Novel Feature Extraction Technique for Fuzzy Relational Clustering of a Flexible Dopamine Reuptake Inhibitor

Milind Misra,[†] Amit Banerjee,[‡] Rajesh N. Davé,[‡] and Carol A. Venanzi*[,†]

Departments of Chemistry and Environmental Science and Mechanical Engineering, New Jersey Institute of Technology, Newark, New Jersey 07102

Received September 28, 2004

This paper describes a novel clustering methodology for classifying over 700 conformations of a flexible analogue of GBR 12909, a dopamine reuptake inhibitor that has completed phase I clinical trials as a treatment for cocaine abuse. The major aspect of the clustering methodology includes an efficient data-conditioning scheme where a systematic feature extraction procedure based on the structural properties of the molecule was used to reduce the associated feature space. This allowed region-specific clustering that focused on individual pharmacophore elements of the molecule. For clustering of the reduced feature set, the fuzzy clustering partitional method was utilized. Due to the relational nature of the feature data, fuzzy relational clustering was employed, and it successfully detected natural groups defined by rotational minima around $N(sp^3)-C(sp^3)$, $O(sp^3)-C(sp^3)$, and $C(sp^3)-C(sp^2)$ bonds. The proposed clustering methodology also employed several cluster validity measures, which corroborated the partitions produced by the clustering technique and agreed with the results of hierarchical clustering using the XCluster program. Representative structures which exhibited a reasonable spread of energies and showed good spatial coverage of the conformational space were identified for use as putative bioactive conformations in a future Comparative Molecular Field Analysis of GBR 12909 analogues. The clustering methodology developed here is capable of handling other computational chemistry problems, and the feature extraction technique can be easily generalized to other molecules.

INTRODUCTION

In ligand-based drug design, the identification of the bioactive conformation of a promising drug candidate is of great interest. The bioactive conformation of a drug is the conformation in which it binds to the protein receptor. In the absence of structural information about the receptor, the prediction of the bioactive conformation of the ligand can be very challenging. There is considerable evidence that drug molecules do not bind to proteins in their vacuum-phase global energy minimum (GEM) conformation.[1−4] Conformational searching techniques can be used to explore the conformational space of a ligand and locate minima on the ligand's potential energy surface. For a flexible molecule the number of minima generated can be very large. This prohibits consideration of every minimum conformation as a putative bioactive conformation. But others have shown the importance of considering conformations other than the GEM in pharmacophore modeling.[5−10] Therefore, selection of a suitable set of representative conformers for analysis is an important first step in 3D QSAR techniques, such as Comparative Molecular Field Analysis (CoMFA).[11] In such cases, it becomes essential to use data reduction techniques such as clustering in order to first identify well-defined groups of conformations and then to select representative structures from each group.

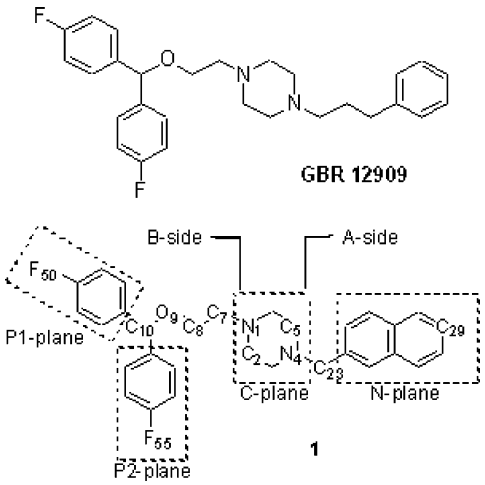

Figure 1. Structures of GBR 12909 and **1**. Atoms C_7, C_8, O_9, C_{10}, C_{23}, C_{29}, F_{50}, and F_{55} and the angles between each pair of planes were used to define the minimal feature set for **1**. Atoms N_1, C_2, N_4, and C_5 were used to define the C-plane. The A-side consists of C_{23} and the N-plane. The B-side consists of C_7, C_8, O_9, C_{10} and the P1- and P2-planes.

This paper presents clustering studies of conformations of **1**[12] (Figure 1), an analogue of 1-[2-[bis(4-fluorophenyl)-methoxy]ethyl]-4-(3-phenylpropyl)piperazine (GBR 12909). Analogues of GBR 12909 belong to a class of dopamine reuptake inhibitors that might be potentially useful in the treatment of cocaine abuse.[13] GBR 12909 has been found to be effective in reducing cocaine self-administration in rhesus monkeys without significantly affecting food-maintained

* Corresponding author phone: (973)596-3596; fax: (973)596-3596; e-mail: venanzi@njit.edu.
[†] Department of Chemistry and Environmental Science.
[‡] Department of Mechanical Engineering.

10.1021/ci049708d CCC: $30.25
Published on Web 04/28/2005

responding[12,14] and has completed phase I clinical trials.[13] The purpose of cluster analysis of the conformations of flexible GBR 12909 analogues is to identify a small number of representative conformations that could aid in understanding the interaction between the analogues and the dopamine transporter (DAT). These representative conformations will be used in a future study as templates for molecular alignment in CoMFA studies of a large set of GBR 12909 analogues.

Data clustering as a means of classification has been used extensively in computational chemistry, and a thorough review of techniques is available.[15] Clustering of molecules in chemical databases has received considerably more attention than clustering of molecular conformations. Attempts to cluster conformations have been based on some sort of proximity measure between pairs of conformers. The most popular clustering techniques for generation of representative conformers are hierarchical techniques such as the single-link clustering and the average-link clustering schemes. The single-link clustering package XCluster[16] clusters conformations based on a root-mean-square (RMS) distance matrix derived either from a set of atom coordinates (with or without rigid body superposition of conformations) or from a set of torsional angles. An average-link clustering technique has been used to demonstrate clustering of 63 conformations of a tripeptide fragment based on a Euclidean distance measure of proximity in a 36-dimensional space.[17] However, as data sets become larger, techniques based on dendrograms are impractical for more than a few hundred patterns.[18] Another problem is that such techniques may tend to find singleton clusters unless carefully selected termination criteria are utilized. Other general disadvantages of hierarchical schemes include sensitivity to outliers, since sources of error and variance are not considered, and since there is no relocation of the objects along the hierarchy, objects once assigned *incorrectly* cannot be reassigned.[19] From an application point of view, hierarchical schemes are particularly useful for detecting sequential levels of clustering, such as in taxonomy and in biological classification.

In contrast, partitional clustering techniques have been developed to uncover a single best partition of the data.[20] Unlike hierarchical clustering, partitional clustering techniques produce a single c-partition of the data based on a clustering criterion, which can either be a global criterion (such as minimization of square-error or maximization of expectation) or a local criterion (such as estimation of regions of high density, a nearest neighbor (NN) criterion). The global clustering criterion is also called an objective-function-based optimization and such schemes represent each cluster by a *prototype* and assign objects to clusters according to the most similar prototype. A variant of the NN-based local clustering criterion, called the nearest single neighbor method,[21] was used to cluster different sets of peptide conformations and was based on a proximity measure derived from RMS distances between pairs of conformations defined by only the peptide backbone structure. While NN techniques have been shown to be useful, they tend to be computationally expensive for large data sets.

Recently, attempts have been made to cluster families of conformations using statistical scaling techniques as cluster analysis tools. For example, families of relatively small or rigid molecules such as dopamine, roseotoxin-B, and cycloheptadecane were clustered by first scaling the higher dimensional data in real space to a reduced 3-D conformational space using both multidimensional and metric scaling techniques. Then either visual inspection or a hierarchical technique applied to a proximity matrix derived from the reduced 3-D data set was used to complete the clustering.[22] Subsequently, the same 3-D data set was clustered using fuzzy clustering.[23] To the best of our knowledge, these are the only instances where partitional clustering schemes have been successfully applied to cluster families of conformations.

Fuzzy *c*-Means[24] and a large number of its derivative algorithms[25] are partitional clustering schemes, which are based on different types of square-error minimization and specifically target detection of cluster prototypes, ranging from a simple prototype that is an n-dimensional point to complex shapes that include nonlinear surfaces and manifolds in n-dimensional space.[26] The motivation to use an objective-function-based partitional scheme in the present work, instead of hierarchical or local criterion-based partitional techniques, comes from the fact that such schemes are most suited for efficient representation and compression of large data sets. They not only inherently search for natural groupings but also find the cluster prototypes, i.e., ideal representatives, which in the present case are the most representative molecular conformations. While there are many advantages of such clustering methods, they suffer from two problems that face all clustering methods, including hierarchical clustering, namely, susceptibility to noise or outliers and difficulty in determining the exact number of clusters within the data. Fortunately, the recent development of robust clustering algorithms[27] and highly meaningful cluster validity measures[28,29] address these two problems. Availability of these algorithms and recent examples of the application of fuzzy clustering by the Feher group[22,23] provide a motivation for utilizing such methods for the present work.

