The "Nearest Single Neighbor" Method—Finding Families of Conformations within a Sample

Doron Chema[†,‡] and Amiram Goldblum[†,*]

Department of Medicinal Chemistry and Natural Products and the David R. Bloom Center for Pharmacy,
School of Pharmacy, Faculty of Medicine, The Hebrew University of Jerusalem, Israel 91120, and
School of Chemistry, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel

Received August 3, 2002

A simple method for self-organization of conformation samples into families is presented. According to this method, any large sample of molecular conformations may be reorganized according to the nearest single root-mean-square displacement (rmsd) neighbor, starting at any chosen "seed" conformation. Following such reordering, conformational families may be determined by a novel process that maximizes family sizes while minimizing family mixing. This process eliminates much of the arbitrariness that was inherent in most of the related methods of conformation clustering. We demonstrate the construction of rmsd matrices and discuss the convergence criteria for the sample size as well as criteria for determining the cutoff value for the definition of families in each sample. The method is invariant to changes of the "seed" conformation. After applying this method, families of conformations may be more easily recognized in graphic matrices. The method has been applied to the analysis of the conformational space of two cyclic peptides. It is also shown that the "organized" conformational space, at least in those specific examples, has an energy topology that reminds of energy basins. The method is general and applicable to molecules of any type.

1. INTRODUCTION

Conformational analysis is an important technique for exploring molecular structure and flexibility, in relation to molecular activity as well as properties.[1−4] Peptide conformations play an important role in determining the specificity and the potency of peptide drugs.[5−7] Conformational analysis can also be used for mapping the conformational landscape of proteins during folding[8] and near the crystal structure.[9,10] Cluster analysis methods, based on geometric similarity, are commonly used for such mapping.[1,3,4,11−13] Cluster analysis methods applied to conformational spaces commonly assume that the geometric partition reflects a "real" physical partition of the conformational space. These methods suffer from several drawbacks.

Hierarchical clustering is a major family of clustering algorithms. These methods are typically applied by first assigning each conformation into a single cluster. Step by step, the most similar cluster pairs are merged into single clusters. The process terminates when all clusters are merged into one cluster, called the root of the *dendogram*. Hierarchical clustering methods differ mainly by the determination of the distance between clusters. The shorter distance between conformations in different clusters is used as a similarity criterion by the *single linkage* method. The longest distance, on the other hand, is used by the *complete linkage* method, while the average similarity criterion is used by the *average linkage*[14] method. After establishing the dendogram, a line is drawn at an arbitrarily chosen level of similarity of the dendogram. Clusters are collected below this line, and therefore different numbers of clusters may emerge at different similarity levels. Additionally, different distance measurements may lead to different dendograms, for the same data set. Specific difficulties might also occur, for example, the single linkage method tends to produce elongated conformational clusters.[11,15]

Nonhierarchical clustering methods are an alternative set of methods.[1,2,4,5,11,12,15,16] Generally, automatic determination of the cluster boundaries is a major advantage of these methods, compared to hierarchical clustering. Nevertheless, a nontrivial parameter setting is usually required, that reflects some prior knowledge of the conformational space topology.[3,15] "Nearest neighbors" methods are commonly used under this title, and "family clustering" is a typical representative. The clustering process of the family clustering method starts at each stage from the current lowest energy conformation in the conformation set. Conformations are then chosen according to the condition that their root-mean square displacement (RMSD) from the other conformations in the cluster is below a certain cutoff. Hence, the generation of high similarity regions is forced to be near the current lowest energy conformation. It could miss, due to the arbitrary value, the "real" families that exist in conformational space. Recently, families collected by this method were compared to energy basins, for Alanine tetra-peptide.[17] Using a RMSD cutoff of 1.0 Å led to a good overlap of energy basins and clusters. But, increasing the cutoff by a small amount, to 1.1 Å, resulted in a disagreement between them. In one algorithm known to us, the determination of the cutoffs depends on the properties of the particular set of conformations,[12] accompanied by an automatic decision process ("separation ratio") for terminating a multilevel clustering procedure. Our approach is somewhat different.

[*] Corresponding author e-mail: amiram@vms.huji.ac.il.
[†] The Hebrew University of Jerusalem.
[‡] Tel Aviv University.

10.1021/ci0255735 CCC: $25.00
Published on Web 11/19/2002

We propose the "nearest single neighbor" (NSN) method and, following the application of NSN, a novel approach for delimiting and defining conformational families. NSN applies a specific pathway to order a sample of conformations. The RMSD matrix, which includes all pairwise RMSD between conformations according to their order of emergence from the conformational search program, is reorganized automatically by the algorithm. Subsequently, by applying a novel RMSD criterion, the desired family topology emerges through maximizing the family size simultaneously with minimizing the "mixing" between families. Since only a minor arbitrariness is inherent in the family determination, this method is an improvement over previous ones.

The method is applied here to the conformational analysis of two cyclic peptides: Cyclic $(Ala)_6$ and Cyclic [GSAGPV]. It is demonstrated that the method can be applied not only for reordering and reorganizing but also for testing the convergence of the sampling methodology. In addition, we compare between the size of conformational families to clusters generated by the family clustering method under several geometric cutoffs. We demonstrate that the specific conformation ordering, which is dictated by the method, results in a specific topology of the energy surface in those cases where the original sampling was done for studying low energy structures.

2. METHOD

The Nearest Single Neighbor Method. I. **Ordering the Conformations.** Given a sample of conformations from a molecular simulation such as Molecular Dynamics (MD) or Monte Carlo (MC), or from any other source, the RMSD matrix between pairs of conformations may be constructed for studying the structural relations between conformations and for assigning conformational families. In peptides and proteins, it is mostly the backbone structure[18] that is used for structural comparisons.

The method is applied by picking the lowest energy conformation r_0 as a starting point. This choice is convenient but could be replaced by any other conformation. The first step is to scan through all the conformations to find the "next nearest neighbor", i.e., the conformation that is the most similar, in terms of RMSD, to r_0. This conformation, r_1, is the next target for finding a "nearest neighbor" conformation. This simple iterative process is summarized in the equation

$$r_{i+1} = \mathrm{Min}\ (\text{backbone RMSD } (r_i, r_j)) \text{ for all } j > i$$

The common (time or frame) order of the conformations based on the sequence from MD or MC is thus replaced by the reorder suggested by this algorithm. In this "reordered" matrix, given some RMSD cutoff criterion, the conformation families appear as "blocks" of the most similar conformations along the diagonal of the matrix. By definition, each block is one family of conformations. The RMSD criterion thus has a crucial meaning for the presentation of families.

The algorithm is very fast: the ordering process requires only a few seconds for a sample of several hundreds to thousands of conformations of medium sized peptides, on a standard PC. The time-consuming part in the analysis is the RMSD calculation between conformations. It naturally grows as N^2, where N is the total number of conformations within the sample, but is done only once for a reordered sample. We have studied samples of sizes 500−1500 of the above molecules by using the CHARMM[19] package.

