Fuzzy Clustering as a Means of Selecting Representative Conformers and Molecular Alignments

Miklos Feher* and Jonathan M. Schmidt

SignalGene Inc., 2-335 Laird Road, Guelph, Ontario, N1G 4P7, Canada

Received September 27, 2002

This paper describes the first application of fuzzy c-means clustering for the selection of representatives from assemblies of conformations or alignments. In case of alignments, their quality is taken into account using a weighted c-means scheme, developed in this work. The performance of fuzzy cluster validity measures, such as compactness, partition function, and entropy, are studied on several examples, but the visual 3D representation of data points is shown to be most beneficial in determining the optimum number of clusters. Fuzzy clustering is expected to perform better than crisp clustering methods in cases where there are a significant number of "outliers", such as in molecular dynamics simulations and molecular alignments.

INTRODUCTION

Clustering molecular conformations and selecting a subset of representative conformers is useful in many areas of computational chemistry, such as 3D-QSAR applications. We have recently described an efficient method to cluster molecular conformations using metric and multidimensional scaling.[1] We have also demonstrated recently[2] that clustering flexible molecular overlays[3] is useful in identifying potential binding modes and explaining partitioning behavior. The clustering process was shown to be robust and automatic, and the results could be displayed visually, enabling rapid determination of the optimum number of clusters in many cases. From the determined cluster memberships it is usually easy to derive the cluster centers, except when the clusters almost overlap. It was shown that the multidimensional scaling process fails to work when the standardized residual sum of squares (STRESS) falls above a certain number (around 0.15). This usually occurs when the data do not cluster well.

The need to select representative conformations or alignments in cases where there are no clear groupings in the data provided the initial motivation to pursue fuzzy clustering. If fuzzy logic is applied, there is no need to assign individual data points to any one cluster. In this paradigm, a data point can belong to more than one cluster. Inclusion in each cluster is quantified by the degree of membership. To our knowledge fuzzy clustering has not previously been applied to cluster molecular conformations or alignments. The aim of the present work was to develop and test such an algorithm in the hope that it will provide a generally applicable clustering approach to deal with a collection of conformers and molecular alignments.

FUZZY C-MEANS CLUSTERING

In this work, the fuzzy c-means algorithm (FCM)[4,5] was applied to cluster conformations. This algorithm is an iterative optimization that minimizes the cost function

$$J = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m ||x_k - v_i||^2 \qquad (1)$$

where n is the number of data points, c is the number of clusters, u_{ik} is the degree of fuzzy membership of the kth data point in the ith cluster, m is the fuzziness index (it was always set to 2 in this work), x_k is the kth data point, and v_i is the ith cluster center. In this formula, the membership function, u_{ik}, is described by the following equation:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{x_k - v_i}{x_k - v_j}\right)^{2/(m-1)}} \qquad (2)$$

In the optimization process, using the desired number of clusters, c, and some initial guess for the cluster centers, v_i, the membership function is calculated and the cost is minimized, resulting in a genuine local minimum or a saddle point.[6] The quality of the solution depends on the initial choice of parameters. In this work, the initial membership matrix was populated using random numbers. Hence for each desired number of clusters, the clustering was repeated several times, and the best solution was chosen based on criteria described below.

Fuzzy c-means clustering of conformational or alignment data was achieved in the following way. First, the distance matrix was calculated. This is essentially a matrix of Euclidean distances obtained after pairwise rigid-body superposition of all conformations or alignments. Usually only symmetry-unique heavy atoms are considered in this distance matrix to avoid artificially increased distances that would be generated for symmetry equivalent atoms.[1] In an optional prefiltering step, duplicates or too similar conformations/alignments were removed from the set as previously described.[1,2] Next metric scaling is performed on the distance matrix. In this process the distance matrix is projected to

* Corresponding author phone: (519)823-9088; fax: (519)823-9401; e-mail: miklos.feher@signalgene.com.

10.1021/ci0200671 CCC: $25.00 © 2003 American Chemical Society
Published on Web 03/27/2003

lower dimensional (3D) space in such a way that each entity (conformation or alignment) is represented by a point and the geometric distance between the projected points remains proportional to the original distances of the projected entities. The technique involves computing the eigenvalues of the so-called F matrix, derived from the original distance matrix by simple transformations.[7] The three largest positive eigenvalues indicate the three dimensions in which the distances were to be projected, with the eigenvectors providing the 3D coordinates. In this process, the proportionality of the original distances is only retained provided that none of the eigenvalues is negative. Hence in all cases it was ensured that none of the normalized eigenvalues were significantly negative (i.e. less than −0.1). The projected 3D coordinates from metric scaling were then used as inputs for the fuzzy clustering. First the fuzzy membership matrix was initialized by random numbers so that the sum of its elements was 1. From the membership matrix the fuzzy centroids were derived, and the fuzzy membership matrix was calculated. The process was repeated until the decrease in the cost function, J, was less than a predefined amount (10^{-5}). The process was repeated using different random seeds for the membership matrix. The fuzzy cluster membership matrix and the cluster representatives were obtained as the outcome of the FCM clustering process. Despite the fact that the 3D representation of the metrically scaled distance matrix often allowed the number of clusters to be assessed directly, the FCM clustering process was repeated for different numbers of clusters. The results could be evaluated visually (i.e. do the cluster representatives appear to be truly representative of a significant group of conformers or alignments) or by using different computed figures of merit.

WEIGHTED FUZZY C-MEANS CLUSTERING

Clustering of flexible alignments was first described in our earlier report.[2] To cluster alignments, a projection scheme was developed that separates the orientational and conformational aspects of the overlays, so that these could then be clustered separately. A detailed description of the scoring function, used to rank alignments for a given molecule pair, has been given.[3] In summary, the overlap scores are expressed as the weighted sum of Gaussian functions with exponential distance dependence. The weighting is based on different properties, such as the presence or absence of atoms, aromatic rings, hydrophobic atoms, H-bond acceptors and donors as well as surface exposure, the value of atomic logP contributions, charge and molar refractivity contributions. In practice, volume, aromatic, and hydrogen bonding contributions were shown to be the most relevant in alignments.[3] The scores can be normalized using simple normalization or a Tanimoto-like scheme.[2] It was shown[2] that these scores are useful when separating alignments with multidimensional scaling, although with the described method this separation could only be achieved using a somewhat arbitrary manual selection process.

