Density-Based Clustering of Small Peptide Conformations Sampled from a Molecular Dynamics Simulation

Minkyoung Kim,[†] Seung-Hoon Choi,[†] Junhyoung Kim,[†] Kihang Choi,[‡] Jae-Min Shin,[§] Sang-Kee Kang,[||] Yun-Jaie Choi,[||] and Dong Hyun Jung*,[†]

Insilicotech Co. Ltd., A-1101, Kolontripolis, 210, Geumgok-Dong, Bundang-Gu, Seongnam-Shi 463-943, Korea, Department of Chemistry, Korea University, Anam-dong, Seongbuk-Gu, Seoul 136-701, Korea, SBScience Co. Ltd., B-1212, Kolontripolis, 210, Geumgok-Dong, Bundang-Gu, Seongnam-Shi 463-943, Korea, and School of Agriculture Biotechnology, Seoul National University, San 56-1, Shilim-Dong, Kwanak-gu 151-742, Korea

Received December 1, 2008

This study describes the application of a density-based algorithm to clustering small peptide conformations after a molecular dynamics simulation. We propose a clustering method for small peptide conformations that enables adjacent clusters to be separated more clearly on the basis of neighbor density. Neighbor density means the number of neighboring conformations, so if a conformation has too few neighboring conformations, then it is considered as noise or an outlier and is excluded from the list of cluster members. With this approach, we can easily identify clusters in which the members are densely crowded in the conformational space, and we can safely avoid misclustering individual clusters linked by noise or outliers. Consideration of neighbor density significantly improves the efficiency of clustering of small peptide conformations sampled from molecular dynamics simulations and can be used for predicting peptide structures.

1. INTRODUCTION

Conformational analysis is an important technique for exploring the structures of peptides and relating them to their properties. Because understanding of peptide conformations provides important insights into the specificity and potency of peptide drugs, structure prediction could help to evaluate possible conformations prior to synthesis, thus avoiding wasteful trial and error experiments on peptide modification.[1−4] Often, however, the number of conformations generated by a conformational sampling method is very large, so it is not feasible to test them all in terms of receptor-binding geometry. Hence, it becomes essential to use data reduction techniques such as clustering first to identify well-defined groups of conformations and then to select representative geometries from each group that can be used as putative bioactive peptide conformations.[5]

The word "clustering" generally implies the identification of groups, such that the similarities within any group are significantly greater than those among groups.[6] A number of methods exist for clustering conformations,[7−14] and the most popular ones for selecting representative conformations are hierarchical clustering[8] and nearest-neighbor Jarvis−Patrick schemes.[1] Simulated annealing with conformational energy as the clustering criterion[15] and, more recently, multidimensional and metric scaling,[6] fuzzy clustering,[12] and nearest single neighbor (NSN)[1] have been used to cluster families of conformations with promising results.

A density-based algorithm has been applied to many data-mining studies because of its efficiency and scalability in clustering data sets. The main idea of this approach is to find regions of high and low data density, such regions being separated. Although a density-based approach can facilitate the discovery of clusters of arbitrary shape and size,[16−18] this algorithm, to the best of our knowledge, has not previously been used to cluster small peptide conformations. In this work, we apply a density-based algorithm to cluster conformations sampled from a long molecular dynamics (MD) trajectory of two small peptide systems. One is alanine dipeptide, and the other is cyclic [GSAGPV]. Alanine dipeptide (*N*-acetyl-L-alanine-*N*′-methylamide, AcAlaNHMe) consists of two alanines connected by a peptide bond; because of its simple structure, it has been studied extensively as a model biomolecule.[19−24] We also apply the clustering algorithm to cyclic [GSAGPV] as a more realistic peptide model and examine the general applicability of the algorithm to the conformational study of small peptides sampled from molecular dynamic simulations.

2. MATERIALS AND METHODS

All molecular modeling tasks in this work were performed with Discovery Studio version 2.0[25] and SciTegic Pipeline Pilot version 6.1.5[26] for molecular dynamics simulations and a standard superimposition protocol for conformations within a cluster. Hierarchical clustering results were obtained by running the complete-linkage Hierarchical Clustering Analysis (HCA) method in Cerius2 version 4.10.[27]

2.1. Similarity Measures. Clustering of molecular systems is based on a distance measure between pairs of conformations, and the root-mean-square deviation (RMSD) is frequently used as the measure of differences between molecular structures. We compared all pairwise frames from a MD simulation trajectory and calculated an $N \times N$ RMSD

* Corresponding author phone: +82-31-728-0443; fax: +82-31-728-0444; e-mail: dhjung@insilicotech.co.kr.
[†] Insilicotech Co. Ltd.
[‡] Korea University.
[§] SBScience Co. Ltd.
[||] Seoul National University.

10.1021/ci800434e CCC: $40.75 © 2009 American Chemical Society
Published on Web 10/28/2009

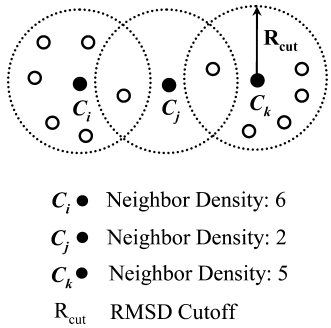


Figure 1. Concept of RMSD cutoff and neighbor density.

matrix for the backbone atom (N, C_α, C, O) group. The atom group was superimposed before the RMSD calculation:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\delta_i^2} \quad (1)$$

where δ is the distance between a pair of equivalent backbone atoms, and N is the number of the pairs.

