of the squares of the differences between distances and disparities divided by the sum of the squares of distances. The computation of

* To whom correspondence should be addressed. Phone: (519) 823-9088. Fax: (519) 823-9401. E-mail: mfeher@nanodesign.com.

10.1021/ci000112+ CCC: $20.00
Published on Web 01/27/2001

STRESS involves the monotonic regression of the fitted distances to the observed ones.[1,7] Multidimensional scaling seeks to find the set of points that minimizes STRESS. The value of STRESS is indicative of the performance of the method. Increasing the dimensionality of the solution generally reduces the value of STRESS. To qualify solutions obtained from multidimensional scaling, it was originally proposed that values below 0.05 are excellent, values between 0.05 and 0.10 are satisfactory, and values between 0.1 and 0.15 are marginally acceptable.[7] However, the proper assessment of the acceptability of these values needs to be made by considering the number of points and the number of dimensions involved, and higher values may also be acceptable.[1]

In this work, scaled distances obtained from metric scaling of the distance matrix were used as initial estimates for multidimensional scaling. In the case of metric scaling, it was always checked whether all eigenvalues are positive. The relative magnitude of the first 2−3 normalized eigenvalues was applied to judge the efficiency of the process. As in principal component analysis, the ratio of the sum of the eigenvalues in the selected dimensions and the original number of dimensions provides a goodness of fit of the representation.[1] However, this figure of merit is no longer applicable when there are negative eigenvalues.

In the examples below, multidimensional scaling was considered to have achieved its goal if the original distance matrix could be reduced to two dimensions with a value of STRESS of less than 0.15. In the case that this value was slightly higher than this, the results were treated qualitatively. In cases in which it was over 0.25, the scaling was considered unsuccessful. Although solutions in higher dimensional space could be acceptable in such cases, this would lead to the loss of the visual advantages of the approach described here and hence was not attempted. It must be noted that the square of the STRESS value is also often used in the literature as an indicator of the quality of the multidimensional scaling process. In this study no significant difference in the quality of clustering conformers was found with the two indicators.

ASSIGNING CLUSTER REPRESENTATIVES FROM MULTIDIMENSIONAL SCALING

As previously mentioned, multidimensional scaling has been identified in the past as a possible method for rapid visual characterization of conformer space.[9] For the method to be used more quantitatively, it was necessary to devise a way to determine the optimum number of clusters, assign cluster memberships, and select cluster representatives.

Determining the optimum number of clusters is a controversial issue in conformational clustering.[4] Methods have been suggested based on reordering entropy[4] and on ANOVA-like analysis of cluster distribution,[4] although both deliver values different from intuitive ones. It has also been shown that if clusters are clearly defined and well separated, the statistics based on the separation ratio performs well.[4] Assuming that the aim is simply to derive a small number of clusters to simplify handling multiple conformations, the solution developed in this work works well. It is based on the visual perception of the separation of clusters.

In the multidimensional scaling process, the distances between clusters are projected onto a lower number of dimensions. In most cases studied in this work, two dimensions were sufficient to provide a reasonable representation of the data. If the STRESS value is zero, the rank order of all the original distances will be preserved; i.e., the closer two points are to each other in reality, the closer they will be on the two-dimensional plot. In this case, applying one of the studied hierarchical clustering methods to the points of this plot would be equivalent to clustering the original dataset. As the STRESS value increases, the validity of this approximation deteriorates, but at the same time the validity of the entire scaling process also decreases.

In view of the above, we performed hierarchical clustering directly on the multidimensional scores. This provides a fast and efficient way of grouping the points in the plot. Clustering on the original data can also be performed and the results displayed in the multidimensional plot. This somewhat slower process delivers identical results in the case of low STRESS values (<0.2). The clustering process is carried out in the following way. First, the number of clusters is selected, and then the hierarchical clustering on the multidimensional scores is performed. These are then displayed, showing cluster boundaries. This process is subsequently repeated with different numbers of clusters until the resulting memberships correspond to those expected intuitively, i.e., the assignment of cluster memberships is in accordance with the visual separation of clusters. Although this process is somewhat subjective, the definition of the cluster (i.e., that similarities within the groups are significantly greater than those between the groups) was always considered when decisions were being made about the quality of the outcome. The calculation for a given number of clusters takes less than 1 s (on a 300 MHz PC), and hence a satisfactory solution can be found quickly. In the present work different hierarchical clustering methods were tested and compared. After the assignment of cluster memberships, cluster representatives are selected on the basis of proximity to the cluster centroid, defined by the arithmetic mean of the corresponding multidimensional scaling scores.