The use of fuzzy memberships has several advantages. First, they are particularly useful when there are no easily identifiable, clear groupings in the data set. Second, having fuzzy memberships is generally helpful in terms of smoother convergence of the numerical algorithm as compared to the use of nonfuzzy (also termed crisp or hard) memberships.[30] Moreover, partitioning schemes provide automatic detection of cluster boundaries. In the case of fuzzy clustering, these cluster boundaries can overlap. Use of fuzzy memberships is therefore advantageous in the present application because overlap of cluster boundaries is expected since the flexibility of the molecule results in a continuous "spectrum" of closely related conformers. For fuzzy clustering, since every individual data entity (a conformer, in the present case) belongs to not one but all clusters with varying degrees of membership, the clustering results can provide a natural interpretation of the goodness of the partition. The fuzzy membership (a numerical quantity between 0 and 1) is directly related to the structure of the partition. Hence almost all the fuzzy cluster validity measures are based on fuzzy memberships and even though they do not take the cluster geometry into account, they often provide meaningful interpretable results.[31,32] Based on these well-known advantages of fuzzy memberships, it was decided to utilize fuzzy clustering algorithms. However, the most important aspect of the present application is that it incorporates the use of an

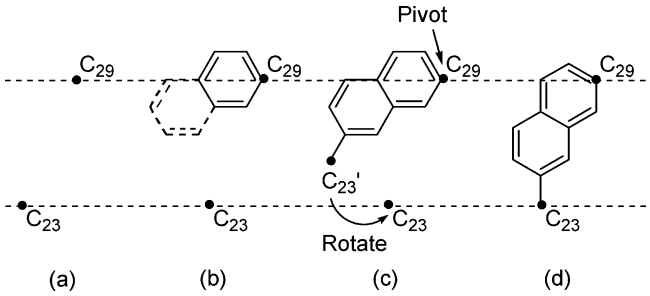

Figure 2. Reconstruction sequence for the A-side. (a) Exact locations of C_{23} and C_{29} are known and fixed on the N-plane. (b) Construction of the naphthalene fragment is possible using information about bond lengths and bond angles to obtain an arbitrary orientation of the fragment. (c) C_{23}' is the arbitrary position of C_{23} obtained after construction of the naphthalene ring. (d) Rotation about C_{29} so that C_{23}' coincides with C_{23} gives the true orientation of the ring on the N-plane.

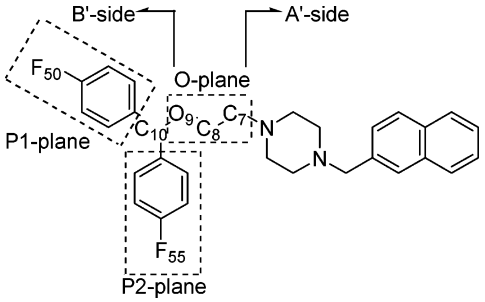

Figure 3. Elements of the modified feature vector for the B′-side only. Atoms C_{10}, F_{50}, and F_{55}, and the angles between each of the phenyl planes and the O-plane were used to define the feature vector for the B′-side. Atoms C_7, C_8, and O_9 were used to define the O-plane. The A′-side consists of the piperazine and naphthalene rings and the connecting carbon atom. The B′-side consists of the bisphenyl group (C_{10}, P1- and P2-planes).

objective-function-based partitional scheme instead of methods such as hierarchical clustering or multivariate data analysis.

Our fuzzy clustering approach differs from that of Feher and Schmidt[23] in several ways. Since there are problems associated with labeling the symmetric heavy atoms in the phenyl ring (discussed further in the next section), the present approach is based on reducing the raw data to a manageable form, taking care of the symmetry issue in the process. A proximity-based distance measure using all available heavy atom information appeared to be more relevant than Feher and Schmidt's approach of scaling down the data by considering only the symmetry-unique atoms. The use of a proximity matrix also meant that the clustering should be performed on the relational data domain instead of the object-space-based clustering used by Feher and Schmidt. Further, the use of a proximity measure facilitates incorporation of heuristic or many popular non-Euclidean similarity measures (for example, L_1 norm, p norm, Manhattan norm, etc.), implying that the relational data is not necessarily based on Euclidean measure. For clustering non-Euclidean relational data, the NERFCM (non-Euclidean relational fuzzy *c*-means),[33] which is the relational dual of fuzzy *c*-means (FCM),[24] is a popular choice. This method was initially utilized in our work and was also compared with the newly developed Fuzzy Relational Clustering (FRC) procedure.[34] Both the methods were found to provide nearly identical results. However, FRC was used for the results reported in this paper, because it does not require the beta-spread transformation that is needed in NERFCM. Hence it is computationally attractive, and its performance can be further improved by using a Seidel iterative scheme. In terms of the overall methodology developed in this paper, the use of FRC is indicated for all future applications because it can be easily utilized in its robust version to handle noisy data, although that is not highly relevant to the case considered here.

Besides the potential advantages listed above, the present approach is novel for several reasons. First, to our knowledge it is the only fuzzy clustering study of a very flexible molecule. Second, region-specific clustering that focused on individual pharmacophore elements of the molecule was made possible by defining feature vectors in terms of the A- and B-side (see Figure 1) or A′- and B′-side (see Figure 3) moieties which contain important chemical features of the pharmacophore. Third, fuzzy relational clustering was performed using a novel feature extraction technique (outlined in the next section). The FRC procedure used a proximity matrix derived from a feature vector that contained real space elements (atom coordinates and angles between planes) that were related to the pharmacophore elements of the molecule. As a result, the feature vectors described in this work are unique to **1** since they were derived from the geometric constraints of this particular molecule. However, as will be illustrated in the Discussion section, the novel feature extraction technique can be generalized to other molecules.

NOVEL FEATURE EXTRACTION

Careful feature extraction is a crucial first step in pattern recognition. Two motivations guided the feature extraction process: (a) reduction of the feature space and (b) handling of the symmetry of each phenyl ring.

The data set consists of conformations of **1**. The molecular structure of **1** is shown in Figure 1. If the Cartesian coordinates of each heavy (non-hydrogen) atom were taken as a feature, the dimensionality of the resulting feature space would be $N\times3$, where N is the number of heavy atoms in the molecule. In general, reduced dimensionality of a large input data matrix is desirable for more easily interpretable results. Moreover, some features are redundant, and retaining redundant data not only makes the feature space high dimensional and cluttered yet sparse but also usually has periodicity that makes data classification and interpretation very difficult. Therefore, it is preferable to use the smallest possible feature vector by replacing the factor $N\times3$ by the minimum number of features necessary to describe the molecule (i.e., the *minimal feature set*). In the Methods section we present a novel feature extraction method for obtaining a reasonably sized feature vector for clustering. We investigate how different molecular superpositions can be used in combination with different minimal feature sets in order to focus the clustering on those parts of the molecule that contain important pharmacophore features.

The problem of phenyl ring symmetry was handled by using molecular planes as part of the feature set. Each of the two phenyl rings of **1** contains symmetry-equivalent atoms that have different atom labels. For example, the 2-position carbon in the P1-plane in Figure 1 is atom number C_{12}, whereas the 2′-position carbon is atom number C_{16}.

Rotation of the phenyl ring of the P1-plane by 180° gives a molecular structure which is indistinguishable from the previous one, yet the labeled atoms are in different positions. Superposition of these two structures would show a perfect fit, yet calculations that are based on atom labels, such as the RMS distance between atoms, would show a large difference. As described in the Methods section, our consideration of the planes on which the phenyl rings lie provides an atom label-independent solution to the symmetry problem. Atom label-independent description of the phenyl rings was achieved by using plane equations, which specify the planar orientations of the phenyl rings, and selected atomic coordinates (of the two fluorine atoms), which specify their exact location.

FUZZY CLUSTERING

As discussed above, an object-space-based fuzzy clustering scheme such as FCM[24] could be used to cluster the data in the reduced feature space. However, to be consistent with the proposed general methodology, it was clear that converting the data in the reduced feature space into a proximity distance matrix would provide a better understanding of the interconformational similarities and allow for introducing measures that are not strictly based on Euclidean distance. Moreover, such a proximity matrix could also handle any subjective similarity information which would be impossible to achieve in an object space, while it would be easily handled through use of NERFCM or FRC. As will be demonstrated in the Methods section, the feature vectors for **1** consist of "mixed" features (i.e., features with different physical units). In this case, each feature vector is a set of atom coordinates and angles between planes. The relational data matrix obtained from this feature vector can easily involve a non-Euclidean measure of dissimilarity. Accordingly, a relational clustering technique capable of handling non-Euclidean data to generate partitions was used.[34]

FRC is a recently developed relational clustering technique and is conceptually attractive because it works directly on the non-Euclidean data without first converting it to a Euclidean measure. The scheme is therefore less constrained than most of the other relational clustering techniques. 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

$$D_{jj} = 0,\ D_{jk} \ge 0,\ D_{jk} = D_{kj},\ 1 \le j, k \le n \tag{1}$$

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.[34] Here n is the number of data objects and c is the number of clusters fixed a priori. The update equations used for **U** and **A** are shown in (2) and (3)

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

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

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

$$v_i = \frac{(u_{i1}^{m}, u_{i2}^{m}, \ldots, u_{in}^{m})^T}{\sum_{k=1}^{n} u_{ik}^{m}} \tag{4}$$

The membership matrix, **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, ϵ. 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 $u_{ij} > u_{kj}$ $(k \ne j)$ for all $1 \le i \le c$, $1 \le j \le n$. The representative conformation is identified as the one with the highest membership value in that particular cluster. 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

Different fuzzy cluster validity indices and measures have been proposed in the literature to characterize the goodness of the partition. The simplest of these is the partition coefficient,[24] which describes the fuzziness of the partition. It is inversely proportional to the average fuzzy overlap between the clusters, and is given by

$$F = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^2 \tag{5}$$

$F = 1$ indicates no overlap between clusters and is the case when FCM degenerates to hard c-means. On the other hand, $F \to 1/c$ is the extreme fuzzy case when all the entities are shared equally between all the clusters. Hence, the partition coefficient can take values between $1/c \le F \le 1$. Normalizing F as shown below can compensate for this dependence on c.

$$F' = \frac{cF - 1}{c - 1} \tag{6}$$

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.

