Feature Extraction Using Molecular Planes for Fuzzy Relational Clustering of a Flexible Dopamine Reuptake Inhibitor

Amit Banerjee,[†,⊥] Milind Misra,[‡,#] Deepa Pai,[§,○] Liang-Yu Shih,[§,▽] Rohan Woodley,[§]
Xiang-Jun Lu,[∥,◆] A. R. Srinivasan,[∥] Wilma K. Olson,[∥] Rajesh N. Davé,[†,◇] and Carol A. Venanzi*[,‡]

Departments of Mechanical Engineering, Chemistry and Environmental Science, and Computer Science, New Jersey Institute of Technology, Newark, New Jersey 07102, and Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854

Received May 15, 2007

Six rigid-body parameters (Shift, Slide, Rise, Tilt, Roll, Twist) are commonly used to describe the relative displacement and orientation of successive base pairs in a nucleic acid structure. The present work adapts this approach to describe the relative displacement and orientation of *any* two planes in an arbitrary molecule—specifically, planes which contain important pharmacophore elements. Relevant code from the 3DNA software package (*Nucleic Acids Res*. **2003**, *31*, 5108−5121) was generalized to treat molecular fragments other than DNA bases as input for the calculation of the corresponding rigid-body (or "planes") parameters. These parameters were used to construct feature vectors for a fuzzy relational clustering study of over 700 conformations of a flexible analogue of the dopamine reuptake inhibitor, GBR 12909. Several cluster validity measures were used to determine the optimal number of clusters. Translational (Shift, Slide, Rise) rather than rotational (Tilt, Roll, Twist) features dominate clustering based on planes that are relatively far apart, whereas both types of features are important to clustering when the pair of planes are close by. This approach was able to classify the data set of molecular conformations into groups and to identify representative conformers for use as template conformers in future Comparative Molecular Field Analysis studies of GBR 12909 analogues. The advantage of using the planes parameters, rather than the combination of atomic coordinates and angles between molecular planes used in our previous fuzzy relational clustering of the same data set (*J. Chem. Inf. Model*. **2005**, *45*, 610−623), is that the present clustering results are independent of molecular superposition and the technique is able to identify clusters in the molecule considered as a whole. This approach is easily generalizable to any two planes in *any* molecule.

INTRODUCTION

There is considerable evidence that drug molecules do not bind to proteins in their vacuum phase global energy minimum (GEM) conformations.[1−4] The importance of considering conformations other than the GEM in pharmacophore modeling studies has been well-documented.[5−11] Therefore, defining a protocol for the selection of a suitable set of representative conformers for input to Three-Dimensional Quantitative Structure Activity Relationship (3D-QSAR) studies, such as Comparative Molecular Field Analysis (CoMFA),[12] is an important first step in ligand-based pharmacophore modeling. This is particularly crucial in the modeling of flexible ligands, such as those in the present study. The use of data reduction techniques such as clustering is essential to separate conformations into groups in order to identify a representative conformer from each group.

The Venanzi laboratory has recently explored the application of hierarchical[13] and fuzzy[14] clustering and singular value decomposition[15] techniques to the grouping of analogues of GBR 12909 1-[2-[bis(4-fluorophenyl)methoxy]ethyl]-4-(3-phenylpropyl)piperazine (Figure 1(a)) for the purpose of identifying representative conformers for 3D-QSAR studies.[16] This class of dopamine reuptake inhibitors appears to be potentially useful in the treatment of cocaine abuse.[17] GBR 12909 effectively reduced cocaine self-administration in rhesus monkeys[18] and has completed Phase I clinical trials.[17] Our previous fuzzy clustering study[14] focused on classifying over 700 conformations of the GBR 12909 analogue, **1**. The approach used a novel feature extraction technique in which specified atom locations and angles between certain molecular planes were combined to construct the feature vector. The results compared well to those from hierarchical clustering of the same data set.[13] However, due to explicit inclusion of atomic coordinates in the feature vector, the clustering was dependent on how the conformers were

* Corresponding author phone: (973)596-3596; fax: (973)596-3596; e-mail: Venanzi@adm.njit.edu.

[†] Department of Mechanical Engineering, New Jersey Institute of Technology.

[‡] Department of Chemistry and Environmental Science, New Jersey Institute of Technology.

[§] Department of Computer Science, New Jersey Institute of Technology.

[∥] Rutgers, The State University of New Jersey.

[⊥] Present address: Evolutionary Computing Systems Laboratory, Department of Computer Science and Engineering, University of Nevada, Reno, NV 89557.

[#] Present address: Computational Biosciences Department, Sandia National Laboratories, Albuquerque, NM 87185.

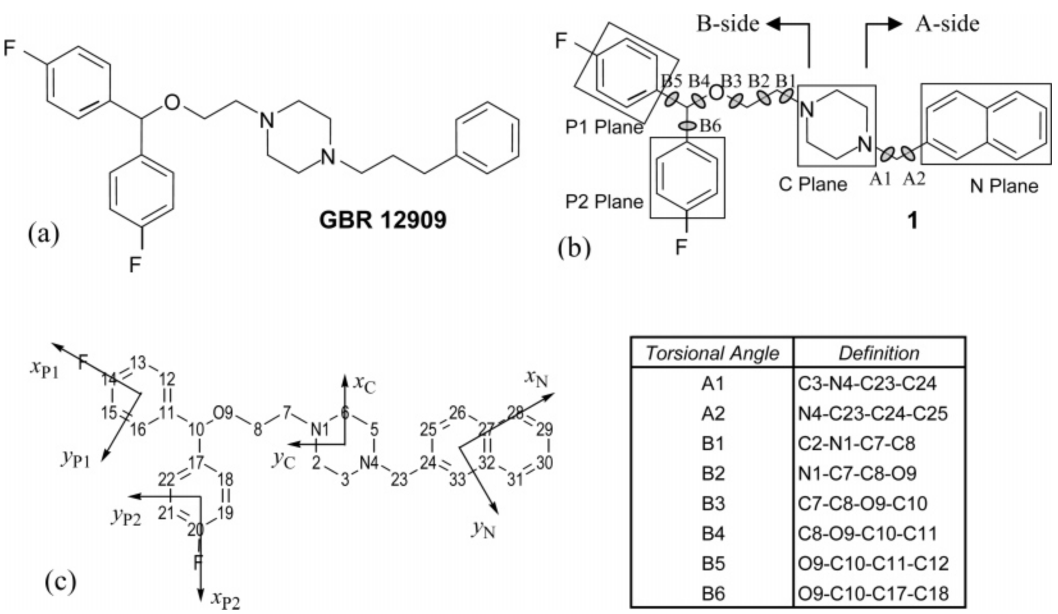
[○] Present address: Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724.

[▽] Present address: Amgen, Inc., Thousand Oaks, CA 91320.

[◆] Present address: Department of Biological Sciences, Columbia University, MC 2442, New York, NY 10027.

[◇] Present address: Department of Chemical Engineering, New Jersey Institute of Technology, Newark, NJ 07102.

10.1021/ci7001632 CCC: $37.00
Published on Web 10/30/2007



Torsional Angle	Definition
A1	C3-N4-C23-C24
A2	N4-C23-C24-C25
B1	C2-N1-C7-C8
B2	N1-C7-C8-O9
B3	C7-C8-O9-C10
B4	C8-O9-C10-C11
B5	O9-C10-C11-C12
B6	O9-C10-C17-C18

Figure 1. Structures, planes, and local coordinate systems: (a) GBR 12909 and (b) identification of molecular fragments (used to define planes) and torsional angles of **1**. The C plane was defined using atoms N1, C2, N4, and C5. (c) Local coordinate systems of molecular fragments of **1**.

superimposed. The clustering was able to uncover natural groups in the data only for certain molecular superpositions (i.e., superposition of all the conformations on a subset of atoms on one side of the molecule or the other). Representative conformers for the molecule as a whole were defined in a hybrid fashion by searching the conformer data set to identify structures that had the combined characteristics of the representative conformers from different superpositions.

In order to identify representative conformers for the molecule as a whole without the necessity of using a hybrid technique, in the present work we define a protocol for constructing a superposition-independent feature vector. We adapt the method used to calculate the rigid-body (translational and rotational) parameters of DNA bases and base pairs[19] to give the equivalent "planes" parameters for any two molecular planes of **1**. The planes parameters are then used to construct a feature vector. Since the relative orientation of the two planes is independent of the absolute Cartesian coordinates of points on those planes, the resulting clustering is superposition-independent. Since the features are pairs of planes, the feature extraction technique (i.e., calculation of the planes parameters) is easily generalizable to any number of planes in any type of molecule. As in our previous work, this approach allows for clustering based on features particular to specific regions of the molecule in order to focus on individual pharmacophore elements.

