Clustering Peptide Structures through Identification of Commonly Exposed Groups

Thy-Hou Lin,*,[†] Jia-Jiunn Lin,[†] Yung-Feng Huang,[†] and Jin-Hwang Liu[‡]

Department of Life Science, National Tsing Hua University, Hsinchu, Taiwan, ROC, and Department of Internal Medicine, Veterans General Hospital and National Yang-Ming University, Taipei, Taiwan, ROC

Received November 24, 1998

A clustering analysis method using the number of commonly exposed groups identified as a clustering criterion for a group of peptide structures generated from an in vacuo molecular dynamics simulation is presented. The number of commonly exposed groups is identified as the number of atoms of the same type which appear on vertices of groups of three dimensional convex hulls computed for groups of structures sampled and collected as blocks. Blocks of structures of high structural similarity are classified as clusters if their corresponding number of commonly exposed groups identified are larger than a preset criterion. Linkages between blocks are provided with the generation of blocks consisting of overlapping structures. However, the linkage can be eliminated by employing a minimal distance criterion for each block generated. The feasibility of this proposed clustering method is tested through a comparison of results obtained from a conventional and a hierarchical clustering method. Since change in fine structural features can be detected as the change in the number of commonly exposed groups identified, we find that the method is superior to the conventional clustering one in partitioning compact and well-separated clusters.

INTRODUCTION

Clustering analysis (CA) is now widely used as a tool for identifying the structural similarity or commonality between different conformers of molecules so that a representative conformer can be extracted to represent each cluster.[1] CA methods are conventionally classified as hierarchical or nonhierarchical ones. In hierarchical methods, one either takes all objects together as one cluster to produce succeeding clusters by dividing some or all of those so far produced or treats *n* objects as *n* separated clusters at the beginning and then joins them into clusters step by step. A hierarchy of partition thus produced is usually represented by a dendrogram. The nonhierarchical methods are classified as those based on the estimation of density or distance between a given number of objects or those based on the graph theory. A set of target criteria is used by these methods to attain a single optimal partition of objects into clusters.

Algorithms designed for clustering molecular conformers often require the computation of a similarity matrix in which each element represents the structural difference between a pair of structures. Levitt[2] calculated such matrices and their projections into two-dimensional Cartesian cdordinates. These projections were then used to sketch out the path of a trajectory and suggest the presence of clusters in conformational space. Adzhubei et al.[3] performed a pairwise superposition for a series of NMR-derived structures using Cα atoms to generate a set of root mean square (rms) distances. After CA based on these distances, a cutoff was used to determine the final membership of clusters and therefore the representative structures. Rooman et al.[4] clustered short peptide fragments into a hierarchical scheme based on the sum of the squared distances of individual elements to the center of mass of clusters. The method was used to identify recurring motifs in broad conformational classes. A novel CA method was proposed by Gordon and Somorjai[5] to group parathyroid hormone fragment conformations into a predetermined number of clusters. Since clusters can be assigned with fuzzy membership, this method does not require one to concede the membership of a cluster while reducing a mass of structures down to a manageable number of representative cluster centers. A CA method using the rms differences of dihedral angles rather than of coordinates as the clustering criterion was proposed by Karpen et al.[6] for clustering 15 000 configurations from a trajectory of a pentapeptide. However, criticism[7] has been drawn to the method due to the fact that structural similarity is not necessarily well-correlated with the difference in internal angles. A clustering scheme in which multiple thresholds were assigned to different clustering criteria which include both dihedral angles and interatomic distances was reported by Bravi el al.[8] This method has the advantage that higher weights can be assigned to the central dihedral angles than those applied to the terminal ones to avoid the leverage effects caused by the structural flexibility.

In general, factors that affect a CA result can be summarized as[8,9] (i) the nature of data sets, (ii) the clustering criteria chosen for comparing objects, and (iii) the mathematical strategy used for grouping objects. The input parameters for obtaining an optimal partition such as the choice of a threshold or the number of clusters for most of the CA methods are rarely known a priori. It is often required to perform a series of trials before an optimal partition can be obtained. To obtain a reliable partition, one also needs to compare all clusters obtained at various conditions or use more than one clustering criterion.[9] In this report, we present a clustering method based on the identification of structural

* Corresponding author. Phone: (886) 03-574-2759. Fax: (886) 03-572-1746. E-mail address: lslth@life.nthu.edu.tw.

[†] National Tsing Hua University.

[‡] Veterans General Hospital and National Yang-Ming University.

10.1021/ci9801623 CCC: $18.00
Published on Web 04/17/1999

convexity for 300 structures of a flexible peptide generated from molecular dynamics. We use a three-dimensional (3D) convex hull computation method published previously[10] to identify the exposed functional groups on the vertices of a convex hull for each structure generated. Since the degree of structural similarity between a series of structures of the same structural topology can be correlated with the identifiable number of commonly exposed groups, we partition the generated structures into different blocks corresponding to different numbers of commonly exposed groups. To measure the compactness of each block partitioned, we compare the averaged distance between the coordinates of a structure and the centroid of a given block with similar distances of all the other blocks partitioned. A block is considered to be compact and accepted for further comparison if none of the former distances calculated is greater than the latter ones for each structure allocated in the block. Blocks of structures of high similarity can be repeatedly classified as clusters by considering a different number of commonly exposed groups as the identifying criterion. The feasibility of our clustering method is tested through comparing our results with those obtained from a conventional clustering method based on the partition of distances, namely, JOINER,[11] and from a hierarchical clustering method implemented in the SYBYL 6.4 package.[12]

MATERIALS AND METHODS

Our convex hull calculation and cluster analysis procedure is based on the structures of a tachykinin peptide generated by the AMBER 4.1 program.[13] The amino acid sequence of the tachykinin peptide studied, namely, substance P,[14] is as follows:

$$\mathrm{Arg}^{1}\text{-}\mathrm{Pro}^{2}\text{-}\mathrm{Lys}^{3}\text{-}\mathrm{Pro}^{4}\text{-}\mathrm{Gln}^{5}\text{-}\mathrm{Gln}^{6}\text{-}\mathrm{Phe}^{7}\text{-}\mathrm{Phe}^{8}\text{-}\mathrm{Gly}^{9}\text{-}\mathrm{Leu}^{10}\text{-}\mathrm{Met\text{-}NH_2}^{11}$$

Using standard AMBER Link-Edit-Parm operation procedures, we generated a random structure composed of united atoms for the peptide. The Prep module was employed to make the special amino acid residue Met-NH_2. A 8.0 Å nonbonded cutoff was used, and bonds involving hydrogen atoms were fixed at their equilibrium values employing the shake algorithm.[15] The structure was equilibrated for about 100 ps in an in vacuo MD simulation run, and then a total of 300 structures were collected using a distance dependent dielectric constant and a time period of about 300 ps. Extensive energy minimization was performed for each collected structure.