The application of Shannon's entropy[35] to fuzzy clustering resulted in another cluster validity measure known as the

partition entropy[24] and is given by

$$H = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ik} \ln u_{ik} \tag{7}$$

A good partition is characterized by a low value of H; it can take values between $0 \leq H \leq \ln c$. Since H varies with ln c, the monotonically decreasing tendency of H with c is not as severe as in the case of F and hence normalizing H has little beneficial effect.

Both the partition coefficient and entropy measure the amount of fuzziness from cluster membership information and do not consider geometric properties such as size, shape, and compactness of the clusters. Gath and Geva[36] proposed using fuzzy volume and fuzzy density of the clusters as cluster validity criteria; a good cluster is characterized by a high value of fuzzy partition density and an accompanying low value of fuzzy hypervolume. The compactness criterion[28] 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 a_{ik}'s from eq 3 by

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

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

Conformational Analysis. The data set of conformations of **1** was obtained by random search of the conformational space using version 6.9 of the SYBYL molecular modeling package.[37] Since ab initio quantum mechanical calculations at the HF/6-31G* level on GBR 12909 showed that protonation on N_4, the piperazinyl nitrogen distal to the bisphenyl group, was favored over protonation on N_1 (W. J. Skawinski, personal communication), and Dutta et al.[38] showed that N_4 is more essential for activity than N_1 in piperidine analogues of GBR 12909, the molecule was protonated on N_4 prior to random search. The piperazine ring was fixed as an aggregate, and the eight torsional angles were randomly altered during the search. The piperazinyl side chains were both maintained in the equatorial position by checking for chirality so that the conformers were not reflected through a plane for comparison against conformers already found by the search. Symmetry was checked to reduce the number of bonds selected for rotation at each iteration. One thousand search iterations were carried out. At each step in the iteration, the eight torsional angles were randomly altered, and the resulting structure was minimized using the Powell minimization method[39] and a convergence threshold of 0.05. The Tripos force field[40] was used along with Gasteiger−Hückel charges and a nonbonded distance cutoff of 8.0 Å. A minimized conformer was "accepted" as a new conformer if it met the following energy and RMS criteria: (1) Its RMS distance difference compared to all other conformers was at least 0.20 Å, and (2) its energy was within 20 kcal/mol of the energy of the conformer identified to have the lowest energy at that particular step in the random search. The random search procedure ended after 1000 steps, and the "accepted" conformations were collected and used for the clustering study. The conformational analysis was performed on a SGI 500-MHz IP35 processor with 512 MB RAM on an IRIX release 6.5 operating system.

Superposition and Feature Extraction. The GBR 12909 analogue, **1**, contains two pharmacophore features that are found in most dopamine reuptake inhibitors: a quaternary nitrogen (N_4 in Figure 1) in close proximity to an aromatic ring (here, the naphthalene ring). It also contains a bisphenyl group which has been shown to be necessary for good binding affinity of GBR 12909 analogues.[12,14,41−43] To aid in the feature extraction process, the molecule was conceptually divided as described below into regions containing these pharmacophore elements. Two different types of superpositions were applied to the data set of molecular conformations. Superposition 1 involved atoms in the region close to the DAT inhibitor pharmacophore elements, while Superposition 2 involved atoms in the region close to the bisphenyl group. Different *minimal features sets* were identified for each superposition. In this way the effect of clustering the conformations using a feature set defined for the molecule as a whole versus using feature sets defined for various fragments could be compared. Because the different superpositions affect the data dissimilarity matrix, **D**, and the distance matrix, **A**, the clustering results are superposition-dependent. The superpositions and the related feature vectors are summarized in Table 1 and are described below.

1. Superposition 1. The data set of molecular conformations was superimposed by a rigid body superposition using atoms N_1, C_2, N_4, and C_5 in the piperazine ring. These four atoms form the central piperazine ring plane, C-plane, shown in Figure 1. The C-plane was fixed in the $y = 0$ plane for all structures. The molecules were translated in space so that N_1 was at the origin of the coordinate system. The molecule was divided into A- and B-sides around the C-plane as shown in Figure 1. The A-side and N_4 contain the DAT inhibitor pharmacophore elements. The B-side contains the bisphenyl group.

If the features were defined by the Cartesian coordinates of each heavy atom, the dimensionality of the resulting feature space would be $35 \times 3 = 105$. However, since the six heavy atoms in the ring have the same coordinates in every conformer in Superposition 1, they can be excluded from the coordinate data matrix. This results in a feature space of size $29 \times 3 = 87$, which is still quite large. Three different feature vectors were constructed using the novel feature extraction method described below in order to further reduce the size of the feature space and to compare the effects of clustering on the full molecule versus the A-side or the B-side.

a. Feature Extraction for A-side Clustering, Superposition 1. Examination of **1** indicates that the A-side of the molecule can be reconstructed using two sets of atom coordinates and one plane equation. The reconstruction sequence for the A-side, using coordinates of atoms C_{23} and C_{29} and the plane equation for the N-plane, is illustrated in Figure 2. Starting with the known position in space of a

Table 1. Summary of Feature Vectors Used for the Two Superpositions

		feature vector		
superposition[a]	clustering side	atoms	angle between planes[b]	best result[c]
1	A	C_{23}, C_{29}	N/C	$c = 3$, $c = 6$
1	B	C_7, C_8, O_9, C_{10}, F_{50}, F_{55}	P1/C, P2/C	$c =$ none
1	full molecule	C_7, C_8, O_9, C_{10}, C_{23}, C_{29}, F_{50}, F_{55}	P1/C, P2/C, N/C	$c =$ none
2	B′	C_{10}, F_{50}, F_{55}	P1/O, P2/O	$c = 9$

[a] 1: Atoms defining the C-plane (N_1, C_2, N_4, C_5). 2: Atoms defining the O-plane (C_7, C_8, O_9). [b] X/Y denotes angle between X-plane and Y-plane. [c] $c =$ none: no natural groups detected.

single atom, C_{29}, it is possible to use bond length and bond angle information to construct the rest of the naphthalene fragment in the plane specified by the known plane equation of the N-plane. Once an arbitrary orientation of the naphthalene fragment is obtained, it is rotated about C_{29} within the N-plane such that C_{23}', the arbitrary location of C_{23}, coincides with the true known coordinates of C_{23}. The resulting fragment fully specifies the A-side. The coordinates of atoms C_{23} and C_{29} and the plane equation for the N-plane form the minimal feature set for the A-side because these features contain the minimum information needed to completely specify the A-side of each conformation. The A-side feature vector used as the input to the fuzzy clustering algorithm was derived from the minimal feature set and consists of coordinates of C_{23} and C_{29} and the angle between the N-plane and C-plane, as summarized in Table 1. Since the C-plane is fixed in the $y = 0$ plane for all conformations, it is excluded from the definition of the minimal feature set and only the equation for the N-plane need be included. The two atoms and the two planes that define the angle between planes are labeled in Figure 1. The dimensionality of the feature space for A-side-only clustering is thus reduced to [2 × 3 coordinates + 1 angle] = 7.

b. Feature Extraction for B-Side Clustering, Superposition 1. Examination of **1** indicates that the B-side of the molecule can be reconstructed using six sets of atom coordinates and two plane equations. The reconstruction sequence for the B-side begins with known coordinates for atoms F_{50} and F_{55} and known equations of the P1- and P2-planes. The two phenyl rings are constructed within the P1- and P2-planes using bond angle and bond length information for atoms in a phenyl ring. Once arbitrary positions for each phenyl ring are obtained, they are rotated about F_{50} and F_{55} within the P1- and P2-planes, respectively, such that C_{10}', the arbitrary location of C_{10}, coincides with the true known coordinates of C_{10}. Further inclusion of the known coordinates of atoms O_9, C_8, and C_7 then completely specifies the B-side of each conformation. Thus, the coordinates of atoms C_7, C_8, O_9, C_{10}, F_{50}, and F_{55} and the equations of P1- and P2-planes form the minimal feature set for the B-side. The feature vector for B-side clustering derived from this minimal feature set is summarized in Table 1, and the required atoms and planes are labeled in Figure 1. The dimensionality of the feature space for B-side-only clustering is thus reduced to [6 × 3 coordinates + 2 angles] = 20.

c. Feature Extraction for Full Molecule Clustering, Superposition 1. Examination of **1** indicates that the combination of the minimal feature sets for the A- and B-sides leads to the minimal feature set for the entire molecule. Since, for all conformations of **1**, the piperazinyl nitrogen N_1 was fixed at the origin and the C-plane was fixed in the $y = 0$ plane, reconstruction of the entire molecule can be fully described using the minimal feature sets of the A- and B-sides. Thus, the molecule can be reconstructed using known coordinates of eight atoms and three known plane equations. These eight atoms and three planes are labeled in Figure 1, and the feature vector derived from this minimal feature set is summarized in Table 1. Compared to a dimensionality of 87 based only on atom coordinates, the dimensionality of the feature space obtained here is significantly reduced to [8 × 3 coordinates + 3 angles] = 27.