PHARMACOPHORE FEATURES AND MOLECULAR PLANES

The GBR 12909 analogue, **1**, contains two pharmacophore features that are common to most dopamine reuptake inhibitors: (1) a basic nitrogen (N4 in Figure 1(c)) in close proximity to (2) an aromatic ring (here, the naphthalene moiety on the A-side of the molecule). An additional pharmacophore feature, which has been shown to be important for the binding of GBR 12909 analogues, is the bisphenyl moiety (on the B-side of the molecule). The presence of a 2-[bis-(4-fluorophenyl)methoxy]ethyl- substituent results in slightly higher affinity for the dopamine transporter (DAT) than a 2-benzhydryloxyethyl- group, whereas the latter leads to better selectivity for the DAT over the serotonin transporter (SERT).[20−24]

Each of these pharmacophore features can be viewed in terms of a molecular plane. As shown in Figure 1(b), the N, P1, and P2 planes are defined by the planar naphthalene and phenyl moieties, respectively. The basic nitrogen is a member of the central piperazine ring, which is generally in a (nonplanar) chair conformation. However, four of the six atoms of the chair fall on a plane and can be used to describe a fourth plane (the C plane). The relative orientation and displacement of these four molecular planes define the relative spatial disposition of the important pharmacophore features. For example, the notation N_P1 or N_P2 expresses the relative orientation and displacement of the N plane and the P1 or P2 planes, respectively. These planes can be used to calculate (translational and rotational) planes parameters for the molecule as a whole since they relate the extremities of the molecule to each other. Clustering based on these parameters is identified as "full molecule" clustering. Similarly, N_C designates the relative orientation and displacement of the C and N planes on the A-side of **1**. Rotational and translational parameters calculated from these planes express relationships between the A-side pharmacophore elements and are used for "A-side" clustering studies. Finally, C_P1 or C_P2 expresses the relative orientation and displacement of the C and P1 or P2 planes, respectively; the associated planes parameters are used for "B-side" clustering studies.

PLANES PARAMETERS: GENERALIZED DNA RIGID-BODY PARAMETERS

For the description of local DNA morphology, six rigid-body parameters (Shift, Slide, Rise, Tilt, Roll, Twist; Figure 2) are commonly used to describe the relative orientation and displacement of the mean planes of successive base pairs in a nucleic acid structure.[25] Similar parameters are used to

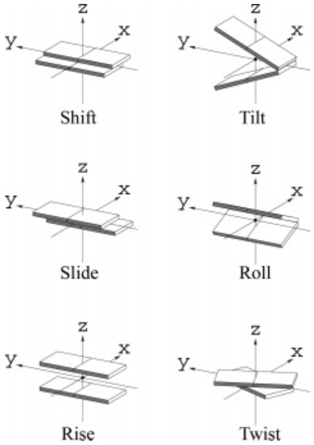


Figure 2. Rigid-body parameters used to characterize the orientation and displacement of base-pair steps in nucleic acid structure analyses.

describe the relative orientation and displacement of complementary bases in a Watson-Crick base pair.[25] Many techniques have been used for the calculation of these parameters yet yield inconsistent descriptions of chain conformation.[26] In particular, the computed parameters depend upon the choice of reference frame upon which they are based.[27,28] The 3DNA software package[19] developed by the Olson group uses a matrix-based scheme, in combination with a standard, base-centered reference frame,[29] to calculate local rigid-body parameters. The symmetric construction of the matrices and the introduction of a "middle frame" (located halfway between the coordinate frames of complementary bases or successive base pairs) insure that computed parameters are independent of chain direction. Importantly, this rigorous approach allows for the reconstruction of molecular models from the derived parameters. Although this approach is described in detail elsewhere,[29] a summary of the main points is given below.

For each pair of planes, the cross product of the normals (*z*-axes) provides the *hinge axis* (or the line of intersection between the two planes). The scalar product of the *z*-axes provides the *net bending angl*e Γ between the two planes. Each plane is rotated about the hinge axis by half of the net bending angle but in opposite directions, aligning the *z*-axes of the two planes such that the two rotated planes are parallel to each other. The *x*-, *y*-, and *z*-axes of the *middle frame* are obtained from the average of the corresponding axes following these two rotational operations. The origin of the middle frame is the average of the coordinates of the origins of the two planes. Parameters defined with respect to the middle frame are internally consistent (or "absolute") in that they are independent of an external frame of reference and therefore of superposition. The three translational parameters are the projections of the vector between the origins of the two planes onto the *x*-, *y*-, and *z*-axes of the middle frame. The rotational parameter, Twist, is the angle between the *y*-axis (or *x*-axis; since it is symmetrical, either axis gives exactly the same result) of one rotated plane and that of the other rotated plane. The sign of Twist is determined by the scalar product of the *z*-axis of the middle frame and the vector resulting from the cross product of the *y*-axes of the two rotated planes. The angle between the hinge axis and the *y*-axis of the middle frame is termed the *phase angle* ϕ. The remaining two rotational parameters are expressed in terms of Γ and ϕ: Roll = Γ cos ϕ, Tilt = Γ sin ϕ. Translational parameters (Shift, Slide, and Rise) are given in Ångstrom units and rotational parameters (Tilt, Roll, and Twist) in degrees.

In the present work, we apply the same approach to describe the relative position and orientation of *any* two planes in a molecule. The resulting planes parameters used as features in the fuzzy clustering approach correspond to the translational (Shift, Slide, and Rise) and rotational (Tilt, Roll, and Twist) parameters for each of the pairs of planes (N_P1, N_P2, C_P1, C_P2, and N_C) of **1**. Due to the flexibility of **1**, the relative orientation of these pairs of planes takes on a much wider range of values than that of consecutive base pairs in the considerably more rigid double helical DNA structure. In order to investigate the sensitivity of the clustering results to the use of translational and/or rotational planes parameters, three types of feature vectors were constructed using the following: (1) all six translational and rotational parameters (T+R), (2) only the three translational parameters (T), and (3) only the three rotational parameters (R). In order to investigate the sensitivity of results to the choice of pairs of planes, three sets of pairwise plane combinations were used for the following clustering studies: (1) A-side clustering, using the proximal planes N and C on the A-side of the molecule; (2) B-side clustering, using planes C and P1 or C and P2; and (3) full molecule clustering, using the planes at the extremities of the molecule, N and P1 or N and P2. Shorthand notation was used to designate each clustering study/feature vector combination, such as $[N\_C]_{T+R}$, $[N\_C]_T$, and $[N\_C]_R$ for A-side clustering.

FUZZY RELATIONAL CLUSTERING USING PLANES PARAMETERS

The use of data clustering techniques in computational chemistry has been recently reviewed.[30] Most applications have involved the clustering of molecules in chemical databases rather than the application proposed here—the clustering of molecular conformations. Hierarchical clustering[8,13,31] has more frequently been applied to the classification of molecular conformations than partitional clustering.[14,32−34] Previously,[14] we described in detail the advantages of partitional over hierarchical clustering techniques and our motivation for using the fuzzy relational clustering (FRC) algorithm.[35] We also defined an easily generalizable feature extraction protocol based on using certain atomic coordinates and molecular planes as features. The associated feature vector consists of the selected coordinates and the angle between each chosen pair of planes.[14] The present work uses the same FRC technique but takes the four molecular planes (C, N, P1, and P2) as features. The associated feature vector consists of the translational and/or rotational planes parameters for the chosen pairs of planes.

As in our previous work,[14] each of the present "T+R" feature vectors consists of "mixed" components, a set of three translational parameters and three rotational parameters, which completely specifies the spatial relationship between the chosen pair of planes. An object-space-based clustering technique such as Fuzzy *c*-Means (FCM) could be used directly on this feature vector or the feature vector could be first transformed into a proximity matrix, which relates pairwise dissimilarity between conformers, and then a relational clustering scheme could be used to cluster conformers over this relational space. To be consistent with our previous approach,[14] a relational clustering scheme was chosen as the partitioning methodology. Converting the data in the planes feature space to a proximity distance matrix also provides a better understanding of the interconformational similarities. Use of a compact representation in a relational space allows not only for a certain amount of freedom in choosing an appropriate quantitative criterion for dissimilarity but also for the inclusion of non-Euclidean information in later stages, as and when a generalized methodology for clustering is developed. The same motivation prompted us to use FRC with the superposition-dependent feature set.[14]

FRC is a recently developed relational clustering technique[35] and is conceptually attractive because it works directly on non-Euclidean data without first converting it to a Euclidean measure. The scheme is therefore less constrained than most other relational clustering techniques which expect the proximity matrix to be Euclidean. The method has been described in detail in our previous publication.[14] The main concepts are summarized below for the convenience of the reader. Given a dissimilarity data matrix, $\mathbf{D} = [D_{jk}]$, $1 \le j, k \le n$, FRC only assumes that its elements are subject to the minimal constraints given below

$$D_{jj} = 0, \quad D_{jk} \ge 0, \quad D_{jk} = D_{kj}, \quad 1 \le j, k \le n \qquad (1)$$

The algorithm then alternates between optimizing the memberships, $\mathbf{U} = [u_{ik}]$, and a related distance matrix, $\mathbf{A} = [a_{ik}]$, $1 \le i \le c$, $1 \le k \le n$, using a successive-substitution method as described by Davé and Sen.[35] Here n is the number of data objects, and c is the number of clusters fixed a priori. The "update" equations used for the iterative calculation of $\mathbf{U}$ and $\mathbf{A}$ are shown in eqs 2 and 3

$$u_{ik} = \frac{\left[\dfrac{1}{a_{ik}}\right]^{1/(m-1)}}{\sum_{w=1}^{c}\left[\dfrac{1}{a_{wk}}\right]^{1/(m-1)}} \qquad (2)$$

$$a_{ik} = \frac{m\sum_{j=1}^{n} u_{ij}^{m} D_{jk}}{\sum_{j=1}^{n} u_{ij}^{m}} - \frac{m\sum_{h=1}^{n}\sum_{j=1}^{n} u_{ij}^{m} u_{ih}^{m} D_{jh}}{2[\sum_{j=1}^{n} u_{ij}^{m}]^{2}} \qquad (3)$$