The algorithm used for computation of a convex hull for each collected structure has been described previously.[10] Briefly, we treated each atom in a structure as a point in space. A triangulation procedure was used next to form a series of triangular facets for the point set. For each triangular facet generated, a series of parallelepipeds were further formed by the facet and a fourth point arbitrarily selected from the rest of the points. The volume of each parallelepiped formed can be negative or positive depending on the angle between the normal of the facet and a vector pointing from one apex of the facet to the fourth point. By sorting all such volumes, a convex hull facet was identified as the one with all such calculated volumes as either negative or positive. The vertices lying on each convex hull facet were identified and collected for further clustering analysis. To utilize these vertices to form a clustering criterion, we first generated a large number of blocks of structures. The number and content of structures in each block were varied from block to block. The minimum number of structures allowed in a block was set at 5. Further, the structure number in each block was arranged in the same numerical order as that collected from the MD run. For all the structures kept in a given block, we compared the atomic numbers of those atoms identified to be convex hull vertices. The aim was to count the number of commonly exposed groups that appeared on the series of convex hulls computed for the block. Blocks were selected for further comparison if the number of commonly exposed groups counted for them were greater than a cutoff. The centroid of each selected block was calculated, and the rms dfference between coordinates of a structure in the block with its centroid was compared with the rms difference between the same structure's coordinates and the centroids of all the other selected blocks. Blocks with the latter distances that were greater than the former were kept for further comparison. This is the so called minimal distance criterion[16] in the clustering literature. Both the block spread and the number of selected blocks were reduced substantially by using this distance criterion. The rms difference in coordinates we computed was actually the averaged Euclidean distance between the coordinates of a pair of similar atoms appearing on a pair of structures. The block spread for each selected block was calculated as the averaged rms difference in coordinates between every pair of structures kept in the block.

The performance of the entire scheme depends on the selection of two cutoffs, namely, the minimum number of structures allowed in a block and the number of common exposed groups identified. To identify clusters of structures that have greater structural similarity, we gradually increased the minimum number of structures allowed in a block from 5 to about 14 while setting the cutoff for the number of common exposed groups at 10. By comparing the content of each block searched and the corresponding calculated averaged block spread, we found that clusters of structures of greater structural similarity were repeatedly identified as blocks of various sizes. We defined the boundary for each cluster identified as the boundary of the block of the largest size, e.g., the block with the most populous members for that cluster. The identification process for several clusters from corresponding blocks generated is schematically demonstrated in Figure 5. A conventional clustering method based on searching the minimum of the sum of the squared Euclidean distances of the cluster members from their centroids, namely, the JOINER program,[11] was also employed for clustering all the structures collected. To perform further comparison, we divided the collected structures into two groups (1−155 and 156−300) and used a hierarchical clustering method implemented in the SYBYL 6.4 package[12] to cluster each of them.

RESULTS AND DISCUSSION

A convex hull computed for a complicated molecule can distinguish atoms inside the hull from the ones on the vertices.[10,17] Only the latter ones have a greater possibility

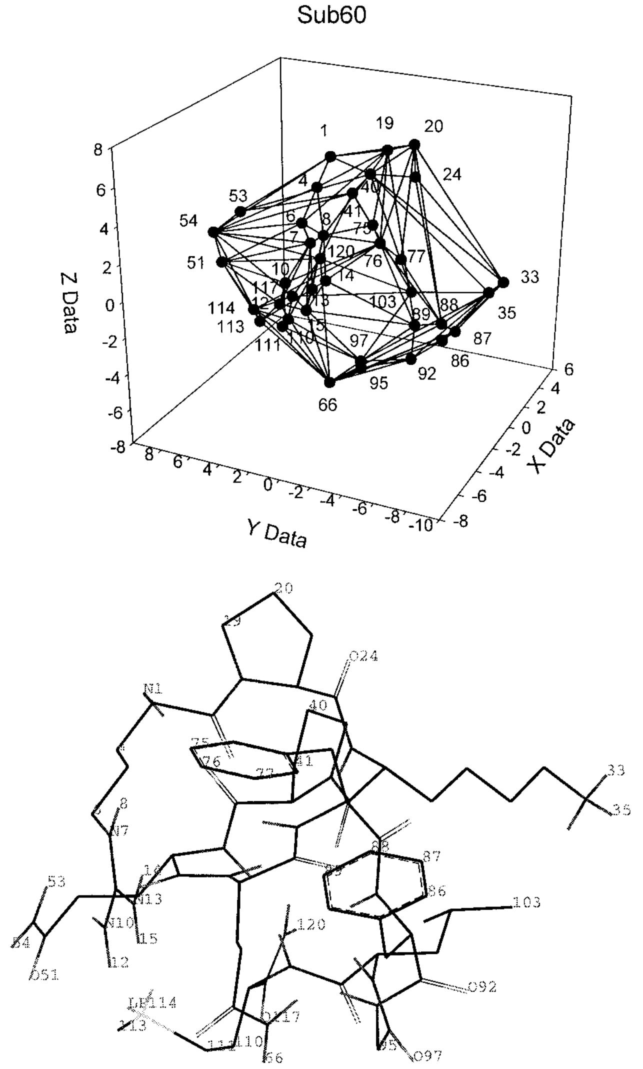


Figure 1. (a) 3D convex hull computed for a structure of peptide substance P collected from an in vacuo MD simulation run. The structure was extensively energy minimized. Atoms identified as convex hull vertices are marked as filled circles, and the corresponding atomic numbers are also marked alongside. The total number of vertices identified for the convex hull is 38. (b) Peptide structure that is used for computation of the 3D convex hull displayed in a. Atoms identified as convex hull vertices for the structure are marked with the corresponding atomic numbers alongside.