2. Superposition 2. To focus the clustering on the side of the molecule containing the bisphenyl group, **1** was divided into an A′- and a B′-side as shown in Figure 3. The molecular conformations of **1** were superimposed on the O-plane formed by atoms C_7, C_8, and O_9. For all structures, this O-plane was fixed in the $z = 0$ plane with the oxygen atom at the origin.

a. Feature Extraction for B′-Side Clustering, Superposition 2. Examination of **1** indicates that the B′-side of the molecule can be reconstructed using three sets of atom coordinates and two plane equations. The minimal feature set for the B′-side consists of coordinates of atoms C_{10}, F_{50}, and F_{55} and the equations of the P1- and P2-planes. The reconstruction sequence for the B′-side begins with known coordinates for atoms F_{50} and F_{55} and known equations of the P1- and P2-planes. The two phenyl rings are constructed within the P1- and P2-planes as above. After arbitrary positions for each phenyl ring are obtained, the rings are rotated about F_{50} and F_{55} within the P1- and P2-planes, respectively, such that C_{10}', the arbitrary location of C_{10}, coincides with the true known coordinates of C_{10}. Since, for all conformations, the O_9 atom is fixed at the origin and the O-plane is fixed in the $z = 0$ plane, atom O_9 and the O-plane are excluded from the definition of the minimal feature set for the B′-side. The feature vector for B′-side clustering is summarized in Table 1, and the required atoms and planes are labeled in Figure 3. The dimensionality of the feature space for the B′-side is [3 × 3 coordinates + 2 angles] = 11.

Feature extraction to generate the various feature vectors from the minimal feature sets was carried out by an in-house C++ program on a Visual C++ compiler on a 32-bit windows desktop environment.

Determining the Angle Between Planes. The angle between two planes is an important part of each feature set. The following protocol describes the procedure for calculating this angle. If two nonparallel planes intersect, their angle of intersection can be characterized by the acute angle θ, where $0 < \theta < \pi/2$, or the obtuse angle ϕ, where $\phi = \pi - \theta$. Whether the acute or obtuse angle is used to define the angle between two planes, each containing a fragment of

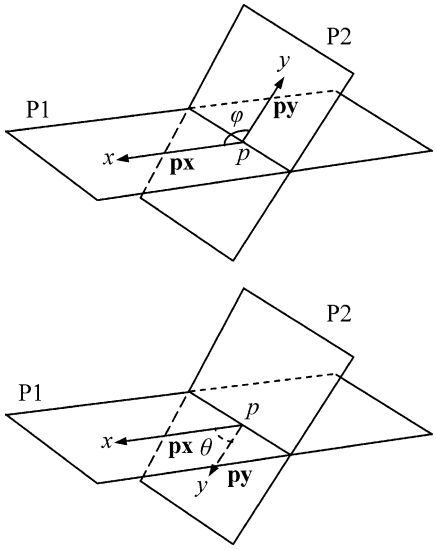

Figure 4. Determination of the angle between two planes.

molecular substructure, depends on which side of the planes the fragments lie. Figure 4 illustrates the two cases of the angle between planes containing molecular fragments. In general the equations of the two intersecting planes, P1 and P2, are known, and hence the line of intersection of the two planes can be determined. Choosing any arbitrary point p on the line of intersection determines vectors **px** and **py**, where x and y are points on the desired side of the planes P1 and P2, respectively. The dot product of the vectors **px** and **py** is $\mathbf{m} = \mathbf{px}\cdot\mathbf{py}$. For nonparallel planes, $\mathbf{m} \neq 0$; for positive **m** the desired angle of intersection is the acute angle θ and for negative **m** it is the obtuse angle ϕ. This scheme ensures a consistency in the way the feature vector is built for all conformations in the data set.

Clustering. Each feature vector was converted into the interconformer dissimilarity matrix **D** and used as the input for FRC. For each value of c ($2 \leq c \leq 12$), the clustering routine was run 10 times with a different random initialization, and the median value of membership was considered for cluster assignments. 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 ϵ (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.

RESULTS

Conformational Analysis. A total of 728 unique conformers were found in the range 11.2−27.9 kcal/mol. Every conformer was found at least once. The measure of completeness of the random search is given by[44]

$$\text{probability of finding all conformers} = 1 - (0.5)^n$$

where n is the number of times each conformer was found.

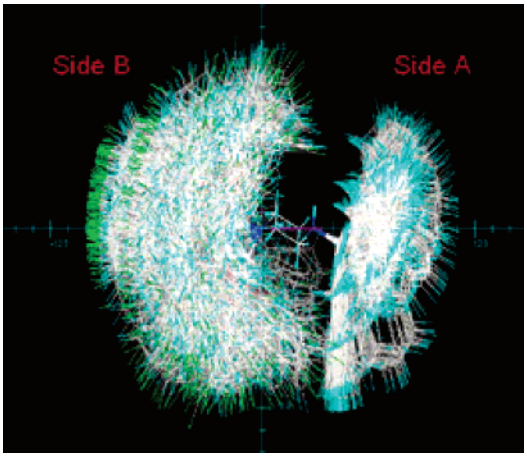

Figure 5. Side view of the 728 conformations of **1**, Superposition 1.

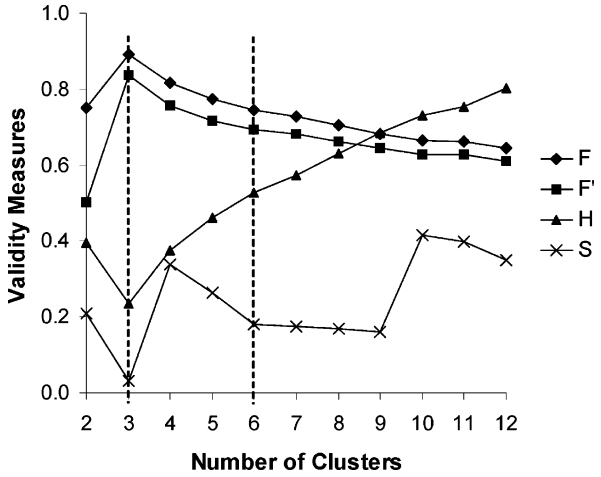

Figure 6. Cluster validity plots for partitions on the A-side.

This suggests that there was a 50% chance that all possible conformations were found during the random search conformational analysis.

Clustering. The optimal number of clusters found for each feature vector and superposition is given in the last column of Table 1. The optimal number of clusters was determined as that value of c for which the cluster validity indices have the following relationship: partition coefficient (F) − high, normalized partition coefficient (F') − high, partition entropy (H) − low, and compactness index (S) − low.

a. Full-Molecule Clustering, Superposition 1. The flexibility of **1** ensured that a large conformational space was covered by the random search protocol, as can be seen by superposition of all 728 conformations in Superposition 1, Figure 5. Clustering of the conformations using the whole-molecule feature vector outlined in Table 1 indicated the absence of natural groupings according to the behavior of the cluster validity indices (not shown). This is perhaps not surprising, given the wide range of positions occupied by the atoms of the B-side in Superposition 1. Figure 5 shows more clearly-defined groups on the A-side of the superimposed conformations due to more limited positions available to the naphthalene ring. Since the piperazine and naphthalene rings contain the pharmacophore features that are found in most DAT inhibitors, the next clustering study used a feature vector defined only in terms of the A-side in order to focus on these pharmacophore features.

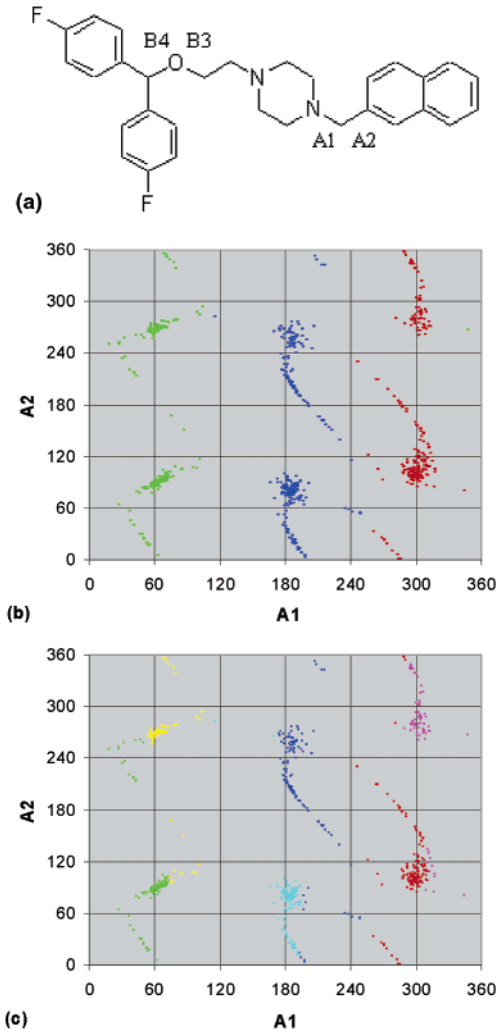

Figure 7. (a) Torsional angles defined for scatter plots. (b) Plot of 728 conformations of **1** in (A1, A2) space for A-side clustering at $c = 3$, Superposition 1. The separate clusters are identified by color (green, blue, and red). (c) Plot of 728 conformations of **1** in (A1, A2) space for A-side clustering at $c = 6$, Superposition 1. The separate clusters are identified by color (yellow, green, blue, cyan, magenta, and red).

b. A-Side Clustering, Superposition 1. The cluster validity results for the A-side partitions for Superposition 1 are shown in Figure 6. All four validity indices attain their first inflection point and their respective optima at $c = 3$ suggesting a good three-cluster partition. The compactness index, *S*, indicates good partitioning for $c = 6$ through $c = 9$ with the other three indices either monotonically increasing or decreasing over that range. This suggests a good second level partitioning at the lower bound, $c = 6$.