To cluster flexible molecular alignments using fuzzy clustering, some changes were also necessary in the FCM method. This need arose because alignment solutions differ in the number and quality of overlapped features, as reflected by the normalized alignment score.[2] To favor grouping of higher quality alignments without the disturbing effects of low quality alignments, a weighting scheme was developed for the fuzzy c-means clustering. In traditional fuzzy c-means clustering the new cluster centers are determined using the following formula:

$$V_i^{centroid} = \frac{\sum_{j=1}^{n} (u_{ij})^m x_j}{\sum_{j=1}^{n} (u_{ij})^m} \tag{3}$$

in our weighted fuzzy c-means scheme we weighted the data matrix with the pth power of the vector of similarity scores, s

$$V_i^{centroid,weighted} = \frac{\sum_{j=1}^{n} (u_{ij})^m x_j s_j^p}{\sum_{j=1}^{n} (u_{ij})^m s_j^p} \tag{4}$$

The similarity scores, s_j, were the normalized alignment scores from flexible alignments, as defined previously.[2] The exponent p determines the extent to which lower quality alignments are considered and its choice is somewhat arbitrary. In case of the estradiol-raloxifene example, the alignments fell into three groups: 0.95−0.98, 0.67−0.81, and 0.37−0.59.[2] In light of the significant difference among these three groups of alignments, the choice of this parameter had only a minor impact on the results of the clustering process. In contrast, in the thiorpan-retrothiorpan example the normalized alignment similarities were nearly continuously distributed between 0.68 and 0.99.[2] The alignments fell into three categories based on the number of overlapping features.[2] Consistent separation between these alignment types could be achieved if the exponent p was chosen to be 8. This value of p was satisfactory for all the examples tested. To help the visual identification of clusters, points of the metrically scaled distance matrix were color-coded according to their normalized similarity. To avoid biases introduced into the color-coding and to allow automatic color assignments, the similarity score was linearly mapped to a color map. In this way, improving alignments range in color from blue to red. This color-coding also simplifies the visual determination of the optimum number of clusters.

FUZZY CLUSTER VALIDITY MEASURES

A variety of different figures of merit or cluster validity criteria are proposed in the literature to characterize the quality of the clustering process. The simplest is the partition coefficient, which describes the fuzziness of the partition[6]

$$F = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} (u_{ij})^2 \tag{5}$$

where the definitions of the quantities in this equation are the same as those given above. The partition coefficient is inversely proportional to the overall average overlap between pairs of fuzzy subsets, with $F = 1$ indicating no membership sharing between any pairs of fuzzy clusters (i.e. sharp

clustering). Finding the maximum of the partition coefficient for different number of clusters is often assumed to produce the optimum number of clusters.[6]

Another often-used cluster validity criterion is the partition entropy, which is related to the previous quantity. It is defined in the following way:[6]

$$H = -\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik} \ln u_{ik} \tag{6}$$

This quantity is essentially the application of Shannon's entropy to fuzzy partitioning. It is often assumed that the partition entropy, calculated for different number of clusters, attains a local minimum for the optimum number of clusters.

It must be mentioned that both the partition coefficient and entropy show a monotonically decreasing tendency with the number of clusters. This arises as these quantities are not normalized: the partition coefficient can take values between $1/c \leq F \leq 1$, whereas the partition entropy ranges between $0 \leq H \leq \ln c$. The dependence of these quantities on the number of clusters can be compensated for by normalizing these functions the following way:[6]

$$F' = \frac{cF - 1}{c - 1} \tag{7}$$

and

$$H' = \frac{H}{1 - \frac{c}{n}} \tag{8}$$

Although these normalization schemes remove the monotonic decrease at high *c*-values (when *c* is close to the number of data points *n*), they have little beneficial effect in the present study in which only a small number of clusters was involved.

The partition function and entropy both measure the amount of fuzziness of the data from the membership functions. What they both ignore is the actual geometric properties of the derived clusters. Cluster validity criteria have been proposed to allow for this deficiency.[6,8] In this work, the "compactness and separation validity function" (it will be referred to as compactness) was applied for this purpose. It is defined by the following equation

$$S = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} u_{ij}^{2}||v_i - x_j||^2}{n \cdot (\min_{i,j}||v_i - v_j||)^2} \tag{9}$$

where *i* and *j* run through the identified cluster centers. Note that the quantity $\min_{i,j}||v_i - v_j||$ in the denominator is the minimum distance between cluster centroids, $||v_i - x_j||$ is the Euclidean distance between center *i* and data point *j*, whereas $u_{ij}||v_i - x_j||$ is their fuzzy deviation. For the FCM algorithm and the *m* value of 2 applied in this work, the formula is simply the following

$$S = \frac{J}{n \cdot (\min_{i,j}||v_i - v_j||)^2} \tag{10}$$

which is easy to calculate since the objective function, *J*, is being minimized by the FCM. The additional factor in the denominator, the minimum distance between cluster centroids, describes the separation of the clusters. Clearly, the more widely separated the clusters are, the larger this distance is and the smaller *S* becomes. Similarly to the other fitness criteria above, *S* will also monotonically decrease when the number of clusters is large. Although functions exist that would eliminate this decrease,[9] as the number of clusters was much smaller than the number of data points[8] this problem was thought to be insignificant and hence no correction was performed in this work.