CALCULATIONS

All modeling work, including conformational searches and molecular dynamics calculations, was performed using the Molecular Operating Environment (MOE) suite of programs.[13]

The molecules studied in this work are displayed in Figure 1. For the smallest molecule, dopamine, a systematic conformational search was applied using the two significant torsional angles and a step size of 60°. The conformers were optimized with the MMFF94 force field[14] using a continuum solvation model.[15,16] The gradient criterion for the optimization was 0.001. For the other molecules, the conformational search was performed using the random incremental pulse search (RIPS) method.[17] The search was terminated after 1000 failures in a row to generate a novel conformer within the specified energy window of 10 kcal/mol above the minimum energy conformer. The conformational search was then repeated to ensure that the conformational space was adequately covered. In all cases, the number of new conformers was below 5%. Clustering was performed on a subset of these conformers, selected in accordance with the literature examples (in the 3−10 kcal/mol range above the

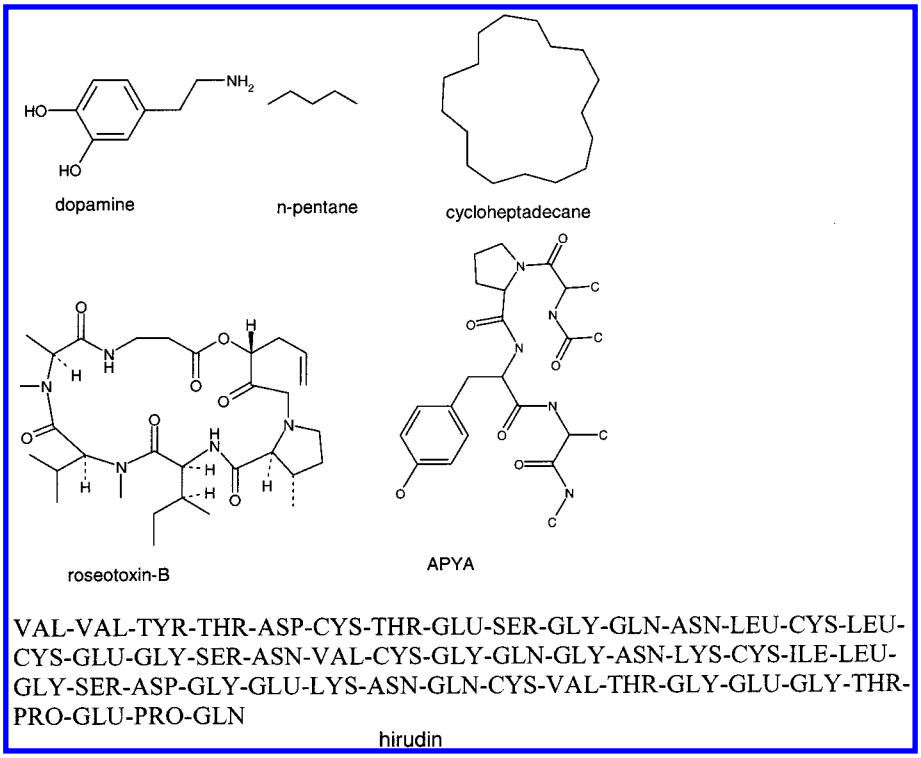

Figure 1. Structures of the molecules used as examples in this study.

global minimum). The MMFF94 force field was used with a continuum solvation model in the geometry optimizations.

Metric and multidimensional scaling, as well as hierarchical clustering, were performed using the NAG statistical add-in for Excel.[18] This tool has recently been shown[19] to produce results identical to those of the NAG libraries and is free of the statistical errors that have been documented for Excel itself.[20] The following hierarchical clustering methods were applied: single-linkage (or nearest-neighbor), complete-link (or furthest-neighbor), group-average (average-linkage), and minimum-variance.

In this work, the distance matrix was used as a starting point for clustering. This was obtained from the pairwise Euclidean rms distance between the same atoms of two conformations after rigid body superposition (i.e., after this rms distance was minimized). Next, the distance matrix was filtered to remove duplicates and similar conformers. During this process, the conformers were first sorted according to their energies. The minimum value in each row of the distance matrix was identified. This value is the rms distance between the given conformer and the one most similar to it and was used to apply cutoff criteria. Generally, conformers that were less than 0.1 Å away from one with lower energy were removed from the set.

A general problem in the analysis of the distance matrix is related to the existence of symmetry-equivalent atoms in the molecule (i.e., local symmetry). For example, if an otherwise nonsymmetrical molecule contains a phenyl group, the paired ortho carbons and the paired meta carbons are equivalent. After the phenyl group is rotated around the bond of attachment by 180°, the conformation of the molecule remains identical. Despite this, there will be a nonzero entry in the distance matrix between the two ortho and also between the two meta "conformers". In this work, symmetry-equivalent atoms were identified by calculating the Cahn−Ingold−Prelog priority number for each atom in the molecule (the aPrioCIP function in SVL[13]) and identifying those atoms with equal priorities. Symmetry-equivalent atoms were then eliminated from the calculation of the distance matrix of the molecule. As a result, the rms distances in the distance matrix were indeed representative of the real difference between two conformers.

RESULTS AND DISCUSSIONS

The performance and utility of the two scaling techniques for conformer clustering was tested on a diverse set of examples in this study. These included simple conformational searches on small molecules and peptides, the analysis of molecular dynamics runs, and the analysis of protein conformations from NMR spectra. These examples will be described next.