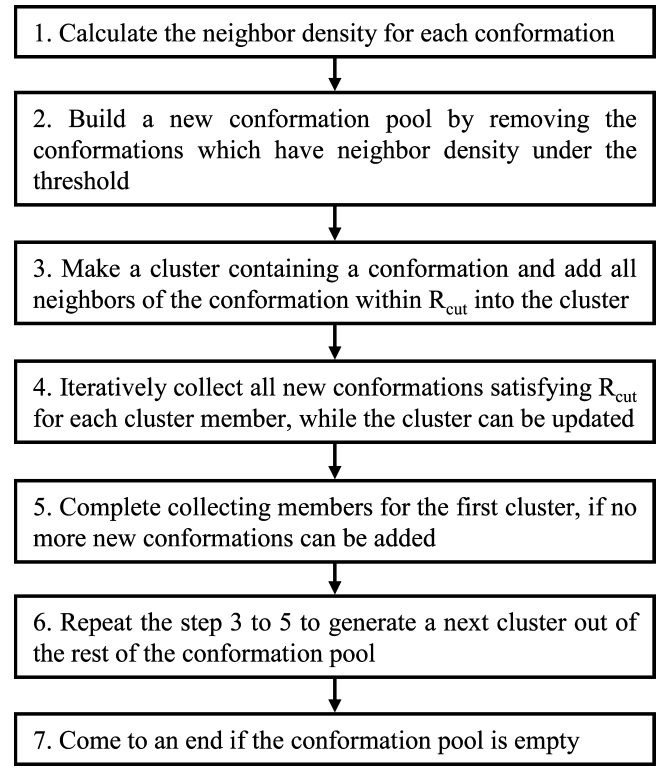
2.2. Generatin of Conformations. All energy minimizations and MD simulations were carried out using the CHARMm 33b1[28,29] program of the Discovery Studio simulation package. An initial structure with the lowest energy was selected from 100 random starting structures, which had been generated by random assignment of main chain torsion angles followed by energy minimization (maximum 5000 steps). A sample of 500 conformations was obtained from a 20 ns MD trajectory at 300 K.

To consider the effect of hydration, the Generalized Born (GB) implicit solvation model[30,31] and an explicit solvation model were applied to the MD simulations. GB solvation energies and forces were calculated by the pairwise sum of the atomic volumes used to calculate the effective Born radii in the Coulomb field approximation. For explicit solvation under periodic boundary conditions (PBC), the peptides were solvated in a cubic box explicitly containing water molecules. The size of PBC is 21.41 Å for alanine dipeptide and 26.87 Å for cyclic peptide. During the simulations, the Particle Mesh Ewald method[32] was employed to account for long-range electrostatic interactions within a periodic system.

2.3. The Density-Based Clustering Algorithm. Conformations within R_{cut} are considered as neighbors of the center point of the same cluster, and the number of neighboring conformations is defined here as the neighbor density. The key idea of the density-based clustering algorithm is that the neighbor density of a conformation C_i must exceed some threshold for C_i to be a cluster member. Otherwise, it corresponds to a noise point in the data set. The concept of R_{cut} and neighbor density is illustrated in Figure 1, in which we can easily and unambiguously detect clusters of points and noise points that do not belong to any of the clusters. When we set the threshold to 4, C_i and C_k have enough neighbor density to be cluster members, but C_j is classified as a noise point because its neighbor density is only 2.

In this work, the density-based clustering algorithm was used to cluster peptide conformations. A flow chart of the procedure is presented in Scheme 1. The algorithm starts by calculating the neighbor density for each conformation (step 1). Any conformations with neighbor density below the threshold value are removed from the conformation pool and

Scheme 1. Flow Chart of the Density-Based Clustering Algorithm



defined as noise or outliers (step 2). To initiate clustering, we make a cluster containing a conformation randomly selected from the new pool, and then all neighbors within the given R_{cut} from the selected conformation are added to it (step 3). Now we have the newly generated first cluster, and all neighbors within the R_{cut} for each conformation in that cluster are retrieved into it without redundancy. To finish making the first cluster, all neighbor conformations of all cluster members that meet the R_{cut} criterion are iteratively collected into it and removed from the conformation pool. The process repeats iteratively while the cluster is updated with new conformations (step 4). If no further conformations can be added, the collection of members of the first cluster is completed (step 5). To identify the second cluster, the algorithm repeats from step 3, building a new cluster with a randomly selected conformation among those remaining in the pool (step 6). When all conformations in the pool are consumed, the algorithm comes to an end (step 7). As mentioned previously, the algorithm was applied to two peptide systems to discover clusters in conformational space and their representative conformations in the MD trajectory.

2.4. Complete-Linkage Hierarchical Clustering Algorithm. The traditional hierarchical clustering algorithm was applied to conformations from a long molecular dynamics trajectory to compare the results with those of the density-based clustering algorithm. The traditional algorithm creates a hierarchical structure of clustering, visually represented as a dendrogram displaying levels of the hierarchy according to values of an objective function. The objective function measures the overall dissimilarity within clusters, and the optimal partition is obtained by minimizing it. In the complete-linkage hierarchical clustering algorithm, the objective function considers the distance between one cluster and another to equal the greatest distance from any member of

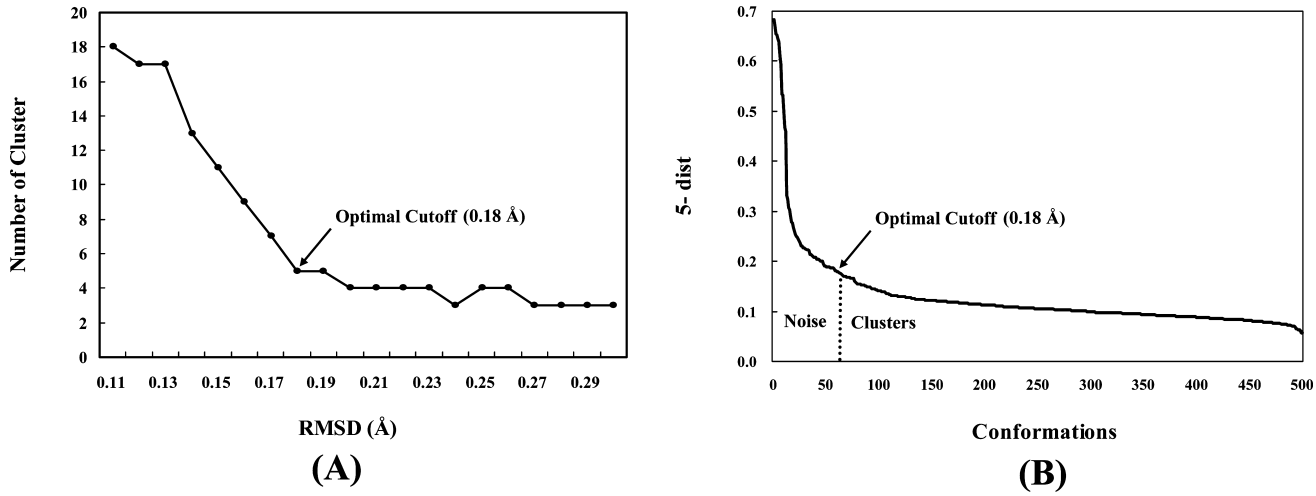


Figure 2. (A) The number of clusters according to RMSD cutoff. (B) The sorted *5-dist* graph for alanine dipeptide conformations retrieved from the MD trajectory using the implicit GB solvation model.

one cluster to any member of the other. The objective function of complete-linkage method is as follows:

$$d(\mathrm{A},\mathrm{B}) = \max_{i\in \mathrm{A}/j\in \mathrm{B}} d_{ij} \qquad (2)$$

where d means distance between clusters A and B.