As mentioned previously, fuzzy clustering was initiated from a random seed and repeated a number of times after which the best solution was determined. It was found that for any given number of clusters, the best solution could be selected based on any of the three cluster validity functions described above. However, there are significant differences in the sensitivity of these functions. In the case of five clusters in the roseotoxin example (vide infra), the difference in the *S*-values for the best and worst solutions (0.0069 and 1.2221) was far greater than the corresponding difference in the entropies (0.158 and 0.372) or the partition coefficients (0.943 and 0.809). Similarly dramatic differences in the *S*-values between solutions of differing quality were found in all other examples described in this work. The rank order of the solutions was generally the same for the three validity functions (apart from small numerical differences in case of similar quality solutions). Nonetheless, in many cases a range of *S*-values corresponded to the same partition coefficient or entropy values for solutions of visually different quality. This situation probably arose because the partition function and entropy only consider the membership matrix but not the actual data points.[6,8] As a result, the compactness function was applied to select the best random solution. Although the best solution was usually found within the first five attempts, fuzzy clustering in all the described examples was run using 100 random attempts.

METHODS

All molecular modeling tasks in this work were performed with the Molecular Operating Environment (MOE) modeling suite.[10] Distance matrices for conformations and flexible alignments were produced in our earlier work[1,2] and were used as starting points for fuzzy clustering. The calculation of the distance matrix, the separation of the conformational and orientational contributions, and the prefiltering steps were implemented in the SVL programming language of MOE.[10] The distance matrix was calculated from the pairwise Euclidean rms distance between the same atoms of two conformations or alignments after their optimal rigid body superposition (i.e. by minimizing the weighted least squares error function). Symmetry equivalent atoms were identified using the previously described procedure[1] and eliminated from the calculation of the distance matrix. An optional prefiltering step was applied to reduce the number of conformers or alignments.[1,2] Our 3D-graphical visualization tools, written using the scripting languages of MOE[10] and MATLAB,[11] allowed rapid assessment of the data.

The fuzzy clustering work was performed within the MATLAB environment.[11] The metric scaling and cluster validity functions were programmed using the MATLAB

language. The fuzzy-clustering algorithm was based on the FCM code in the Fuzzy Logic Toolbox but was modified as described above. The weighted FCM scheme was also developed within MATLAB.

Calculations were performed on a 1.4 GHz Intel Pentium IV processor with 256 MB RAM running Windows NT. The time taken by the fuzzy clustering process was linearly dependent on the number of random trials. For example, for 20 random trials it took 1−2 min to test cluster sizes between 2 and 20, calculate the cluster validity functions for all, find the optimum, and plot the results. Clustering could then be repeated for a specified number of clusters in case it was thought necessary after visual observation of the results, this typically took a few seconds.

RESULTS AND DISCUSSION

To characterize the quality of alignments with fuzzy clustering and to gauge potential improvements in comparison to our earlier work, all examples were taken from our multidimensional scaling publications[1,2] and clustered using techniques developed in this work. The first three examples exemplify the application of fuzzy clustering to conformations, whereas the remaining two demonstrate their use in flexible alignments. For simple comparison, the sets of conformers and alignments had to be identical to the ones generated for multidimensional scaling and these were taken from our earlier works.[1,2] Hence issues relating to the process of conformer and alignment generation are only briefly discussed here. The starting point in this work in each case is the distance matrix that describes the similarity of all possible conformer/alignment pairs. As in our previous work, the objective was to establish similarities/differences and find groupings in the generated sets of conformations and alignments.

Clustering the Conformations of Roseotoxin-B. As the first example, the results of the conformational analysis of the cyclic peptide, Roseotoxin-B, was studied. The RIPS conformational search on this molecule yielded 106 conformers in a 3 kcal/mol window.[1] The distance criteria ensured that no conformer pairs were closer to each other (in rms distance) than 0.1 Å.

The first three eigenvalues in the metric scaling are 0.53, 0.20, and 0.06, indicating that these three dimensions account for almost 80% of the variability of the data. The clustering of conformations is shown by the corresponding metric scaling plot (see Figure 1). From the 3D plot in Figure 1 five distinct conformer families can be identified. Cluster representatives, obtained by fuzzy clustering for the choice of five clusters, are indicated by asterisks in Figure 1. In this example fuzzy memberships are converted easily to "crisp" clusters, since all conformers have a minimum of 95% membership in their respective clusters.

Cluster validity measures were calculated for different number of clusters in order to evaluate their performance. It was hoped that these functions will clearly and objectively indicate the optimum number of clusters. As discussed previously, we would expect minima in the compactness and entropy curves and maxima in the partition coefficient curves at the optimum number of clusters. The results are displayed in Figure 2. As we can see, the three types of validity functions display some common behavior but there are also differences. The separation of the data into five clusters is decisive in this case and indeed the partition function has

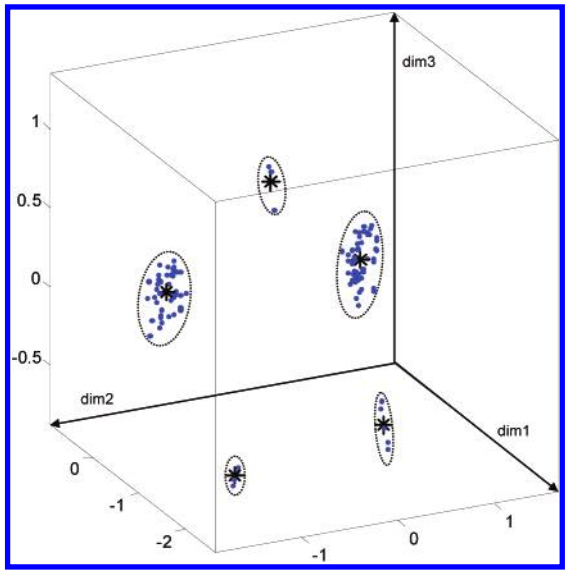

Figure 1. Metrically scaled representation of the conformers of roseotoxin from a stochastic conformational search. There are five well-separated clusters. The asterisks mark cluster representative conformations from fuzzy c-means clustering. The dotted lines encircle points with over 50% membership in the given cluster. The axes represent the dimensions of the scaled coordinates and are marked as *dim1*, *dim2*, and *dim3*. See text for further details.