to interact with the receptor atoms.[10] By treating a convex hull as a set of correspondences between matching points, it is possible to utilize it as a tool to preferentially screen conformational space for aligning structures generated for rather large and flexible molecules such as peptides.[18] In principle, structures with exact similarity could give rise to exactly similar convex hulls computed. Therefore, the number of exposed groups that are common between several convex hulls computed for several structures can be used as a guide to measure the structural similarity between them. A convex hull computed for a collected structure is presented in Figure 1a. The number of vertices identified on this convex hull is 38. Atoms identified as vertices are labeled, and the corresponding peptide structure from which the convex hull is computed is depicted in Figure 1b. It is feasible to treat a set of commonly exposed atoms identified from computation of convex hulls as a set of correspondences for aligning structures and then derive a reasonable 3D quantitative structure−activity relationship (QSAR).[18] Since the commonly exposed atoms exhibit globular distribution on a series of dissimilar structures compared, the reliability of these as a correspondence criterion for similar structures and the fitness of the structures for most of the side chains are better than that obtained from the conventional superposition using the backbone atoms as a set of correspondences. To highlight this point, we compare in Figure 2a,b a group of structures superposed by using a set of commonly exposed atoms identified from computed convex hulls or by fitting them using the coordinates of the backbone atoms as correspondences, respectively. The set of commonly exposed atoms identified and used as the set of correspondences for fitting the group of structures are 2, 4, 11, 12, 15, 20, 35, 40, 51, 54, 66, 76, 84, 85, 86, 87, 88, 89, 90, and 104. The locations of these atoms are on backbone (2 and 4) and side chain (11, 12, and 15) of Arg^1, backbones (20 and 40) of Pro^2 and Pro^4, and side chains (35, 51, 54, 66, 76, 84−90, and 104) of Lys^3, Gln^5, Gln^6, Phe^7, Phe^8, and Leu^{10}.

To find the number of commonly exposed groups or the number of similar vertices appearing on a series of computed convex hulls, we count the frequency of each atom, identified to be a vertex on each convex hull, together and keep only those that are greater than a cutoff. Note that most of these commonly exposed groups are not necessarily polar ones. To utilize these commonly exposed groups as a clustering criterion, we randomly select structures from the collection of 300 structures to form a large number of blocks of structures. Structures kept in each block are arranged in the same numerical order as they are collected from the MD run. There is some overlapping of structures kept in some blocks, but no redundancy is allowed. Further, unless the number of structures selected for a block is smaller than a cutoff, the number of structures kept in blocks can be varied as large as up to 100. The number of commonly exposed groups are then searched for and compared for all the blocks generated. At this stage, a further cutoff is applied to choose only blocks that have a larger number of commonly exposed groups or greater structural similarity. Table 1 presents the distribution of the number of commonly exposed groups over the averaged block spread which was calculated for a series of blocks generated using a cutoff of 5 for the block size. These data show that the averaged block spread calculated decreases with the increase in the number of commonly exposed groups. The degree of structural similarity for structures kept in blocks with a smaller number of commonly exposed groups is in general worse than that kept in blocks with a larger number of commonly exposed groups. However, there are some blocks with large block spreads that also happen to have a larger number of commonly exposed groups. To eliminate those blocks with some diverse structures, we calculate and compare the rms difference in coordinates between a member of a block with the centroid of the block and with those of all the other blocks. Figure 3 presents the fraction of blocks that are kept after the elimination using the minimal distance criterion[16] against the number of commonly exposed groups identified for blocks with three numbers of minimum members allowed. These data indicate that only a small fraction of blocks meets the minimal distance criterion[16] when the corresponding number of commonly exposed groups identified is smaller than a criterion around 16. On the other hand, most of the blocks