3. RESULTS AND DISCUSSION

To evaluate and characterize the quality of our density-based clustering algorithm comprehensively, we applied the method to alanine dipeptide and cyclic [GSAGPV] peptide conformations sampled from the MD trajectory under the implicit GB or the explicit hydration condition at 300 K.

3.1. Clustering Alanine Dipeptide Conformations. Alanine dipeptide has long served as a major model for theoretical studies of the conformational space of proteins and peptides. It has many structural features of polypeptide backbones including flexible Φ and Ψ angles, a peptide group (NHCO), and methyl groups attached to C_α positions. Calculations on this system have therefore been used to prove the validity of various conformational analysis techniques.[19−24]

The performance of the density-based clustering algorithm depends strongly on the RMSD cutoff. To determine an optimal RMSD cutoff for the data set, we used two different approaches. First, the number of clusters was monitored with increasing RMSD cutoff values. Figure 2A shows a gradual decrease in the number of clusters, and after a cutoff value of 0.18 Å a slightly rugged plateau begins to form. In the second approach, we used a simple but effective heuristic method previously developed by Martin Ester et al.[16] In brief, for a given neighbor density k, each point in the data set was mapped to the distance from the point itself to its kth nearest neighbor (*k-dist* value), and the points were sorted in descending order of *k-dist* values. The RMSD cutoff was determined from a point near the first "valley" of the sorted *k-dist* graph. Figure 2B, for example, shows the *5-dist* graph for the conformation set used in the first method. All points with higher *5-dist* values (left of the optimal cutoff) are considered noise and are discarded because the conformations are regarded as low density regions according to the density distribution plot. All other points (right of the optimal cutoff) are assigned as cluster members. In the case of alanine dipeptide, the *k-dist* graphs for k values of 3−8 show no significant differences (data not shown), so we set the neighbor density threshold to 5. We found that these two approaches led to the same RMSD cutoff value, 0.18 Å.

We first examined the conformations generated from the MD simulation by the implicit GB solvation model. The results of the density-based algorithm for clustering peptide conformations are summarized using a Ramachandran plot[33] (Figure 3), which projects the conformational space onto the Φ−Ψ 2D plane. When the conformations are clustered without considering neighbor density (neighbor density threshold (ND) = 0), most of them are assigned to a single cluster C-1-0 (Figure 3A; each cluster is labeled C-*i*-*j*, where i is an index of the cluster and j is the neighbor density threshold). When the neighbor density threshold is set to 5, however, two separate clusters (C-1-5 and C-2-5) are generated, and some noise points in the boundary region have disappeared (Figure 3B). These clusters correspond to the two preferred secondary structures of the dipeptide (C-1-5, β-sheet; C-2-5, right-handed α-helix).

In Figure 4, the backbone conformations of alanine dipeptide are superimposed to compare the clustering results further. The conformations within each cluster generated with a neighbor density threshold of 5 superimpose well, showing good overall structural similarity (Figure 4B). In contrast, the conformations within cluster C-1-0, generated without considering neighbor density (Figure 4A), do not superimpose well mainly because of significant variations in the dihedral angle Ψ. The results clearly indicate that the introduction of neighbor density improves the performance of the clustering method and discriminates conformations with high similarity from others. When neighbor density is considered, the algorithm easily detects and filters out the noise points that do not belong to any clusters, thus preventing the creation of too many small clusters.

For the explicit solvation model of the alanine dipeptide system, the distribution of the sampled conformations was more diffuse, and the RMSD cutoff was determined as 0.25 Å using the two methods previously discussed. As shown in Figures 5 and 6, the clustering results are similar to those from the implicit solvation model. The explicit solvation model generates the conformational space with enough diversity as compared to the implicit model and good representation of additional cluster (C-2-0 and C-3-5; left-

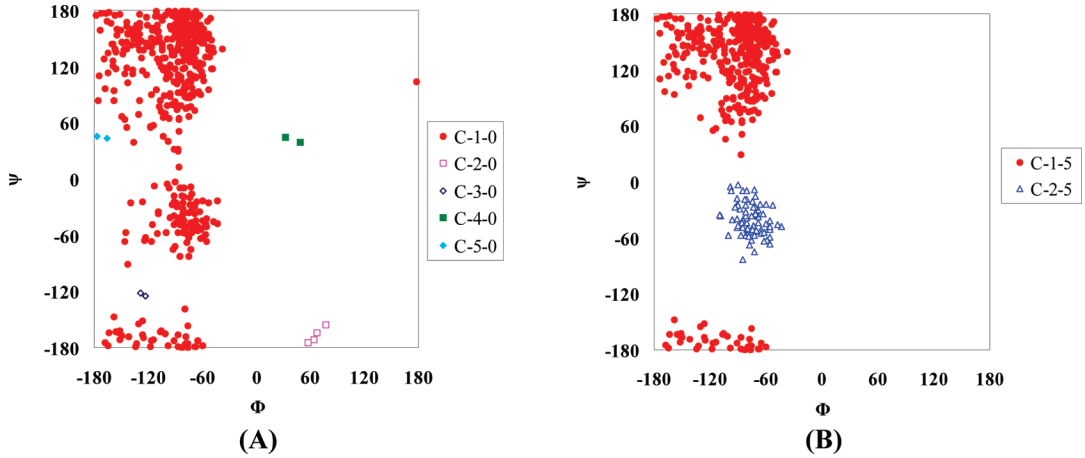


Figure 3. Ramachandran plot of alanine dipeptide clustered with RMSD 0.18 Å and neighbor densities 0 (A) and 5 (B) from the MD trajectory using the implicit GB solvation model. For the C-*i*-*j* code, *i* is the cluster index and *j* is the neighbor density threshold.