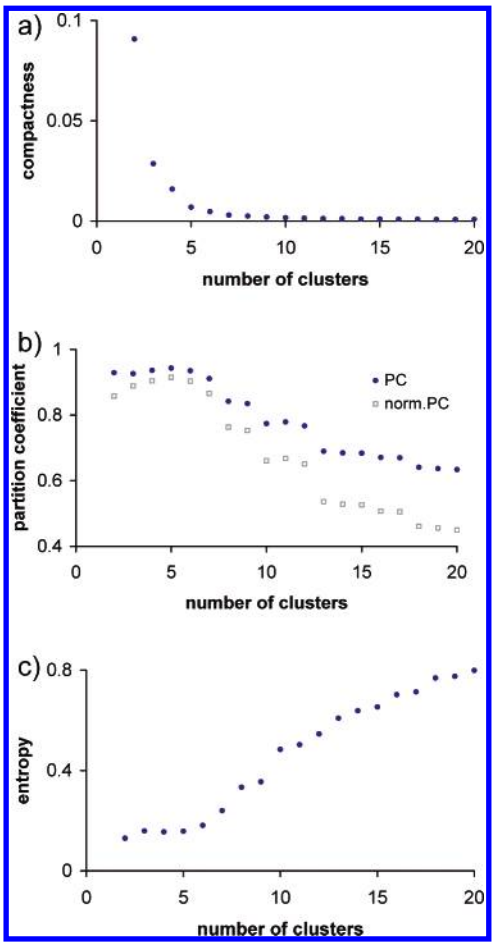

Figure 2. Figures of merit displayed as a function of the number of clusters in the fuzzy c-means clustering of roseotoxin-B conformations (a) compactness (b) partition coefficient and normalized partition coefficient (c) entropy.

its global maximum and the entropy a local minimum at five clusters. Surprisingly, the compactness function displays no minimum at this point despite the fact that, as described above, for any given number of clusters a higher *S* always corresponded to a higher or equal partition coefficient and a lower or equal entropy. The compactness function is maximal for two clusters and decreases monotonically with increasing number of clusters. The entropy function reaches its global minimum at two clusters, and it has local minima at 5, 9, 11, and 15 clusters. In case of the partition coefficient there is a global maximum at five clusters and well-defined local maxima at 2, 9, 11, 15, and 17 clusters. The characteristic point at two clusters reflects the observation that the data in Figure 1 can indeed be grouped visually into two major groups. Even in this relatively simple case the three cluster validity functions predict different optimum number of clusters and the visual inspection of the distribution of data points provides invaluable insight.

Similar results were obtained by Shenkin[12] who used the single-link hierarchical method to cluster Roseotoxin conformations. The optimum number of clusters was two, as determined using the separation ratio.[12] These two clusters are essentially centered where the two biggest clusters are shown in Figure 1. Although a distance map allowed some visualization of the data,[12] the relative distance of identified clusters and the existence of smaller separate clusters could not be determined.[12] The advantage of our scaling method is obvious in this situation: the graphical rendering of the groups makes it possible to visualize the distances between and relationships among clusters.

Clustering of Experimental NMR-Derived Structure. We have previously described the application of multidimensional scaling to the clustering of experimentally determined NMR structures of hirudin.[1] The same data set has also been analyzed using the average linkage algorithm for clustering.[13] The result of applying fuzzy clustering on this data set is shown in Figure 3. Five natural clusters can be readily identified by visual inspection of the plot. The corresponding cluster memberships can be deduced from the membership matrix. The majority of points have fuzzy memberships exceeding 0.5 in the cluster they are identified with in Figure 3. The exception, indicated by an arrow in Figure 3, has a membership of 0.37 in the cluster nearest to it and about 0.2 in the other two neighboring clusters. This is not surprising since in 3D space that point is only marginally closer to its identified cluster representative than to the other two neighboring cluster representatives. It was tested whether any choice in the number of clusters would lead to all memberships being greater than 0.5. For different number of clusters up to 20, there are always at least two conformers that have memberships less than 0.5 and only in the case of three clusters are all memberships above 0.5. However, for the selection of five clusters, crisp cluster memberships can be easily produced from the fuzzy membership function by assigning each conformer to the cluster in which it has the highest membership. In this case the results concur with visual observation.

The dependence of cluster validity functions on the number of clusters is shown in Figure 4. In this example the compactness function clearly indicates that five is the optimum number of clusters, in agreement with visual observation. The partition function has a local maximum at three clusters, whereas the entropy function has a small inflection at three clusters.

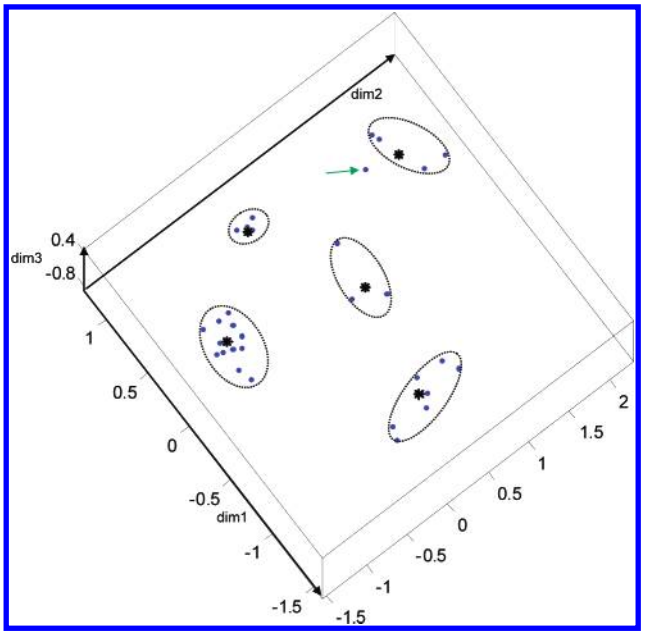

Figure 3. Metrically scaled representation of the experimental NMR conformations of hirudin, using the α-carbons to define the distance matrix. The asterisks mark cluster representative conformations from fuzzy c-means clustering. The dotted lines encircle points with over 50% membership in the given cluster. The arrow identifies the point with a fuzzy membership of 0.37 in the cluster it is closest to. The axes represent the dimensions of the scaled coordinates and are marked as *dim1*, *dim2*, and *dim3*. See text for further details.

