Recursive Partitioning Analysis of a Large Structure−Activity Data Set Using Three-Dimensional Descriptors[1]

Xin Chen,[†,‡] Andrew Rusinko III,[‡] and S. Stanley Young*[,‡]

Laboratory for Molecular Modeling, School of Pharmacy, University of North Carolina at Chapel Hill,
Chapel Hill, North Carolina 27599, and Chemoinformatics Group, Research Information Systems,
Glaxo Wellcome Inc., Five Moore Drive, Research Triangle Park, North Carolina 27709

Received May 17, 1998

Large chemical data sets are becoming available from high throughput screening of corporate collections and chemical libraries. There is a growing need to develop three-dimensional pharmacophores from these large data sets to guide database screening, chemical library design, and lead optimization. Recursive partitioning (RP) is a statistical method that can be used to analyze very large data sets; data sets of over 100 000 observations and over 2 000 000 descriptors pose no computational problems. Our idea is to encode the three-dimensional features of chemical compounds into bit strings and use RP to determine the important features that statistically correlate to the biological activities of these compounds. This kind of structure−activity relationship analysis (SAR) can be considered as the first step to the goal of pharmacophore identification for large chemical data sets. We report here our RP work that for the first time successfully retrieved 3D SARs from a large, heterogeneous data set of 1650 monoamine oxidase (MAO) inhibitors, which indicates the feasibility of 3D analysis of a few thousand compounds.

INTRODUCTION

The recent progress of combinatorial chemistry[2,3] and high throughput screening techniques has brought a revolution in the drug discovery process in the pharmaceutical industry. It is now feasible to obtain biological activity data for thousands to hundreds of thousands of chemical compounds in a short period of time. How to manage and utilize these large data sets is becoming a major challenge in the field of chemoinformatics, because the traditional QSAR methods[4,5] are no longer suitable to determine the structure−activity relationships (SAR) from such large data sets.

Large data sets inherently require that any method designed to analyze them must have the following properties. First, it has to be fast enough to satisfy the practical requirement of the drug discovery process. Second, it must be able to deal with the multiple mechanisms that very probably exist in a large data set. Multiple mechanisms imply that there are several different structure−activity relationships simultaneously existing in a large data set. Traditional QSAR and statistical methods cannot handle this situation. Third, it must also be able to model nonlinear structure−activity relationships and do so even in the face of strong interactions between the chemical descriptors. The simple linear model assumption, the prerequisite of many traditional QSAR methods, is thus unlikely to hold for large data sets with heterogeneous compounds.

Several methods designed to analyze large data sets have been reported in recent years.[6−12] Among them, recursive partitioning (RP) is shown to be a powerful approach to decode the complex structure−activity relationships hidden in a large chemical data set.[10−12] We have written a special version of RP that deals with very large data sets, over 100K observations and over 2M descriptors.[13] RP method overcomes the difficulties of handling nonlinear relationships and strong interactions in the large SAR data set by generating a dendrogram or tree diagram in which the statistically best chemical features are used to split the large data set into smaller and more homogeneous subsets. It is inherently fast when compared with other methods for grouping compounds, such as clustering, etc.[14−16] It can detect the multiple mechanisms or multiple binding modes by separating the chemical compounds with different mechanisms into the different arms and terminal nodes of the dendrogram. Furthermore, the tree display of the SAR results is clear, easy to understand, and often suggestive of new hypotheses.

Although RP methods have been successfully applied to the large chemical data sets,[10−13] the previous work was done using two-dimensional (2D) descriptors. Atom pairs,[17] topological torsions,[18] and atom triples[19] were generated from molecular topology and used to represent the structural features of chemical compounds. Then the statistically most significant features among them were detected by RP to classify the whole data set. Traditionally, chemists believe chemical compounds exert their biological effects three-dimensionally by specific active conformations that complementarily interact with the biological target. Therefore, the study of 3D-SAR is expected to give a better and deeper understanding of the relationships between the chemical structure and biological activity. Attractiveness notwithstanding, derivation of 3D-SAR has proven to be much more difficult and time consuming than 2D-SAR. A major reason is that the active conformations or binding modes of the chemical compounds in a practical data set are almost always

* Corresponding author: e-mail: ssyO487@glaxowellcome. com. fax: 919-483-2494.
[†] University of North Carolina at Chapel Hill.
[‡] Glaxo Wellcome Inc.