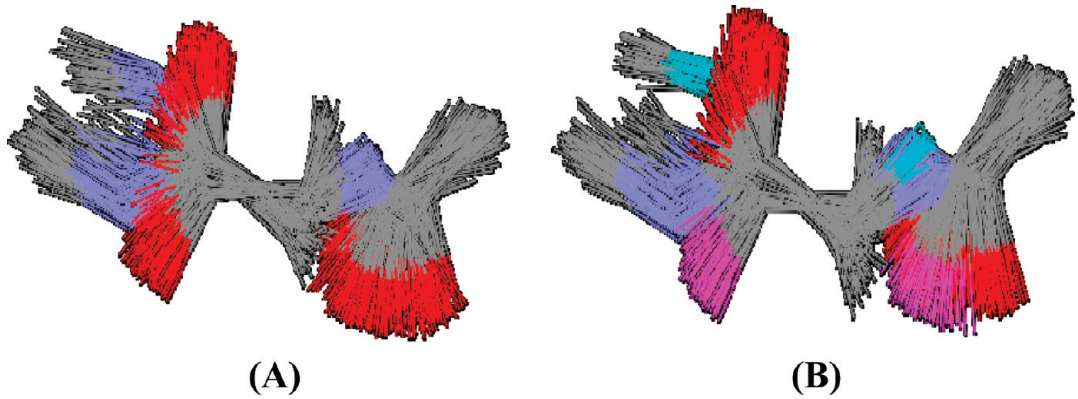


Figure 4. Superimposition of backbone conformations of alanine dipeptide clustered with RMSD 0.18 Å and neighbor densities 0 and 5 from the MD trajectory using the implicit GB solvation model. (A) C-1-0, (B) C-1-5 and C-2-5, with the following colors: oxygen (red), nitrogen (blue), and carbon (gray) in C-1-0 and C-1-5; oxygen (pink), nitrogen (sky blue), and carbon (gray) in C-2-5.

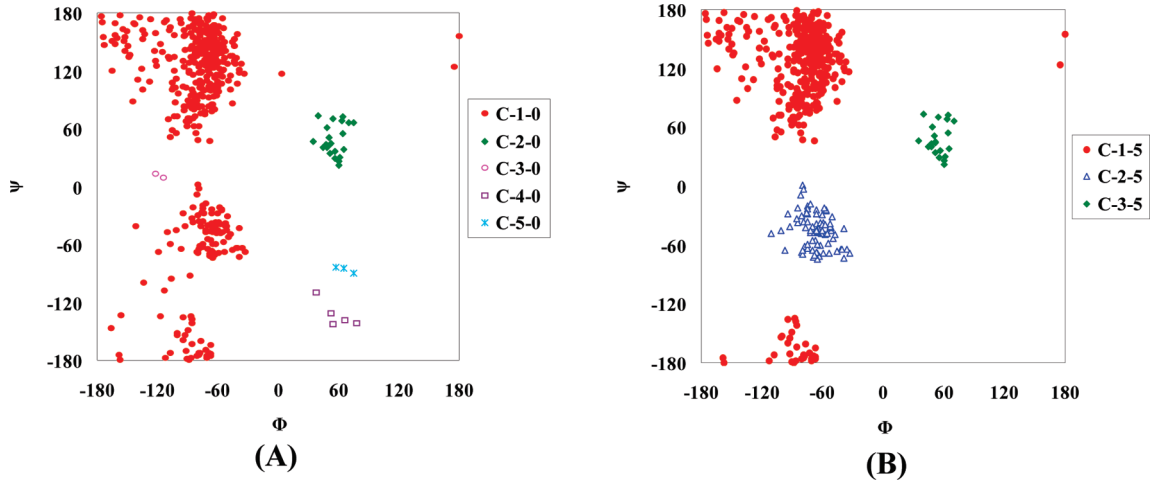


Figure 5. Ramachandran plot of alanine dipeptide clustered with RMSD 0.25 Å and neighbor densities 0 (A) and 5 (B) from the MD trajectory using the explicit water solvation model.

handed α-helix). In Figures 5B and 6B, we can observe three well-separated clusters (C-1-5, *β*-sheet; C-2-5, right-handed α-helix; C-3-5, left-handed α-helix). The *β*-sheet and α-helix conformations, clustered together in C-1-0 (Figures 5A and 6A), are separated into two clusters by considering the neighbor density (ND = 5). When the neighbor density was not considered in making clusters, noise points could play a role to link two separate clusters, resulting in one cluster. On taking account of the neighbor density, we can identify the noise points, and by removing them the two clusters can be successfully divided. The clusters C-3-0, C-4-0, and C-5-0 disappear because they can not satisfy the threshold (ND = 5). The results clearly show that the density-based clustering algorithm is helpful for finding peptide conformation families under both solvation conditions.

3.2. Clustering Cyclic [GSAGPV]. We next investigated the cyclic peptide GSAGPV as an example of another type of peptide molecule. Cyclic peptides found in nature range

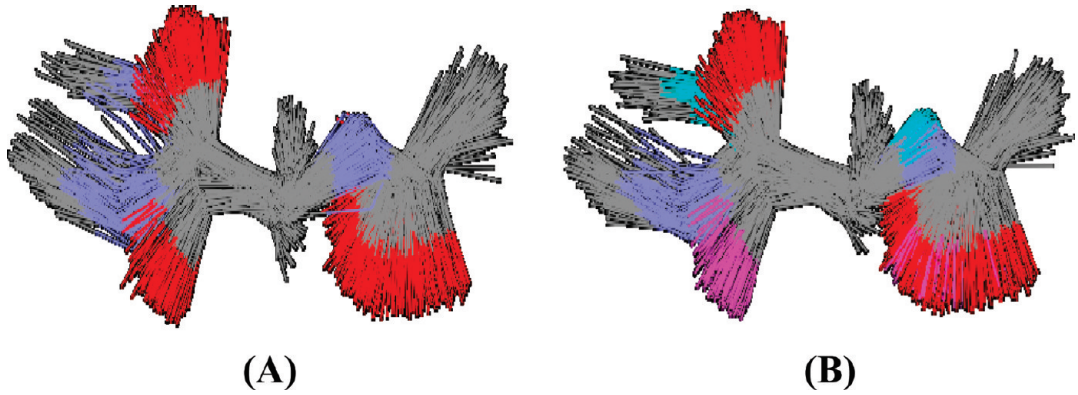


Figure 6. Superimposition of backbone conformations of alanine dipeptide clustered with RMSD 0.25 Å and neighbor densities 0 and 5 from the MD trajectory using the explicit hydration model. (A) C-1-0, (B) C-1-5 and C-2-5, with the following colors: oxygen (red), nitrogen (blue), and carbon (gray) in C-1-0 and C-1-5; oxygen (pink), nitrogen (sky blue), and carbon (gray) in C-2-5.