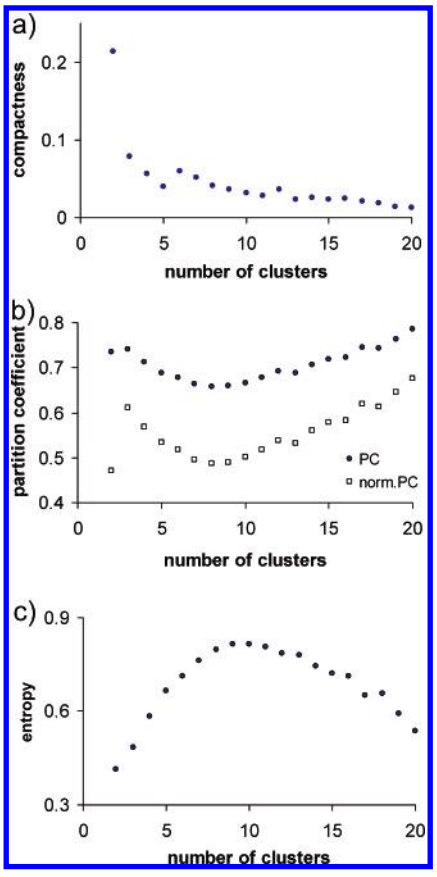

Figure 4. Figures of merit displayed as a function of the number of clusters in the fuzzy c-means clustering of experimental NMR conformations of hirudin (a) compactness (b) partition coefficient and normalized partition coefficient (c) entropy.

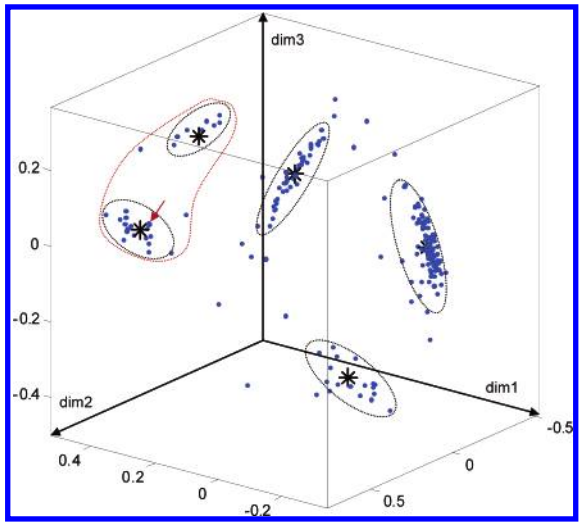

Figure 5. Metrically scaled representation of the conformations from a molecular dynamics simulation of pentane. The asterisks mark cluster representative conformations from fuzzy c-means clustering. The black dotted lines encircle points with over 50% membership in the given cluster. The red dotted line encircles the two clusters that merge when data is clustered into four instead of five clusters, and the red arrow points to the new cluster representative of this merged cluster. See text for further details.

The results obtained using fuzzy clustering were essentially identical to our earlier results with multidimensional scaling.[1] It is also interesting to compare the current results to those obtained by Kelley et al. using the average linkage algorithm.[13] From a penalty function that is the sum of the normalized average spread and the number of clusters, the optimum was found at four clusters.[13] However, without the visualization of cluster distances, this optimum is difficult to verify. Apart from this, the classification of conformations essentially agrees with our results.

Clustering Conformations from Molecular Dynamics Simulations for Pentane. The molecular dynamics (MD) simulation of pentane was investigated using fuzzy clustering, an example that had already been characterized using multidimensional scaling.[1] The conditions of the simulation (300 K temperature, 20 ns simulation time, sampling every 100 ps) were identical to our earlier work[1] and to those used by Torda et al.[14]

From the metric scaling of the distance matrix, the three largest eigenvalues were 0.58, 0.20, and 0.16, and the first three dimensions account for over 94% of the variability in the data. The 3D plot of the clustered conformations is presented in Figure 5. The separation of the data points into five well-separated clusters with a few outlier points is easily visible in this figure. The cluster representatives, selected by fuzzy clustering, are indicated by asterisks. The existence of outliers results from the fact that conformers in molecular dynamics simulation are essentially snapshots along the trajectory of the simulation and are not necessarily close to local minima. Nonetheless, as can also be seen in Figure 5, most simulation time is indeed spent around the minima. When the geometries of the 200 conformers in this work were optimized, they collapsed into only four "real" conformers. These four conformers corresponded to the selected cluster representatives in those four clusters (rmsd 0.4 Å). We also tested how the selection of only four clusters affects the results. In this case, the two clusters in the top left corner of Figure 5 essentially merged as shown by the red dotted line and the cluster representative of this merged cluster remained similar to representative of the bigger one. The other three clusters remained unaffected by this change.

Clustering conformations from molecular dynamics simulations is an example of a situation in which only some points are expected to fall into well-defined clusters. In this case, the outliers are conformations that arise from taking the snapshot at a transition step. These will simply have similar memberships in two or more clusters. In this respect they are not outliers in the fuzzy scheme, although they would be in a crisp clustering scheme. If, however, we would like to turn these fuzzy clusters into crisp ones, each conformer can be assigned to the cluster in which it has over 50% membership. It can be seen from the cluster boundaries in Figure 5 that in this case there are conformations outside these clusters. Assigning conformers to clusters in which they have the highest memberships does not work in this case because it leads to large, elongated, and diffuse clusters.

In comparison to our previous results using multidimensional scaling, cluster representatives in the present study were about 20% closer in their rmsd to the minimized structures. Also, "intermediate" conformations (i.e. those that arose in the simulations when moving from one minimum to another) could be recognized easily from the membership matrix. Our results essentially correspond to the four clusters identified using hierarchical clustering based on the torsional angles of pentane,[14] although a more accurate comparison would require that the identical simulation be clustered in both cases. Again, the hierarchical clustering work would have benefited from a graphical representation of the conformer distribution like the one presented here, as the actual time evolution and its relation to the distribution of conformers could then be easily followed.