The c-mean vectors, $\mathbf{V} = [\upsilon_i]$, $1 \le i \le c$, are scaled n-tuples of memberships

$$\upsilon_i = \frac{(u_{i1}^{m}, u_{i2}^{m}, ..., u_{in}^{m})^{T}}{\sum_{k=1}^{n} u_{ik}^{m}} \qquad (4)$$

The membership matrix, $\mathbf{U}$, is initialized randomly. The number of clusters, c (>1), and the "fuzzifier", m (>1), are fixed. The algorithm then iterates between eqs 2 and 3, until the change in memberships in two successive iterations falls below a certain prefixed threshold. Termination of the algorithm indicates that a local minima partition is achieved. In every iteration, the c-mean vectors are updated using eq 4 after all the membership values have been updated. After the algorithm converges, the membership information is "defuzzified" by assigning the conformation j to the cluster i if $u_{ij} > u_{kj}$ $(k \neq i)$ for all $1 \le k \le c$. The representative conformation is identified as the one with the highest membership value in that particular cluster, i.e., for cluster i, the representative conformation is defined as the conformation l if $u_{il} > u_{ij}$, $(l \neq j)$ for all $1 \le j \le n$. This process is carried out for a range of values for c. The clustering results are then evaluated by cluster validity analyses as described in the next section.

FUZZY CLUSTER VALIDITY MEASURES

Since the number of clusters in a data set is seldom known a priori, clustering is performed on a wide range of plausible numbers of clusters. For all the data sets in this study, clustering was performed over $2 \le c \le 14$. Cluster validity indices were then used to validate the results—these indices are indicative of the goodness of a certain value of c compared to other values on a numerical scale. The four cluster validity indices used in our previous work[14] were used in this study as well. They are briefly described in this section. (Note that typographical errors in the formulas for F and H in our previous work have been corrected in eqs 5 and 7 below.)

The partition coefficient[36] measures the fuzziness of the partition (or grouping) and is inversely proportional to the average fuzzy overlap between the clusters. The coefficient F and its normalized version F' are shown in eqs 5 and 6. The normalization helps in compensating for the dependence of F on c.

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

$$F' = \frac{cF - 1}{c - 1} \qquad (6)$$

The partition entropy[36] is the application of Shannon's entropy to quantify uncertainty. The form used as a cluster validation criterion is shown in eq 7.

$$H = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij} \ln u_{ij} \qquad (7)$$

A high value of F (and F') indicates a better partition, where clusters are compact and well separated, as opposed to a low value which indicates almost equal sharing of all entities among all the clusters. A good partition is characterized by a low value of H.

The compactness criterion[37] considers cluster compactness and separation as a measure of cluster validity. This criterion is also sometimes referred to as the Xie-Beni index, and a modified version for use in relational clustering is given in terms of the a_{ik}'s from eq 3 by

$$S = \frac{\sum_{i=1}^{c} \sum_{k=1}^{n} {u_{ik}}^2 a_{ik}}{n[\min_{1 \le i,j \le c, i \ne j} \|v_i - v_j\|^2]} \tag{8}$$

While the numerator describes the compactness of clusters in the partition, the factor in the denominator describes the separation of the clusters. A low value of S indicates a good partition.

METHODS

Data Set of Conformations. A data set of 728 conformers of **1** was collected by random search conformational analysis using version 6.9 of the SYBYL molecular modeling package.[38] The Tripos force field[39] along with Gasteiger-Hückel point charges were used for the calculation. The molecule was protonated on N4 prior to random search. The piperazine ring was fixed in the chair conformation by treating it as an "aggregate". Side chains were attached to the piperazine ring in the equatorial position. The eight torsional angles (A1,...,B6; Figure 1(b)) were randomly altered during the search. Full details of the conformational analysis are given in our previous publication.[14]

Definition of Standard Rings. DNA structures that are determined experimentally may contain bases that are not perfectly planar. For this reason 3DNA defines a set of standard (or ideal) base structures which are fit to the structures of the bases in the experimental data set. Each base is assigned a local coordinate system. As described above, rotation and translation of these local coordinate systems into a common reference frame allows for the calculation of the six rigid-body parameters that describe the relative position and orientation of each base in a base pair or successive base pairs in a particular DNA structure. The present work follows the same approach by defining standard structures for each ring system in **1**, fitting the standards to the corresponding molecular fragments of **1**, and using the equivalent 3DNA algorithm for calculating the six planes parameters that describe the relative position and orientation of any two planes in a particular conformation of **1**.

Figure 1(b) shows the four planes that were chosen for **1**: the two phenyl ring planes (P1 and P2 planes, defined by six atoms each), the central piperazine ring plane (C plane, defined by atoms as described below), and the naphthalene plane (N plane, defined by ten atoms). However, there are at least 250 other analogues of GBR 12909, many of which have a benzene-like substituent in place of the naphthalene ring of **1**. In order to make the protocol as general as possible and to make possible the comparison of the planes parameters of different GBR 12909 analogues in future studies, only the benzene-like part of the naphthalene ring closest to the piperazine ring was used here in the calculation of the planes parameters. Therefore only piperazine and benzene ring standards were required. These standards were defined as the piperazine and benzene rings of the SYBYL fragment library. Four (in the case of piperazine, see Figure 3(a)) and six (for benzene, see Figure 3(b)) atoms were selected to define a plane. The Cartesian coordinates of a "pseudo"-atom centroid were calculated as the average of the coordinates of the atoms used to define the respective planes. A right-handed coordinate system was attached to each standard ring as shown in Figure 3. The x- and y-axes lie in the respective ring planes with the direction of the z-axis defined by the right-hand rule as the cross product of the **x** and **y** unit vectors: $\mathbf{z} = \mathbf{x} \times \mathbf{y}$, where $\mathbf{x} = (1, 0, 0)$, $\mathbf{y} = (0, 1, 0)$, and $\mathbf{z} = (0, 0, 1)$, relative to the centroid at (0, 0, 0).

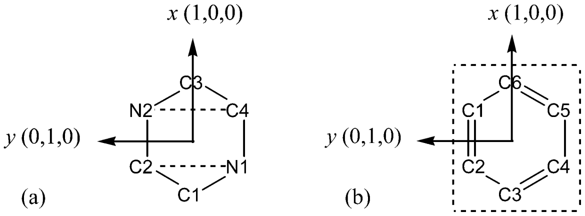


Figure 3. Assignment of coordinate reference frame to standard rings: (a) piperazine standard (schematic) and (b) benzene standard.

Table 1. Mapping Scheme for Standard Rings and Corresponding Molecular Fragments

ring standard numbering scheme[a] benzene	molecule numbering scheme[b] N plane	P1 plane	P2 plane
C3	24	11	17
C4	25	12	18
C5	26	13	19
C6	27	14	20
C1	32	15	21
C2	33	16	22

ring standard numbering scheme[a] piperazine	molecule numbering scheme[b] C plane
N2	1
C2	2
N1	4
C4	5

[a] From Figure 3. [b] From Figure 1(c).

Mapping of Standard Rings onto Molecular Planes. A crucial step in the calculation of the planes parameters is the definition of a convention for mapping the benzene and piperazine standard rings onto the corresponding molecular fragments for each conformer in the data set. This consists of a prescription for orienting the x- and y-axes of each standard with respect to the "local" x- and y-axes of each corresponding molecular fragment of **1**. The convention used in the present work for defining these local axes is illustrated in Figure 1(c). As in 3DNA, each standard coordinate frame was "mapped" onto the corresponding molecular fragment of each conformer of **1** by performing a unit quaternion-based least-squares fitting procedure[40] that superimposes the atoms that define the planes. The mapping scheme is given in Table 1. This procedure was repeated for all the conformers in the data set to give the reference coordinate frame of the planes for each conformation.