II. Least-Arbitrary Determination of Conformational Families. Once the conformations were reordered by the next nearest neighbor (NSN) principle, the boundaries of conformation families should be determined. Most other methods performed this step in the past by using arbitrary values. We propose a more rational approach, as follows:

1. The RMSD matrix is reconstructed based on the new order of conformations, in the upper triangular matrix.

2. Start the calculation of families with a very small RMSD criterion of 0.1 Å. Begin on the first row and compare RMSD (i,j) (where $j \geq i$) to the RMSD criterion along a row i (upper triangular) before moving to row $i+1$.

3. Any conformation which has [RMSD (i,j) < RMSD criterion] is included in the (contiguous) family of conformations at row i. The last conformation that is part of this contiguous set is being assigned, for our purposes, j_0.

4. Initialize a counter of "family mixing" or "blending", $BC_0 = 0$. If a conformation $j > j_0$ has [RMSD (i,j) < RMSD criterion], increment the "mixing value" counter by one. This counter displays how many times there is blending or mixing—i.e., how many times do more distant conformations occur, that are close in RMSD to the one at the diagonal of the current row i.

6. Add the mixing value for all the rows in the upper triangular matrix.

7. Increase the RMSD criterion by some small increment and repeat steps 2−6.

8. Continue steps 2−7 until the counter is raised and then reduced back to a very small number.

It is quite clear that in the reordered matrix, given a very low RMSD criterion, there will be very few conformations that belong to the first formed family in each row. Apart from a few conformations near the diagonal, the rest would be different than the first by larger RMSD values, and it is unlikely that the very low RMSD criterion would be met again along a row. This is displayed in Figure 1a for a "reordered" sample that includes 500 conformations of Cyclic $(Ala)_6$. This figure shows that families are mostly restricted to the region that is very close to the diagonal and there is a negligible amount of "mixed families", i.e., those that have a colored component off the diagonal. Figure 1b-e shows the change of the "mixing" with increasing the RMSD criteria (to 0.2 Å, 0.4 Å, 0.8 Å and 1.5 Å respectively). With a very large RMSD criterion, there will be just one large family encompassing the whole matrix, and the mixing value should be zero. For a RMSD criterion of 2.0 Å, we find just a single family. The variations in the mixing value with the RMSD resolution are shown in Figure 1f. The mixing value goes through a maximum around RMSD = 1.5 Å and is minimal for low RMSD (many small families) and for large RMSD (one large family).

To choose an optimal RMSD value, we begin with the lowest RMSD (0.1Å) and scan to higher values until the ratio of the mixing value for two successive RMSD values reaches its maximum. The value of the RMSD criterion to be used for family definition is where this maximum occurred. In addition to an initial value of 0.1 Å and integral multiples of this value, we also tested much smaller ones, of 0.01 Å and 0.005 Å. These values did not add any further insight to the similarity or family definition issues, compared

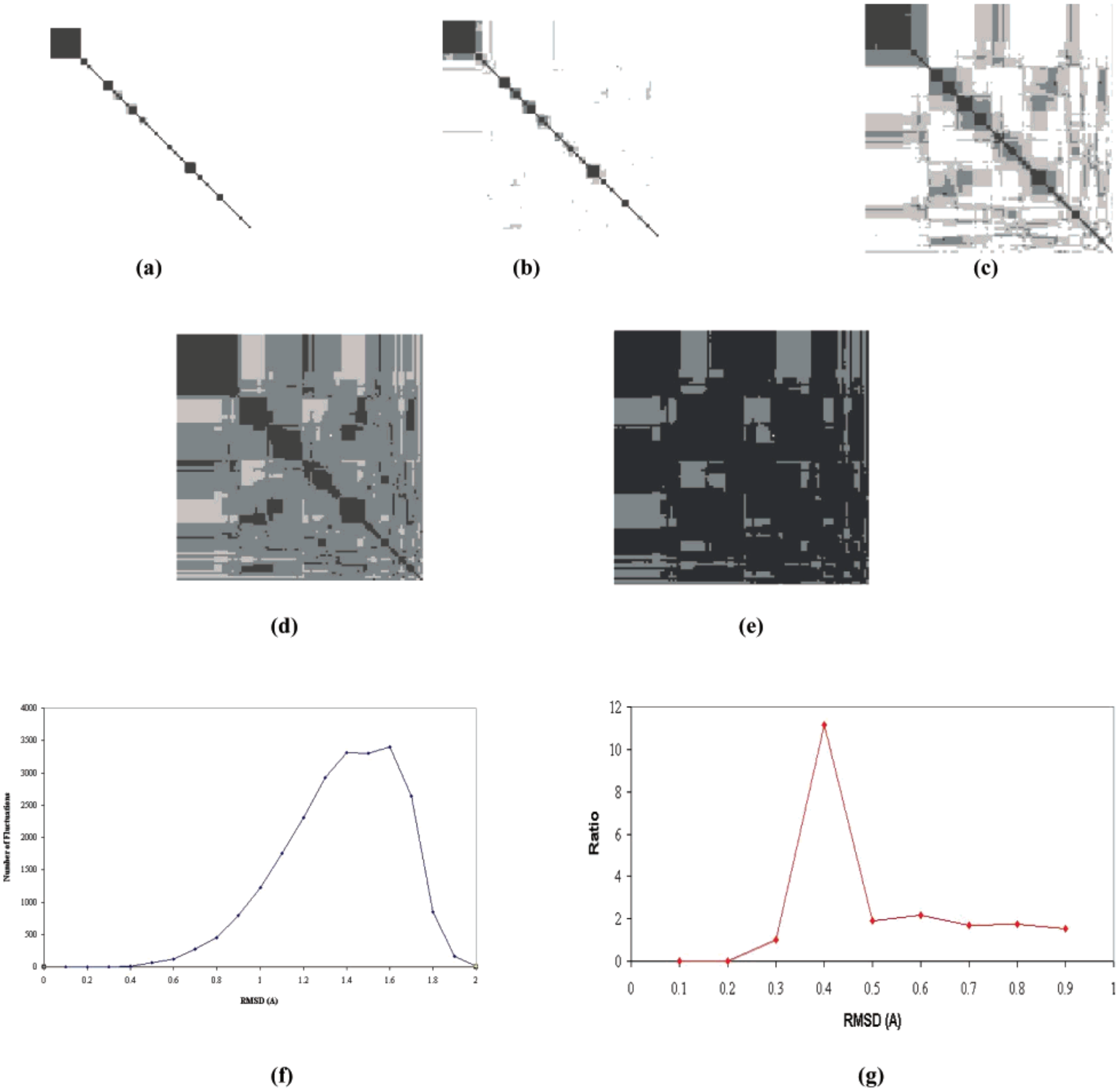


Figure 1. The RMSD matrix of Cyclic $(Ala)_6$ with 500 conformations, after being reordered by the nearest single neighbor algorithm at resolutions of (a) 0.1 Å (b) 0.2 Å (c) 0.4 Å (d) 0.8 Å (e) 1.5 Å. In (f), the off diagonal mixing values (Number of fluctuations, y axis) versus the resolution (RMSD) are plotted. The graph 1f includes also values for which the full graphs are not presented. In (g) the ratio value between sequential mixing values, taken from (f), are plotted.