The cluster validity measures for this example are displayed in Figure 6. It can be seen that although the compactness measure rapidly declines and practically reaches its asymptote by six clusters, there is no clear-cut minimum at any number of clusters. The other two graphs show that the partition coefficient has a maximum at four clusters. Although this is a reasonable selection, five appear to be better based on visual selection. The difference between the four and five cluster solutions is that the two clusters in the top left corner of Figure 5 (encircled in red) are merged into one, and the cluster representative for this merged group (shown with a red arrow) is shifted toward the center of gravity of the merged cluster. The entropy function has a small inflection at around four and five clusters but no minimum. Therefore in this example only the partition function and its normalized version provide an unambiguous indication as to where the optimum lies, although even in this case visual inspection probably provides a more straightforward solution.

Clustering the Flexible Alignments of Estradiol and Raloxifene. We have previously described the application of multidimensional scaling to the clustering of the flexible alignments of raloxifene and estradiol.[2]

The result of applying metric scaling to the alignment of estradiol and raloxifene is shown in Figure 7. The first three eigenvalues are 0.48, 0.15, and 0.10, and hence three dimensions account for about 73% of the variability in the data. When viewing all alignments without considering their

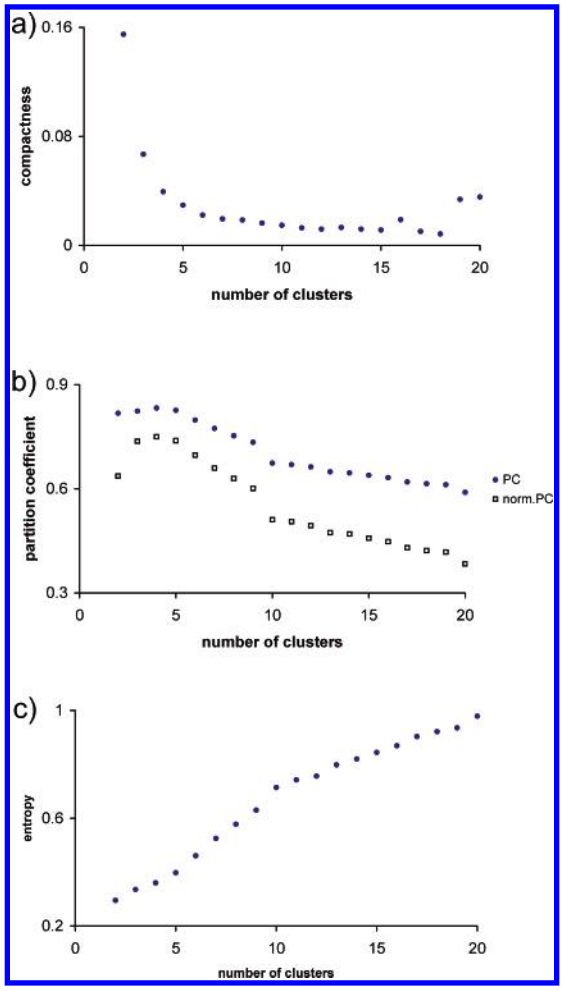

Figure 6. Figures of merit displayed as a function of the number of clusters in the fuzzy c-means clustering of pentane conformations from a molecular dynamics simulation of (a) compactness (b) partition coefficient (c) entropy.

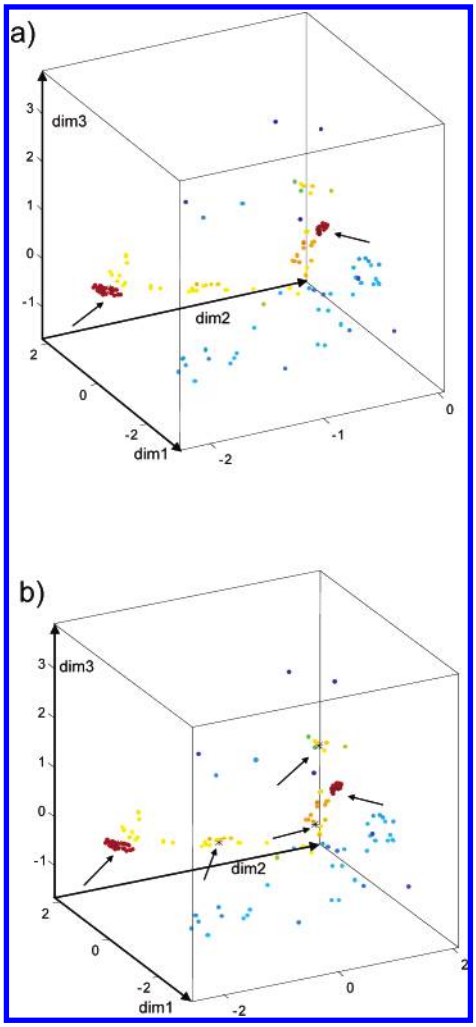

Figure 7. 3D representation of the metrically scaled distance matrix obtained for the flexible overlays of estradiol and raloxifene. The color-coding was obtained by linearly mapping the vector of normalized alignment similarities to the color map. The orange spots correspond approximately to the range 0.65−0.80; the red points are better alignments ($s > 0.95$), whereas the blue and green points are poor alignment solutions ($s < 0.60$). The asterisks and arrows mark the cluster representatives. (a) Two clusters are produced, and these essentially correspond to the best quality alignments (shown as red dots). (b) Five clusters are produced—the obtained clusters correspond to the better half of the alignments (shown as red and orange dots). The axes represent the dimensions of the scaled coordinates and are marked as *dim1*, *dim2*, and *dim3*.

quality, the corresponding data points diffusely fill a large volume and clustering is rather inefficient. However, when the quality of alignments (shown by color-coding) is also considered some patterns can be identified. In particular, if only the best alignments (shown in red) are considered, only two very tight clusters are seen, containing 29 and 21 alignment solutions.

The results of weighted fuzzy c-means clustering on the data are displayed in Figure 7. If the number of clusters is set to 2 as shown in Figure 7a, the method only considers the two highest quality clusters, and the obtained cluster centroids are in the middle of these two clusters. If the number of clusters is increased to 5, the two best groups of alignments are effectively considered by the method, as shown in red and orange in Figure 7b. The cluster centroids of these five clusters are also indicated. As is shown by this example, in weighted FCM clustering the only user decision is the determination of the number of clusters, the quality of alignments is automatically taken into account in this scheme.