Calculation of Planes Parameters. After the above mapping procedure, the six translational and rotational planes parameters were calculated for every possible pairwise combination of planes in each molecular conformation. The procedure was repeated for all the conformations in the data set. A MATLAB (version 6.0, available from The Mathworks, Inc., Nantick, MA) program derived from the base-pair step section of the 3DNA source code was written by

R. Woodley to identify planes in **1**, map the standard rings onto the planes, and calculate the planes parameters.

Proximity Matrices. For each pair of planes, each feature vector (T+R, T, or R) defines a distinct proximity (or dissimilarity) matrix for the following: (a) proximity defined by all six (translational plus rotational) planes parameters, (b) proximity defined by the three translational parameters, and (c) proximity defined by the three rotational parameters, respectively. The results of clustering using the three different types of proximity matrices were compared in order to evaluate the separate contributions of the translational and rotational components to the observed clustering. For proximity matrices involving mixed (rotational and translational) feature vectors, the distance between any two conformers k and j is defined as

$$D_{kj} = \left[\sum_{p=1}^{3} (t_{pk} - t_{pj})^2 + s \sum_{p=1}^{3} (r_{pk} - r_{pj})^2 \right]^{1/2} \qquad (9)$$

where t_{pk} and t_{pj} are the three translational parameters for k and j respectively, and r_{pk} and r_{pj} are the three rotational parameters for k and j respectively, *for* $1 \le p \le 3$. The difference between the rotational parameters, $r_{pk} - r_{pj}$, was adjusted to be always less than π in order to account for angle circularity. If $r_{pk} - r_{pj}$ was greater than π, the difference was reset to $2\pi - (r_{pk} - r_{pj})$. A judicious choice for the scaling factor, s, is the ratio of the absolute squared differences between the maximum and minimum of the translational parameters and the rotational parameters over the entire data set

$$s = \frac{(t_{\max} - t_{\min})^2}{(r_{\max} - r_{\min})^2} \qquad (10)$$

Such a scaling scheme is known as *range-based scaling*.[41] This was done prior to computing the proximities using the Euclidean distance norm. For feature vectors consisting of only the translational or the rotational parameters, no scaling was required.

Fuzzy Clustering. The clustering routine for every proximity matrix was performed for $2 \le c \le 14$, and, for every value of c, the routine was run 20 times with a different random initialization of memberships. The partition that minimized the FRC objective functional, J,[35] shown in (11), was used for membership and cluster assignments.

$$J = \sum_{i=1}^{c} \frac{\sum_{j=1}^{n} \sum_{k=1}^{n} u_{ik}^m u_{ij}^m D_{jk}}{2 \sum_{t=1}^{n} u_{it}^m} \qquad (11)$$

Conformers were assigned to a cluster based on the largest value of their memberships over the c clusters. The representative structure for each cluster was defined as the conformation with the highest membership value in that cluster. The two user-defined parameters used for FRC were $m = 2$, and the termination condition (change in memberships of successive iterations) $= 10^{-5}$. (The clustering results, however, are not very sensitive to these parameters.) The output of the clustering was used as input to the validity procedures. The FRC and cluster validity procedures were implemented by C programs developed in-house on a Sun Blade 1500 workstation running a 1-GHz 64-bit Ultrasparc III processor.

Table 2. Summary of Cluster Calculations[a]

type of clustering	proximity matrix	optimal number of clusters, c
full molecule	$[\text{N\_P2}]_{\text{T+R}}$	5
full molecule	$[\text{N\_P2}]_{\text{T}}$	5
full molecule	$[\text{N\_P2}]_{\text{R}}$	8
B-side	$[\text{C\_P2}]_{\text{T+R}}$	4
B-side	$[\text{C\_P2}]_{\text{T}}$	4
B-side	$[\text{C\_P2}]_{\text{R}}$	10
A-side	$[\text{N\_C}]_{\text{T+R}}$	6
A-side	$[\text{N\_C}]_{\text{T}}$	7
A-side	$[\text{N\_C}]_{\text{R}}$	5

[a] See text for explanation of proximity matrix notation.

RESULTS

For the 728 conformers of **1**, Table 2 summarizes the results of each type of clustering study for each pair of planes. Note that the behavior of the two phenyl rings in **1** is correlated due to their coupled rotational barriers.[42] This means that clustering studies with the $[\text{N\_P1}]_{\text{T+R}}$, $[\text{N\_P1}]_{\text{T}}$, $[\text{N\_P1}]_{\text{R}}$, $[\text{C\_P1}]_{\text{T+R}}$, $[\text{C\_P1}]_{\text{T}}$, and $[\text{C\_P1}]_{\text{R}}$ proximity matrices investigate regions of space that are somewhat redundant with their [N_P2] and [C_P2] counterparts. For this reason, the clustering studies involving P1 are not summarized in Table 2 nor are these results discussed in detail below. Figures S1 and S2 of the Supporting Information plot the results of $[\text{N\_P1}]_{\text{T+R}}$ (full molecule) and $[\text{C\_P1}]_{\text{T+R}}$ (B-side) clustering. The figures are typical in that, although the P1 studies give somewhat different cluster sizes and cluster memberships than the P2 studies discussed below, they do not provide significantly different information.

Visualization of the clustering results is possible in the cases where the feature vectors consist of *either* translational *or* rotational planes parameters. In those cases, the clustering is shown in two-dimensional (2D) or three-dimensional (3D) translational (Shift, Slide, Rise) or rotational (Tilt, Roll, Twist) space, using the planes parameters as coordinate axes. Note that visualization of higher dimensional data sets is not possible without resorting to data scaling and/or coordinate projection.

Full Molecule Clustering. Clusters were sought using the full molecule proximity matrix, $[\text{N\_P2}]_{\text{T+R}}$, and distinct partitions were obtained in the range of $2 \le c \le 6$. This is substantiated by the Xie-Beni index, S (Figure 4), which takes low values over that range. (The partition entropy, H, is seen to be monotonically increasing, and F and F' are monotonically decreasing over the entire range $2 \le c \le 14$.) In the low range of $2 \le c \le 6$, S takes on very similar values indicating that there are several possible partitions. However, the minimum of S is at $c = 5$, and, hence, $c = 5$ can be reasonably assumed to be the best partition (shown with a dashed line in Figure 4).

This separation into five clusters is best visualized in the three-dimensional (Shift, Slide, Rise) translational space and in the two-dimensional (Slide, Rise) plane as shown in Figure 5(a),(b). In Figure 5 and all subsequent plots, the conformers are color-coded by cluster; the translational parameters are

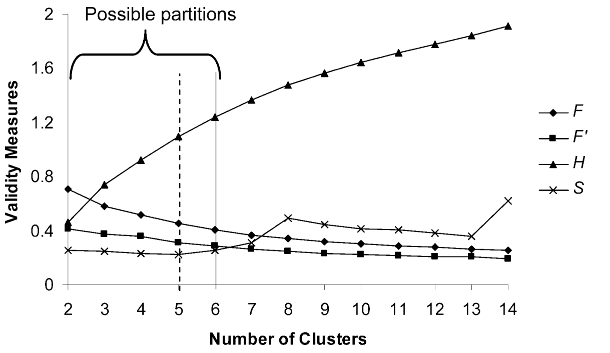


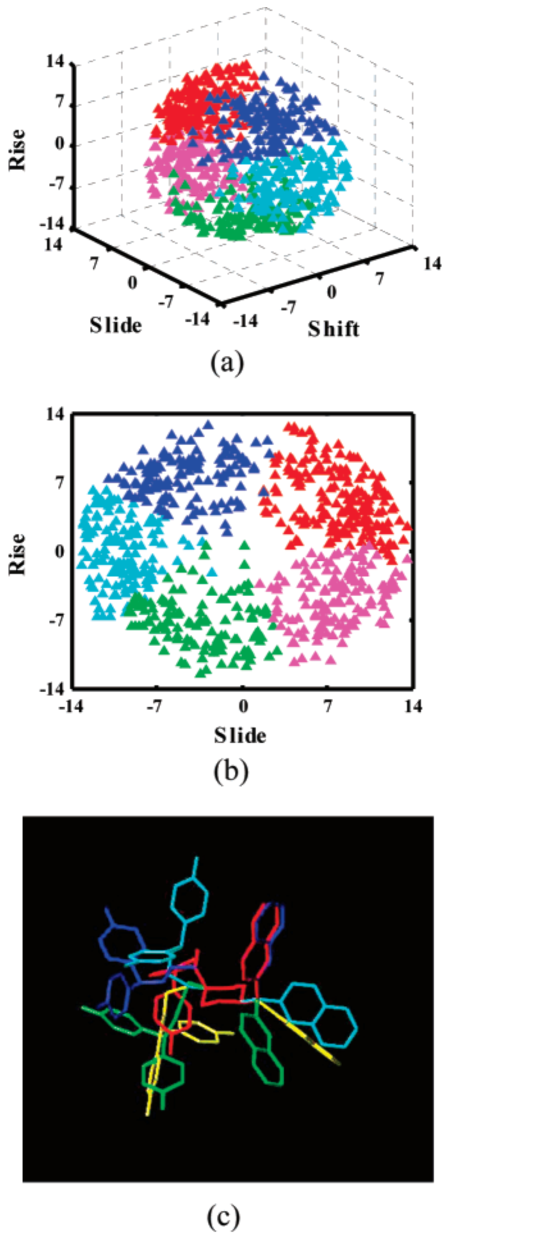
Figure 4. Cluster validity plots for clustering using the full molecule proximity matrix, [N_P2]$_{\mathrm{T+R}}$.

given in Ångstroms (Å), and the rotational parameters are given in degrees. The five representative conformers, one from each cluster, are identified by their torsional angles in Table 3 and are shown in Figure 5(c). The conformers appear to be representative of the regions of space occupied by **1**.