Conformations of Dopamine. This example was selected because the small size of the molecule enabled us to make detailed comparisons with other clustering methods. The systematic conformational search produced 22 conformations under the given conditions. The results of metric and multidimensional scaling are shown in Figure 2. In the metric scaling the three highest eigenvalues were 0.64, 0.20, and 0.09. These eigenvalues show that the first two dimensions account for about 85% of the variability of the data, and Figure 2 is plotted along these dimensions.

Differentiation between clusters was only slightly better with multidimensional scaling (STRESS = 0.085); results of the latter are shown in Figure 2b. The cluster memberships shown in this figure are for the selection of five clusters and the single-linkage approach. The cluster memberships obtained were identical to those from the single-linkage method on the original dataset with the same number of clusters. Increasing the number of clusters to six had the effect of splitting up cluster 3, which is probably an equally acceptable

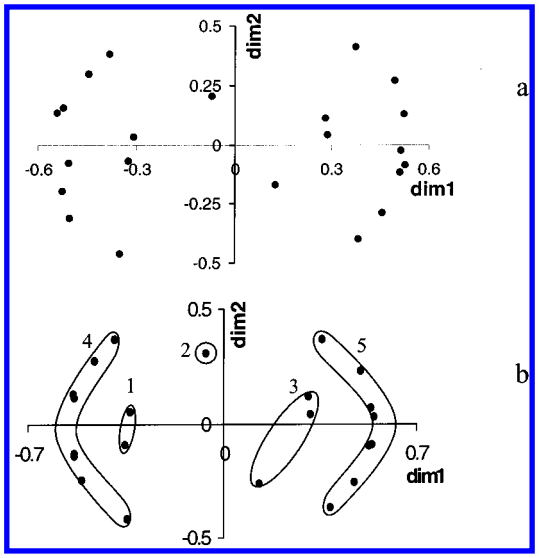

Figure 2. Clustering of the conformations of dopamine obtained from a systematic conformational search using (a) metric scaling (the three highest eigenvalues were 0.64, 0.20, and 0.09) with the first two dimensions plotted against each other and (b) multidimensional scaling (STRESS = 0.085). The legend is the cluster membership from hierarchical cluster analysis using the single-link method.

solution. Increasing the number further will split clusters 4 and 5 first, whereas fewer clusters becomes intuitively acceptable again when the number is two (i.e., differentiating between positive and negative abscissa values in Figure 2b or, in other words, between conformers differing in the torsional angle around the C−C substitution of the phenyl ring, which is a central angle and thus has greater effect on the conformation than the other angle).

The single-linkage method is known to lead to elongated clusters, in which pairs of very dissimilar units may occur.[1] This arises because a unit can join a group on account of its similarity with just one existing member of that group. Although this problem is not immediately apparent in this example, other clustering methods were also tested. The group-average and the minimum-variance methods also provided intuitively acceptable results. In comparison to the results from the single-linkage method (Figure 2b), clusters 1 and 2 were joined, whereas molecules in cluster 4 with positive and negative values in the second dimension end up assigned to different clusters. In contrast, the complete-link method provided much less satisfactory solutions. In these, the points with positive values in the second dimension in clusters 1 and 4, as well as in clusters 3 and 5, are merged and separate from those with negative values. It is important to emphasize that for all methods and at each clustering level, the cluster memberships using the original distance matrix were identical to those obtained from the scaled distances.

Clustering the Conformations of Roseotoxin-B. As a second example, a cyclic peptide, roseotoxin-B, was selected. The RIPS conformational search on this molecule yielded 106 conformers in a 3 kcal/mol window. The distance criteria ensured that no conformer pairs were closer to each other (in rms distance) than 0.1 Å.

Metric scaling (Figure 3a) reveals the basic pattern in the data, and some groupings can be readily recognized. The first three eigenvalues are 0.53, 0.20, and 0.06, indicating that the plot shown in Figure 3a accounts for almost 75% of