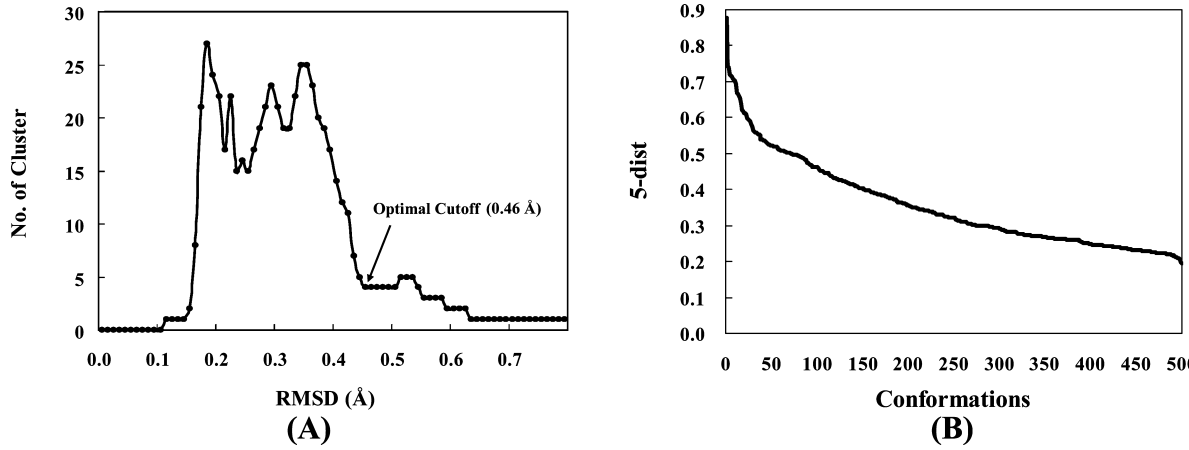


Figure 7. (A) The number of clusters according to RMSD cutoff. (B) The sorted *5-dist* graph for cyclic [GSAGPV] conformations retrieved from the MD trajectory using the explicit hydration model.

in size from a few to hundreds of amino acids. Because they tend to be extremely resistant to digestion, they are considered for use as scaffold structures for protein-based drugs that can be delivered orally.[34]

The conformations of cyclic [GSAGPV] were retrieved from the MD trajectory under the explicit hydration condition. The RMSD cutoff parameter was determined as 0.46 Å following the methods used for alanine dipeptide. The number of clusters fluctuated in the low RMSD cutoff region (Figure 7A), but after 0.35 Å it showed a steep decrease and reached a plateau beginning from 0.46 Å. In contrast, the sorted *5-dist* graph does not give the optimal position for determining the starting point of the "valley" (Figure 7B).

Figure 8 displays the superimposed conformations in each cluster collected using four different neighbor density thresholds (ND = 0, 5, 6, and 8). The result obtained with a threshold of 7 is not shown because it is very close to that obtained with the value of 6. From Figure 8, we can clearly see that the clusters may be separated further by increasing the neighbor density threshold (from the top to the bottom). The 464 conformations of cluster 1 obtained without considering neighbor density (C1000) are divided into two groups, C1100 and C1200, when the neighbor density threshold is set to 5. Cluster C1100 is divided into two further groups, C1110 and C1120, at ND 6; and, finally, clusters C1121 and C1122 bifurcate from cluster C1120 at ND 8.

To verify the structural differences between the clusters (Table 1), we performed a quantitative analysis by examining

Table 1. Structural Differences between Cyclic [GSAGPV] Clusters by Complete-Linkage Hierarchical Clustering

cluster	no. of conformations	average RMSD value (Å)	standard deviation (Å)
C-1-30	84	0.442	0.1218
C-2-30	169	0.397	0.1576
C-3-30	41	0.547	0.2306
C-4-30	23	0.497	0.2462
C-5-30	9	0.302	0.2078
C-6-30	5	0.503	0.3788
C-7-30	51	0.54	0.1935
C-8-30	33	0.418	0.2164
C-9-30	43	0.535	0.1656
C-10-30	2	0.492	0.6961
C-11-30	40	0.316	0.1014
C-1-40	294	0.537	0.2259
C-2-40	88	0.682	0.2818
C-3-40	76	0.683	0.217
C-4-40	42	0.366	0.2354

the average RMSD between cluster members calculated by eq 3 and the standard deviation thereof:

$$\text{average RMSD} = \min\left(\frac{\sum_{j,j\neq 1}^{n}\text{RMSD}_{1j}}{n-1}, \frac{\sum_{j,j\neq 2}^{n}\text{RMSD}_{2j}}{n-1}, \ldots, \frac{\sum_{j,j\neq n}^{n}\text{RMSD}_{nj}}{n-1}\right) \quad (3)$$

where RMSD_{ij} is the RMSD between conformations i and j, and n is the number of cluster members. Because dissimilar