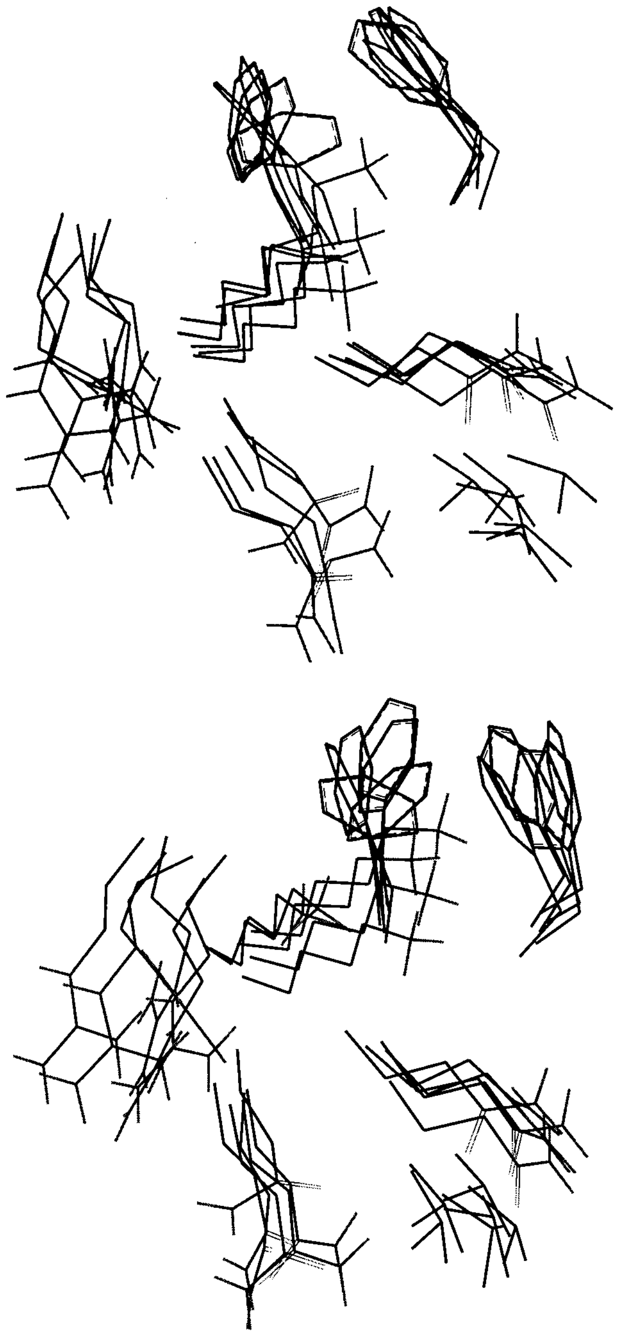


Figure 2. (a) Superposition of a group of six structures collected from the MD simulation run using the SYBYL 6.4 program. The commonly exposed groups identified, namely, 2, 4, 11, 12, 15, 20, 35, 40, 51, 54, 66, 76, 84, 85, 86, 87, 88, 89, 90, and 104, are used as a set of correspondences in the superposition process. The rms value of the fit is 1.28. Side chain atoms selected for the display are 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 (from Arg[1]); 27, 28, 29, 30, 31, 32, 33, 34, 35 (from Lys[3]); 47, 48, 49, 50, 51, 52, 53, 54 (from Gln[5]); 59, 60, 61, 62, 63, 64, 65, 66 (from Gln[6]); 71, 72, 73, 74, 75, 76, 77, 78 (from Phe[7]); 83, 84, 85, 86, 87, 88, 89, 90 (from Phe[8]); and 101, 102, 103, 104 (from Leu[10]); respectively. (b) Same group of structures displayed in a superposed using the coordinates of backbone atoms as the set of correspondences by the SYBYL 6.4 program. The rms value of the fit is 1.11. Side chain atoms selected for display are those described in a.

with a large number of commonly exposed groups meet the criterion.

Table 1. Distribution of the Number of Commonly Exposed Groups Against the Averaged Block Spread Calculated for Each Block Using the Minimum Number of Structures Allowed in a Block To Be 5

	10[a]	11	12	13	14	15	16	17	18	19	20	21	22
1.4[b]	0	1	0	0	1	1	1	0	1	1	0	0	0
1.5	2	0	0	3	3	3	3	0	2	1	1	0	0
1.6	0	1	3	6	1	1	3	2	1	0	1	0	2
1.7	2	2	5	7	3	6	7	4	3	1	2	0	0
1.8	5	3	6	10	4	8	5	4	1	0	0	0	0
1.9	3	14	9	10	9	11	6	0	4	2	0	0	0
2.0	12	8	7	10	9	6	2	1	4	0	0	0	0
2.1	5	9	9	8	9	2	3	3	2	1	2	0	0
2.2	11	5	15	10	8	3	1	1	3	0	0	0	0
2.3	11	14	7	7	6	5	2	2	1	1	1	0	0
2.4	10	8	9	8	2	4	3	3	1	1	0	0	0
2.5	6	3	9	6	4	4	3	0	0	0	0	0	0
2.6	8	11	6	4	5	1	1	1	0	0	0	0	0
2.7	8	12	7	4	4	1	1	0	0	0	0	0	0
2.8	8	5	2	4	0	1	0	0	0	0	0	0	0
2.9	4	1	0	2	1	1	1	0	0	0	0	0	0
3.0	3	3	1	0	0	0	0	1	0	1	0	1	0
3.1	6	2	1	1	0	0	0	1	0	0	1	0	0
3.2	0	0	0	0	1	0	0	0	0	0	0	0	0
3.3	1	1	0	0	0	1	0	0	0	0	0	0	0
3.4	0	0	1	0	0	0	0	0	1	0	0	0	0
3.5	0	0	0	1	0	0	0	0	0	0	0	0	0
3.6	0	0	0	0	0	0	1	0	0	0	0	0	0
3.7	1	0	1	0	0	0	0	0	0	0	0	0	0

[a] The number of commonly exposed groups. [b] The averaged block spread calculated is in units of angstroms.

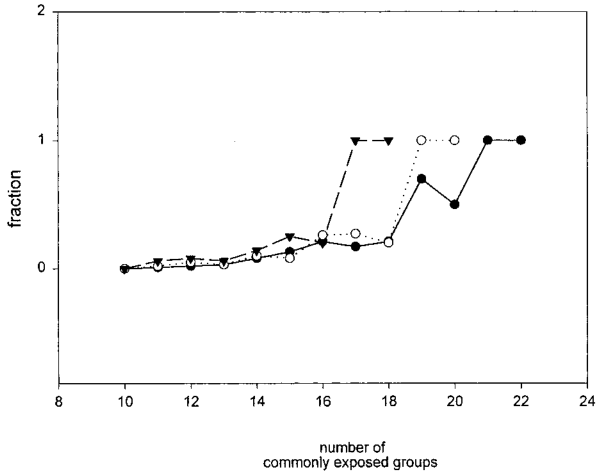


Figure 3. Fraction of blocks that meet with the minimal distance criterion for blocks generated using three different minimum numbers of structures allowed for each as a function of the number of commonly exposed groups. Filled circles, open circles, and filled triangles are respectively for blocks generated using 5, 6, and 7 as the minimum number of structures allowed for the generation of a block.

To obtain a reliable clustering result using the conventional clustering method with a single clustering criterion, one needs to perform the clustering process at various conditions before a final decision can be made.[8,9] However, linking clusters using only distance as a criterion could be a dangerous try. Distances used for measuring structural similarity between complicated molecular structures are usually averaged, and they are not precise enough to distinguish some fine structural differences. For example, we present in Table 2 a comparison of the calculated averaged block spread and the corresponding sets of commonly exposed groups for three different blocks. While the number of commonly exposed groups for these three blocks vary substantially from 11 (block-11) to 15 (block-15) or to 19 (block-19), the averaged block spreads calculated for them only varied from 1.44 to 1.48 Å. The

Table 2. Comparison of the Block Spread and the Number of Commonly Exposed Groups for Three Blocks

no. of commonly exposed groups	avd block spread (Å)	structure members	atomic no. of each set of commonly exposed groups
11	1.44[a] (0.79[b])	102−106	20, 31, 34, 40, 63, 75, 76, 88, 89, 113, 120
15	1.45 (0.77)	57−61	15, 34, 35, 53, 54, 76, 88, 89, 97, 104, 106, 110, 111, 113, 120
19	1.48 (0.93)	159−163	2, 11, 12, 15, 20, 35, 40, 51, 54, 66, 76, 87, 88, 89, 94, 95, 97, 104, 113

[a] The averaged block spread calculated for a block. [b] The coordinates of structures in each block have been fitted using the coordinates of the backbone atoms as a correspondence set through the SYBYL 6.4 program before the block spread is calculated.