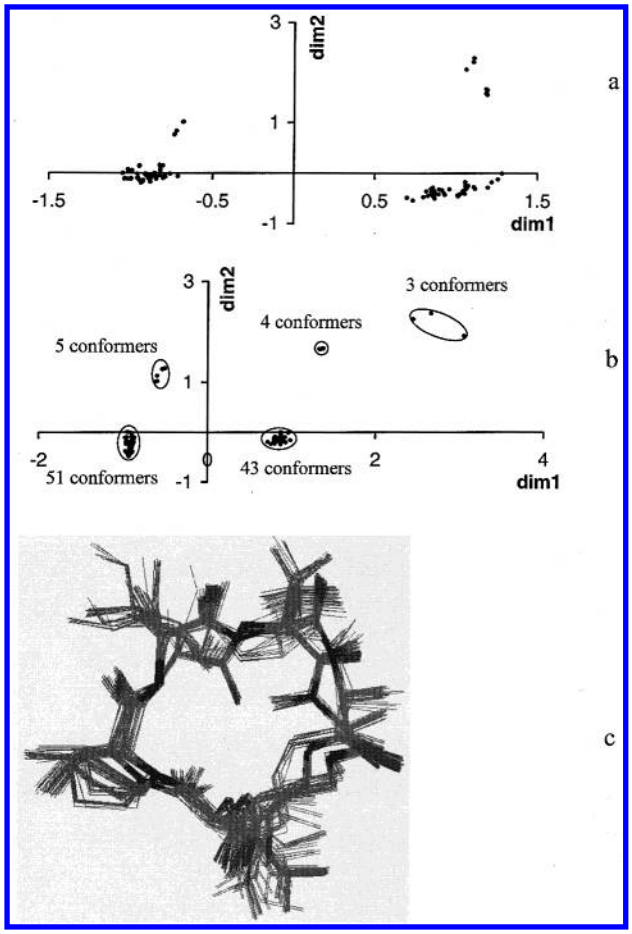

Figure 3. Clustering the conformations of the cyclic peptide roseotoxin-B using (a) metric scaling (the three highest eigenvalues were 0.53, 0.20, and 0.06) and (b) multidimensional scaling (STRESS = 0.041). Cluster memberships were determined using the group-average method on the scaled distances. (c) Overlay of roseotoxin-B conformers.

the variability of the data. The separation into five conformer families is well distinguishable from the plot. However, as can be seen in Figure 3b, multidimensional scaling (STRESS = 0.041) produces an image with much crisper resolution in grouping the conformers into clusters. We can clearly recognize five groups of conformers, two of them containing the majority of observed conformations.

For comparative purposes, the distance matrix of this peptide was also analyzed directly using hierarchical clustering. It was found that if the number of clusters was defined as five (an obvious choice on the basis of the plot), the cluster memberships with the four hierarchical clustering methods were identical to each other and to those obtained by clustering the distances using multidimensional scaling. If the number of clusters was changed from this optimum, differences between the four hierarchical clustering methods arose. When the number of clusters was decreased, the four methods all correctly identified membership in the two major groups (those with 51 and 43 members in Figure 3b), but differences occurred in membership composition of the smaller groups.

Clustering of the same molecule was also investigated by Shenkin[4] using single-link hierarchical clustering with essentially similar results. In that work, two major groupings were identified, similar to those found in our study, with

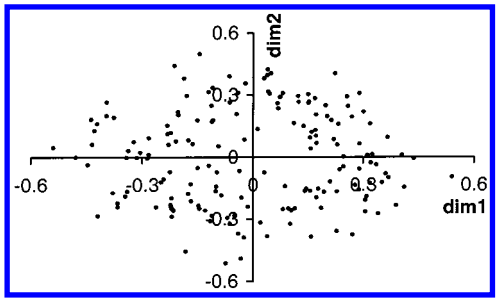

Figure 4. Multidimensional scaling of the conformers of cycloheptane (STRESS = 0.47).

some subgroups differing in the position of the side chains. How these groups relate to each other could not be determined. The advantage of the scaling method is obvious in this situation: the graphical rendering of the groups helps to determine the optimum number of clusters unambiguously and makes it simple to assign cluster memberships.

Cycloheptadecane Conformations. Conformations of the 17-membered alkyl ring were studied as the next example. This molecule is highly flexible, and its conformers have been shown to exhibit little or no clustering.[4] It was hoped that this example would demonstrate how our model behaves in such situations.

The RIPS conformational search (which was performed in vacuo because no polar interactions were expected) produced 5600 conformers within a 10 kcal/mol window, 339 conformations of which were retained within a 3 kcal/mol window on the basis of an rms tolerance criterion of 0.1 Å. These conformers are indeed substantially different from each other: even when the rms distance cutoff between two conformers was raised to 0.6 Å, 171 conformers remained.

The metric scaling plot displayed no conspicuous pattern. Some of the eigenvalues in the process were negative, although these negative values were close to zero (the most negative value was −0.000 23). The existence of small negative eigenvalues implies that metric scaling introduces some small distortions, which nonetheless cast no doubt on the validity of the process.[1] On the other hand, the highest eigenvalues are small (the largest values were 0.100, 0.087, and 0.085), indicating that the first two dimensions explain only an insignificant portion of the variability of the data. This has the effect that the separation of data points with metric scaling is insufficient in two dimensions.

Unfortunately, the situation is not much different with multidimensional scaling. The obtained plot is shown in Figure 4. Although some structure can be recognized in this plot, the value of STRESS for the process is 0.47. Hence, we can conclude that the scaling process introduces unacceptable distortions into the representation of the data and clustering with multidimensional scaling was unsuccessful in this example.

Clustering the Results from Molecular Dynamics Simulations: Pentane. To check the utility of the scaling techniques as classifiers in molecular dynamics (MD) runs, the MD simulation of pentane was studied in this work. Pentane was chosen so that the conclusions can be compared to those obtained with hierarchical clustering.[4] The conditions of the simulation were similar to those of the second simulation in ref 4 (300 K temperature, 20 ns simulation time, saving one structure every 100 ps).