The cluster validity indices (not shown) for clustering using the full molecule translational component proximity matrix, [N_P2]$_{\mathrm{T}}$, behave similarly to those for the [N_P2]$_{\mathrm{T+R}}$ proximity matrix in Figure 4. As in the [N_P2]$_{\mathrm{T+R}}$ case, they also identify five clusters as the optimal partition from a range of closely related possible partitions. The plot of conformers in 3D translational space for the [N_P2]$_{\mathrm{T}}$ translational proximity matrix (Figure S3 of the Supporting Information) appears to be almost identical to that for the [N_P2]$_{\mathrm{T+R}}$ translational plus rotational proximity matrix (Figure 5(a)). This is to be expected since the number of conformers in each cluster is approximately the same. This similarity seems to indicate that the translational parameters may be the chief determinant for clustering conformations in the full molecule case, at least for molecules with planes separated by a distance on the order magnitude of (or greater than) that between the N and P2 planes in **1**. This is supported by two additional observations. First, the plot (not shown) of conformers in 3D *rotational* space for the [N_P2]$_{\mathrm{T+R}}$ proximity matrix clustering study shows no separation of conformations into clusters. Also, the validity plot (not shown) for clustering using the rotational proximity matrix, [N_P2]$_{\mathrm{R}}$, identifies eight clusters, which is dissimilar to the [N_P2]$_{\mathrm{T+R}}$ and [N_P2]$_{\mathrm{T}}$ ($c = 5$) results shown in Table 2.

B-Side Clustering. The cluster validity indices for the B-side proximity matrix, [C_P2]$_{\mathrm{T+R}}$ (Figure S4 of the Supporting Information), are not as definitive as for the full molecule case. The Xie-Beni index, *S*, behaves well over $2 \leq c \leq 10$, after which it takes unnaturally large values for all $c > 10$. In other words, good clusters are arbitrarily subdivided into artificial overlapping clusters for all $c > 10$. This prompted a search for a good partition in the range $2 \leq c \leq 10$. In this range, *S* attains its lowest value at $c = 4$, and the normalized partition coefficient, *F*′, also attains its maximum value. The other two indices, *F* and *H*, are nonindicative for $2 \leq c \leq 10$.

Figure 6 shows the conformers plotted in 3D (Shift, Slide, Rise) translational space for $c = 4$ for the [C_P2]$_{\mathrm{T+R}}$ proximity matrix study. The conformers separate well in translational space, and this can be seen particularly in the (Slide, Rise) and (Shift, Rise) plots (Figure S5 of the Supporting Information).



(a)

(b)

(c)

Figure 5. Conformers plotted in (a) 3D translational space and (b) 2D Slide vs Rise plane for clustering with the full molecule proximity matrix, [N_P2]$_{\mathrm{T+R}}$, at $c = 5$. Translational units are in Ångstroms. In this and subsequent figures, conformers are color-coded by cluster. Number of conformers in each cluster in (a): blue - 141, red - 187, green - 122, magenta - 138, and cyan - 140. (c) Representative conformers of Table 3 displayed by superimposing the structures by the atoms of the piperazine ring. Color-coding of the conformers matches that of the clusters which they represent. Conformation number (membership values): blue - #378 (0.967), red - #709 (0.898), green - #713 (0.910), yellow (the magenta cluster in Figure 5(a) is displayed as yellow here for easier viewing) - #332 (0.915), and cyan - #371 (0.910).

As noted in Table 2, the cluster validity plots (not shown) for clustering with the [C_P2]$_{\mathrm{T}}$ and [C_P2]$_{\mathrm{R}}$ proximity matrices identify $c = 4$ and $c = 10$, respectively, as the optimal number of clusters. The plot of conformers in 3D translational space for the [C_P2]$_{\mathrm{T}}$ translational proximity matrix (Figure S6 of the Supporting Information) appears to be almost identical to that for the [C_P2]$_{\mathrm{T+R}}$ translational plus rotational proximity matrix (Figure 6). As in the full molecule case, this is to be expected since the number of conformers in each cluster is approximately the same. The

Table 3. Torsional Angles[a] and Relative Energies[b] of Representative Conformers

conformer[c]	A1	A2	B1	B2	B3	B4	B5	B6	relative energy
332 yellow	264	209	161	299	273	184	173	73	13
371 cyan	297	92	211	69	278	293	153	32	9
378 blue	61	89	158	314	187	74	195	272	9
709 red	65	94	42	42	62	279	306	12	8
713 green	185	272	78	60	169	290	142	120	11

[a] Angles are in degrees. See Figure 1(b) for definition. [b] Energy relative to that of the GEM conformer, in kcal/mol. [c] Identification number of conformer out of 728 conformers. Conformer color refers to the color of the conformers displayed in Figure 5(c) and to the colors of the clusters in Figure 5(a) that they represent. Note that magenta in Figure 5(a) is represented by yellow in Figure 5(c) and in this table.

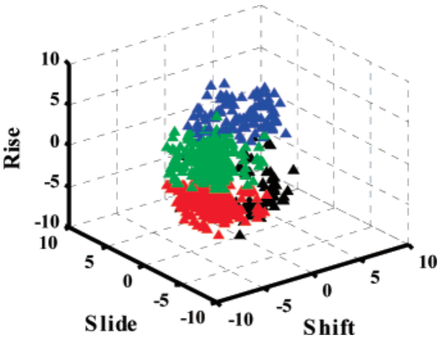


Figure 6. Conformers plotted in 3D translational space for clustering with the B-side proximity matrix $[C\_P2]_{T+R}$ at $c = 4$. Number of conformers in each cluster: red - 287, green - 257, blue - 100, and black - 84.

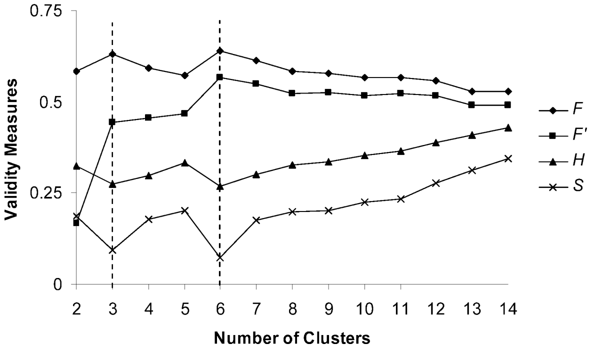


Figure 7. Cluster validity plots for clustering with the A-side proximity matrix, $[N\_C]_{T+R}$.

fact that clustering with the $[C\_P2]_{T+R}$ and $[C\_P2]_T$ proximity matrices results in the same number and the same size of clusters supports the behavior noted with full molecule clustering: translational rather than rotational features dominate clustering based on planes that are relatively far apart.

A-Side Clustering. Figure 7 shows the cluster validity plots for the A-side over the range $2 \leq c \leq 14$ for clustering with the $[N\_C]_{T+R}$ proximity matrix. At $c = 6$ and $c = 3$, F and F' take their maximum values, and H and S take their minimum values. Unlike the full molecule and B-side cases, all four validity measures seem to be in agreement.

Figure 7 shows that $c = 6$ is a slightly more optimal partition than $c = 3$. (For $c = 3$ results, see Figure S7 of the Supporting Information). Conformers plotted in the 3D (Shift, Slide, Rise) translational and 3D (Tilt, Roll, Twist) rotational space for $c = 6$ are shown in Figure 8(a),(b). The separation of conformations into six clusters is clearly visible in both translational and rotational space. Compared to the full molecule and B-side clustering results, both the translational and rotational parameters appear to play a role in separating the conformations into clusters. This may be because the N and C planes are, for most of the conformations in this study,[42] much closer in space than the N and P1 or P2 planes or the C and P1 or P2 planes. A complete analysis of the conformational profile of **1** will be given in a separate publication.[42] The proximity of the N and C planes means that their relative rotation as well as their relative separation is important to the clustering. This can be seen in Table 2 which shows that the optimal number of clusters is different for A-side clustering with the $[N\_C]_{T+R}$ and $[N\_C]_T$ proximity matrices ($c = 6$ and 7, respectively), whereas for full molecule or B-side clustering, the equivalent proximity matrices result in the same optimal number of clusters.

Figure 8(c) displays the clusters of Figure 8(a),(b) using atomic coordinates rather than translational or rotational planes parameters. Figure 8(c) shows that the six clusters cannot be clearly visualized in atomic coordinate space. This is not unexpected, because the features used in the clustering were defined as the translational and rotational planes parameters rather than the atomic coordinates. This illustrates that clustering is best visualized in the space of the features from which the proximity matrices were defined.

DISCUSSION

The study presented here proposes a novel protocol for feature extraction and clustering, one that is intuitive and easily generalizable. A conformer is represented by translational and rotational parameters derived from pairs of planes defined by molecular substructure. Although the analysis was applied to conformers of **1**, such pairs of planes can be easily identified in molecules of *any* size and the corresponding parameters calculated. A numerical measure of dissimilarity between two conformers could then be calculated based on these parameters and fuzzy relational clustering carried out to classify a data set of conformations into groups. The advantage of using the planes parameters, rather than the combination of atomic coordinates and angles between molecular planes used in our previous fuzzy relational clustering of the same data set,[14] is that the present clustering results are independent of molecular superposition, and the technique is able to identify clusters in the molecule considered as a whole. This approach is easily generalizable to any two planes in *any* molecule.