magnitudes of these spreads are somewhat reduced when SYBYL 6.4[12] is used to superpose the structures kept in each block using the coordinates of the backbone atoms as a set of correspondences. As shown in the table, the variation in the spreads after superposition is not changed significantly. This is in contrast to the set of corresponding commonly exposed groups listed in the same table. Larger differences in these commonly exposed groups reveal that differences in structural features for each block of structures are actually larger than those one can expect from only comparing the distances. If we take the first structure kept in block-11 as a target and then use all the structures kept in block-15 to superpose on it through the SYBYL 6.4 program,[12] the rms difference in coordinates obtained is 1.23, which is much larger than those listed in Table 2 for each of these two original blocks through the same mean. The superposed backbone structures and some superposed and selected side chain groups of the two selected structures are presented in Figure 4a,b, respectively. The superposed structures selected are the target from block-11 and the first structure selected from block-15. These two plots show that while backbone atoms fit well to an extent, most of the side chains do not. Therefore, this further indicates that fine differences in structural features can be detected by the number of commonly exposed groups.

By gradually increasing the minimum number of structures allowed for a block or by increasing the block size using the minimal distance criterion, we find that clusters of structures can be repeatedly detected by blocks of various sizes (Figure 5). In fact, clusters of structures detected by several smaller blocks can be linked together by larger blocks (Figure 5). Table 3 presents the identification of a cluster by comparing the members of structures generated from several blocks. Since structures are kept in numerical order in each block, we show only the boundary numbers of structure members kept in each block. These data show that a cluster of structures approximately from 157 to 172 is detected by 14 blocks with a different number of commonly exposed groups. However, we set the boundary for the cluster detected to be from 159 to 172, which is the boundary of the largest block generated for the region. These data also show that smaller clusters detected by smaller blocks can be linked together by the larger blocks generated.

A complete listing of all the clusters detected for the 300 structures collected are presented in Table 4. The total number of clusters detected is 26. The population of these clusters detected is varied from 5 to 14 members. Most of these clusters detected are well-separated compact clusters. Several energetically favored clusters, namely, b2, c3, c4, and d4, are detected. Clusters of structures of higher energies, namely, a2, c6, d2, and d3, are also detected. It also appears that structures collected much later in the simulation time are not clustered with those collected in the early time. This implies that no recurring conformations are clustered during the period of the MD simulation run. A comparison of our clustering result with those obtained using a distance partition algorithm, the JOINER program,[11] and a hierarchical clustering method implemented in the SYBYL 6.4 package[12] is presented in Figure 6. Structures for clustering using the JOINER program[11] are not superposed in advance, while those for clustering using the SYBYL 6.4 package[12] are done so using the coordinates of backbone atoms as a set of correspondences. Structures for clustering by the SYBYL 6.4 program[12] are divided into two groups, namely, from 1 to 155 and from 156 to 300, respectively. In this figure,

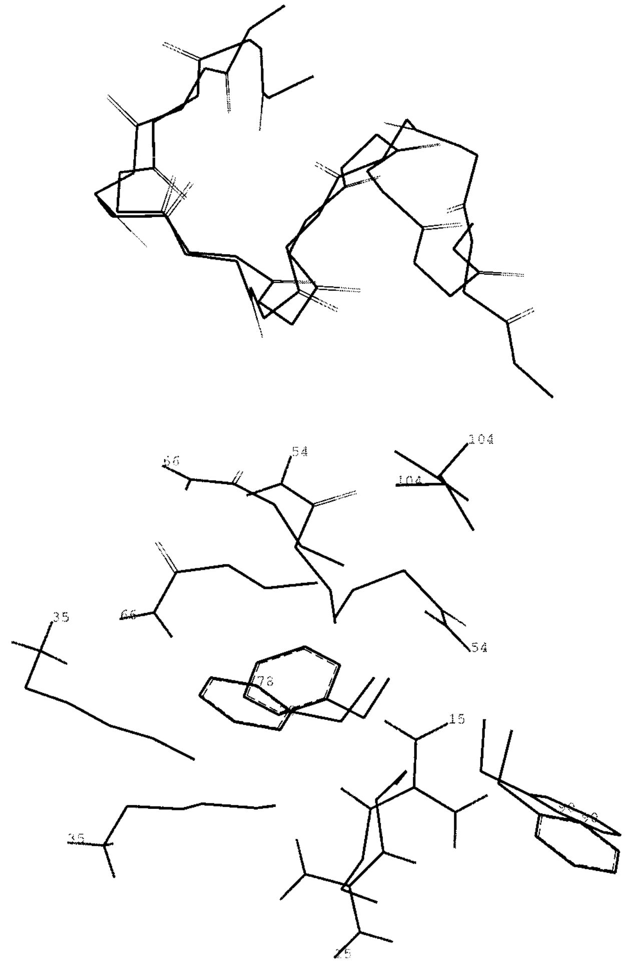


Figure 4. (a) Superposition of the first structure taken from block-15 (see discussion in the text) against the first structure taken from block-11 using the coordinates of the backbone atoms as a set of correspondences and the SYBYL 6.4 program. The rms value of the fit is 1.37. Only the superposed backbone atoms from both structures are displayed. (b) Superposition of the first structure taken from block-15 (see discussion in the text) against the first structure taken from block-11 using the coordinates of the backbone atoms as a set of correspondences and the SYBYL 6.4 program. The rms value of the fit is 1.37. Only some side chain atoms are selected for display (see legend of Figure 2a for the set of side chain atoms selected). To distinguish these two superposed structures, some atoms selected for display are marked with their corresponding atomic numbers alongside.