10.1021/ci980089g CCC: $15.00 © 1998 American Chemical Society
Published on Web 10/24/1998

unknown, except for the rare ligand−receptor complexes whose 3D structures happen to have been determined in X-ray or NMR.[20] More careful considerations and more subtle computational techniques are always required to obtain a satisfying 3D-QSAR model.[21] Therefore, most of the attempts to derive 3D SAR or QSAR models are limited to small data sets, wherein the number of compounds is usually less than 100.[22,23] We present here a report on the application of RP to analyze a large chemical data set using 3D descriptors. To our knowledge, this is the first time that 3D SARs are successfully derived from a large data set containing multiple mechanisms of action. The data set studied in this work comprises 1650 monoamine oxidase (MAO) inhibitors from Abbott Laboratories,[24] containing 290 active compounds.

METHODS

Conformation Database. Both single and multiple, low-energy conformations were used to model the 3D structures of the compounds in MAO data set. Thus, a single-conformation database and a multiple-conformation database were created, respectively. The generation of multiple conformations for each compound was expected to include the correct active conformations, and we hoped that the following RP analyses could identify these active conformations from the inactive ones.

A single conformation for each compound was generated using program CONCORD.[25] Of the 1650 compounds in MAO data set provided by Abbott Laboratories, 1644 were successfully converted from the 2D structures (in the form of SD-file)[26] to the 3D structures (in the form of MOL2-file).[27] 3D structures could not be generated for six compounds. Five of them contain one or more alkyl side chains with uncertain lengths according to the original records in MAO data set, and the other one contains an unusual arsenic atom. All of these six compounds are inactive, and they were excluded from the analysis of the data set.

Multiple conformations for each compound were generated by the random search of the whole conformation space, using the CONCORD-generated structures as the starting points. All the conformation searches were done on the Sybyl platform[27] using internally developed SPL (Sybyl Programming Language) code. Internal and Cartesian coordinate representations were used to search the chain conformations and the ring conformations, respectively. A flexible ring atom is determined as the nonfused atom in a flexible ring, and its spatial position was changed by randomly flipping it relative to the average plane of the ring it belonged to. For each nonterminal single bond, the torsion angle around it was randomly assigned a value from [0°, 360°], while for each nonterminal double bond, its torsion angle was randomly chosen from the regions of [−20°, 20°] and [160°, 200°]. After these geometrical manipulations, 10 steps of energy minimization were used to relieve any extremely close contacts between the nonbonded atoms. There was no attempt to further optimize the conformational structures to the local energy minimums, since the active conformations do not have to locate at any position corresponding to the local energy minimum in vacuum. Next, the relative conformational energy was calculated. The global energy minimum was recorded as the lowest conformational energy that was found in the whole process of random search. If the conformational energy of a generated structure was greater than 20 kcal/mol above the global energy minimum, it was deleted. The choice of 20 kcal/mol as the cutoff value was based on the work of Nicklaus et al.[28] They compared the experimentally determined bound conformations with the global energy minimum conformations calculated by the molecular mechanism for 33 small compounds and found that most of the receptor bound conformations existed within about 20 kcal/mol above the calculated in vacuum global energy minimum. All the molecular mechanism calculations were done using the Tripos force field.[29] Since the final goal of our project is to rapidly analyze large SAR data sets, explicit electrostatic interactions were not taken into account to reduce the time needed to build a 3D multiple-conformation database. The number of the conformations generated for each compound is dependent on its flexibility and ranges from 1 (for a totally rigid structure) to 800 (for a highly flexible structure). A total of 362 977 conformational structures were saved in the multiple-conformation database for the 1 644 MAO inhibitors.