to the coarser values. To detect a "critical" resolution value we calculate, for each RMSD value, the ratio between the next mixing value and the present one. In calculating the ratio, to avoid division by zero and skewed results, we use the following conventions:

1. for $BC_{i+1} = 0$ and $BCi = 0$, ratio = 0
2. for $BC_{i+1} = m$ and $BCi = 0$, ratio = 1

By checking a few other criteria for increasing family sizes while minimizing the mixing of families, we conclude that the calculated ratio of two adjacent "mixing values" is the most sensitive criterion. One of the other tested criteria has been the numeric derivative of the change in mixing with respect to the change in RMSD value, i.e., Δ"mixing"/ΔRMSD. The mixing value with the appropriate ratios and the numeric derivatives for the Cyclic $(Ala)_6$ data are presented in Table 1. A clear peak is found (last column, line 4) where the ratio of adjacent mixing value (11.2) is displayed. The numeric derivative on that line does not display such a characteristic maximum. A change in the choice of the initial ("seed") conformation for the construction of the RMSD matrix does not affect the size of families or their mixing, as we show in the results section.

III. Determining the Sizes of Families. Families appear along the diagonal of the matrix. Once the RMSD cutoff was determined, individual families may be collected, as follows:

```
For each row (i) in the "ordered" matrix:
If RMSD $(i+1, i) < $ cutoff
Then Add conformation $i+1$ to the current cluster
Else start a new cluster
```

IV. Uniting Split Families. In those rare cases that families were "split" during the re-ordering step, they may be joined according to the following:

For family i

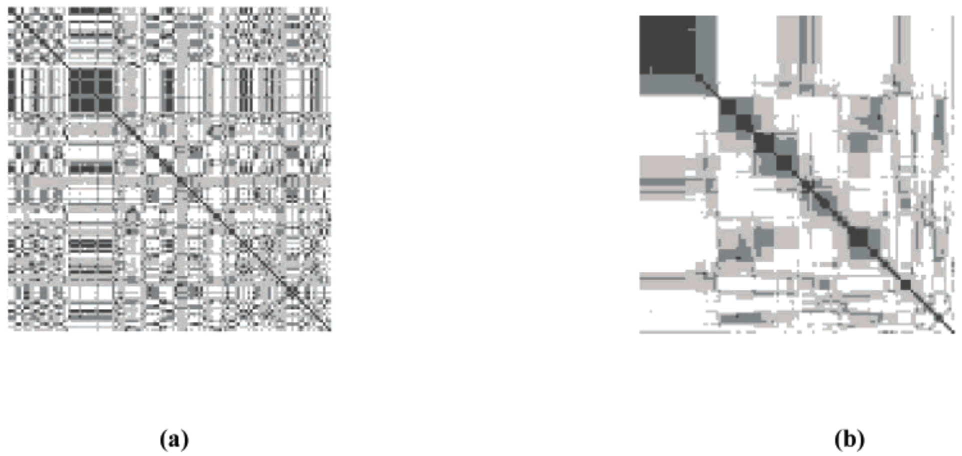


Figure 2. The RMSD matrix of Cyclic $(Ala)_6$ with 500 conformations, reordered by the nearest single neighbor algorithm (a) before and (b) after reordering. The gray scale (4 colors) represents the following backbone rmsd values: 0−0.4 (black), 0.4−0.8, 0.8−1.2, and above 1.2 Å (white).

Table 1. Mixing Value Ratios Compared to Numeric Derivative

RMSD values in Å	mixing value	numeric derivative Δ*mixing value]/0.1*	ratio of mixing values (from the first column)
0.1	0	0	0
0.2	0	0	0
0.3	0	60	1.0
0.4	6	610	11.2
0.5	67	590	1.9
0.6	126	1510	2.2
0.7	271	1810	1.7
0.8	452	3460	1.8
0.9	798	4330	1.5
1.0	1231	5220	1.4
1.1	1753	5610	1.3
1.2	2314	6180	1.3
1.3	2932	3850	1.1
1.4	3317	−130	1.0
1.5	3304	990	1.0
1.6	3403	−7570	0.8
1.7	2646	−17950	0.3
1.8	851	−6830	0.2
1.9	168	−1620	0
2.0	6		

If Average RMSD (family i, family j) < cutoff

Then join families i and j

Samples. (a) *Cyclic [GSAGPV]* − A sample of 1000 conformations was obtained from a 500 picoseconds (ps) high-temperature (1000 K) trajectory of molecular dynamics. Each high-temperature conformation was then gradually cooled to 300 K. At each step, the temperature was reduced by 100 K and the simulation continued for 0.8 ps. Reaching 300 K, each conformation was minimized to the nearest local minimum, using 300 steps of Steepest Descent followed by 1000 steps of Adopted Basis Newton−Raphson (ABNR). The simulations were performed using the molecular dynamics program CHARMM[19] and the CHARMM all-atom force field[20] using 2 fs time steps. A 15 Å cutoff was applied to nonbonding interactions (VDW and electrostatic), and the SHAKE constraints were applied to all bonds to hydrogen atoms.

(b) *Cyclic $(Ala)_6$* - A sample of 500 conformation of the peptide, was available from previous work.[21] The conformations were generated according to the same protocol as that for *Cyclic [GSAGPV]*.

(c) Additional samples of the *Cyclic $(Ala)_6$* containing 1000 and 1500 conformation were generated, using the same protocol described above.

3. RESULTS

In the following examples we use the above methods to analyze the conformational space of a few peptides. The potential energy surfaces of peptides and proteins contain multiple energy minima that are separated by energy barriers.[10,22] Each of these regions of minima is also characterized by a unique backbone structure, which is used to define families.[17] Peptides are more flexible than proteins and may adopt many different conformational states, thus they constitute a challenging system for analysis.

Cyclic $(Ala)_6$. The accumulated conformations of Cyclic $(Ala)_6$ were ordered by the nearest single neighbor method starting from the global energy minimum. Then, the families' determination process described in the Methods was followed. RMSD values starting from 0.1 Å up to 1.0 Å were applied, in intervals of 0.1 Å. The ratios were plotted against RMSD in Figure 1g. The decision to use a cutoff of RMSD = 0.4 Å was done according to this graph, as a substantial change to more "mixing" takes place between the RMSD cutoffs of 0.4 Å and 0.5 Å. The original RMSD matrix, as it emerged from the MD simulation, is shown in Figure 2a. In the reordered matrix (Figure 2b), in which the different RMSD values were colored by a gray scale, a clear pattern of conformational families is displayed. Families of conformations appear in the matrix as "blocks" with variable sizes, and the matrix also includes information about similarity between families, represented as off-diagonal blocks of high similarity.

The Energy Profile of the Ordered Conformations. The algorithm operates on geometrical similarity considerations alone. However, it is important to question whether the formation of families based on RMSD criteria has any energy equivalent. If we begin the process with the structure that is the global energy minimum, do the other conformations in its "family" have similar or different energies? Starting from the same global minimum, we have drawn the energy levels of each conformation after re-ordering according to NSN.