Superposition-Independent Feature Vector. A major contribution of this work is the development of a superposition-independent feature vector. This feature vector is fundamentally different from the one proposed in our previous work,[14] and, although the underlying data set is the same in both cases, the feature sets, and hence the feature vectors, are not directly comparable. Since the feature extraction procedure in its present form only considers a pair of planes, the results are independent of the intervening molecular structure between the planes. Our previous work[14] used specific molecular planes *and* the Cartesian coordinates of certain atoms to define a minimal (superposition-dependent) feature set from which the molecule could be reconstructed. In that work, a proximity matrix for the full molecule was derived from eight atom locations and three

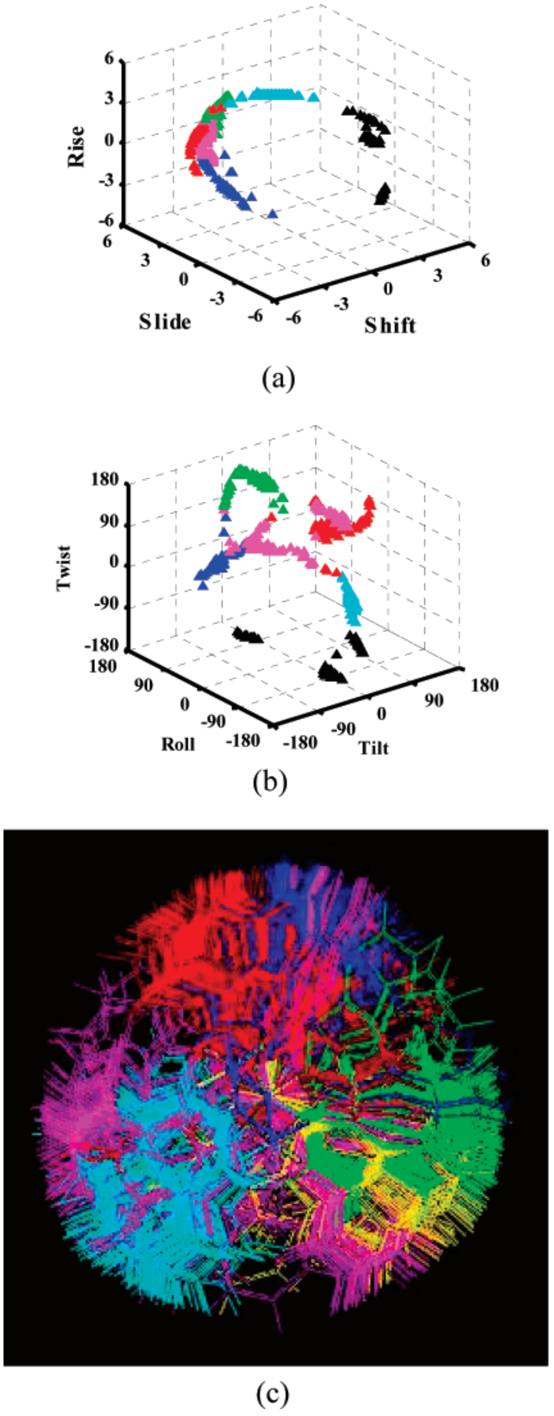


Figure 8. Conformers plotted in (a) 3D translational and (b) 3D rotational space for clustering with the A-side proximity matrix, $[N\_C]_{T+R}$, at $c = 6$. Translational parameter units are Ångstroms; rotational parameter units are in degrees. (c) Conformers plotted in 3D atomic coordinate space for $c = 6$. Conformers were superimposed by aligning the atoms of the central piperazine ring of each conformer with those of the other conformers. Since the piperazine ring was held fixed in a chair conformation during the random search, all conformers superimpose perfectly only on that part of their substructure. The present view is of the A-side only and shows the range of values taken on by the angles A1 and A2 in the set of conformers. The view is a 90° rotation from the view shown in Figure 1, with the A-side facing outward and the B-side and piperazine ring deleted for clarity. Clusters are color-coded as in (a) and (b). Number of conformers in each cluster: blue − 82, green − 126, yellow (the conformers in the black cluster of Figures 8(a) and 8(b) are given in yellow here for easier viewing) − 62, magenta − 166, cyan − 125, and red − 167.

sets of angles between planes. In contrast to our previous superposition-dependent approach, the size of the feature vector in the present study does not depend on the span of the molecule. In other words, the feature vector is always fixed (either six or three elements, depending on whether or not all orientational parameters are considered).

Based on the evidence provided by the cluster validity indices, full molecule clustering with the previous superposition-dependent feature vector proved inconclusive. For that reason, it was necessary to adopt a hybrid approach wherein the results of A-side and B-side clustering were combined together to obtain representative structures for the full molecule. In contrast, the present approach was able to *directly* identify clusters in the molecule considered as a whole (i.e., full molecule clustering). Yet both approaches allow for clustering on subsets of the full molecule framework (A-side or B-side) in order to focus on certain areas of the molecule which may contain specific pharmacophore moieties of interest.

Generalization of the Feature Extraction Method. The protocol presented here can be generalized to include more than just a pair of planes. The full molecule feature vector could alternatively be defined by 12 planes parameters, 6 from the N_C planes and 6 more from the C_P2 planes. In other words, the full molecule can be considered to be a combination of the A-side and the B-side feature vectors in their entirety. The planes protocol presented here is flexible in that it leaves room for additional generalizations such as this. In addition, an alternative feature vector (one that might be redundant from an information point of view) could be constructed by combining the two approaches—the superposition-independent approach discussed here and the reconstruction-based approach presented in our previous work, where a pair of planes was represented by the angle at which the planes intersected. In addition, specific heavy atom locations were also considered as part of the feature vector (this is what made the feature vector dependent on superposition). In the present work, a pair of planes is represented by six parameters rather than the single parameter used previously—the angle between the planes. A new combinatorial feature vector could be constructed which combines atomic locations and the six parameters representing a pair of planes.

Relative Contributions of Rotational and Translational Parameters. As summarized in Table 2, for planes that are generally far apart (such as the N and P1 or P2 planes of the full molecule clustering study or the C and P1 or P2 planes of the B-side clustering study), their relative rotational orientation is of lesser significance to clustering than their distance of separation. As a result, the optimal number of clusters and the relative distribution of conformers between clusters determined by the $[N\_P2]_{T+R}$ proximity matrix ($c = 5$) is equal to that from the $[N\_P2]_T$ calculation. The same is true for the results from the $[C\_P2]_{T+R}$ ($c = 4$) and $[C\_P2]_T$ proximity matrix calculations. In contrast, when the two planes are very close and held in a somewhat rigid orientation by the molecular substructure, such as the C and N planes of the A-side clustering study, the optimal number of components determined by the (T+R) and (T) proximity matrices are different, indicating that the rotational orientation of the planes play a larger part in describing their relative orientation. For example, the optimal number of clusters is

$c = 6$, 7, and 5 for the $[N\_C]_{T+R}$, $[N\_C]_T$, and $[N\_C]_R$ calculations, respectively.

Clustering Parameters and Validity Measures. The Fuzzy Clustering section of the Methodology section introduced two user-defined parameters: the fuzzifier, *m*, in eq 11 and the termination condition (change in the memberships of successive iterations). In the present work, they were set equal to 2 and 10^{-5}, respectively. It was stated that the clustering results are not very sensitive to these parameters. The fuzzifier is a weighting factor for the distances between input vectors and prototype vectors, used in the computation of membership values, **U**. It essentially controls the fuzziness of the resulting *c*-partition. The fuzzifier influences the convergence rate of the algorithm and, to a small extent, the cluster validity of the *c*-partition. Moreover, in the limit of *m* tending to zero, it will produce “hard” or “crisp” partitions, i.e., memberships of only 0 or 1, and in the limit of *m* tending to infinity, it will produce completely fuzzy, yet meaningless partitions. Many attempts have been made to optimize the value of *m*; however, there is no theoretical foundation for the optimal choice of *m*. The appropriate value is a heuristic and is set via experimentation.[43−45]

In the Fuzzy *c*-Means clustering research community, $m = 2$ is widely accepted, but it has long been established that the choice of *m* could largely depend on the data set in question.[36] In the present work using $m = 1.5$, 2, or 2.5 gave no significant variation in the partition produced, which is typical of most real-life data. A possible reason may be the inherent fuzziness of the data set in this study.

The termination condition controls the accuracy of the partition resulting through the convergence of the algorithm. Any clustering algorithm is said to have produced satisfactory results if the change in cluster memberships remains unchanged across two successive iterations of the algorithm. In practice this is realized when the membership difference across two successive iterations is negligible. The smaller the value of the termination condition, the higher is the computational time (with no significant improvement in performance). Using a termination condition of 10^{-5} gave a reasonable balance between computation time and performance. This is also a typical value that works in most real-life situations.