It is not possible to visualize the clusters in a feature vector space that consists of only an angle (between the C- and N-planes) and two points (C_{23} and C_{29}). A qualitative way of visualizing the clusters is to show their relationship to easily identifiable physical features of the molecule, such as the A1 and A2 torsional angles (see Figure 7(a)). These angles are important because they determine the relative orientation of the C- and N-planes. However, it should be noted that because the feature vector for A-side clustering was not defined specifically in terms of A1 and A2, representation of the clusters in (A1, A2) space is not equivalent to representation of the clusters in feature vector space. As a result, clusters which are well-separated in feature vector space may appear to overlap somewhat in (A1, A2) space. This is similar to the effect seen if clusters that are well-separated along the *x*-, *y*-, and *z*-axes in three-dimensional Cartesian space are projected onto a plane (such as defined by the *x*- and *y*-, *x*- and *z*-, or *y*-and *z*-axes). In other words, clusters that appear to mix and have poor separation could be an artifact of viewing the results in (A1, A2) space.

Torsional angles A1 and A2 have physical significance because they determine how the conformations form natural groups on the A-side of the molecule. The 728 conformations of **1** were identified by the random search procedure to be minima on the conformational potential energy surface of the molecule. This means that each conformation has values of A1 and A2 (as well as B1−B6) that were determined by minimizing the conformational energy for rotation around these angles. The A1 and A2 torsional angles contain $N(sp^3)-C(sp^3)$ and $C(sp^3)-C(sp^2)$ bonds, respectively. Therefore the A1 and A2 torsional angles output by the Random Search technique should be close to the values of the torsional angles found in staggered conformations of compounds such as aminomethane and methylbenzene, which can be considered to be models for the A1 and A2 torsional angle rotational barriers, respectively.

To visualize these relationships, the 728 conformations were plotted in (A1, A2) torsional angle space with each conformation color-coded by the color of the cluster to which it was assigned by the FRC procedure. Figure 7(b) is a scatter plot of the 728 conformations for the $c = 3$ cluster level. The figure shows that the three clusters (identified by green, blue, and red points) are located about 120° apart on the A1 axis (at approximately A1 = 60°, 180°, and 300°, respectively). The location of these clusters corresponds to rotational minima around the $N_4(sp^3)-C_{23}(sp^3)$ bond in A1 and is typical of the rotational minima in aminomethane. The three clusters appear to be well-defined by differences in the A1 values of the conformations. This shows that the fuzzy clustering technique has identified the natural groups based on the minimum in the conformational energy for the A1 torsional angle.

Figure 7(b) also shows that most of the conformations are found clustered along the A2 axis at approximately A2 = 90° and 270°. Quite a few conformers are also found "spread out" along the A2 axis at intermediate values of the angle. The location of these clusters corresponds to rotational minima around the $C_{23}(sp^3)-C(sp^2)$ bond in A2. This complex pattern of rotational minima is due to the effect of large substituent groups on the carbons in the $C_{23}(sp^3)-C(sp^2)$ bond in A2.

Figure 7(c) shows that the six clusters identified for $c = 6$ (shown in green, yellow, blue, cyan, magenta, and red) are directly related to the three clusters of Figure 7(b). The yellow and green clusters of Figure 7(c) contain the same points as the green cluster of Figure 7(b). A similar relationship holds for the blue and cyan/blue and magenta and red/red clusters of Figure 7(c)/7(b), respectively. Figure 7(c) shows that the clusters are well-separated by their A1 values and fairly well-separated by their A2 values. There is, however, some apparent mixing between the groups of colored points based on the value of A2. Some green points are found in the yellow cluster centered around (A1 = 60°,

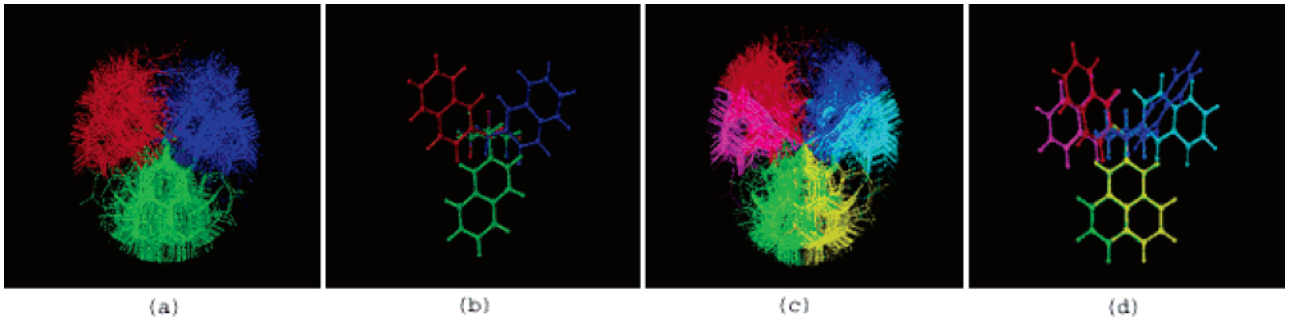

Figure 8. Results for the A-side clustering at $c = 3$ and $c = 6$, Superposition 1. The view was obtained by a 90° clockwise rotation about the central plane of Figure 5. For clarity, the B-side and the superimposed piperazine ring were located behind the clusters and were not displayed. (a) For $c = 3$, three distinct clusters. Number of conformations: red − 229, blue − 270, and green − 229. (b) For $c = 3$, representative structures. Conformation number (membership value): red − #62 (0.998), blue − #251 (0.997), green − #96 (0.999). (c) For $c = 6$, six distinct clusters. Number of conformations: magenta − 77, red − 153, blue − 128, cyan − 142, yellow − 82, and green − 146. (d) For $c = 6$, representative structures. Conformation number (membership value): magenta − #154 (0.990), red − #531 (0.994), blue − #428 (0.991), cyan − #248 (0.995), yellow − #232 (0.996), green − #177 (0.998).

A2 = 260°). Similarly some yellow points are found in the green cluster located near (A1 = 60°, A2 = 60°). Similar mixing is seen in the blue and cyan clusters as well as the magenta and red clusters. Since the cluster validity indices in Figure 6 indicate good cluster separation, the apparent mixing is probably due to viewing the scatter plot in (A1, A2) space rather than feature vector space.

The FRC results suggest that grouping the conformers by the value of the A1 angle provides a clear separation into three clusters, if only the A-side of the molecule is considered. However, separation of the conformations into a larger number of clusters appears to be more complex than simply basing the grouping on the value of A2. The results of the cluster validity tests show that while separation at the $c = 3$ level is obvious, separation at the $c = 6$ level is less so. This indicates the complexity involved in separating conformers of a very flexible molecule into many clusters. However, Figure 7(c) shows that at cluster level $c = 6$ the FRC technique has identified natural groups related to minimum in the conformational energy for combinations of the A1 and A2 torsional angles.

Figure 8 shows the molecular conformations that correspond to the scatter plots in Figure 7. The view depicted is a 90° clockwise rotation about the central plane of Figure 5 such that the A-side naphthalene rings are presented frontally (the piperazine ring and B-side are not shown). The molecules are oriented in a somewhat off-center view along the $C(sp^3)-C(sp^3)$ bond of the A1 torsional angle. Figure 8(a) shows three clusters of 229, 270, and 229 conformations each, which correspond to the red, blue, and green clusters, respectively, of Figure 7(b). Figure 8(b) shows the representative structure for each cluster identified as random search conformations #62 (red; A1 = 305°, A2 = 96°), #251 (blue; A1 = 174°, A2 = 268°), and #96 (green; A1 = 69°, A2 = 101°). Thus, the representative structures have A1 values that differ by about 120°. The clusters at the $c = 3$ level appear to separate well in terms of the A1 torsional angle.

Figure 8(c) shows six clusters of 77, 153, 128, 142, 82, and 146 conformations each, which correspond to the magenta, red, blue, cyan, yellow, and green clusters, respectively, of Figure 7(c). The figure illustrates the apparent cluster mixing noted in Figure 7(c). As in Figures 7(b) and 7(c), comparison of the conformers in the clusters in Figures 8(a) and 8(c) shows that, in general, the $c = 6$ magenta and red clusters form the $c = 3$ red cluster, the $c = 6$ blue and cyan clusters form the $c = 3$ blue cluster, and the $c = 6$ yellow and green clusters form the $c = 3$ green cluster. Figure 8(d) shows the representative structure for each cluster identified as random search conformations #154 (magenta; A1 = 301°, A2 = 293°), #531 (red; A1 = 298°, A2 = 128°), #428 (blue; A1 = 192°, A2 = 191°), #248 (cyan; A1 = 183°, A2 = 68°), #232 (yellow; A1 = 62°, A2 = 265°), and #177 (green; A1 = 62°, A2 = 90°), respectively. Note that the A1 values of the representative structures of the color-related (yellow/green, red/magenta, blue/cyan) clusters at the $c = 6$ cluster level have very similar values to those in the color-related representative structure at the $c = 3$ level. For example, the A1 value of the representative structure for the $c = 3$ red cluster (305°) is close to those of the representative structures for the $c = 6$ magenta (301°) and red (298°) clusters. Figure 8(d) illustrates that the representative structures at the $c = 6$ and $c = 3$ levels have similar A1 values that differ by about 120°. Figure 8(d) also shows that the A2 values of the color-related representative structures differ by various amounts: 165° (magenta/red), 123° (blue/cyan), and 175° (yellow/green). The fact that the A2 values of the blue and cyan representative structures differ by much less than 180° could be due to incomplete searching of the conformational space. As noted above, there is only a 50% chance that all possible conformations were found during the random search conformational analysis.

c. B-Side Clustering, Superposition 1. Clustering using the B-side feature vector indicated the absence of natural groups. This is consistent with the fact that the B-side of **1** is much more flexible than the A-side due to the presence of six rotatable bonds on the B-side versus two on the A-side. The B-side of **1** can access a much wider range of conformational space than the A-side, as shown in Figure 5. None of the validity indices provides a reason to believe that there is an underlying structure on the B-side (Figure 9). The compactness index is not plotted because the results were not considered to be sufficiently consistent, indicating a lack of substructure. The normalized partition coefficient, F', takes values very close to zero ($cF \rightarrow 1$) and hence the results at all levels of clustering are too fuzzy to be of any significance.