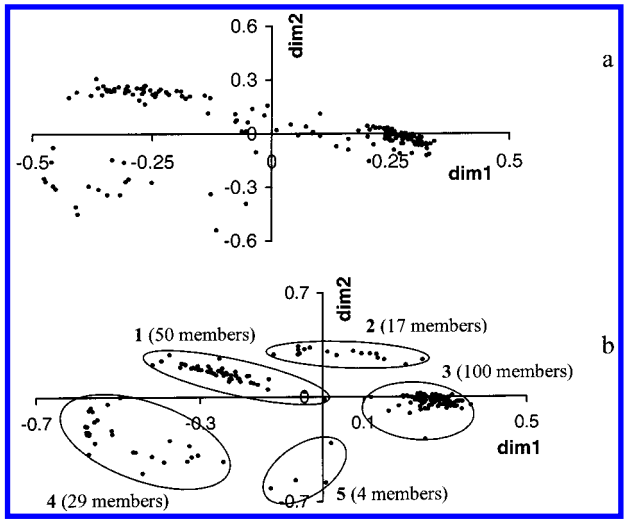

Figure 5. Clustering the results of a molecular dynamics simulation of pentane using (a) metric scaling (the three largest eigenvalues are 0.58, 0.20, and 0.16) and (b) multidimensional scaling (STRESS = 0.2). Cluster memberships were determined using the group-average method on the scaled distances.

The result of the clustering with the metric scaling method is presented in Figure 5a. The three largest eigenvalues are 0.58, 0.20, and 0.16. Although the first two dimensions explain over 75% of the variability in the data, the separation of clusters is still relatively inefficient. Nonetheless, the clustering of points in certain areas can clearly be recognized.

In contrast to metric scaling, separation into groups with multidimensional scaling was better (see Figure 5b). The value of STRESS in this process was 0.20, indicating that the distortions introduced by the scaling are high but still acceptable. In Figure 5b five groups can be recognized. Figure 6 displays an example of the high quality of the conformer separation, presenting the original conformational assembly from the simulation (Figure 6a) and the conformers in cluster 1 (Figure 6b).

Using the torsional angles of pentane,[4] it has been shown that a total of four minima are expected if symmetry is taken fully into account. However, it should be borne in mind that conformers in the molecular dynamics simulation are essentially snapshots along the trajectory of the simulation and are not necessarily close to local minima. This is the source of the distribution of conformers in Figure 5. When the geometries of the 200 conformers in this work were optimized, they collapsed into only 4 "real" conformers. When these 4 real conformers were added to the set of 200 above and were displayed in the same diagram as Figure 5b, they were located within the displayed boundaries of clusters 1−4. In this sense, the major clusters identified in this work indeed correspond to the four identified using hierarchical clustering.[4] A better comparison of the groupings would require that multidimensional scaling is performed on the actual data in ref 4. Obviously, in the original hierarchical clustering work[4] a graphical representation of the conformer distribution was unavailable, although it would have helped to identify conformers that lay further away from the cluster centers.

It is interesting to compare the performance of different hierarchical clustering algorithms, using the multidimensional plot only as an aid for visualizing the results. The most

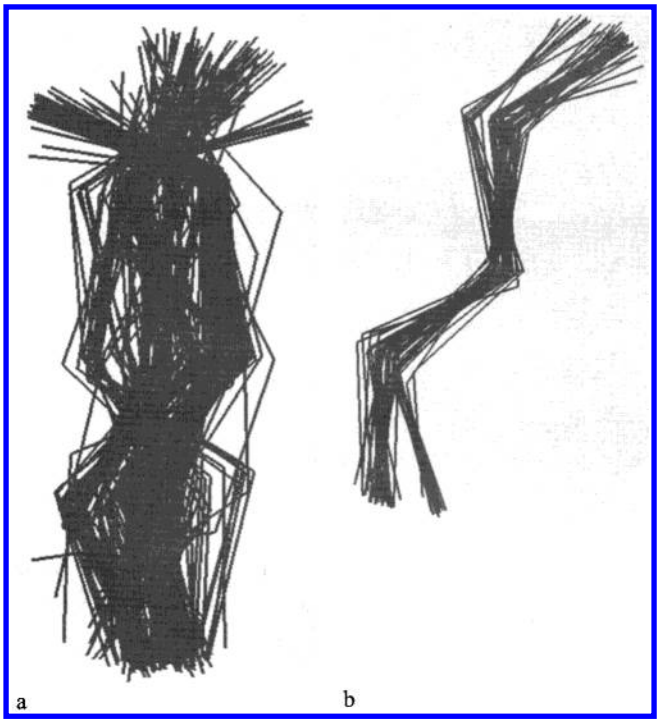

Figure 6. Pentane structures obtained from a molecular dynamics run (see the text for details): (a) all structures overlapped; (b) structures selected using multidimensional scaling, displaying the 50 conformers in cluster 1 in Figure 5b.

intuitive results were obtained with the group-average method. When this is used on the original dataset with five clusters specified, group memberships identical to those shown in Figure 5b are obtained. Other clustering methods provided different results. Most similar to the above were the groups with the minimum-variance method where some of the conformers that were previously in cluster 4 and were nearest to cluster 5 change membership to the latter. The other two methods generated far worse results. When the complete-link method was applied, the major change with respect to Figure 5b was the merger of clusters 2 and 3, whereas cluster 4 was separated into two groups. More disturbing in this case are individual conformers that should seemingly belong to one of the above clusters but end up in a more distant grouping. Finally, the single-linkage method yielded the worst results: clusters 1−3 were merged, containing 166 of the 200 conformers. Cluster 4 was cut into two groups, whereas three conformations of cluster 5 (those closest to the bottom of Figure 5b) each formed a cluster of their own. Identical conclusions concerning the inapplicability of the single-linkage method for clustering conformations from molecular dynamics were reached by Torda and van Gunsteren.[5]