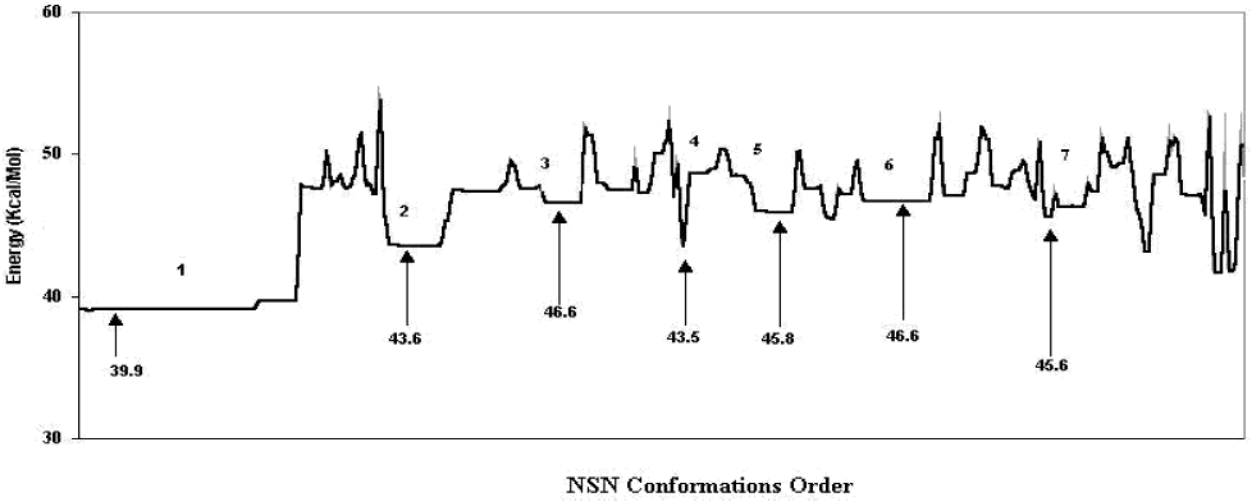


Figure 3. The potential energy of the ordered Cyclic (Ala)$_6$ conformations. Integers indicate families along the diagonal in Figure 2b.

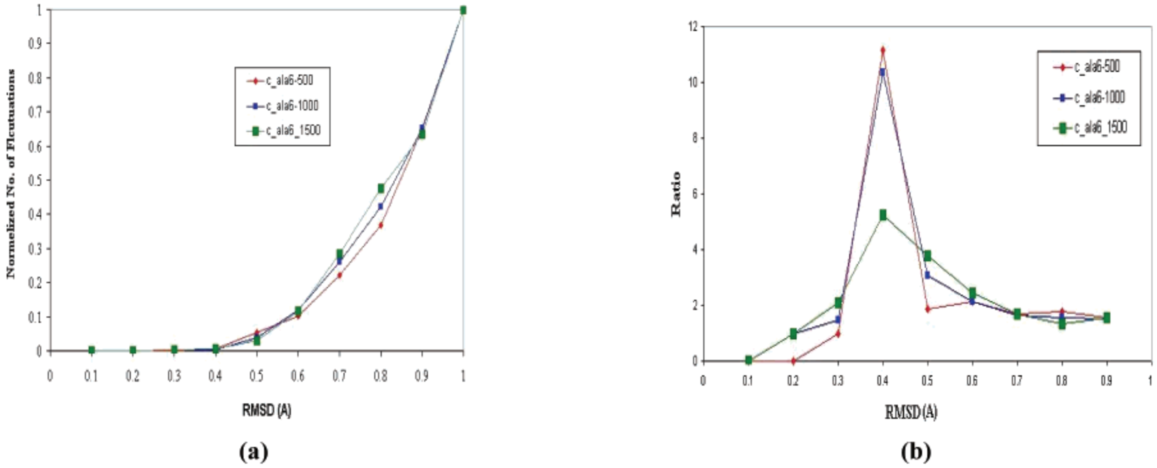


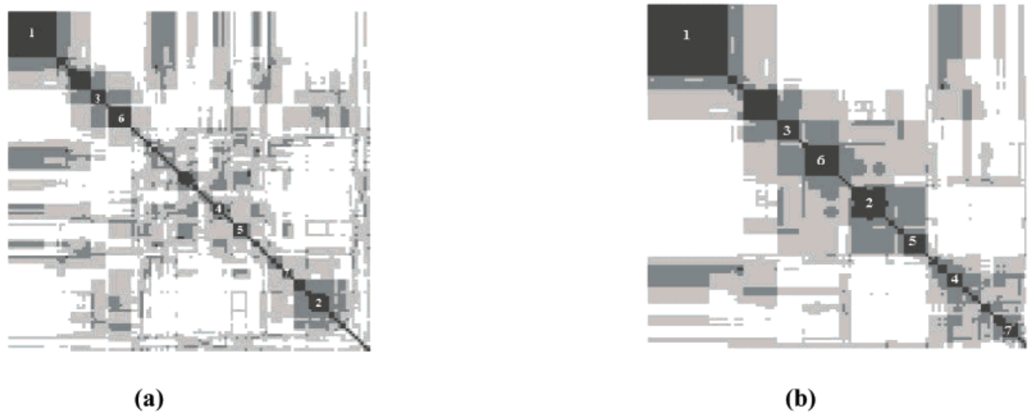
Figure 4. (a) The mixing value versus RMSD values for different sample sizes of Cyclic (ala)$_6$: 500 conformations (diamond), 1000 (small square), and 1500 (rectangle). For comparison, the numbers are normalized. (b) The ratio between mixing values for the samples of 4a.

The results are shown in Figure 3, where the largest families (numbered) coincide with the larger energy basins**.** It is clear that some smaller families exhibit a basin topology along this graph. There are also some cases where a family has a relatively small size. The exact “boundaries” of the basin-families are not clearly determined from the energy profile in many cases. In these cases, geometric considerations become helpful.

Convergence Criteria. The influence of the sample size, 500 conformations initially, on the results was examined by simulating additional samples of 1000 and 1500 conformations of Cyclic (Ala)$_6$. As shown in Figure 4, the three samples have the same characteristic cutoff value (0.4 Å). Figure 4a presents the change in the mixing value as a function of the resolution, while Figure 4b displays the ratio as a function of resolution, for the three samples.