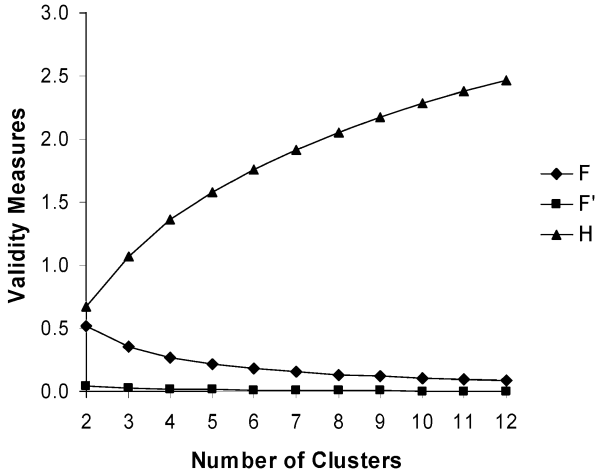

Figure 9. Cluster validity plots for partitions on the B-side.

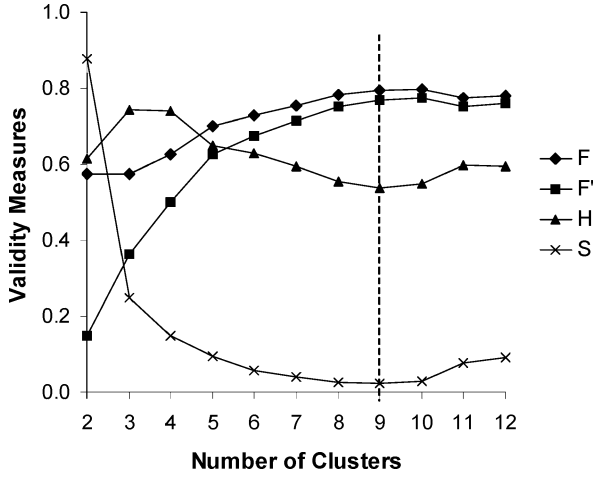

Figure 10. Cluster validity plots for partitions on the B′-side.

d. B′-Side Clustering, Superposition 2. The cluster validity indices plotted in Figure 10 suggest nine optimal clusters for the B′-side. The compactness index, *S*, has its lowest value for $c = 9$. The other indices support this partition, indicating well-separated and compact clusters.

As above, one way to visualize the clusters is by scatter plots of the conformations in torsional angle space. Comparison of Figure 3 with Figure 1 shows that the B′-side of **1** contains only three rotatable bonds instead of the six bonds on the B-side. Comparison of Figures 7(a) and 3 shows that torsional angles B3 and B4 in part control the orientation of the bisphenyl group with respect to the O-plane and the rest of the molecule. Since Superposition 2 is based on the O-plane, it allows the grouping on the B′-side of the superimposed conformations to be observed. One way of visualizing the clusters is to plot the conformational minima in (B3, B4) space. However, it should be emphasized here, as above, that the B′-side feature vector was not defined directly in terms of B3 and B4 but rather in terms of the related angles between the O- and P1- and P2-planes. Therefore, plotting the data in (B3, B4) space, although an obvious way to visualize the physical data, is not exactly equivalent to plotting the results in the B′-side feature vector space.

Figure 11 shows the nine distinct clusters on the scatter plot of the conformations in (B3, B4) torsional angle space. This corresponds to the nine rotational minima that result

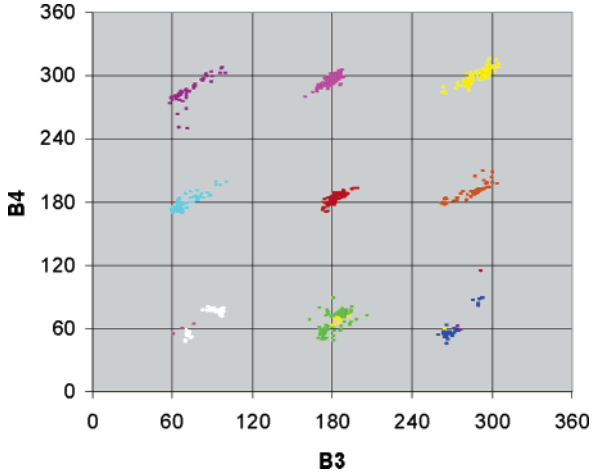

Figure 11. Plot of 728 conformations of **1** in (B3, B4) space for B-side clustering at $c = 9$, Superposition 2. The separate clusters are identified by color (purple, cyan, white, magenta, red, green, yellow, orange, and blue).

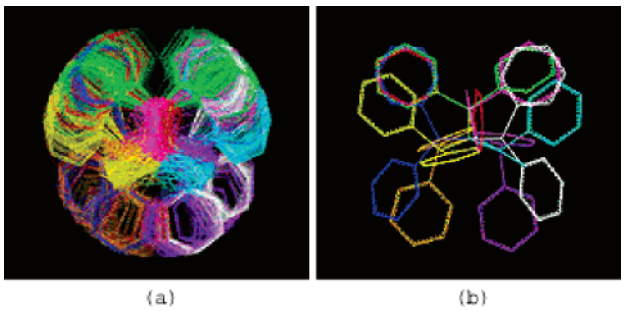

Figure 12. Clustering results for the B′-side at $c = 9$. The view is obtained by a 90° rotation of Figure 5 to show the bisphenyl group end-on. For clarity, the superimposed O-plane and the A′-side, located behind the bisphenyl clusters, are not displayed. (a) Nine distinct clusters. Number of conformations: red − 99, orange − 48, magenta − 146, blue − 31, white − 49, cyan − 87, purple − 52, green − 114, and yellow − 102. (b) Nine representative structures. Conformation number (membership): red − #213 (0.999), orange − #72 (0.994), magenta − #307 (0.999), blue − #207 (0.995), white − #108 (0.997), cyan − #402 (0.998), purple − #150 (0.995), green − #692 (0.978), and yellow − #716 (0.992).

from a combination of the three rotational minima for staggered conformations around the C(sp^3)−O(sp^3) bond in B4 with the three rotational minima for staggered conformations around the O(sp^3)−C(sp^3) bond in B3. As shown in the figure, the clusters are located at about 120° intervals along the B3 and B4 axes. Apparent mixing of some clusters is due to representation of the clusters in (B3, B4) space rather than feature vector space.

Figure 12 shows these nine B′-side clusters as well as the representative conformations from each cluster. Each cluster is formed by the bisphenyl group on the B′-side (the A′-side is not shown). Each phenyl ring of the bisphenyl group in a cluster occupies two different regions in space. For example, three clusters (blue, green, and white) have both of their phenyl rings located out on the edge, and six clusters (red, magenta, purple, cyan, orange, and yellow) have one phenyl ring located out on the edge and the other located in the center, coming out of the plane of the figure. Since no two colors appear in the same region, the clusters are distinct. For example, while one phenyl ring of both the orange and the yellow clusters seems to be overlapping in the center, the other phenyl ring of the orange cluster lies on the bottom

Table 2. Torsional Angles (deg) and Relative Energy (kcal/mol) for Full-Molecule Representative Structures

color	conformation no.	A1	A2	B3	B4	energy[a]
white	723	303	290	181	296	3
blue	232	62	265	177	294	6
yellow	638	63	92	180	182	6
green	391	61	89	292	297	7
red	394	60	90	266	57	8
cyan	710	61	89	181	295	10

[a] Relative to the global energy minimum.

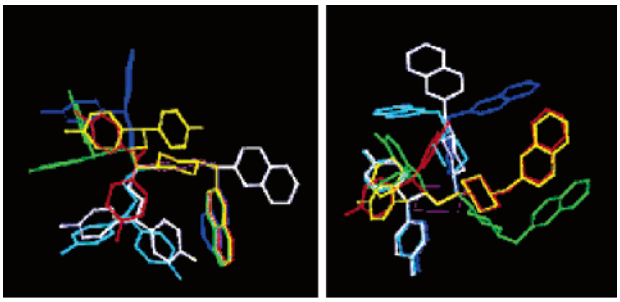

Figure 13. Full-molecule representative structures that will be used for CoMFA of GBR 12909 analogues. Conformers aligned using (a) Superposition 1 and (b) Superposition 2.

left and that of the yellow cluster lies on the top left. Thus, the orange and yellow clusters are distinct and do not overlap.

Identification of Full-Molecule Representative Structures. Since the full-molecule clustering suggested the absence of natural groups, the superposition-based and region-specific clustering results obtained above were used to identify putative representative structures for **1**. The A1 and A2 values of the six cluster representatives from the A-side clustering were combined with the B3 and B4 values of the nine representatives from the B′-side clustering to construct 54 "ideal" combinations of the four torsional angles. Then a search through the data set of 728 conformations using a tolerance of ±2.5° on each torsion angle produced six matches (given in Table 2). The table ranks the structures by their energy relative to that of the GEM conformation. It is interesting to note that the GEM conformer is not one of the representative structures. Since ligands have been shown to bind to proteins in conformers that have energies over 10 kcal/mol above the GEM,[4] the representative structures appear to be reasonable in terms of energy.