Atom Pair Descriptor. The three-dimensional (3D) atom pairs were used as the molecular descriptors in our work. Each atom pair descriptor was composed of the atom types of the two component atoms and the “binned” Euclidean distance between these two atoms. The term “atom” here is used in a generalized sense. It is much like the feature point in a pharmacophore model, which is usually composed of the key chemical or physical feature points and the spatial relationships between them. For example, the atom can represent a real atom in the chemical structure, such as hydrogen, carbon, or oxygen, or the center of some special chemical functionality, such as aromatic ring, double bond, triple bond, hydrogen bond donor, etc. The use of atom pair as molecular descriptor is expected to reflect a component of a pharmacophore model: two key features and their spatial relationship. In this work, 17 different atom types were defined. Some of them are the feature points that are often used in pharmacophore models, such as positive charge center, negative charge center, hydrogen bond acceptor, and aromatic ring center. Others are explicit atom types like polar hydrogen, nonpolar hydrogen, carbon, oxygen, nitrogen, etc. We also included several special atom types that are not generally considered, such as triple bond center and double bond center. The details on the definitions of all the 17 atom types are listed in Table 1. Following these definitions, an atom can give rise to more than one atom type. For example, the oxygen atom in carbonyl group is classified to both type 3 (hydrogen bond acceptor) and type 12 (oxygen atom). Including more atom types than usually used in pharmacophore models and allowing an atom to have more than one atom type is based on the observation that RP can select the most significant features from a vast pool of descriptors.

The distance between two atoms was measured and assigned into one or two distance bins. The width of each distance bin was chosen as 1.0 Å in this work. Since it was also designed to let the adjacent bins have 10% overlap with each other, the actual length of each distance bin is 1.2 Å. Any distance located in the overlap region was assigned into both of the bins. This “fuzzy distance” concept was designed to alleviate the possible unfavorable boundary effects of the

Table 1. Definitions of the Atom Types Used in This Work

atom type	definition
1	negative charge center: including carboxylic group, sulfinic group, phosphinic group, etc.
2	positive charge center: all the nitrogen in primary, secondary, and tertiary amines
3	hydrogen bond acceptor: all the nitrogen, oxygen, and sulfur with at least one available lone pair electron
4	polar hydrogen atom: all the hydrogen atoms linked on the nitrogen, oxygen, sulfur or the terminal of a triple bond
5	nonpolar hydrogen atom: all the hydrogen atoms linked on the carbon
6	hydrogen atom: including both polar and nonpolar hydrogen atoms
7	triple bond center
8	double bond center
9	aromatic ring center
10	carbon atom
11	nitrogen atom
12	oxygen atom
13	sulfur atom
14	phosphor atom
15	fluorine atom
16	chlorine, bromine, or iodine atom
17	other element

distance bins. For example, with strict boundary conditions, a distance of 2.05 Å would be assigned only to bin no. 2, but it could be reasonably argued that it is almost as close to the upper half of bin no. 1 as to bin no. 2. With fuzzy boundary conditions, 2.05 Å belongs to both bin no. 1 and bin no. 2 allowing a possible match to either. All the distances larger than 20 Å were assigned into the last bin. Thus, 0.0−1.1 Å is the no. 0 bin, 0.9−2.1 Å is the no. 1 bin, 1.9−3.1 Å is the no. 2 bin, ..., and 19.9−∞ Å is the no. 20 bin.

When the CONCORD-generated single conformation was used to represent the 3D structure of a compound, it was straightforward to use all the atom pairs existing in this conformation to describe that compound. When the multiple conformations were used to represent a compound, all the atom pairs that exist in any of the conformations were used to describe that compound. This is akin to a Boolean OR operation on all the conformations of a compound, creating a virtual mixture of conformations. This 3D descriptor is similar to the 3D flexible descriptor used by Brown and Martin in their structure-based clustering work.[24] It is expected that this kind of descriptor can adequately retain the major information content in the multiple-conformation database: whether or not a specific atom pair can exist in one of the reasonable 3D structures. However, the information on whether several atom pairs can simultaneously exist in a reasonable 3D structure has been lost.

Both the “atom” type identification and the “binned” distance measure were done using internally developed C codes.

SCAM/RP. All the RP analyses were done using the SCAM (Statistical Classification of Activities of Molecules),[13] a novel computer program recently developed in our group with the special intent to compute a SAR from a large chemical data set. SCAM can handle very large data sets; over 100 000 compounds and over 2 000 000 descriptors are easily handled. SCAM was optimized with respect to speed and memory utilization.

SCAM reads three input files: a data file containing the compound names and potencies, a descriptor dictionary file containing the text explanation of each descriptor, and a binary file containing a record for each compound that indicates the presence of each descriptor in this compound. In this work, the binary file contained the bit string that indicated the presence of the 3D atom pairs in each compound. A 1 indicated that a specific atom pair existed in any of the conformations representing a compound, while a 0 indicated that that atom pair was not found in any of the compound’s conformations. The data file contained the four graded potencies of MAO inhibitors: 0 indicates no activity, while 1, 2, and 3 indicate increasing potencies.