The number and frequency of revisiting different conformational families during the course of the molecular dynamics simulation is a potentially useful piece of information. We can visualize this by plotting the cluster code of the given conformation, such as in Figure 7a. This figure shows that all major conformational families have been revisited at least once. A much simpler way to arrive at the same conclusion is to plot the time evolution of the first component from multidimensional scaling, as displayed in Figure 7b. As can be seen in this figure, the information content is essentially similar to that in Figure 7a.

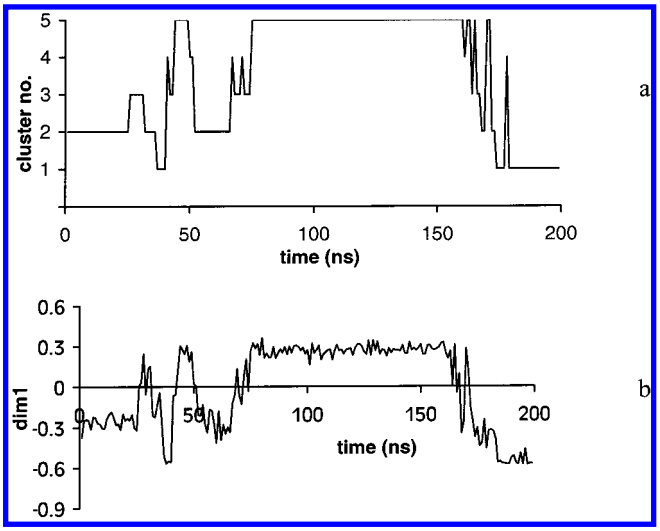

Figure 7. Time evolution of the molecular dynamics run for pentane. The frequency of jumping from one conformational family to another and the time spent for each family can be conveniently deduced from this plot. (a) The cluster code, obtained from multidimensional scaling, is displayed as a function of time. (b) The first component from the multidimensional scaling is plotted as a function of time.

Conformations of the Tetrapeptide N-Acetyl-Ala-Pro-Tyr-Ala. The conformational situation for this peptide has been studied by a random search technique, followed by principal component clustering.[21] It was concluded that a PCA-based technique is useful in mapping the conformations of this peptide (especially because this approach allows the conformer distribution to be inspected graphically). The clustering was performed on the 12 torsional angles of the peptide, and the authors identified 5 clusters as the optimum number.[21]

The RIPS-based conformational search in this work identified 845 conformers within a 10 kcal/mol window. (This can be compared with 2000 conformers in ref 21 using no energy window, vacuum calculations, and the ECEPP force field.) Clustering was performed on the 209 conformations with less than 5 kcal/mol of energy above the lowest energy conformation.

As described above, metric scaling under the conditions of the present study is equivalent to PCA. In this work, metric scaling was performed, yielding maximum eigenvalues of 0.34, 0.18, and 0.11. The low magnitude of these eigenvalues implies that the separation of the components is not very good and higher dimensions may play a significant part in the proper separation of conformers. These conclusions are similar to those obtained by the PCA study[21] in which the first three components accounted for about 60% of the variance. The separation of families was also similar between metric scaling on the current conformational set and those with PCA in ref 21. The conformers form two relatively well separated families. As in the PCA analysis,[21] this separation can be recognized when the first component is plotted against the second or the third, but not when the second and third are plotted against each other (see Figure 2 in ref 21). The two families seem to differ principally in the main chain torsional angles.

As was the case for the previous examples, multidimensional scaling (see Figure 8) substantially improved the separation of clusters. The STRESS value was 0.25, which

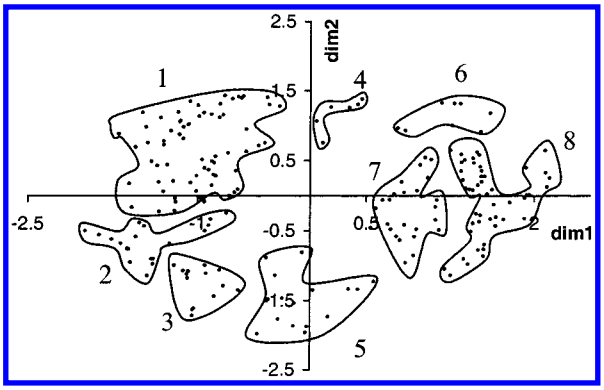

Figure 8. Clustering of the conformations of the tetrapeptide *N*-acetyl-Ala-Pro-Tyr-Ala (APYA) using multidimensional scaling (STRESS = 0.25). Cluster memberships were determined using the group-average method on the scaled distances.

is quite high. The assignment of groups shown in Figure 8 corresponds to clustering the scaled data using the group-average method. The results were essentially similar when the group-average method was applied on the original distance matrix specifying eight clusters, although the cluster boundaries were somewhat fuzzier in this case: about 5% of the data points changed cluster memberships compared to those shown in Figure 8. This is probably because of the distortions introduced by scaling (see high STRESS value). It must be noted that clustering with the group-average method proved to be the best of the studied hierarchical clustering methods in this work. The minimum-variance method led to similar but somewhat fuzzier separations, whereas the complete-link and single-link methods led to completely undefined boundaries in the multidimensional plot.