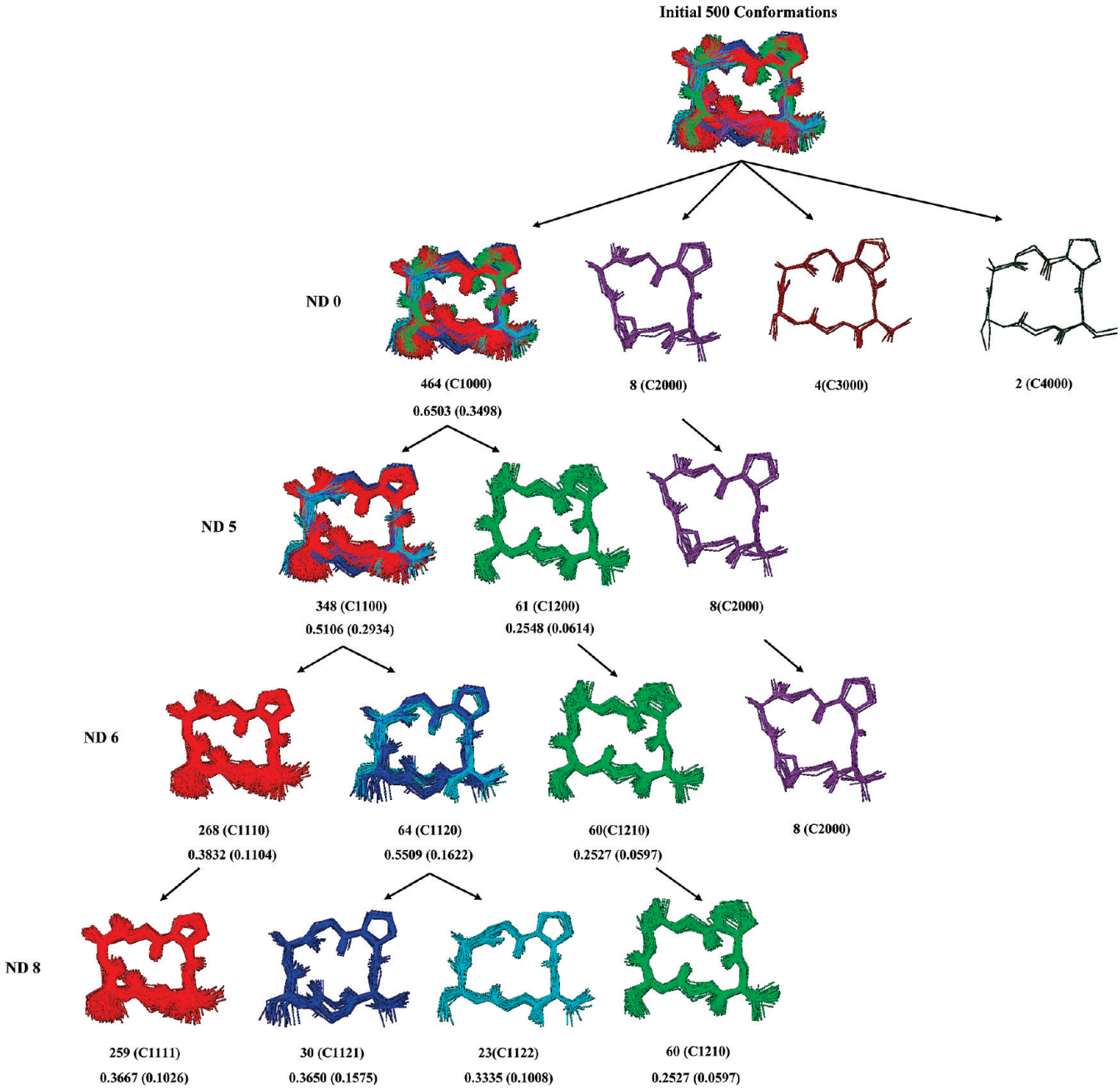


Figure 8. Superimposed cyclic [GSAGPV] conformations in each cluster at different neighbor density thresholds with RMSD cutoff 0.46 Å. The first symbol, *i*(*j*), immediately below the peptide conformations, represents the number of conformations superimposed (*i*) and an index (*j*) of the corresponding cluster. For example, cluster 1 at neighbor density level (ND) 0 has cluster index C1000; it branches into two clusters, C1100 and C1200, at ND 5. The second symbol, *k*(*l*), represents the average RMSD value *k* and the standard deviation *l* calculated for each cluster. The clusters at ND 8 are drawn in different colors, and those at other ND levels are drawn in the color of the corresponding conformation at ND 8.

objects are pulled into other clusters with increasing neighbor density threshold, the similarity between the remaining members becomes greater, so the average RMSD and standard deviation values gradually decrease (Figure 8). One exception is that the average RMSD of C1120 at ND 6 (0.5509) is slightly higher than that of C1100 at ND 5 (0.5106). This may be because C1120 includes two or more distinct clusters, and, as expected, it can be split again into C1121 and C1122 at neighbor density 8. Finally, all clusters at ND 8 have average RMSDs less than the RMSD cutoff value (0.46 Å) previously determined. Thus, comparison between the average RMSD of the members of a cluster and the RMSD cutoff may be a criterion for deciding whether the cluster members have a common structure.

From these results, we find that the density-based clustering method easily detects noise points with relatively low neighbor density and, as a result, can determine the distinct families in the molecular conformation space. We also find that the clustering results can be verified easily by examining the superimposed structures and their quantitative analysis results.

3.3. Comparison with the Hierarchical Clustering Method. To validate the quality of our clustering procedure, the hierarchical clustering method was applied to alanine dipeptide conformations from an implicit solvation MD and to cyclic [GSAGPV] conformations from an explicit solvation MD, and the results were compared to those from the

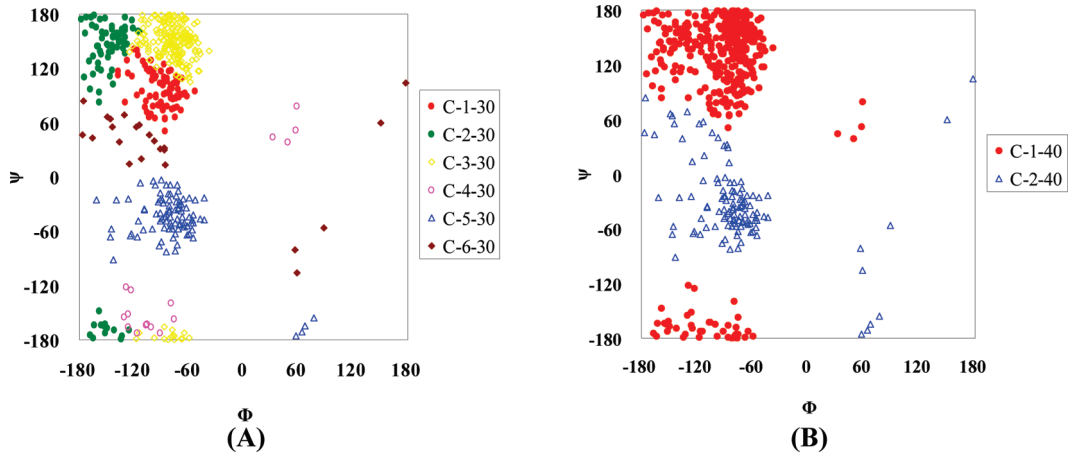


Figure 9. Ramachandran plot of alanine dipeptide clustered by the complete-linkage hierarchical method with objective function thresholds 30 (A) and 40 (B) from the MD trajectory using the implicit GB solvation model. For the C-*i*-*j* code, *i* is the cluster index and *j* is the objective function threshold.