The Student’s t-test is used to recursively partition the whole data set into smaller and more homogeneous subsets, until each subset could no longer be split.[30] SCAM checked all the existing atom pairs sequentially and split the data set into two subsets according to whether or not that atom pair was in the structure. Then the Student’s t-test was computed for these two subsets to calculate the t-value according to

$$t = \frac{\dfrac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{1}{M} + \dfrac{1}{N}}}}{\sqrt{\dfrac{\mathrm{SSX} + \mathrm{SSY}}{M + N - 2}}} \qquad (1)$$

where

$$\mathrm{SSX} = \sum_{i=1}^{M} (X_i - \bar{X})^2, \quad \bar{X} = \sum_{i=1}^{M} X_i/M$$

$$\mathrm{SSY} = \sum_{i=1}^{N} (Y_i - \bar{Y})^2, \quad \bar{Y} = \sum_{i=1}^{N} Y_i/N$$

X_1, X_2, ..., X_M are the potencies in the first subset, and Y_1, Y_2, ..., Y_N are the potencies in the second subset. M and N are respectively the numbers of compounds in these two subsets. The atom pair that gave the smallest p-value was chosen as the best descriptor for the split in the output SAR tree. In this analysis process, a large number of statistical tests are performed. With many statistical tests, there is an increased probability of a false split if the number of tests is not taken into consideration. The Bonferroni adjustment multiplies the raw Student t-test p-value by the number of variables under consideration to give an adjusted p-value that takes this multiplicity into account.[31] The default Bonferroni adjusted p-value of 0.01 was used as the termination criterion,