Several validity indices can be used to validate the clustering results. However, a standing criticism of the partition coefficient (*F* and *F*′ in eqs 5 and 6, respectively) and the partition entropy (*H* in eq 7) is that they are functions of **U** (memberships) alone and, as a result, are only explicitly a function of the data set, **X**. In other words, they do not use the underlying structure in the data to test cluster validity. But their ease of formulation, coupled with the fact that on well-defined data sets they perform as well as any other measure, makes them quite attractive to use (at least, as a first pass filter). They also have other known disadvantages.[46] The other set of validity measures designed for fuzzy clustering uses three components—**U**, **V: X**, where **V** is a vector of *c*-prototypes of the data set **X**. This set of measures uses information from both the partition and the data set. Apart from the Xie-Beni index (*S* in eq 8), they include Gunderson’s separation coefficient[47] and the Fukayama-Sugeno index.[48] Since, these indices use more explicit information, they are not as intuitive as such measures as the partition coefficient and partition entropy. In a quantitative comparison of these validity measures on two very distinct data sets over a range of *m* (fuzzifier) values, it was found that the Xie-Beni index produced the most discriminating results.[49]

The Xie-Beni index provides a reasonable indication of the goodness of a *c*-partition only when the compactness of the clusters is of a comparable order of magnitude to the separation of the clusters. However, with increasing values of *c*, it is highly probable that the fuzzy clustering algorithm (FRC in the present case) finds the same cluster over and over again. Under these conditions, the denominator of eq 8 tends to zero as *c* increases. As a result, the index takes abnormally large values. This is illustrated by the case of the B-side proximity matrix $[C\_P2]_{T+R}$ (Figure S4 of the Supporting Information), where *S* is well behaved only in the region $2 < c < 10$ and takes on very large values for $c > 10$.

Clustering Space versus Visualization Space. In almost all visualization plots shown here, with the exception of 3D plots for the A-side clustering, the data set appears to be continuous (as opposed to being discrete and widely separated). A natural question is—why search for clusters in a continuous data set. However, the data could appear to be continuous as an artifact of the visualization space. For example, for full molecule clustering, the data set of conformers, which separates into five clusters in six-dimensional (6D, translational plus rotational) space, was viewed due to necessity in a 3D translational (Shift, Slide, and Rise) space. Although the five clusters appear to be random and continuous in the 3D translational space, this does not necessarily mean that in the higher 6D space the data are continuous as well. Even if the data are continuous in the higher-dimensional clustering space, the objective of clustering is to locate widely dissimilar representative patterns. The identified clusters cover a distinct region within this continuous space, and, hence, a representative conformer (a mean-located conformer within such a distinct region) is truly dissimilar to other identified representatives. This makes this technique suitable for identifying representative conformers for use in 3D-QSAR applications, such as CoMFA, as described below.

Relevance of Fuzzy Clustering to 3D-QSAR Studies. Rigid molecules have only a small number of conformers available at low energy. As a result, rigid molecules give important clues as to the binding (or bioactive) conformation. If a rigid molecule with only one possible low-energy conformation is shown experimentally to have high affinity for the receptor protein, then it is a reasonable assumption that its conformation is the bioactive conformation. Flexible molecules, however, may take on a large range of conformers that are very similar in energy. There is considerable evidence that they do not bind to receptors in their lowest energy conformation. The question arises, therefore, as to what conformation(s) of a flexible molecule should be used in 3D-QSAR studies. It should be noted that 3D-QSAR methods such as CoMFA do not require that the user know the binding conformation. Rather, CoMFA provides a statistical model that relates changes in molecular structure and properties to changes in biological activity. There is considerable evidence in the literature that a stable and predictable CoMFA model can be derived without exact knowledge of the binding conformation. For example, the Venanzi group used a superposition-dependent hierarchical clustering approach[13] to determine a set of representative conformers of **1** and a piperidine analogue of **1**. These representative conformers were used as templates for the construction of about 50 related analogues. Separate CoMFA studies were carried out on each set of analogues superimposed on a different representative conformer. Each study resulted in a CoMFA model. The most stable and predictive model was used to successfully interpret the DAT/SERT selectivity of the analogues.[16] This demonstrates that, for flexible molecules, using representative conformers from individual clusters as templates for CoMFA can result in a useful 3D-QSAR model.

The objective of the present work was to define a novel protocol for clustering conformations of flexible molecules using planes parameters. The protocol had the advantage of being independent of molecular superposition and could easily be generalized to any two planes in any molecule. The protocol was able to determine clusters in the molecule as a whole—an advantage over our previous superposition-dependent fuzzy clustering approach. A representative conformer for each cluster was determined. The procedure makes no claim that the binding conformation of the molecule is actually one of the representative conformers. In fact, it remains to be seen whether the representative conformers from the present work or those from our previous superposition-dependent fuzzy clustering work[14] would probe a sufficiently different part of conformational space as to lead to significantly different (and improved) CoMFA models than those based on the hierarchical clustering approach. This forms the basis of future work.

ACKNOWLEDGMENT

This work was supported in part by grants DA018153 to C.A.V. and R.N.D. and GM20861 to W.K.O. from the National Institutes of Health. The authors would like to thank Deepangi Pandit for assistance with some of the figures.

Supporting Information Available: Additional two- and three-dimensional conformer plots in translational and rotational space as well as validity plots. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Nicklaus, M. C.; Wang, S.; Driscoll, J.; Milne, G. W. A. Conformational Changes of Small Molecules Binding to Proteins. *Bioorg. Med. Chem.* **1995**, *3*, 411−428.

(2) Veith, M.; Hirst, J. D.; Brooks, C. L., III Do Active Site Conformations of Small Ligands Correspond to Low Free-Energy Solution Structures? *J. Comput.-Aided Mol. Des.* **1998**, *12*, 563−572.

(3) Debnath, A. K. Comparative Molecular Field Analysis (CoMFA) of a Series of Symmetrical Bis-Benzamide Cyclic Urea Derivatives as HIV-1 Protease Inhibitors. *J. Chem. Inf. Comput. Sci.* **1999**, *38*, 761−767.

(4) Perola, E.; Charifson, P. S. Conformational Analysis of Drug-like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization Upon Binding. *J. Med. Chem.* **2004**, *47*, 2499−2510.

(5) Guarnieri, F.; Weinstein, H. Conformational Memories and the Exploration of Biologically Relevant Peptide Conformations: An Illustration for the Gonadotropin-Releasing Hormone. *J. Am. Chem. Soc.* **1996**, *118*, 5580−5589.

(6) Hopfinger, A. J.; Tokarski, J. S. Three-Dimensional Quantitative Structure-Activity Relationship Analysis. In *Practical Application of Computer-Aided Drug Design*; Charifson, P. S., Ed.; Marcel Dekker: New York, 1997; pp 105−164.

(7) Barnett-Norris, J.; Guarnieri, F.; Hurst, D. P.; Reggio, P. H. Exploration of Biologically Relevant Conformations of Anandamide, 2-Arachinodylglycerol, and Their Analogues Using Conformational Memories. *J. Med. Chem.* **1998**, *41*, 4861−4872.

(8) Barnett-Norris, J.; Hurst, D. P.; Lynch, D. L.; Guarnieri, F.; Makriyannis, A.; Reggio, P. H. Conformational Memories and the Endocannabinoid Binding Site at the Cannabinoid CB1 Receptor. *J. Med. Chem.* **2002**, *45*, 3649−3659.

(9) Greenidge, P. A.; Merette, S. A. M.; Beck, R.; Dodson, G.; Goodwin, C. A.; Scully, M. F.; Spencer, J.; Weiser, J.; Deadman, J. J. Generation of Ligand Conformations in Continuum Solvent Consistent with Protein Active Site Topology: Application to Thrombin. *J. Med. Chem.* **2003**, *46*, 1293−1305.

(10) Bernard, D.; Coop, A.; MacKerell, A. D., Jr. 2D Conformationally Sampled Pharmacophore: A Ligand-Based Pharmacophore to Differentiate δ Opioid Agonists from Antagonists. *J. Am. Chem. Soc.* **2003**, *125*, 3101−3107.

(11) Bernard, D.; Coop, A.; MacKerell, A. D., Jr. Conformationally Sampled Pharmacophore for Peptide Delta Opiod Ligands. *J. Am. Chem. Soc.* **2005**, *48*, 7773−7780.

(12) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(13) Gilbert, K. M.; Venanzi, C. A. Hierarchical Clustering Analysis of Flexible GBR 12909 Dialkyl Piperazine and Piperidine Analogs. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 209−225.

(14) Misra, M.; Banerjee, A.; Davé, R. N.; Venanzi, C. A. Novel Feature Extraction Technique for Fuzzy Relational Clustering of a Flexible Dopamine Reuptake Inhibitor. *J. Chem. Inf. Model.* **2005**, *45*, 610−623.

(15) Fiorentino, A.; Pandit, D.; Gilbert, K. M.; Misra, M.; Dios, R.; Venanzi, C. A. Singular Value Decomposition of Torsional Angles of Analogs of the Dopamine Reuptake Inhibitor GBR 12909. *J. Comput. Chem.* **2006**, *27*, 609−620.

(16) Gilbert, K. M.; Boos, T. L.; Dersch, C. M.; Greiner, E.; Jacobson, A. E.; Lewis, D.; Matecka, D.; Prisinzano, T. E.; Zhang, Y.; Rothman, R. B.; Rice, K. C.; Venanzi, C. A. DAT/SERT Selectivity of Flexible GBR 12909 Analogs Modeled Using 3D-QSAR Methods. *Bioorg. Med. Chem.* **2007**, *20*, 1146−1159.