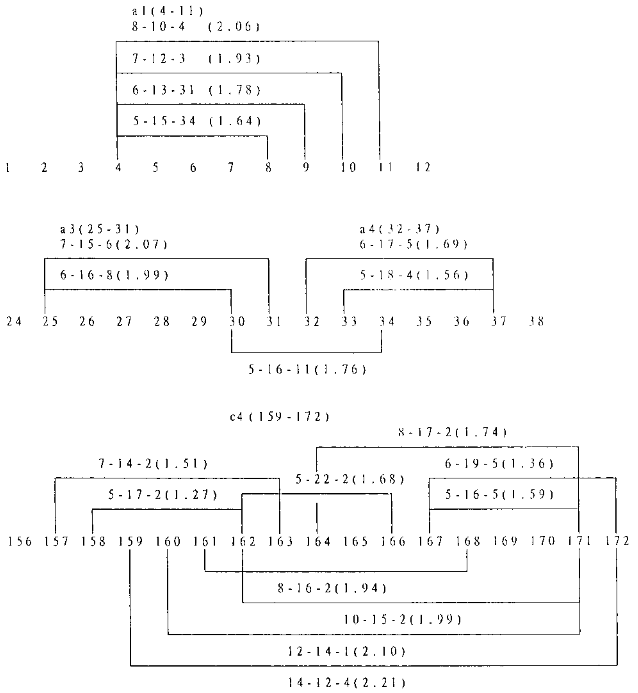
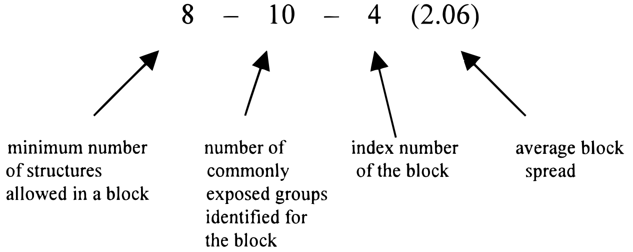


Figure 5. Identification of clusters a1, a3, a4, and c4 from their corresponding blocks searched. Structures are numerically ordered in the scheme. Each block searched is characterized by three digits of numbers followed by the averaged block spread in angstroms calculated for it as follows, for example:


8 – 10 – 4 (2.06)
minimum number of structures allowed in a block
number of commonly exposed groups identified for the block
index number of the block
average block spread

The boundary of each block is defined as the boundary of the block of the largest size for the same group of structures searched. A larger cluster such as c4 shown in the scheme is formed through extensive linking between blocks.

Table 3. Identification of Cluster Members for Cluster c4 (159−172)

minimum no. of structures allowed	no. of commonly exposed groups identified	structure member in the blocks	avd block spread (Å)
5	17	158−162	1.27
5	22	161−166	1.68
5	16	167−172	1.59
6	20	161−166	1.73
6	16	167−172	1.59
7	14	157−163	1.51
7	18	162−168	1.89
7	17	164−171	1.74
8	13	157−164	1.61
8	16	161−168	1.94
8	17	162−171	1.74
10	15	162−171	1.99
12	14	160−171	2.10
13	12	159−172	2.21

clusters obtained by our method are underlined and marked with the name of a cluster, while those obtained by the JOINER[11] or by the SYBYL 6.4[12] programs are represented by an upper and a lower series of numerical numbers, respectively. Apparently, there are eight clusters, namely, a1, a4, b2, b4, c6, c7, c8, and d2, which are unanimously detected by all three methods used. Although there are slight variations in the boundary of the clusters detected, clusters a3, b5, b6, c3, c4, d5, and d6 detected are agreeable with those detected by the JOINER program,[11] while clusters b3, c2, c5, d1, and d3 detected are in accord with those detected by the SYBYL 6.4 package.[12] However, there are more links between clusters detected by the SYBYL 6.4 package[12] since structures are superposed before the SYBYL[12] partition. For example, the following groups of clusters are completely or partly linked by the SYBYL program,[12] (1) a2, a3, and a4; (2) b1 and b2; (3) b7 and c1; (4) c7, c8, d1, d2, d3, d4, and d5; and (5) d7 and c4. However, it is found that there are at least two faulty links given by the JOINER program[11] between two groups of clusters, namely, b5 and b6 or c4 and c5, as compared with those obtained from the SYBYL package.[12] These comparisons clearly indicate that faulty links between clusters can be avoided by our partition even when structures subjected for clustering are not superposed in advance. The SYBYL[12] hierarchical CA attempts to find groupings within a set of data. At the beginning of the analysis, each row of the data table is a cluster. Using the method of complete linkage, the nearest pair of clusters is merged and then the next nearest and so forth until there is only one cluster containing all the rows. Since structures are divided into two groups for our analyses using the SYBYL program,[12] there are several cluster levels generated with outstanding relative distances calculated in between for each of them during the analyses. At this moment, our convex hull clustering result is used as a guide to pick out some meaningful results for comparison using the SYBYL program.[12]

Table 4. Cluster Members and Mean, Maximum, and Minimum Conformational Energy for All Clusters

cluster	members of cluster	E_{mean} (kacl/mol)	E_{max} (kcal/mol)	E_{min} (kcal/mol)
a1	4−11	−10.1	−4.4	−14.5
a2	20−24	−6.6	−5.1	−8.5
a3	25−31	−10.7	−8.1	−13.6
a4	32−37	−10.5	−5.8	−13.4
b1	39−50	−10.0	−4.9	−15.3
b2	59−70	−14.5	−9.3	−17.8
b3	74−78	−8.2	−2.8	−18.2
b4	82−87	−9.5	−7.7	−14.2
b5	92−101	−8.6	−2.0	−12.9
b6	103−109	−7.7	−4.3	−11.4
b7	111−120	−10.1	−7.8	−14.9
c1	122−128	−10.3	−5.5	−15.0
c2	131−138	−8.4	−5.5	−10.2
c3	139−151	−12.6	−2.1	−17.6
c4	159−172	−15.9	−11.9	−20.4
c5	177−183	−11.4	−7.5	−15.4
c6	191−195	−5.3	−2.9	−8.6
c7	199−205	−9.5	−6.8	−15.5
c8	206−211	−10.6	−6.7	−16.8
d1	212−220	−10.4	−4.0	−16.8
d2	231−237	−6.2	−1.3	−10.0
d3	238−246	−6.5	−0.8	−13.5
d4	249−256	−12.2	−4.6	−16.5
d5	266−274	−8.7	−4.8	−11.6
d6	277−281	−8.5	−4.1	−11.9
d7	288−297	−8.2	−2.1	−18.1

Recently, the rms differences in dihedral angles has been used as a clustering criterion for clustering structures of short peptides.[6] The reincarnation of this approach is that proposed