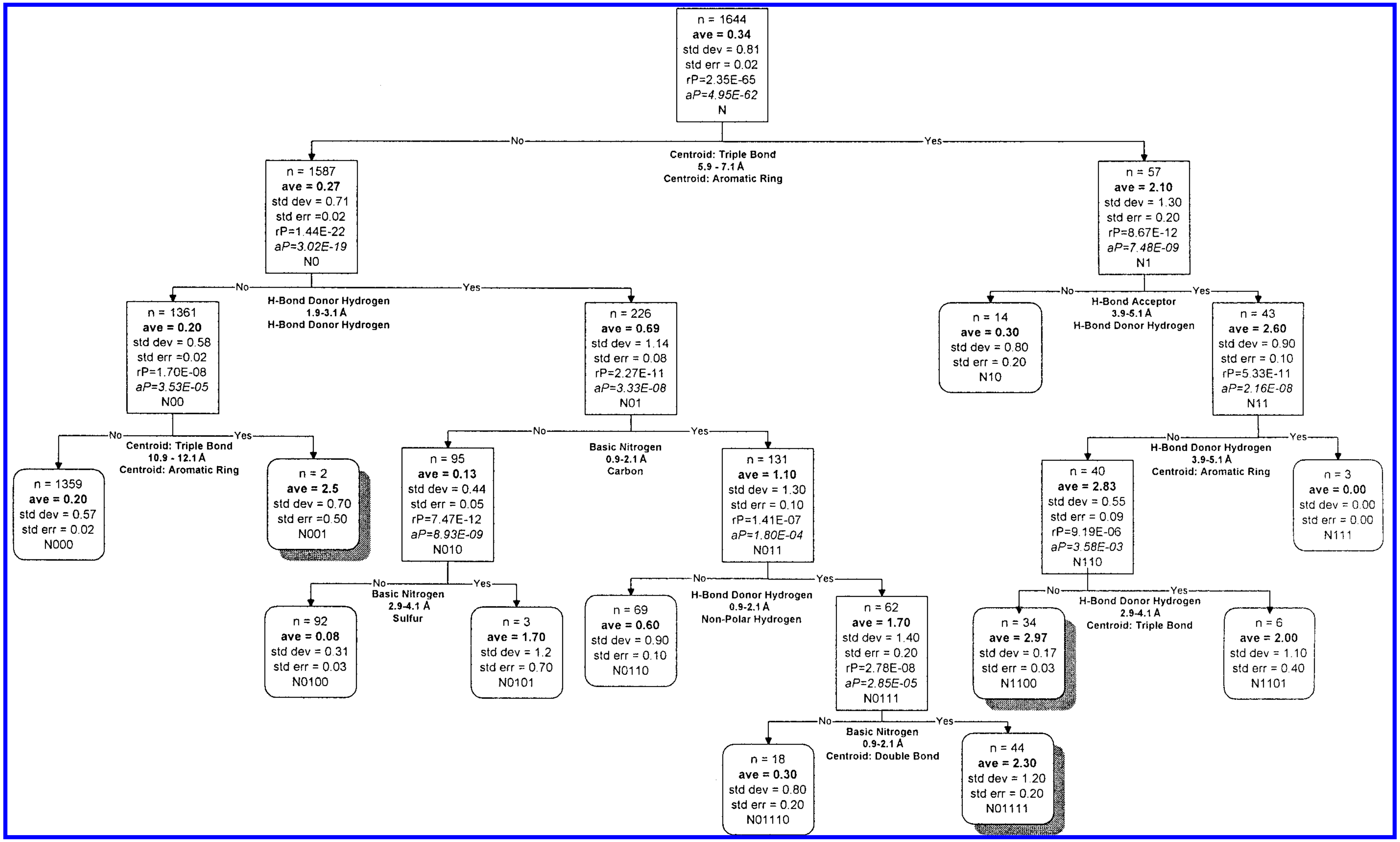

Figure 1. The SAR tree of MAO data set generated by SCAM/RP using single-conformation representation.

stopping the splitting process when this *p*-value was larger than 0.01.

Our descriptors are binary, 0/1, and can be perfectly correlated. In examination of an analysis we note where perfectly correlated variables would make the same split, to alert the medicinal chemist that alternative features would split the data equally well.

RESULTS

SAR Tree Derived from the Single-Conformation Database. This SAR tree is shown in Figure 1. This figure is read in the following way. In the topmost node, there are 1644 compounds with an average potency of 0.34. The variability of these compounds is measured with the standard deviation of the individual compound and the standard error of the mean of the compounds. The best feature for splitting is determined, written below the node. Two *p*-values are given: the raw Student t-test *p*-value, rP, and the Bonferroni adjusted *p*-value, aP. Compounds with the best feature are split to the right, "yes"; compounds without this feature are split to the left, "no". Each node is named with a bit string indicating the splitting process. The rules tracing to a terminal node give the features that define a class of compounds. The quality of a tree can be globally evaluated by examination of the variability across the terminal nodes relative to the variability within the terminal nodes, using an F-test. An F-value of 125.71 was obtained for Figure 1 using the SAS system,[32] indicating the high quality of this tree.

There are three active nodes in this SAR tree, whose average potencies are greater than 2.5. These nodes are marked by gray shadow effect. The examples of the compound structures in nodes N001, N01111, and N1100 are illustrated in Figure 2. Among these three active nodes, node N1100 contains 34 compounds and node N01111 contains 44 compounds. Since these compounds are in terminal nodes that have very different key feature combinations, it is very likely that they indicate the existence of two different mechanisms of action in this MAO data set. This postulation is supported by the previous experimental work. Various hydrazide MAO inhibitors (e.g. compound AL16432 in Figure 2) can be hydrolyzed to acetylhydrazines that act as the irreversible inhibitors,[33] while propargylamines (e.g. compound AL19120 in Figure 2) are themselves suicide inhibitors, which irreversibly inhibit MAO through covalent attachment to its flavin cofactor.[34] This is a clear demonstration that SCAM/RP analysis has the ability to detect multiple mechanisms of action coexisting in a large 3D chemical data set.