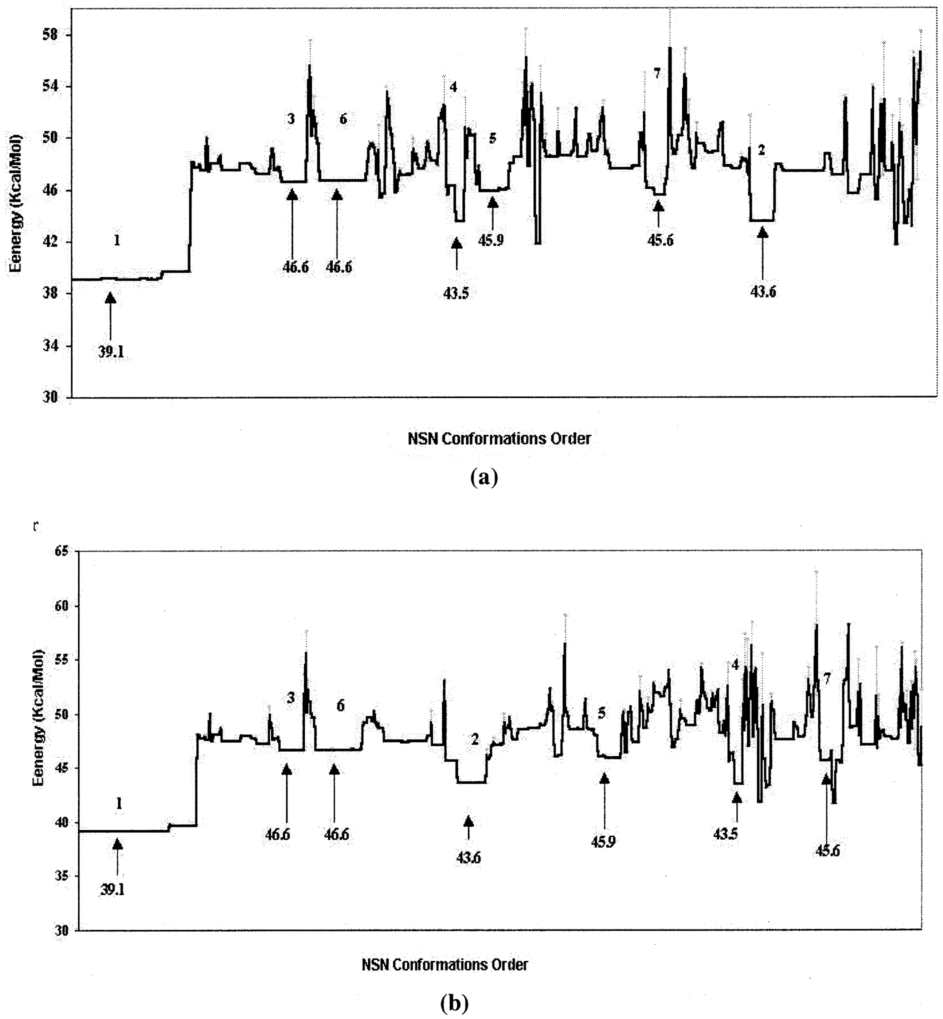
The RMSD matrices of the 1000 and 1500 conformations are shown in Figure 5a and b, respectively. Major families that were found at the 500 conformations’ sample were identified also in the larger samples, by comparing the families’ lowest energy minima and RMSD (Figure 6a,b for 1000 and 1500 conformations, respectively). Additional few (but scarcely populated) families, that are higher in energy in most cases, were identified in the larger samples. We have also compared the results for smaller samples of 100 and 250 conformations, in which some major families are populated significantly differently than in the larger samples. A comparison between the major family sizes (percentage of total) in the different samples is summarized in Table 2. The relative population of the seven most populated families changes little along the series of 500−1500 conformations (40%-45%). We have not increased further the number of accumulated conformations, as the current ones seem to achieve convergence already for 500 conformations, and any increase affects the length of RMSD calculations. The reordering helps to perceive the effects of increasing numbers of conformations, which would otherwise be very complex to analyze.

One may notice that the order of the families changes between similarity matrices of different sample sizes (Figure 2b for 500, 5a for 1000 and 5b for 1500). Starting from the original reordered matrix (Figure 2b), in which the numbers are sequential, this order changes with the larger samples. The reason for this change in family order stems from the “automatic” switch from one family to another according to structural criteria, where small variations of structures might raise the energy considerably and the shift to the next family could take different directions in different samples. Only a



(a) (b)

Figure 5. The RMSD matrices for larger samples of cyclic (Ala)$_6$, after being ordered by the nearest single neighbor algorithm, of (a) 1000 and (b) 1500 conformations, using a gray scale similar to Figure 2. Families similar to those in Figure 3a are given the same indexes.



(a)

(b)

Figure 6. The potential energy of the larger samples of cyclic (Ala)$_6$: (a) 1000 and (b) 1500 conformations. The families index, are the same as in Figure 5.

few of these high-energy conformations are statistically found in a sample (see discussion).

Different Choice of "Seed Conformation". To study the effect of changing the "seed conformation" from the lowest energy one to some other conformation, we decided to pick conformations from a few of the families that were identified and presented above. Toward that goal, conformations from each of the families 3, 5 and 6 were tested. The RMSD matrices for each of the three different starting positions are drawn, and the results also include the decisions for the cutoff value for each of these matrices, separately. It may be seen in Figure 7a−c that the numbers and sizes of families in these matrices are similar to those of Figure 5a, in which the seed conformation was the one with a lowest energy. Not only the sizes but also the actual "members" (numbered conformations) of these families are identical. In Figure 7d the ratio of sequential mixing counters is plotted, showing a similar cutoff value (0.4 Å) as in the previous cases. There