This example demonstrates that multidimensional scaling is clearly helpful in finding conformer families even in situations when these families are not too well separated. In contrast, clustering based on PCA (and hence also metric scaling) does not lead to a well-defined separation of groups (see, e.g., the spread of cluster 5 after PCA clustering in Figure 2 of ref 21).

Clustering Experimental NMR-Derived Structures: Hirudin. NMR-derived protein structures in the Protein Data Bank are often available as an ensemble containing many structures. An automatic procedure to rapidly cluster these structures based on the hierarchical clustering of the distance matrix using the average-linkage (i.e., group-average) method has been described.[22] As with all the previous examples, the rapid and automatic allocation of conformers to clusters, coupled with the graphical representation of the shape, size, and relative distances among these groups make the application of multidimensional scaling highly appropriate. The example selected to demonstrate the capabilities of the method was the same as that described in ref 22: the experimental conformations of hirudin (4HIR), as taken from the Protein Data Bank.[23]

The three highest eigenvalues for the metric scaling were 0.33, 0.15, and 0.09. This implies that the first two components account for about only half of the variability of the data and that metric scaling is not useful in this case. The results from multidimensional scaling are displayed in Figure 9a. The STRESS value was 0.16, indicating that the

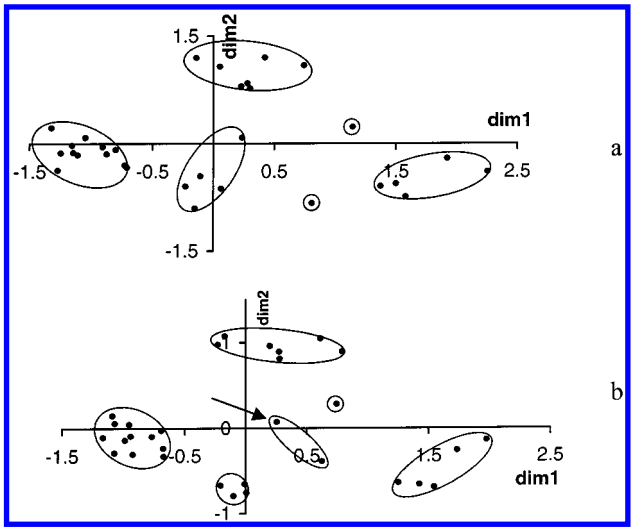

Figure 9. Clustering of the NMR conformations of hirudin (4HIR) using multidimensional scaling and the group-average method on the scaled distances: using all atoms to calculate the distance matrix (STRESS = 0.16); using only the α-carbons to define the distance matrix (STRESS = 0.09). The arrow points to the single conformation, the membership of which has changed as a result.

rank order of the distances would be preserved by using only two components. When six clusters were specified in the published automated process using the group-average method, four major clusters and two singletons (i.e., clusters containing one member each) were obtained.[22] Not surprisingly, if the group-average method with six clusters is applied on the scaled data, an identical distribution, shown in Figure 9a, is obtained. The additional advantage of the multidimensional scaling procedure is that the decision on the number and relative separation of clusters can be verified by visual inspection.

In the analysis of protein conformations, consideration of only the position of the α-carbons is often applied as a simplification, as was done, for example, in the analysis and clustering of the NMR-derived conformations of the Sox-5 HMG-box protein.[24] To evaluate the effect of such an approximation in the current example, multidimensional scaling was repeated on the distance matrix, which was derived using only α-carbon distances. The results are shown in Figure 9b. The higher eigenvalues of the first two components (0.44, 0.22, 0.07) in metric scaling and the lower value of STRESS in the multidimensional scaling (0.09) demonstrate the higher statistical reliability in this case. (The better statistics were obviously achieved at the cost of neglecting major interactions.) As can be seen in Figure 9b, the results are qualitatively similar to those from clustering on all atoms, except for the membership of a single conformer, suggesting that in this case it is a valid approximation to use only α-carbons.

CONCLUSIONS

The application of metric and multidimensional scaling for simplifying the treatment of conformer ensembles is demonstrated in this work. Both scaling processes are fast and allow the visual inspection of conformer maps. These maps usually allow the direct determination of the optimum number of clusters. In all examples in this work, multidimensional scaling was found to be superior to metric scaling.