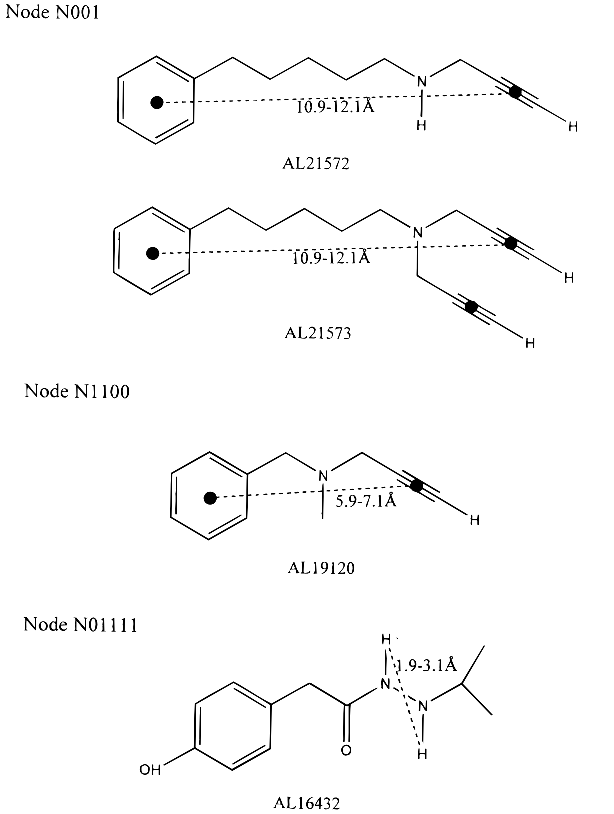

Figure 2. Examples of the chemical structures of the compounds located in the three active nodes of single-conformation SAR tree (cf. Figure 1). Some key features found by SCAM/RP are indicated.

From the SAR tree in Figure 1, some key features for the MAO inhibition can also be detected easily. For example, the feature of "aromatic ring center—triple bond center" atom pair, shown at the first split point, is the structural characteristic of pargyline (see AL19120 in Figure 2), a well-known MAO inhibitor. To visualize the key structural features for these two active nodes and also to confirm the results of our RP analysis, the DISCO pharmacophore mapping module[35] in SYBYL program package was used to independently derive the pharmacophore models for these two groups of active compounds, respectively. First, a utility program for SCAM, Pachinko,[13] was used to identify the compounds in the nodes N1100 and N01111. Individual structures cascaded down the SAR tree according to the key feature(s) listed at each split point. A compound would drop to the right if it possessed the key feature(s) listed at that split point, otherwise it would drop into the left node on the lower level. Then, only the most active compounds, whose potencies are equal to 3, were picked out for the DISCO pharmacophore mapping. To obtain results comparable with SCAM/RP, all the site point features, which were originally designed in DISCO to represent the possible key features on the receptor binding-site, were deleted, and some new ligand point features such as triple bond center (cf. Table 1) were added. The tolerances of pair wise distances between any two kinds of feature points were set at 0.5 Å. The final DISCO-generated pharmacophore models for 33 most active compounds in node N1100 and 29 most active compounds in node N01111 are illustrated in Figures 3 and 4, respectively. In Figure 3, the pharmacophore distances are 6.48 ± 0.5 Å between the hydrophobic center and the triple bond center, 2.90 ± 0.5 Å between the triple bond center and the H-bond acceptor center, and 3.67 ± 0.5 Å between the H-bond acceptor center and the hydrophobic center. This model is consistent with the MAO pharmacophore proposed by Johnson.[36]

In Figure 4, the second pharmacophore is actually composed of an amide group with its adjacent nitrogen atom that has an available lone pair. The pharmacophore distances are 2.66 ± 0.5 Å and 2.22 ± 0.5 Å between carbonyl oxygen and two nitrogen atoms, respectively, and 1.42 ± 0.5 Å between two nitrogen themselves. Both pharmacophore models are consistent with the SAR tree in Figure 1, and all the key features for the nodes N1100 and N0111 are