Weighting the data for alignment quality has a strong effect on the compactness criterion, as is shown in Figure 8. Without weighting, there is a dip at three clusters. The cluster representatives correspondingly do not reflect the quality of alignment and therefore have little chemical relevance. Using the quality of alignment for weighting, the distribution in Figure 8b is quite different from the nonweighted solution, as there is a major discontinuity in the curve. At the beginning of the curve there is a clear minimum between four and five clusters. At six clusters the curve switches to a much higher compactness value (i.e. less cluster validity) and after which the value decreases smoothly with increasing number of clusters. The cluster representatives for five clusters are shown in Figure 7b, and they correspond to expectations (the clustering process is dominated by the higher quality data points, shown in red and orange). Selecting four instead of five clusters essentially merges the two clusters in the bottom right of Figure 7b, with the cluster representative remaining in the dense region. None of the other cluster validity criteria displayed any features that would provide hints as to the optimum number of clusters and hence are not shown here.

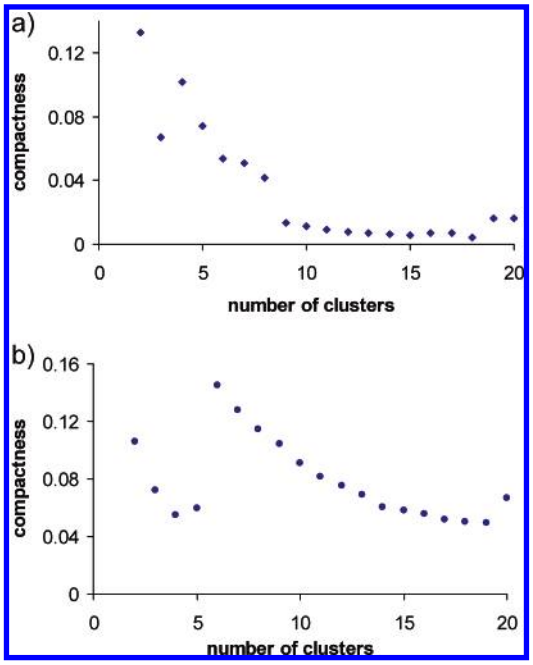

Figure 8. Compactness as a function of the number of clusters from the fuzzy clustering of estradiol-raloxifene alignments (a) c-means clustering with no weighting (b) weighted c-means clustering.

Clustering the Flexible Alignments of Thiorpan and Retrothiorpan. The alignment of these two molecules is described in ref 3, and the distance matrix for fuzzy clustering was taken from this earlier study. As in the previous example, it is important to consider the quality of alignments during the clustering process. Tests have shown that full similarity and pharmacophore similarity perform similarly for this pair of molecules,[2] and the results for the latter are demonstrated here using fuzzy clustering. The results from weighted c-means clustering for setting three clusters is displayed in Figure 9. The best alignments (shown in red and orange colors and corresponding to normalized similarities over 0.9) are concentrated in three areas: in a highly concentrated cluster and two more dispersed ones. The cluster representatives are visibly close to the geometric center of the distribution of these higher quality points.

Turning the fuzzy membership matrix into crisp cluster memberships is unambiguous for the majority of the points. Without imposing weighting in the clustering process, only 10 points out of 250 have less than 50% membership in any one cluster. These cannot easily be assigned to a single cluster. If the points are weighted according to their quality during the clustering process, all but six out of the 250 alignments can be allocated to a single cluster.

Cluster validity functions with and without weighting were calculated for different cluster sizes. Because of the distribution of different quality points in this example, the weighted and nonweighted curves have similar properties. The partition functions displayed in Figure 10 show the optimum at three clusters most clearly. This agrees with the visual inspection of the alignments as described above. This characteristic point can also be observed from the compactness and entropy functions as a slight inflection, although far less readily than from the partition function.

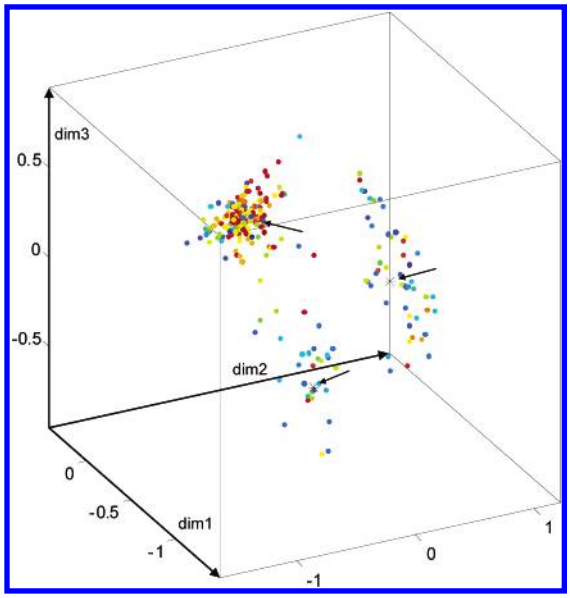

Figure 9. The metrically scaled distance matrix for the alignments of thiorpan and retrothiorpan. The fuzzy c-means clustering was weighted by the pharmacophore similarity of the alignments. The color-coding was obtained by linearly mapping the vector of normalized similarities to the color map and represents the quality of alignments. The orange spots correspond approximately to the range 0.65−0.80; the red points are better alignments ($s > 0.95$), whereas the blue and green points are poor alignment solutions ($s < 0.6$). The asterisks and arrows mark the cluster representatives with the number of clusters specified as three. The axes represent the dimensions of the scaled coordinates and are marked as *dim1*, *dim2*, and *dim3*.