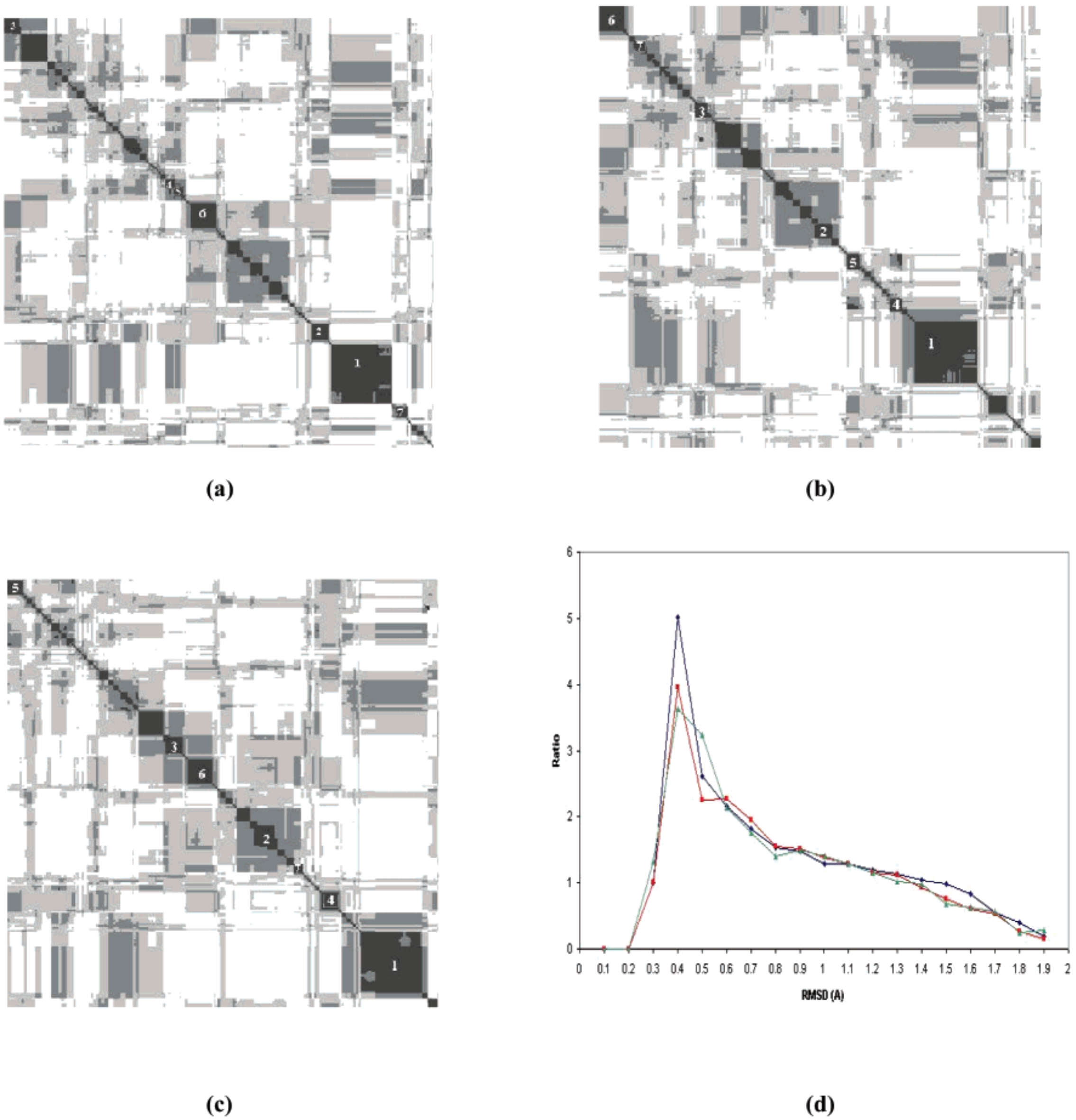


Figure 7. The rmsd matrix of the 1000 conformations' sample of cyclic $(Ala)_6$ generated by the nearest single neighbor method with different seed conformations, each belonging to a different family. The families' numbers are the same as in Figure 5. Seed conformations were taken from family 3 (a), family 6 (b), and family 5 (c). (d) The ratio between mixing value for the sample order in 7a (highest peak), for the sample order in 7b (lowest peak), and for the sample order in 7c.

Table 2. Percentage of Families in Samples of Different Sizes for Cyclic $(Ala)_6$[a]

family minima	size of sample (no. of the conformations) (%)				
(kcal/mol)	100	250	500	1000	1500
39.9	12.0	18	14.4	13.5	13.3
43.5		2	3.2	3.6	3.7
43.6	6	3	4.8	4.0	3.7
45.6		3	5.6	6.0	5.9
45.8			6.4	6.8	6.4
46.6	7	6	4.8	4.4	4.3
46.6	15	8	5.6	6.4	5.9

[a] Only families with size ≥ 2% from the total conformation space are listed, with their $\mathbf{E^0}$ value in the range of 39−47 kcal/mol.

are only minor variations between the graphs in Figure 7d. These data suggest that the results are invariant to the choice of the seed conformation.

Cyclic [GSAGPV]. The RMSD matrix for cyclic [GSAGPV] (after reordering) is also composed of major conformational families (Figure 8). However, a slightly different cutoff value (0.3 Å compared to 0.4 Å) distinguishes this peptide from cyclic $(Ala)_6$ (Figure 9). The energy profile of the

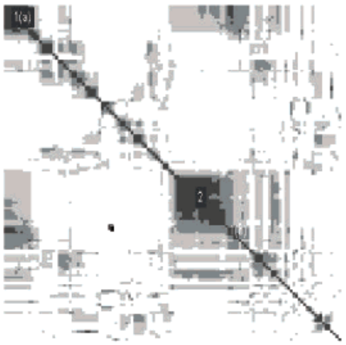


Figure 8. The RMSD matrix of cyclic [GSAGPV] for 1000 conformations after reordering. The gray scale represents the following backbone rmsd values: 0−0.3 (black), 0.3−0.6, 0.6−0.9, and above 0.9 Å (white).

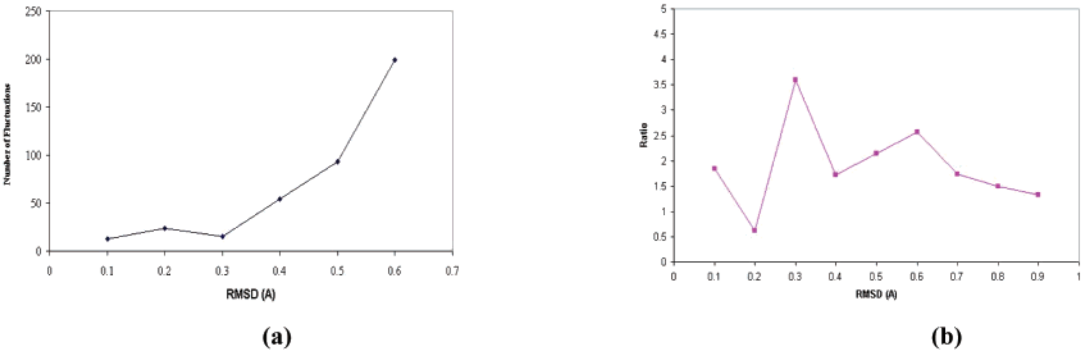


Figure 9. (a) Mixing values versus the RMSD values for the conformations of cyclic [GSAGPV]. The numbers are restricted to the 0.1-0.6 Å, which is the most important region for the cutoff determination. (b) The ratio of mixing values versus RMSD values.

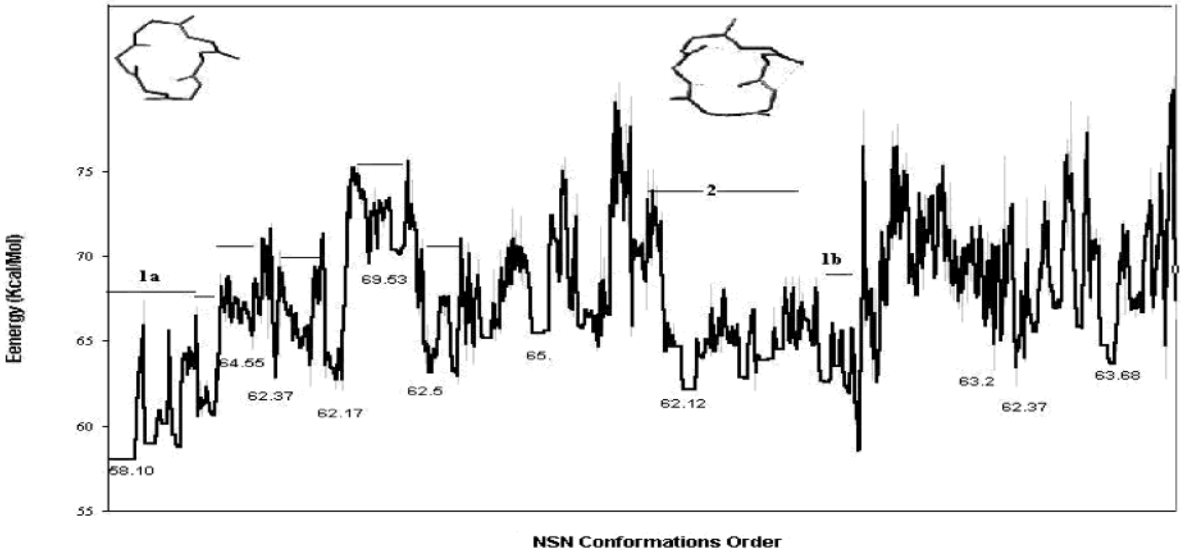


Figure 10. The potential energy of [GSAGPV] after reordering. The family's indexes are similar to Figure 7.