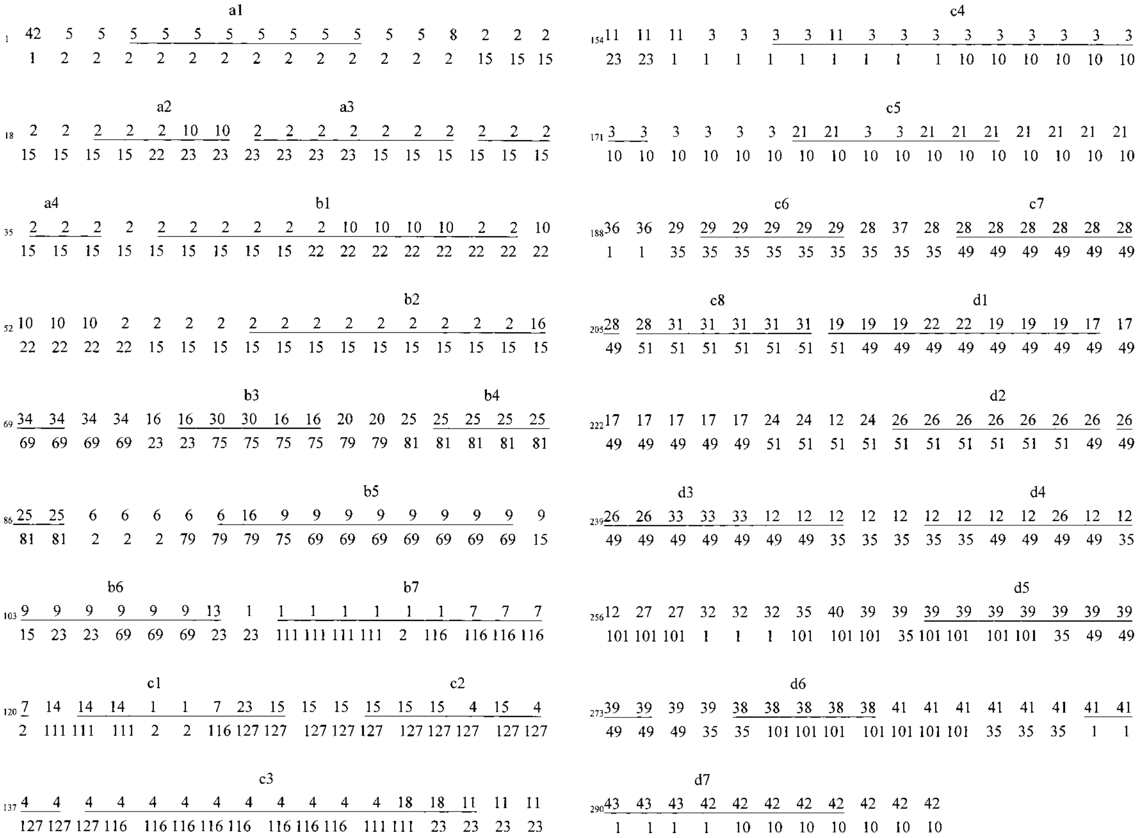


Figure 6. Comparison of the clustering result for the 300 MD generated structures using (i) the number of commonly exposed groups as a clustering criterion, (ii) the JOINER program,[11] and (iii) a hierarchical clustering program implemented in the SYBYL 6.4 package.[12] Clusters obtained by our convex hull method are underlined and marked with the name of a cluster on it while those obtained by the JOINER[11] or by the SYBYL 6.4 program[12] are represented as an upper or a lower series of numerical numbers, respectively. Note that structures for clustering by the SYBYL 6.4 program[12] have been divided into two groups, namely, 1−155 and 156−300.

Table 5. Response of the Convex Hull Vertices to a Gradual Change in the ψ Dihedral Angle of a Structure Collected from the MD Simulation Run

ψ (deg)	convex hull vertices identified	total no. of changes in vertices
19,86,19,114,114, 126,126,19,163, 94	1,2,3,8,11,12,14,15,19,20,21,23,27,29, 31,34,35,39,40,41,54,66,75,76,87, 88,89,103,104,113,114,115,119,120	
49,86,**49**,114,114, 126,126,19,163, 94	1,2,3,8,11,12,14,15,19,20,21,23,31,34, 35,39,40,41,54,75,76,87,88,89,103, 104,113,114,115,119,120	3
19,**116**,19,114,114, 126,126,19,163, 94	1,2,3,8,11,12,14,15,19,20,21,22,23,24, 27,29,31,34,35,39,40,41,54,75,76,87, 88,89,103,104,113,114,115,119,120	4
19,86,19,**144**,**144**, 126,126,19,163, 94	1,2,3,8,9,10,11,12,14,15,19,20,21,23, 27,29,31,34,35,66,75,76,87,88,89, 103,104,113,114,115,119,120	5
19,86,19,114,114, **156**,**156**,19,163, 94	1,2,3,8,9,10,11,12,14,15,19,20,21,23, 27,29,31,34,35,54,66,75,87,88,89, 103,104,113,114,115,119,120	6
19,86,19,114,114, 126,126,**49**,163, 94	1,2,3,7,8,11,12,14,15,19,20,21,23,27, 29,31,34,35,39,40,41,54,66,75,87, 88,89,103,104,113,114,115,119,120	2
19,86,19,114,114, 126,126,19,−**167**, 94	1,2,3,8,11,12,14,15,19,20,21,23,27,29, 31,34,35,39,40,41,54,75,76,87,88, 89,103,104,113,115,119,120	2
19,86,19,114,114, 126,126,19,163, **124**	1,2,3,8,11,12,14,15,19,20,21,23,27,29, 31,34,35,39,40,41,54,66,75,76,87, 88,89,104,113,114,115,119,120	1

by Bravi et al.[8] using multiple thresholds of the dihedral angles for clustering several peptide structures. Since different thresholds or different weights can be used to assess the difference in structural features in some particular regions of structures, this method is deemed to be more effective than the simple rms one in classifying flexible structures. To examine whether the change of the convex hull vertices computed is sensitive to a small change in the backbone dihedral angles, we gradually change each of the backbone ψ dihedral angles by 30° of a selected structure. The total number of structures generated by such a change is 7 since there are two Pro, Gln, and Phe residues on the peptide. Vertices of the corresponding convex hulls and the series of ψ dihedral angles of the generated structures and the original structure are presented in Table 5. These data show that the most dramatic change in structures, as revealed by the largest change in convex hull vertices, is caused by the change in the two central dihedral angles, namely, between Gln^6-Phe^7 and Phe^7-Phe^8. As also revealed by the smaller change in the number of convex hull vertices, the change in structures is apparently less affected when a change in dihedral angle is moved from the central to either end of the peptide. The data also show that small changes in the two central dihedral angles cause the larger structural change than the local structural change.