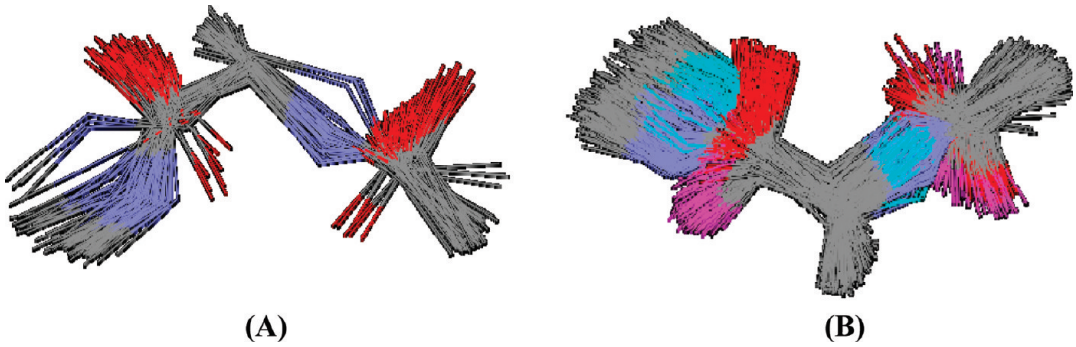


Figure 10. Superimposition of backbone conformations of alanine dipeptide clustered by the complete-linkage hierarchical method with objective function thresholds 30 (A) and 40 (B) from the MD trajectory using the implicit GB solvation model. (A) C-5-30, (B) C-1-40 and C-2-40, with the following colors: oxygen (red), nitrogen (blue), and carbon (gray) in C-5-30 and C-1-40; oxygen (pink), nitrogen (sky blue), and carbon (gray) in C-2-40.

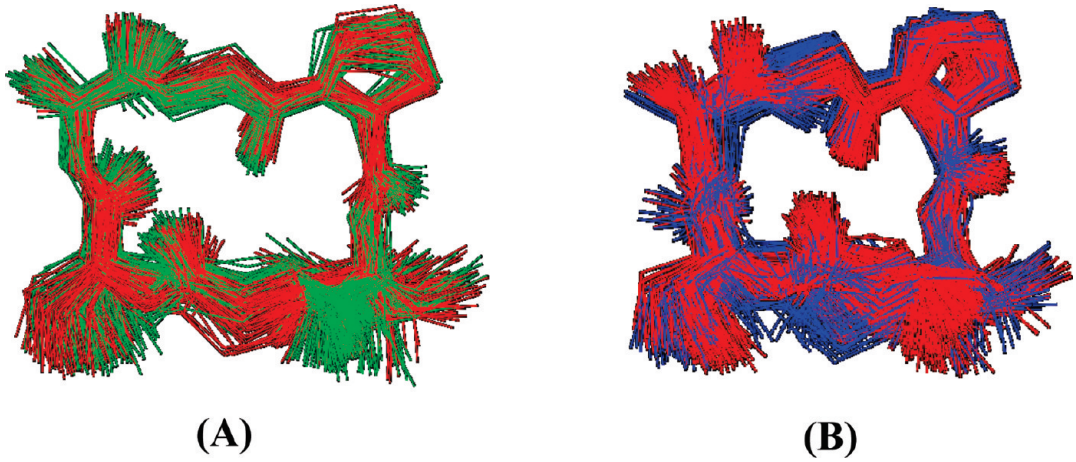


Figure 11. Superimposition of backbone conformations of (A) C-1-30 (red) and C-2-30 (green), and (B) C-1-40 (red) and C-2-40 (blue).

density-based method. The clustering results were investigated at two different values of objective function threshold, 30 and 40. A clustering dendrogram displaying each cluster merge and a graph of the objective function are available as Supporting Information.

The results in Figures 9 and 10 display the failure of sophisticated clustering in the hierarchical method with the alanine dipeptide conformations. As shown in Figure 9A, for objective function threshold 30, the β-sheet region is divided into three different clusters (C-1-30, C-2-30, and C-3-30). Three clusters in the β-sheet region shown in Figure 9A seem not to be a reliable clustering result because the members of each cluster so highly diffuse in other clusters that the discrimination from each other is difficult. The preceding studies[24,35] show that β-sheet regions comprise the three or four distinct $\Phi-\Psi$ sections, but the free energies of conformations are not much different from the lowest energy structure ($\Delta E < 0.9$ kcal/mol). Therefore, it is expected for the sampled conformations to have relatively uniform distribution through the β-sheet region in MD simulation. Eventually, conformations in β-sheet region sampled by MD simulations at 300 K could not be separated

by our method. In this regard, we will discuss in more detail later. One more thing we want to point out is that with both object function thresholds, 30 and 40, the clustering results show unsatisfactory discrimination between the β-sheet and α-helix conformations. C-2-40 seems to include several members that should have been in the β-sheet conformation clusters (Figure 9B). Figure 10A and B shows the conformational spread of members within the cluster C-5-30 and between C-1-40 and C-2-40 by molecular superimposition.

The hierarchical clustering method was also applied to cyclic [GSAGPV] with the same objective function thresholds, and the statistical analysis and superimposition results are given for comparison. The conformational diversity observed among members of a cluster is significantly greater than that obtained from density-based clustering. The degree of conformational spread is also confirmed in Figure 11.

3.4. Limitation of our Method. Because a conformational space is explored continuously during a molecular dynamics simulation, the sets of conformations obtained present a near-continuum in which a low-density region can be considered as noise. In particular, the cyclic peptide system is a relatively rigid and conformationally restricted model. Ring closure restricts the possible range of conformations as compared to a linear peptide, so it induces deep wells in the conformational potential surface. Under such circumstances, the density-based clustering algorithm is applicable and gives satisfactory results. However, low and uniform conformation density is intrinsic to systems with relatively flat potential surfaces such as linear polymer chains and molecules with free rotatable bonds. In such cases, our approach may eliminate too many meaningful conformations and generate too few clusters for consideration. Therefore, another approach should be introduced to deal with those systems.

4. CONCLUSION

To demonstrate the applicability and efficiency of a density-based algorithm for clustering peptide systems, we applied it to data sets consisting of 500 conformations collected from MD simulation trajectories. When conformations from MD simulations are clustered, those between two high density regions may distort the results strongly: members in high density regions connected by noise points are often merged and assigned to a single cluster. To improve clustering performance by eliminating such distortions, we introduced the concept of neighbor density into our clustering algorithm.

The results for the two model systems, alanine dipeptide and cyclic [GSAGPV], demonstrate that the density-based algorithm is very effective in clustering peptide conformations by filtering out noise. Inspection of the clusters identified using Ramachandran plots, structure superimposition, and quantitative analysis confirms the improved performance of this algorithm. This efficient and reliable method for detecting clusters of conformational ensembles is expected to find applications in the prediction of structures of peptides, proteins, and other related compounds.