(17) Prisinzano, T.; Rice, K. C.; Baumann, M. H.; Rothman, R. B. Development of Neurochemical Normalization ("Agonist Substitution") Therapeutics for Stimulant Abuse: Focus on the Dopamine Uptake Inhibitor, GBR12909. *Curr. Med. Chem. CNS Agents* **2004**, *4*, 47−59.

(18) Glowa, J. R.; Fantegrossi, W. E.; Lewis, D. B.; Matecka, D.; Rice, K. C.; Rothman, R. B. Sustained Decrease in Cocaine-maintained Responding in Rhesus Monkeys with 1-[2-[bis(4-fluorophenyl)-methoxy]ethyl]-4-(3-hydroxy-3-phenylpropyl)piperazinyl Decanoate, a Long-acting Ester Derivative of GBR 12909. *J. Med. Chem.* **1996**, *39*, 4689−4691.

(19) Lu, X.-J.; Olson, W. K. 3DNA: A Software Package for the Analysis, Rebuilding and Visualization of Three-dimensional Nucleic Acid Structures. *Nucleic Acids Res.* **2003**, *31*, 5108−5121.

(20) Matecka, D.; Rothman, R. B.; Radesca, L.; de Costa, B. R.; Dersch, C. M.; Partilla, J. S.; Pert, A.; Glowa, J. R.; Wojnicki, F. H. E.; Rice, K. C. Development of Novel, Potent, and Selective Dopamine Reuptake Inhibitors Through Alteration of the Piperazine Ring of 1-[2-(diphenylmethoxy)ethyl]- and 1-[2-[bis(4-fluorophenyl)methoxy]ethyl]-4-(3-phenylpropyl)piperazines (GBR 12935 and GBR 12909). *J. Med. Chem.* **1996**, *39*, 4704−4716.

(21) Matecka, D.; Lewis, D.; Rothman, R. B.; Dersch, C. M.; Wojnicki, F. H. E.; Glowa, J. R.; DeVries, A. C.; Pert, A.; Rice, K. C. Heteroatomic Analogs of 1-[2-(diphenylmethoxy)ethyl]- and 1-[2-[bis-(4-fluorophenyl)methoxy]ethyl]-4-(3-phenylpropyl)piperazines (GBR 12935 and GBR 12909) as High-affinity Dopamine Reuptake Inhibitors. *J. Med. Chem.* **1997**, *40*, 705−716.

(22) Lewis, D. B.; Matecka, D.; Zhang, Y.; Hsin, L. W.; Dersch, C. M.; Stafford, D.; Glowa, J. R.; Rothman, R. B.; Rice, K. C. Oxygenated Analogues of 1-[2-(diphenylmethoxy)ethyl]- and 1-[2-[bis(4-fluorophenyl)methoxy]ethyl]-4-(3-phenylpropyl)piperazines (GBR 12935 and GBR 12909) as Potential Extended-action Cocaine-abuse Therapeutic Agents. *J. Med. Chem.* **1999**, *42*, 5029−5042.

(23) Hsin, L. W.; Dersch, C. M.; Baumann, M. H.; Stafford, D.; Glowa, J. R.; Rothman, R. B.; Jacobson, A. E.; Rice, K. C. Development of Long-acting Dopamine Transporter Ligands as Potential Cocaine-abuse Therapeutic Agents: Chiral Hydroxyl-containing Derivatives of 1-[2-[bis(4-fluorophenyl)methoxy]ethyl]-4-(3-phenylpropyl)piperazine and 1-[2-(diphenylmethoxy)ethyl]-4-(3-phenylpropyl)piperazine. *J. Med. Chem.* **2002**, *45*, 1321−1329.

(24) Lewis, D.; Zhang, Y.; Prisinzano, T.; Dersch, C. M.; Rothman, R. B.; Jacobson, A. E.; Rice, K. C. Further Exploration of 1-{2-[bis-(4-fluorophenyl)methoxy]ethyl}piperazine (GBR 12909): Role of N-aromatic, N-heteroaromatic, and 3-oxygenated N-phenylpropyl Substituents on Affinity for the Dopamine and Serotonin Transporter. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1385−1389.

(25) Dickerson, R. E.; Bansal, M.; Calladine, C. R.; Diekman, S.; Hunter, W. N.; Kennard, O.; von Kitzing, E.; Lavery, R.; Nelson, H. C. M.; Olson, W. K.; Saenger, W.; Shakked, Z.; Sklenar, H.; Soumpasis, D. M.; Tung, C.-S.; Wang, A. H.-J.; Zhurkin, V. B. Definitions and Nomenclature of Nucleic Acid Structure Parameters. *J. Mol. Biol.* **1998**, *208*, 787−791.

(26) Werner, M. H.; Gronenborn, A. M.; Clore, G. M. Intercalation, DNA Kinking, and the Control of Transcription. *Science* **1996**, *271*, 778−784.

(27) Lu, X.-J.; Olson, W. K. Resolving the Discrepancies Among Nucleic Acid Conformational Analyses. *J. Mol. Biol.* **1999**, *285*, 1563−1575.

(28) Lu, X.-J.; Babcock, M. S.; Olson, W. K. Overview of Nucleic Acid Analysis Programs. *J. Biomol. Struct. Dyn.* **1999**, *16*, 833−843.

(29) Olson, W. K.; Bansal, M.; Burley, S. K.; Dickerson, R. E.; Gerstein, M.; Harvey, S. C.; Heinemann, U.; Lu, X.-J.; Neidle, S.; Shakked, Z.; Sklenar, H.; Suzuki, M.; Tung, C.-S.; Westhof, E.; Wolberger, C.; Berman, H. M. A Standard Reference Frame for the Description of Nucleic Acid Base-pair Geometry. *J. Mol. Biol.* **2001**, *313*, 229−237.

(30) Downs, G. M.; Barnard, J. M. Clustering Methods and Their Uses in Computational Chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 2002; Vol. 18, pp 1−40.

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

(32) Chema, D.; Goldblum, A. The Nearest Neighbor Method - Finding Families of Conformations Within a Sample. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 208−217.

(33) Feher, M.; Schmidt, J. M. Metric and Multidimensional Scaling: Efficient Tools for Clustering Molecular Conformations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 346−353.

(34) 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.

(35) Davé, R. N.; Sen, S. Robust Fuzzy Clustering of Relational Data. *IEEE Trans. Fuzzy Syst.* **2002**, *10*, 713−727.

(36) Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*; Plenum Press: New York, 1981.

(37) Xie, X. L.; Beni, G. A Validity Measure for Fuzzy Clustering. *IEEE Trans. Pattern Anal. Machine Intelligence* **1991**, *13*, 841−847.

(38) SYBYL 6.9; Tripos Inc.: 1699 South Hanley Rd., St. Louis, MO 63144, U.S.A.

(39) Clark, M.; Cramer, R. D., III; van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10*, 982−1012.

(40) Horn, B. K. P. Closed-form Solution of Absolute Orientation Using Unit Quaternions. *J. Opt. Soc. Am.* **1987**, *4*, 629−642.

(41) Wishart, D. k-Means Clustering with Outlier Deletion, for Data Mining with Mixed Variables and Missing Values. In *Exploratory Data Analysis in Empirical Research*; Schwaiger, M., Opitz, O., Eds.; Springer: Berlin, 2002; pp 216−226.

(42) Pandit, D.; Roosma, W. A.; Misra, M.; Gilbert, K. M.; Skawinski, W. J.; Venanzi, C. A. Conformational Analysis of Piperazine and Piperidine Analogs of GBR 12909: Stochastic Approach to Evaluation of the Effect of Force Field and Solvent. *J. Molec. Model.* Submitted for publication.

(43) Cannon, R. L.; Bezdek, J. C. Efficient Implementation of the Fuzzy *c*-Means Clustering Algorithms. *IEEE Trans. Pattern Anal. Machine Intelligence* **1986**, *8*, 248−255.

(44) Chung, F. L.; Lee, T. Fuzzy Competitive Learning. *Neural Net.* **1994**, *7*, 539−551.

(45) Bezdek, J. C.; Tsao, E.; Pal, N. R. Fuzzy Kohonen Clustering Networks. *Pattern Recognit.* **1994**, *27*, 757−764.

(46) Davé, R. N. Validating Fuzzy Partitions Obtained Through *c*-Shells Clustering. *Pattern Recognit. Lett.* **1996**, *17*, 613−623.

(47) Gunderson, R. Applications of Fuzzy ISODATA Algorithms to Star-tracker Printing Systems. *Proceedings of the 7th Triennial World IFAC Congress*; Helsinki, Finland, 1978; Pergamon: pp 1319−1323.

(48) Fukuyama, Y.; Sugeno, M. A New Method of Choosing the Number of Clusters for the Fuzzy *c*-Means Method. *Proceedings of the 5th Fuzzy Systems Symposium*: Kobe, Japan, 1989; Japan Society for Fuzzy Sets and Systems: pp 247−250.

(49) Pal, N. R.; Bezdek, J. C. On Cluster Validity for the Fuzzy *c*-Means Model. *IEEE Trans. Fuzzy Syst.* **1995**, *3*, 370−379.

CI7001632