In this work an automated process was described, by which scaled distances can be rapidly clustered and group memberships can be assigned. The quality of this assignment can be inspected visually for different numbers of clusters and different clustering algorithms. For STRESS values below about 0.20, clustering the original distance matrix or the scaled data delivered identical results; at higher STRESS values differences started to emerge. In general, low STRESS values (<0.2) indicated that conformer families were well separated from each other, whereas high STRESS values (>0.3) signified that conformers cannot be separated well. In other words, the value of STRESS proved to be a good figure of merit and gave an immediate indication as to the expected quality of conformer clustering.

The possibility of visually inspecting conformer maps using multidimensional scaling provided a unique opportunity to compare the performance of different hierarchical clustering methods. Among the four methods tested, the group-average method always provided highly intuitive results. The minimum-variance method gave similar, although somewhat less intuitive results. It is also our experience that single-link clustering has a tendency to chain together clusters that would intuitively be regarded as distinct. The opposite was found to be true for the complete-link method: it often separates closely lying clusters, which could be regarded intuitively as belonging to the same group. In this respect, on the basis of the hierarchical clustering schemes examined, we found the group-average method as the method of choice.

ACKNOWLEDGMENT

Dr. Harry Zuzan is acknowledged for first mentioning the technique of multidimensional scaling to the authors.

REFERENCES AND NOTES

(1) Krzanowski, W. J. Principles of Multivariate Analysis; Clarendon Press: Oxford, 1988.
(2) Willett, P. Similarity and Clustering in Chemical Information Systems; Research Studies Press: Letchworth, 1987.
(3) Murray-Rust, P.; Raftery, J. Computer Analysis of Molecular Geometry, Part VI: Classification of Differences in Conformation. *J. Mol. Graph.* **1985**, *3*, 50−59.
(4) Shenkin, P. S.; McDonald, D. Q. Cluster Analysis of Molecular Conformations. *J. Comput. Chem.* **1994**, *15*, 899−916.
(5) Torda, A. E.; Gunsteren, W. F. Algorithms for Clustering Molecular Dynamics Configurations. *J. Comput. Chem.* **1994**, *15*, 1331−1340.
(6) Vesterman, B.; Golender, V.; Golender, L.; Fuchs, B. Conformer Clustering Algorithm and its Application for Crown-Type Macrocycles. *J. Mol. Struct.: THEOCHEM* **1996**, *368*, 145−151.
(7) Schiffman, S. S.; Reynolds, M. L.; Young, F. W. Introduction to Multidimensional Scaling: Th*eory, Methods, and Applications*; Academic Press: New York, 1981.
(8) Everitt, B. S. Graphical Te*chniques for Multivariate Data*; North-Holland: New York, 1978.
(9) Taylor, R. The Cambridge Structural Database in Molecular Graphics: Techniques for the Rapid Identification of Conformational Minima. *J. Mol. Graph.* **1986**, *4*, 123−131.
(10) Glunt, W.; Hayden, T. L.; Raydan, M. Molecular Conformations from Distance Matrices. *J. Comput. Chem.* **1993**, *14*, 114−120.
(11) Shortle, D.; Simons, K. T.; Baker, D. Clustering of Low-Energy Conformations Near the Native Structures of Small Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 11158−11162.
(12) Ward, D. J.; Finn, P. W.; Griffith, E. C.; Robson, B. Comparative Conformation-Activity Relationships for Hormonally- and Centrally-Acting TRH Analogues. *Int. J. Pept. Protein Res.* **1987**, *30*, 263−274.
(13) Molecular Operating Environment, Version 2000.02, Chemical Computing Group Inc., Montreal, Quebec, Canada.
(14) Halgren, T. A. The Merck Molecular Force Field. I. Basis, Form, Scope, Parametrization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(15) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.
(16) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *J. Phys. Chem. A* **1997**, *101*, 3005−3114.
(17) Ferguson, D. M.; Raber, D. J. A New Approach to Probing Conformational Space with Molecular Mechanics: Random Incremental Pulse Search. *J. Am. Chem. Soc.* **1989**, *111*, 4371−4378.
(18) NAG Statistical Add-Ins for Excel, Release 1.1, The Numerical Algorithm Group Ltd., Oxford, England, 1999.
(19) Girvan, R.; Grant, F. From al-Khwarizmi to the Foundation of Computing. *Sci. Comput. World* **2000**, *52*, 31.
(20) McCullough, B. D.; Wilson, B. On the Accuracy of Statistical Procedures in Microsoft Excel 97. *Comput. Stat. Data Anal.* **1999**, *31*, 27−37.
(21) Michel, A. G.; Jeandenans, C. Multiconformational Investigations of Polypeptidic Structures, Using Clustering Methods and Principal Components Analysis. *Comput. Chem.* **1993**, *17*, 49−59.
(22) Kelley, L. A.; Gardner, S. P.; Sutcliffe, M. J. An automated Approach for Clustering an Ensemble of NMR-Derived Protein Structures into Conformationally Related Subfamilies. *Protein Eng.* **1996**, *9*, 1063−1065.
(23) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(24) Adzhubei, A. A.; Laughton, C. A.; Neidle, S. An Approach to Protein Homology Modelling Based on an Ensemble of NMR Structures: Application to the Sox-5 HMG-Box Protein. *Protein Eng.* **1995**, *8*, 615−25.

CI000112+