These six conformations are shown in Figure 13 and form the final set of representative structures of **1** that will be used as templates for future CoMFA studies of GBR 12909 analogues. The conformers appear to be representative of the regions of space occupied by **1**.

DISCUSSION

Generalization of the Feature Extraction Method. Although the feature extraction method was developed keeping the structure of **1** in mind, the method can be generalized to other molecules. Intelligent selection of features combined with a general adherence to these guidelines can produce a useful feature set, consisting of atom coordinates and planes, which can be used to reconstruct the molecule. The associated feature vector, consisting of atom coordinates and angles between pairs of planes, can then be used as input for fuzzy clustering of molecular conformations.

The process begins with the identification of planes that can be used in reconstructing the molecule. Symmetric planar rings (such as phenyl rings) could be selected to avoid symmetry-related problems, as described in the Methods section. Planes that contain important structural moieties, such as pharmacophore elements, would be chemically sensible. To reconstruct the molecule, one atom per selected plane needs to be included in the feature set (for example, atom C_{29} on the N-plane in **1**, see Figure 2). Thus, selection of large planar rings, like naphthalene, enables significant reduction in feature space. Selection of planes at the extremities of the molecule could also be useful in the reconstruction process. The clustering technique being superposition-dependent, a substructure should be identified that will be used to superimpose all conformations. Since this substructure will be common to all conformations, atoms that lie on it need not be included in the feature set. The guidelines for constructing a feature vector for a general molecule are summarized below:

1. Identify planes.
2. Identify all (heavy) atoms that do not lie on the planes.
3. Identify a superimposable substructure.
4. Remove from the set identified in step 2 any atom that is part of the substructure identified in step 3.
5. Include in the set one suitable atom on each plane (not necessary if plane is part of the superimposable substructure).
6. Select angles between pairs of planes such that the overall structure of the molecule can be "captured".

Applying the guidelines to **1**:

1. Planes C, N, P1, and P2.
2. Atom set: Atoms C_7, C_8, O_9, C_{10}, and C_{23}.
3. Superimposable substructure: Piperazine.
4. Atom set: Atoms C_7, C_8, O_9, C_{10}, and C_{23} (remove atoms: none).
5. Atom set: Atoms C_7, C_8, O_9, C_{10}, C_{23}, C_{29}, F_{50}, and F_{55} (include atoms F_{50} (P1-plane), F_{55} (P2-plane), and C_{29} (N-plane)).
6. Angles between pairs of planes: N/C, P1/C, and P2/C.

Applying the guidelines to cocaine (see Figure 14):

1. Planes X, Y, and Z.
2. Atom set: Atoms 2, 3, 4, 6, 7, 8, 9, 10, and 11.
3. Superimposable substructure: Tropane ring.
4. Atom set: Atoms 2, 3, 4, 6, and 7 (remove atoms 8, 9, 10, and 11).
5. Atom set: Atoms 1, 2, 3, 4, 5, 6, and 7 (include atoms 1 (plane X) and 5 (plane Z)).
6. Angles between pairs of planes: X/Y and Z/Y.

This example shows that the novel feature extraction technique can be easily generalized to other molecules.

Relational Data Clustering versus Object-Based Clustering. As discussed in the Introduction, it is convenient to construct a proximity matrix so that all relevant information about the features is incorporated. This was also done by Feher and Schmidt,[23] who constructed such a matrix and subsequently employed metric scaling to map conformations to a reduced dimensional (i.e., 3-D) space for use in object-data-based clustering. In their case, it is natural to utilize clustering in object space because the main purpose appears

Table 3. Feature Vector/Superposition Combinations and Distance Matrix Qualitative Results from XCluster Studies

option	feature vector[a]	superposition[b]	distance map appearance
a	all heavy atoms	□ center ring[c]	many small clusters
b	all heavy atoms	all heavy atoms	many small clusters
c	□ center ring ○ A-side key atoms	□ center ring[c]	five or six large clusters
d	□ center ring ○ A-side key atoms	□ center ring ○ A-side key atoms	six or seven large clusters
e	A1 and A2	five heavy atoms defining A1 and A2[d]	six large clusters
f	□ center ring ◊ B-side key atoms	□ center ring[c]	many small clusters
g	□ center ring ◊ B-side key atoms	□ center ring ◊ B-side key atoms	two large, many small clusters
h	□ center ring ◊ B-side key atoms	oxygen and neighboring carbons[c]	two large clusters and seven small clusters
i	□ center ring ◊ B-side key atoms ● B-side plane atoms	oxygen and neighboring carbons[c]	fifteen small clusters
j	B1 through B6	nine heavy atoms defining B1 through B6[d]	many small clusters

[a] Feature vector: The atoms or angles used to calculate the intermolecular distances. Key atoms (identified by symbols noted) and torsional angles are shown in Figure 15. RMSD calculations for the distance matrix were carried out on the feature set atoms in the corresponding alignment. [b] Superposition performed by XCluster unless otherwise noted. [c] Superposition performed in SYBYL. [d] Superposition of conformers for the torsional angle studies is not necessary to calculate RMSD values, but was carried out to properly visualize clusters.

to be dimensionality reduction. If that is not the main purpose in constructing the proximity matrix, then one must be careful in terms of utilizing relational data-space. It is clear that if the proximity data is derived using strictly Euclidean norms, then it is better to directly utilize object-data-based clustering. That is because it can be shown[33,34] that in such cases, the clustering in the object-space provides identical results as clustering in the relational-space, and conversion to relational space and subsequent application of relational clustering method is a wasted computational effort. In contrast to Feher and Schmidt,[23] in our work judicious feature selection was utilized for the dimensionality reduction, yet clustering was directly performed on the relational data, because our goal is to develop a very flexible clustering methodology where example-specific non-Euclidean proximity measures can be gainfully employed as distinguishing features for more effective classification. Hence in situations when some of the proximity measures are non-Euclidean, conversion to the object space, which in itself requires additional computations, is not necessary. Another useful outcome of our approach is that the validity indices were consistent in identifying the optimum number of clusters or a lack of natural groups, whereas in Feher and Schmidt's work, the cluster validity measures used worked well in certain cases but were not found to be consistently predictive.[23]

Comparison to Hierarchical Clustering Using XCluster. In a separate publication[45] we will present the results of hierarchical clustering of the same data set of 728 conformers of **1** using XCluster.[16] Several feature vector and superposition options were explored and are summarized in Table 3. They range from full molecule clustering (options *a* and *b*) to A-side clustering (options *c*, *d*, and *e*) to B-side clustering (options *f* - *j*). Since XCluster allows for clustering using Cartesian coordinates of atoms *or* torsional angles, whereas our novel feature extraction procedure uses atomic coordinates and angles between planes, the feature vectors are not exactly the same. However, both techniques allow the user to focus on features related to full molecule, A-side or B-side clustering with different superpositions, so the results of the studies are comparable. For example, the "center ring"

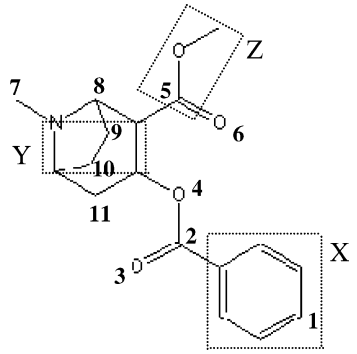

Figure 14. Identification of reduced feature set for cocaine.

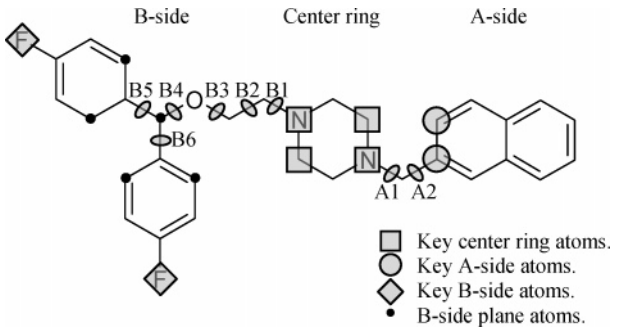

Figure 15. Atomic and torsional angle features chosen for various feature vector/superposition combinations in XCluster studies.

superposition of Table 3 is the same superposition 1 of Table 1. The "oxygen and neighboring carbons" superposition of Table 3 is slightly different than Superposition 2 of Table 1 because it involves C_{10}, O_9, and C_8, whereas Superposition 2 uses O_9, C_8, and C_7 (see Figure 1). The "all heavy atom" feature vector in Table 3 (options *a* and *b*) is related to the "full molecule" feature vector in Table 1. Similarly the A-side feature vectors in options *c*, *d*, and *e* of Table 3 are closely related to the A-side feature vector in Table 1. Of the B-side feature vectors in Table 3, option *j* is the closest to the B-side feature vector of Table 1. The last column of Table 3 summarizes the appearance of the XCluster distance maps for the various feature vector/superposition options and can

be compared to the last column of Table 1. For full molecule clustering, the FRC and XCluster techniques give the same results because no obvious large clusters are detected. Both methods also agree for A-side clustering. XCluster options *c* - *d* all give six large clusters, which is the same as the FRC result. Display of the XCluster A-side results (not shown) shows the same type of separation noted in Figures 7(c) and 8(c). This is because the cluster memberships are a result of natural groupings determined by the (A1, A2) values that correspond to minima in the conformational potential energy surface for rotation around the $N(sp^3)-C(sp^3)$ and $C(sp^3)-C(sp^2)$ bonds in A1 and A2. XCluster options *f*, *g*, and *j* all result in many small clusters and agree with the FRC result shown in Figure 9 that there is no underlying structure on the B-side using these superpositions. Although the "oxygen and neighboring carbons" superposition of option *h* is slightly different than Superposition 2 of the B′-side clustering in Table 1, both these FRC and XCluster studies result in detection of nine clusters. This is because both feature vectors focus on the region near (B3, B4). As discussed in the FRC results section, this is because cluster memberships are the result of natural groupings due to the combination of the three conformational minima for rotation around the $C(sp^3)-O(sp^3)$ bond in B4 with the three conformational minima for rotation around the $O(sp^3)-C(sp^3)$ bond in B3 (see Figure 11). In summary, the novel feature extraction technique presented here combined with the fuzzy relational clustering methodology gives similar results to the hierarchical clustering approach implemented in XCluster when similar feature vectors and superposition options are used.