ACKNOWLEDGMENT

This work was supported by the Korea Science and Engineering Foundation (KOSEF) NRL Program grant funded by the Korean government (MEST) (No. R0A-2008-000-20024-1). We thank Accelrys Korea for provision of modeling software.

Supporting Information Available: Dendrogram and objective function graph of cyclic [GSAGPV] and alanine dipeptide conformations by complete-linkage hierarchical clustering method. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Chema, D.; Goldblum, A. The "nearest single neighbor" methods finding families of conformations within a sample. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 208–217.
(2) Veber, D. F.; Holy, F. W.; Paleveda, W. J.; Nutt, R. F.; Bergestrand, S. J.; Torchiana, M.; Glitzer, M. S.; Saperstein, R.; Hirshmann, R. Conformational restricted bicycle analogues of somatostatin. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *262*, 2636–2640.
(3) 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.
(4) 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. Pept. Protein Res.* **1991**, *37*, 241–251.
(5) Banejee, A.; Misra, M.; Venanzi, C. A.; Dave, R. N. Fuzzy clustering in drug design: Application to cocaine abuse. Fuzzy information, 2004. Processing NAFIPS '04. *IEEE Annu. Meeting* **2004**, *1*, 308–313.
(6) Feher, M.; Schmidt, J. M. Metric and multidimensional scaling: Efficient tools for clustering small peptide conformations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 346–353.
(7) Murray-Rust, P.; Raftery, J. Computer analysis of molecular geometry, Part VI: Classification of differences in conformation. *J. Mol. Graphics* **1985**, *3*, 50–59.
(8) Shenkin, P. S.; McDonald, D. Q. Cluster analysis of small peptide conformations. *J. Comput. Chem.* **1994**, *15*, 899–916.
(9) Torda, A. E.; Gunsteren, W. F. Algorithms for clustering molecular dynamics configurations. *J. Comput. Chem.* **1994**, *15*, 1331–1340.
(10) Vesterman, B.; Golender, V.; Golender, L.; Fuchs, B. Conformer clustering algorithm and its application for crown-type macrocycles. *J. Mol. Struct. (THEOCHEM)* **1996**, *368*, 145–151.
(11) Rayan, A.; Senderowitz, H.; Goldblum, A. Exploring the conformational space of cyclic peptides by a stochastic search method. *J. Mol. Graphics Modell.* **2004**, *22*, 319–333.
(12) Feher, M.; Schmidt, J. M. Fuzzy clustering as a means of selecting representative conformers and molecular alignments. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 810–818.
(13) Donate, L. E.; Rufino, S. D.; Canard, L. H.; Blundell, T. L. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. *Protein Sci.* **1996**, *5*, 2600–2016.
(14) Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *J. Chem. Theory Comput.* **2007**, *3*, 2312–2334.
(15) Perkins, T. D. J.; Dean, P. M. An exploration of a novel strategy for superposing several flexible molecules. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 155–172.
(16) Sander, S.; Ester, M.; Kriegel, H. P.; Xu, X. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Min. Knowl. Disc.* **1998**, *2*, 169–194.
(17) Pei, T.; Zhu, A. X.; Zhou, C.; Li, B.; Qin, C. A new approach to the nearest-neighbour method to discover cluster features in overlaid spatial point processes. *Int. J. Geogr. Inf. Sci.* **2006**, *20*, 153–168.
(18) Daszykowski, M.; Walczak, B.; Massart, D. L. Looking for natural patterns in data Part 1. Density-based approach. *Chemometr. Intell. Lab.* **2001**, *56*, 83–92.
(19) Okumura, H.; Okamoto, Y. Multibaric-multithermal molecular dynamics simulation of alanine dipeptide in explicit water. *Bull. Chem. Soc. Jpn.* **2007**, *80*, 1114–1123.
(20) Gould, I. R.; Kollman, P. A. Ab initio SCF and MP2 calculations on four low-energy conformers of *N*-acetyl-*N*′-methylalaninamide. *J. Phys. Chem.* **1992**, *96*, 9255–9258.
(21) Apostolakis, J.; Ferrara, P.; Caflisch, A. Calculation of conformational transitions and barriers in solvated systems: Application to the alanine dipeptide in water. *J. Chem. Phys.* **1999**, *110*, 2099–2108.
(22) Smith, P. E. The alanine dipeptide free energy surface in solution. *J. Chem. Phys.* **1999**, *111*, 5568–5579.

(23) Mackerell, A. D., Jr.; Feig, M.; Brooks, C. L., III. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Comput. Chem.* **2004**, *25*, 1400–1415.
(24) Chekmarev, D. S.; Ishida, T.; Levy, R. M. Long-time conformational transitions of alanine dipeptide in aqueous solution: Continuous and discrete-state kinetic models. *J. Phys. Chem. B* **2004**, *108*, 19487–19495.
(25) *Discovery Studio, version 2.0*; Accelrys Inc.: San Diego, CA, 2007.
(26) *SciTegic Pipeline Pilot, version 6.1.5*; SciTegic Inc.: San Diego, CA, 2007.
(27) *Cerius2, version 4.10*; Accelrys Inc.: San Diego, CA, 2005.
(28) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.
(29) MacKerell, A. D., Jr.; Brooks, B.; Brooks, C. L., III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. CHARMM: The Energy Function and Its Parameterization with an Overview of the Program. In *The Encyclopedia of Computational Chemistry 1*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R. E., Eds.; John Wiley & Sons: Chichester, UK, 1998; pp 271−277.
(30) Dominy, B.; Brooks, C. L., III. Development of a generalized Born model parameterization for proteins and nucleic acids. *J. Phys. Chem.* **1999**, *103*, 3765–3773.
(31) Nina, M.; Beglov, D.; Roux, B. Atomic Born radii for continuum electrostatic calculations based on molecular dynamics free energy simulations. *J. Phys. Chem. B* **1997**, *101*, 5239–5248.
(32) Darden, T.; York, D.; Pederson, L. Particle mesh Ewald: an Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
(33) Lovell, S. C.; Davis, I. W.; Arendale, W. B., III; de Bakker, P. I.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. Structure validation by C-alpha geometry: phi, psi and C-beta deviation. *Proteins* **2003**, *50*, 437–450.
(34) Craik, D. J. Seamless proteins tie up their loose ends. *Science* **2006**, *311*, 1563–1567.
(35) Zimmerman, S. S.; Pottle, M. S.; Nemethy, G.; Scheraga, H. A. Conformational analysis of the 20 naturally occurring amino acid residues using ECEPP. *Macromolecules* **1977**, *10*, 1–9.

CI800434E