families in this case (Figure 10) appears rugged compared to the relatively smooth energy profile of Cyclic (Ala)$_6$ (Figure 3). This may be due to different side chains in the two peptides. The family around the lowest minimum is split here into two regions, indexed in the matrix as 1a and 1b (Figure 10). The average RMSD between the families' members is smaller than the cutoff and therefore, they may be "united", as described in Methods.

Comparison with the Family Clustering Method. Clustering conformations of a peptide into disjoint families is a common conformational analysis tool described in the Introduction. Examining the families of peptides (Figures 2b, 5a,b, 8), one finds that clustering under a cutoff of 1 Å, which is a typical value used in cluster analysis, may include more than one family. To illustrate this point, the family clustering method was applied and its resulting clusters were compared to those of the NSN method. The family clustering method starts from the global minimum conformation, and collects conformations whose RMSD comparison to other cluster members is below some predetermined cutoff. The collected clusters are then eliminated from the conformation set to be clustered. This process is repeated while starting at each step from the current lowest minimum conformation. Here, we applied three cutoff values to the backbone RMSD: 0.5, 1.0, and 1.1 Å while clustering the Cyclic [GSAGPV] conformations, and three other cutoff values of 0.5, 0.8 and 0.9 Å while clustering the Cyclic (Ala)$_6$. This range of cutoff values is frequently applied to medium sized peptides and also to proteins.[11,23] Tables 3 and 4 present a comparison between major clusters as extracted at the different cutoffs, and major families identified by NSN. As expected, the cluster sizes change according to the cutoff used, while in extreme cases, clusters vanish and reappear at different cutoffs. This situation emphasizes the advantages of the NSN method in forming a clear single option to divide the conformational space into families.

Table 3. Comparison of the Nearest Single Neighbor (NSN) Results to Those of the Family Clustering (FC) for 1000 Conformations of Cyclic [GSAGPV][a]

family minima (kcal/mol)	NSN	FC (0.5 Å)	FC (1.0 Å)	FC (1.1 Å)
62.1	152	164	174	185
58.0	128	135	152	152
64.6	47	0	0	52
69.5	47	44	43	44
62.4	38	0	0	0
62.5	37	17	33	47
65.6	36	0	0	0
63.7	35	44	25	27
62.2	22	70	94	87
62.4	13	38	18	18
63.3	12	0	37	0

[a] Numbers indicate conformations within a family. Three different cutoffs were used to cluster the conformations: 0.5, 1, and 1.1 Å. Only major populations are listed.

Table 4. Comparison of the Nearest Single Neighbor (NSN) Results to Those of the Family Clustering (FC) for 1000 Conformations of Cyclic $(Ala)_6$[a]

minima (kcal/mol)	NSN	FC (0.5 Å)	FC (0.8 Å)	FC (0.9 Å)
39.9	135	160	223	234
41.7	11	0	0	58
43.6	40	0	142	77
45.6	60	40	64	62
45.8	68	0	0	43
46.6	44	47	74	0
46.6	64	62	67	64
47.0	0	0	44	0
47.0	0	0	0	64
47.2	0	57	0	0
47.4	56	0	0	0
47.6	48	40	0	0

[a] Numbers indicate conformations within a family. Three different cutoffs were used to cluster the conformations: 0.5, 0.8, and 0.9 Å. Only major populations are listed.

4. DISCUSSION

The nearest single neighbor method introduced in this paper is an attempt to simplify the analysis and presentation of molecular simulations or conformational searches that result in large numbers of structures. It is also our intention to simplify the process of decision about family boundaries through RMSD cutoff values. We have shown that the method is applicable to the results of a large conformational search and also provided data to display the sample size effects and the detailed process of decision about a RMSD cutoff value for defining and delimiting families of conformations. We have also demonstrated the invariance of the results to changes of the “seed conformation”, the conformation that is chosen to begin the RMSD reordering. Such invariance is another aspect of this algorithm and adds to its robustness. Naturally, a different sampling procedure, such as molecular dynamics at a different temperature, would result in different clustering and different families of conformations. Also, our reordering is heuristic and cannot guarantee that the above displayed invariance would hold in all cases.

One of the main interests in such a large set of conformations is to find how and if they are structurally and energetically related. The reorganization of the conformations according to the next smallest RMSD positions the most closely related conformations next to each other along the new sequence. We have shown that this reordering forms an excellent basis for defining families. We have also proposed a novel approach to an unbiased and least-arbitrary definition of the boundaries for a conformational family. This is applied by producing family boundaries through applying successive RMSD cutoff values from a very small one to larger ones while searching for the best balance between “family size”’(larger are better) and “mixing” of families (larger is worse). Unlike most of the clustering methods that were described in the Introduction, the nearest single neighbor (NSN) method does not enforce conformations into families. Therefore, what is it that “arranges” conformations into families by this algorithm?

Considering any conformation in a sample, it belongs to some family. Following the general characteristics of a family, the closest conformation to this conformation is from the same family. The energy barriers between families ensure that, following the Boltzmann distribution, only very few conformations (if at all) will be found at the top of the barriers and therefore could belong to more than one family. Such conformations could potentially shift the algorithm search from one family into another, before fully spanning the current family. Following these arguments, the algorithm tends to span an entire family before moving into another. Only a small number of “barrier conformations” is expected to be encountered due to the statistical insignificance of these barrier energies compared to the prevalence of the lower energy structures in the “basins”. This is typical to conformational sampling that is accompanied by a minimization procedure, as in our examples. It should also prevail in simulations of low temperatures, including room temperature, of many biologically relevant molecules. Assuming a modest energy barrier of 3.0 kcal/mol between families, the probability to sample a conformation exactly at the top of this barrier is 0.7%. In a sample of 500 conformations, for instance, considering a large family that includes 20% of the total, such conformations will be encountered with low probability. Increasing the size of the sample also increases the probability to find such conformations. Following that, the probability that the algorithm would shift from one family to another, before spanning an entire family, increases. Such a shift is easily detected, by comparing the RMSD within families to its values between pairs of families. A “split” family (Figure 10) would be represented as two or more distinct blocks that are very similar, as shown for cyclic [GSAGPA]. However, the method is not only applicable to samples in which minimal energy conformations were collected, but to any conformational search method with a large collection of conformers. The geometric families are interesting even if the low energies, “valleys” or “basins” were not the focus of research.