Use of Torsional Angles as Feature Vectors? Since Figure 7 shows that the conformations appear to cluster well in (A1, A2) space, the question may arise as to why torsional angles were not used in the FRC feature vector. While this might work well in cases like A-side clustering, Superposition 1, where only two torsional angles can completely specify the conformations allowed for the A-side of the molecule, its application is perhaps less useful in describing regions of the molecule that have many rotatable bonds. In fact the novel feature extraction technique was developed for the express purpose of going beyond torsional angle analysis to a more global approach that employs features of the whole molecule or molecular fragment as a basis for classification. The purpose of the novel feature extraction process was to identify a *minimal feature set* that could be used to classify the conformations into groups. The usefulness of that approach has been demonstrated in A-side and B-side clustering which focused on the important pharmacophore features of **1**.

Different Superpositions. Whereas an analysis such as one based on torsional angles would be independent of superposition of the conformations, the technique presented here is superposition-dependent. A different superposition might necessitate a redefinition of the minimal feature set from which the input feature vector is derived. However, even if the input feature vector is found to be the same for a new superposition, the new positions of the conformations would lead to a change in the distance matrix. This necessitates careful application of this technique when selecting optimal alignment rules for conformations. On the other hand, using different superpositions enables one to focus on specific pharmacophore regions of the molecule and provides clustering results that may not be uncovered without such data reduction. Future work will explore the use of superposition-independent feature vectors involving the relative orientation of planes containing important pharmacophore features.

CONCLUSIONS

Fuzzy relational clustering of a large number of conformations of the flexible GBR 12909 analogue, **1**, resulted in compact clusters that could be attributed to conformational energy minima for rotation around $N(sp^3)-C(sp^3)$, $O(sp^3)-C(sp^3)$, and $C(sp^3)-C(sp^2)$ bonds. These natural groups were identified by superposition-dependent clustering based on segregating important pharmacophore elements of the molecule. The study presented here highlights the importance of using feature extraction that is unique to the problem at hand. The feature extraction process was molecule-specific and led to well-defined clusters that were extracted from numerous conformations of a flexible molecule yet was shown to be easily generalized to other molecules. The partitions found by fuzzy clustering were validated by various standard cluster validity measures and agreed with the results of hierarchical clustering using XCluster. The six full-molecule representative structures obtained by combining results from the A- and B′-side clustering exhibited a reasonable spread of energies and showed good spatial coverage of the conformational space. The FRC method combined with the molecule-specific novel feature extraction technique appears to be a reasonable way to select representative conformations of a flexible molecule for 3D QSAR studies.

ACKNOWLEDGMENT

The authors would like to thank Dr. Denis Blackmore, Department of Mathematical Sciences, New Jersey Institute of Technology, for his valuable suggestions during the feature extraction process, and Kathleen M. Gilbert for providing the XCluster results.

Note Added after ASAP Publication. In Table 1 there was a minor error in the footnote. In Table 3 correct symbols were added. This paper was posted April 28, 2005. The corrected version was posted April 29, 2005.

REFERENCES AND NOTES

(1) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. Conformational changes of small molecules binding to proteins. *Bioorg. Med. Chem.* **1995**, *3*, 411−428.
(2) Veith, M.; Hirst, J. D.; Brooks III, C. L. 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. Three-dimensional quantitative structure−activity relationship study on cyclic urea derivatives as HIV-1 protease inhibitors: Application of Comparative Molecular Field Analysis. *J. Med. Chem.* **1999**, *42*, 249−259.
(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 relation analysis. In *Practical Application of*

Computer-Aided Drug Design; Charifson, P. S., Ed.; Marcel Dekker: New York, 1997; Vol. 105−164.
(7) Barnett-Norris, J.; Guarnieri, F.; Hurst, D. P.; Reggio, P. H. Exploration of biologically relevant conformations of anandamide, 2-arachidonylglycerol, 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.; Makriyanis, 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.; 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 delta opioid agonists from antagonists. *J. Am. Chem. Soc.* **2003**, *125*, 3101−3107.
(11) Cramer III, R. D.; 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.
(12) 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-hydroxyl-3phenylpropyl)piperazinyl decanoate, a long-acting ester derivative of GBR 12909. *J. Med. Chem.* **1996**, *39*, 4689−4691.
(13) 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.
(14) 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.
(15) 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.
(16) Shenkin, P. S.; McDonald, D. Q. Cluster analysis of molecular conformations. *J. Comput. Chem.* **1994**, *15*, 899−916.
(17) Murray-Rust, P.; Raftery, J. Computer analysis of molecular geometry, part IV: Classification of differences in conformation. *J. Mol. Graph.* **1985**, *3*, 50−59.
(18) Jain, A. K.; Dubes, R. C. *Algorithms for Clustering Data*; Prentice Hall, NJ, 1988.
(19) Johnson, R. A.; Wichern, D. W. *Applied Multivariate Statistical Analysis*, 4th ed.; Prentice-Hall: NJ, 1998.
(20) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2nd ed.; John Wiley: New York, 2000.
(21) Chema, D.; Goldblum, A. The nearest neighbor method − Finding families of conformations within a sample. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 208−217.
(22) Feher, M.; Schmidt, J. M. Metric and multidimensional scaling: Efficient tools for clustering molecular conformations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 346−353.
(23) 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.
(24) Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*; Plenum Press: New York, 1981.
(25) Bezdek, J. C.; Keller, J.; Krishnapurum, R.; Pal, N. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*; Kluwer Academic Publishers: Norwell, MA, 1999.
(26) Davé, R. N.; Bhaswan, K. Adaptive fuzzy C-shells clustering and detection of ellipses. *IEEE Trans. Neural Net.* **1992**, *3*, 643−662.
(27) Davé, R. N.; Krishnapurum, R. Robust clustering methods: A unified view. *IEEE Trans. Fuzzy Syst.* **1997**, *5*, 270−293.
(28) Xie, X. L.; Beni, G. A validity measure for fuzzy clustering. *IEEE Trans. Patt. Anal. Machine Intell.* **1991**, *13*, 841−847.
(29) Davé, R. N. Validating fuzzy partitions obtained through c-shells clustering. *Pattern Recognit. Lett.* **1996**, *17*, 613−623.
(30) Tucker, W. T. Counterexamples to the convergence theorem for the fuzzy c-means clustering algorithms. In *Analysis of Fuzzy Information*; Bezdek, J. C., Ed.; CRC Press: Boca Raton, FL, 1987; Vol. III, pp 109−121.
(31) Höppner, F.; Klawonn, F. Obtaining interpretable fuzzy models from fuzzy clustering and fuzzy regression. *Proc. of the 4th Int. Conf. on Knowledge-Based Intelligent Engineering Systems and Allied Technologies (KES)*: Brighton, U.K., 2000; pp 162−165.
(32) Höppner, F.; Klawonn, F.; Eklund, P. Learning indistinguishability from data. *Soft Comput. J.* **2002**, *6*, 6−13.
(33) Hathaway, R. J.; Bezdek, J. C. NERF c-means: Non-Euclidean relational fuzzy clustering. *Pattern Recognit.* **1994**, *24*, 429−437.
(34) Davé, R. N.; Sen, S. Robust fuzzy clustering of relational data. *IEEE Trans. Fuzzy Syst.* **2002**, *10*, 713−727.
(35) Shannon, C. E. A mathematical theory of communication. *Bell Syst. Technol. J.* **1948**, *27*, 379−423.
(36) Gath, I.; Geva, A. B. Unsupervised optimal fuzzy clustering. *IEEE Trans. Patt. Anal. Machine Intell.* **1989**, *11*, 773−781.
(37) SYBYL 6.9 Tripos Inc.; 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA.
(38) Dutta, A. K.; Meltzer, P. C.; Madras, B. K. Positional importance of the nitrogen atom in novel piperidine analogues of GBR 12909: Affinity and selectivity for the dopamine transporter. *Med. Chem. Res.* **1993**, *3*, 209−222.
(39) Powell, M. J. D. Restart procedures for the conjugate gradient method. *Math. Program.* **1977**, *12*, 241−254.
(40) Clark, M.; Cramer III, R. D.; van Opdenbosch, N. Validation of the general purpose Tripos 5.2 force field. *J. Comput. Chem.* **1989**, *10*, 982−1012.
(41) 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 analogues 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.
(42) 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.
(43) 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.
(44) Saunders: M. Stochastic exploration of molecular mechanics energy surfaces. Hunting for the global minimum. *J. Am. Chem. Soc.* **1987**, *109*, 3150−3152.
(45) Gilbert, K. M.; Venanzi, C. A. Unpublished results.

CI049708D