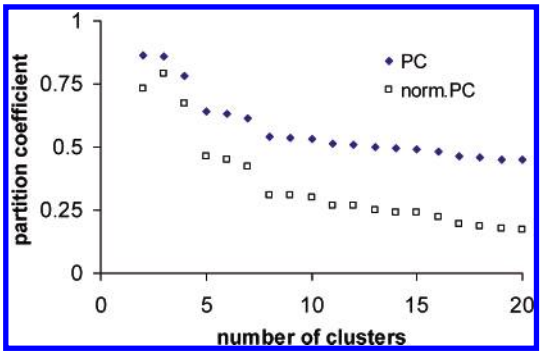

Figure 10. Dependence of the partition function on the number of clusters from the weighted fuzzy c-means clustering of the alignments of thiorpan and retrothiorpan.

CONCLUSIONS

A robust method was developed for the selection of representative conformers or molecular alignments. This method uses data from a previous conformational search or flexible alignment and provides a 3D representation of the relative conformer distances using metric scaling. The rotation of the graphical representation allows observation from any direction and enables visual determination of the optimum number of clusters. Because the distances between points in the metrically scaled distance matrix are proportional to the Euclidean dissimilarity of the conformers, clustering of the metrically scaled matrix should reliably reflect clustering in the original data. The cluster representatives are selected using fuzzy c-means clustering. In the case of alignments, the process can optionally incorporate the quality of alignments using the weighted fuzzy c-means

scheme developed in this work. In fuzzy clustering, information about cluster memberships is contained in the fuzzy membership matrix. Conformers or alignments may belong to more than one cluster with varying degrees of membership. In many cases, the membership matrix can be used to obtain information on individual conformers/alignments or it can be turned into a crisp membership using arbitrary cutoff-criteria.

Cluster validity measures described in the literature were tested in this work. The 3D visualization of the relationship among the identified clusters provides a unique opportunity to judge the performance of these functions. Unfortunately, none of these was found to be consistently predictive, although they performed well in certain examples. The major problem with these functions is that they often display only weak and ambiguous features (such as inflections) in the proximity of the optimum. In conformer selection for 3D QSAR purposes, the selection of representative conformations is often a subjective decision. It is nevertheless the conclusion of the authors that visual inspection of the clustered solutions using a 3D viewer following metric scaling should provide more satisfactory results in determining the optimum number of clusters than any of the calculated validity measures. It must be mentioned that a number of novel algorithms have been suggested recently for dealing with the issue of the optimum number of subclusters, but these were not tested in this work.[15−18]

In comparison with other conformer clustering methods, the obvious advantage of the fuzzy clustering presented in this work is that a visual and undistorted representation of the relative differences among conformers is available. Also, as a result of the fuzzy approach, visual outliers have little impact on the success rate of the clustering process. This was found to be most significant in two areas. When clustering conformations from molecular dynamics, the transition conformations between two equilibrium positions introduce a strong distortion into the clustering process. In clustering alignments using the crisp process,[2] lacking a weighting scheme these had to be manually pregrouped according to their quality. In the fuzzy clustering scheme such conformers/alignments are not outliers, and thus both situations are automatically taken care of. In addition, the automatic selection of cluster representatives in the process is invaluable in drug design applications, such as generating QSAR models or templates for *de novo* design.

REFERENCES AND NOTES

(1) Feher, M.; Schmidt, J. M. Metric and Multidimensional Scaling: Efficient Tools for Clustering Molecular Conformations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 346−353.
(2) Feher, M.; Schmidt, J. M. Identifying potential binding modes and explaining partitioning behavior using flexible alignments and multidimensional scaling. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1065−1083.
(3) Labute, P.; Williams, C.; Feher, M.; Sourial, E.; Schmidt, J. M. Flexible Alignment of Small Molecules. *J. Med. Chem.* **2001**, *44*, 1483−1490.
(4) Bezdek, J. C. Cluster Validity with Fuzzy Sets. *J. Cybernetics* **1974**, *3*, 58−71.
(5) Dunn, J. A fuzzy relative of the ISODATA process and its use in detecting compact, well separated clusters. *J. Cybernetics* **1974**, *3*, 32−57.
(6) Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*; Plenum Press: New York, 1981; pp 43−93.
(7) Krzanowski, W. J. *Principles of Multivariate Analysis*; Clarendon Press: Oxford, 1988.
(8) Xie, X. L.; Beni, G. A Validity Measure for Fuzzy Clustering. *IEEE Trans. Pattern Anal. Mach. Intellig.* **1991**, *13*, 841−847.
(9) Dunn, J. C. Indices of partition fuzziness and detection of clusters in large data sets. In *Fuzzy Automata and Decision Processes*; Elsevier: New York, 1977.
(10) *Molecular Operating Environment*, Version 2000.02; Chemical Computing Group Inc.: Montreal, Quebec, Canada.
(11) *MATLAB*, version 6.1; The MathWorks Inc.: Natick, MA.
(12) Shenkin, P. S.; McDonald, D. Q. Cluster Analysis of Molecular Conformations. *J. Comput. Chem.* **1994**, *15*, 899−916.
(13) Kelley, L. A., Gardner, S. P., Sutcliffe, M. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *J. Prot. Eng.* **1996**, *9*, 1063−1065.
(14) Torda, A. E.; Gunsteren, W. F. Algorithms for Clustering Molecular Dynamics Configurations. *J. Comput. Chem.* **1994**, *15*, 1331−1340.
(15) Tung, W. L.; Quek, C. DIC: A novel discrete incremental clustering technique for the derivation of fuzzy membership functions. *Lect. Notes Comput. Sci.* **2002**, *2417*, 178−188.
(16) Devillez, A.; Billaudel, P.; Lecolier, G. V. A fuzzy hybrid hierarchical clustering method with a new criterion able to find the optimal partition. *Fuzzy Sets Syst.* **2002**, *128*, 323−338.
(17) Turhan, M. Genetic Fuzzy Clustering by Means of Discovering Membership Functions. *Lect. Notes Comput. Sci.* **1997**, *1280*, 383−393.
(18) Nascimento, S.; Moura-Pires, F. A Genetic Approach to Fuzzy Clustering with a Validity Measure Fitness Function. *Lect. Notes Comput. Sci.* **1997**, *1280*, 325−337.

CI0200671