The method was applied to analyze conformational spaces from simulations of two cyclic peptides, Cyclic $(Ala)_6$ and [GSAGPV]. A general family like topology was revealed in the two cases, and larger conformational samples displayed a similar behavior. Since the algorithm does not force the conformational space into a certain topology, one may think of the results as an additional explicit proof to the existence of this topology in the conformational space of these molecules. Additionally, we were able to show that the energy profile of each family, as revealed by the algorithm, resembles the energy profile of energy basins. However, an exact correlation between basins and families requires a more elaborate study by performing transition state calculations between all conformers, which is beyond the scope of this paper. In a study of a cyclic hexapeptide, Shenkin and McDonald[12] concluded that “...the molecule exhibits small groups of substructures with similar energies and similar ring geometries but varying side-chain geometries”. In our case, the link between structures and energies has been found to be much stronger.

The convergence of the sampling methodology is related to the sample size and may be detected by studying successively increasing sample sizes of the molecule. The ability to find the family partition was shown to be useful for examining this stability. Examining several samples of conformations, each of different size, identified the same major clusters in most of the larger samples, and a minor effect of the sample size on the families’ population was observed. However, it is easy to use the successive sample test as we did in this paper to probe the convergence limits.

An explicit comparison was made between the families of conformations found by the nearest single neighbor method to those found by the family clustering method. The results emphasize a well-known problem of using cutoff values within the frame of clustering conformations. One can never be sure if a cutoff correctly describes the families in the conformational space. Consequently, the results depend heavily on the cutoff used.

The nearest single neighbor proposed could be further used to explore several questions related to molecular conformational preferences. Since the method depends only on the coordinates of the molecule under investigation, it could be applied to compare between conformational families generated via different solvation models. The method can also be applied to analyze structural problems. One such question is the relevance of the molecular structure of a ligand in the free state and in bound conformations. Finally, since this method is not necessarily restricted to molecular structures, it may be found useful for analyzing other systems that need to be compared on the basis of their topology.

ACKNOWLEDGMENT

We thank the referees for illuminating the important issue of invariance to the choice of seed conformation as well as for raising other points to strengthen this paper. D.C. thanks Jacob Levy and Oren M. Becker for making their Cyclic (ala)$_6$ data available to us. D.C. also thanks Prof. Mordechai Bixon and Dr. Andrea Zaliani for helpful discussions and suggestions.

REFERENCES AND NOTES

(1) Leach, A. R. In *Reviews in Computational Chemistry*; Lipkowitz, K. B.; Boyd, D. B., Eds.; VCH: New York, 1991; Vol. 2, pp 1−55.

(2) Hampel, J. C.; Fine, R. M.; Hassan, M.; Ghoul, W.; Guaragna, A.; Koerber, S. C.; Li, Z.; Hagler, A. T. Conformational analysis of endothelin-1: Effects of solvation free energy. *Biopolymers* **1995**, *36*, 282−301.

(3) Bravi, G.; Gancia, E.; Zaliani, A.; Pegna, M. SONHICA (simple optimized Nonhierarchical cluster analysis): A new toll for analysis of molecular conformations. *J. Comput. Chem.* **1997**, *18*, 1295−1311.

(4) a) Feher, M.; Schmidt, J. M. Metric and multidimensional scaling: efficient tools for clustering molecular conformations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 346−53. (b) Gordon, H. L.; Somoraji, R. L. Fuzzy cluster analysis of molecular dynamic trajectories. *Proteins* **1992**, *14*, 249−264.

(5) Veber, D. F.; Holy, F. W.; Paleveda, W. J.; Nutt, R. F.; Bergestrand, S. J.; M., T.; Glitzer, M. S.; Saperstein, R.; Hirshmann, R. Conformational restricted bicycle analogues of somatostatin. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *262*, 2636−2640.

(6) Pierschbacher, M. D.; Rouslahti, E. Influence of stereochemistry of the sequence Arg-Gly-Asp-Xaa on binding specificity in cell adhesion. *J. Biol. Chem.* **1987**, *262*, 17294−17298.

(7) Shenderovich, M. D.; Nikiforovich, G. V.; Golbraikh, A. A. Conformational features responsible for the binding of cyclic analogues of enkephalin to opioid receptors. *Int. J. Peptide Protein Res.* **1991**, *37*, 241−251.

(8) (a) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, pathways, and energy landscape of protein folding: A synthesis. Proteins **1995**, *21*, 167−195. (b) 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*, 11158−62.

(9) Caves, L. S. D.; Evanseck, J. D.; Karplus, M. Locally accessible conformations of proteins: Multiple dynamics simulations of Crambin. *Protein Sci.* **1998**, *7*, 649−666.

(10) Elber, R.; Karplus, M. Multiple conformational states of proteins: A molecular dynamic analysis of myoglobin. *Science* **1987**, *235*, 318−321.

(11) Torda, A. E.; Van Gunsteren, W. F. Algorithm for Clustering Molecular Dynamic Configurations. *J. Comput. Chem.* **1994**, *15*, 1331−1340.

(12) Shenkin, P. S.; Mcdonald, D. Q. Cluster Analysis of molecular conformations. *J. Comput. Chem.* **1994**, *15*, 899−916.

(13) Howard, A. E.; Kolmann, P. A. An analysis of current methodologist for conformational searching of complex molecules. *J. Med. Chem.* **1988**, *31*, 1669.

(14) Verkhiver, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering common failures in molecular docking of ligand-protein complexes. *J. Comput. Aid. Mol. Design* **2000**, *14*, 731−751.

(15) Allen, F. H.; Doyle, M. J.; Taylor, R. Automated conformational analysis from crystallographic data 1. A symmetry-modified single-linkage clustering algorithm for three-dimensional pattern recognition. *Acta. Crystallogr.* **1991**, *B47*, 29−40.

(16) Jarvis, R. A.; Patrick, E. A. clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Comput*. **1973**, *c-22*, 1025−1034.

(17) Becker, O. M. geometric versus topological clustering: an insight into conformation mapping. *Proteins* **1997**, *27*, 213−226.

(18) Vlijmen, H. v.; Karplus, M. PDB-based protein loop prediction: parameters for selection and methods for optimization. *J. Mol. Biol.* **1997**, *267*, 975−1001.

(19) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamic calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.

(20) MacKerrel, A.; Bashford, J. D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom empirical potential for molecular modeling and dynamic studies of proteins *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(21) Levy, Y.; Becker, O. M. Effect of conformational constraints on the topology of complex potential energy surface. *Phys. Rev. Lett.* **1998**, *81*, 1126−1129.

(22) Noguti, T.; Go, N. Structural Basis of hierarchical Multiple substates of a protein. *Proteins* **1989**, *5*, 97−103.

(23) Brooks, C. L.; Case, D. A. Simulations of peptide conformational dynamics and thermodynamics. *Chem. Rev.* **1993**, *93*, 2487−2502.

CI0255735