CONCLUSIONS

Given its sensitivity to fine structural changes, we think that the convex hull vertices computed can be used as a molecular descriptor or clustering criterion to classify vast conformations generated for some flexible molecules such as peptides. To cluster molecules with close similarity in structures, one need only to count the number of commonly exposed groups or the number of similar vertices on a series of convex hulls computed for the group of structures to be clustered. The similarity between collected structures is revealed by the number of exposed groups which are common for the members of the cluster. Based on the number of commonly exposed groups, what one searches for in a cluster of molecules is the parts of corresponding molecular structures that are also commonly exposed. There is a certain advantage for this kind of identification especially in the field of drug design.[17] One of the salient features in this method is the efficient sampling and generation of blocks of structures since the number of commonly exposed groups counted depends on how structures are sampled and collected in blocks. This problem is addressed by generating a large number of blocks of structures consisting of regularly or randomly selected structures. To find linkages between clusters, some overlapping in structures is allowed. Two cutoffs that one has to choose in advance for this method are the minimum number of structures allowed in a block and the minimum number of commonly exposed groups identified for a block. Unlike some other clustering methods,[8,19] no specific rules are required for choosing these two cutoffs. Structures inherently conform to the selection criteria if one performs the clustering from more stringent (e.g., higher minimum number of structures allowed and higher number of commonly exposed groups identifiable in a block) to more relaxed (e.g., lower minimum number of structures allowed and lower number of commonly exposed groups for a block) conditions. On the basis of these characteristics, we believe that the method is best suited for the exploratory type of clustering where no information about data clustering tendency is available. Depending on the methods[20] used and the purpose to link clusters, the linkage between clusters can be extensive or restrictive. Linkages between clusters searched by the method presented here are more restricted since the minimal distance criterion[16] is used to assist the selection of compact blocks of structures. It seems that without using such a criterion,[16] our aim to identify compact and well-separated clusters will not be achieved quite well. There are two clustering criteria, namely, the number of identifiable commonly exposed groups and the calculated minimal distances that are used in this method. However, depending on the purpose of clustering, the latter can be set as optional.

ACKNOWLEDGMENT

This work is supported in part by a grant from the National Science Council, ROC (Grant NSC 87-2213-E-007-019).

REFERENCES AND NOTES

(1) Massart, D. L.; Kaufman, L. *The Interpretation of Analytical Chemistry Data by the Use of Cluster Analysis*; Wiley: New York, 1983; and references cited therein.

(2) Levitt, M. J. Molecular Dynamics of Native Protein II. Analysis and Nature of Motion. *J. Mol. Biol*. **1983**, *168*, 621−657.

(3) 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*, 615−625.

(4) Rooman, M. J.; Rodriguez, J.; Wodak, S. J. Relations between Protein Sequence and Structure and Their Significance. *J. Mol. Biol.* **1990**, *213*, 337−350.

(5) Gordon, H. L.; Somorjai, R. L. Fuzzy Cluster Analysis of Molecular Trajectories. *Proteins* **1992**, *14*, 249−264.

(6) Karpen, M. E.; Tobias, D. J.; Brooks, C. L., III. Statistical Clustering Techniques for the Analysis of Long Molecular Dynamics Trajectories: Analysis of 2.2-ns Trajectories of YPGDV. *Biochemistry* **1993**, *32*, 412−420.

(7) Torda, A. E.; van Gunsteren, W. F. Algorithms for Clustering Molecular Dynamics Conformations. *J. Comput. Chem*. **1994**, *15*, 1331−1340.

(8) Bravi, G.; Gancia, E.; Zaliani, A.; Pegna, M. SONHICA (Simple Optimized Non-Hierarchical Cluster Analysis): A New Tool for Analysis of Molecular Conformations. *J. Comput. Chem*. **1997**, *18*, 1295−1311.

(9) Shenkin, P. S.; McDonald, D. Q. Cluster Analysis of Molecular Conformations. *J. Comput. Chem.* **1994**, *15*, 899−916.

(10) Lin, T. H.; Peng, W. J.; Lu, Y. J. Identification of Convexity as a Common Structural Feature for Structures Generated for Two Short Peptides. *Comput. Chem*. **1998**, *22*, 309−320.

(11) Green, P. E.; Carmone, F. J. *Multidimensional Scaling and Related Techniques in Marketing Analysis*; Allyn & Bacon: Boston, 1970.

(12) *SYBYL 6.4*; The Tripos Associates: St. Louis, MO.

(13) Pearlman, D. A.; Case, D. A.; Caldwell, J. C.; Seibel, G. L.; Singh, C.; Weiner, P.; Kollman, P. A. *AMBER 4.1*; University of California: San Francisco, 1995.

(14) Dutta, A. S. *Small Peptides: Chemistry, Biology and Clinical Studies*; Elsevier: New York, 1993.

(15) van Gunsteren, W. F.; Berendsen, H. J. C. Algorithms for Macromolecular Dynamics and Constraint Dynamics. *Mol. Phys*. **1977**, *34*, 1311−1327.

(16) Friedman, H. P.; Rubin, J. On Some Invariant Criteria for Grouping Data. *J. Am. Stat. Assoc*. **1967**, *62*, 1159−1178.

(17) Crippen, G. M. Intervals and the Deduction of Drug Binding Site Models. *J. Comput. Chem.* **1995**, *16*, 486−500.

(18) Lin, T. H.; Peng, W. J.; Lu, Y. J. A CoMFA Study on Several Bioactive Peptides using the Alignment Rules Derived from Identification of Commonly Exposed Groups. *Biochim. Biophys. Acta* **1999**, *1429*, 476−485.

(19) Jarvis, R. A.; Patrick, E. A. Clustering Using a Similarity Measure Based on Shared Near Neighbors. *IEEE Trans. Comput.* **1973**, *C-22*, 1025−1034.

(20) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*; Wiley: New York, 1990.

CI9801623