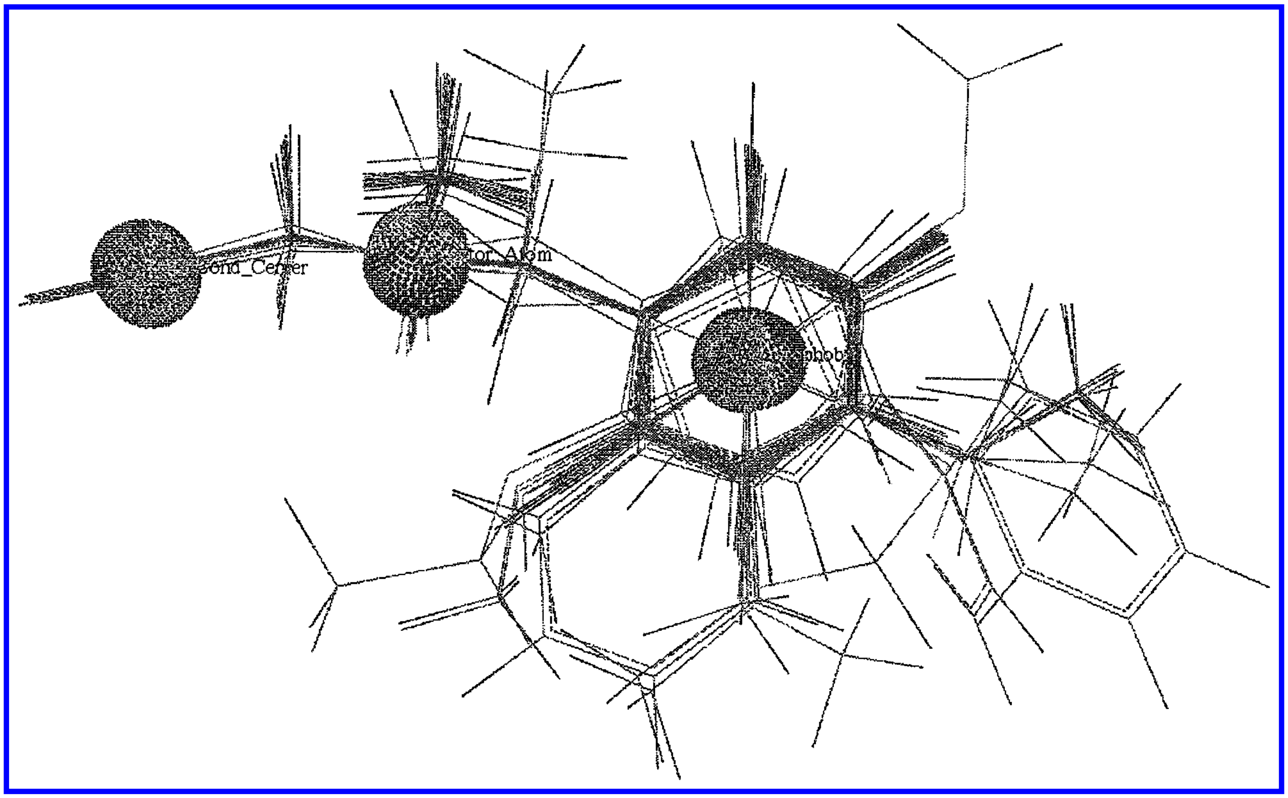

Figure 3. Pharmacophore model for the node N1100 of single-conformation SAR tree (cf. Figure 1) generated by the SYBYL/DISCO.

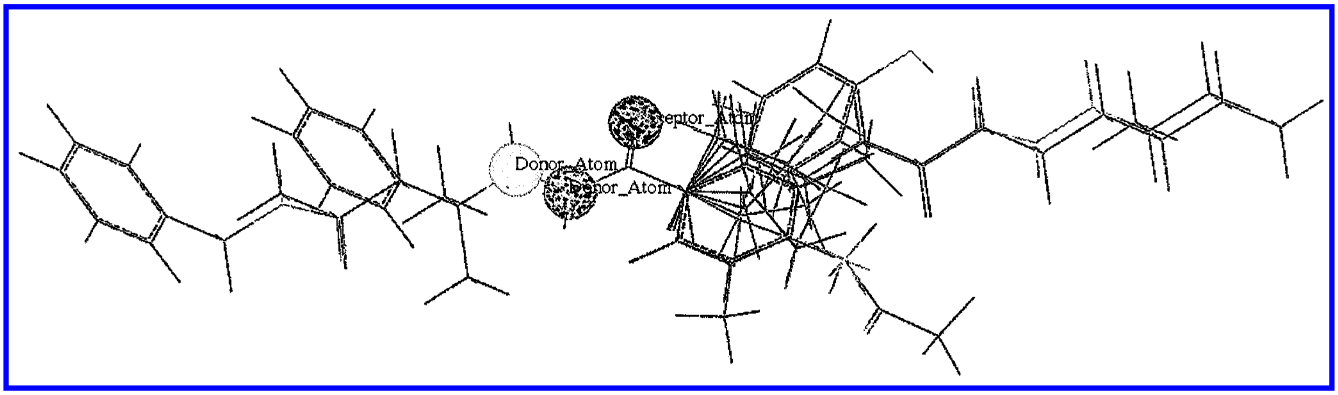

Figure 4. Pharmacophore model for the node N01111 of single-conformation SAR tree (cf. Figure 1) generated by the SYBYL/DISCO.

confirmed from the Figures 3 and 4. This demonstrates that SCAM/RP analysis can provide similar SAR results as other pharmacophore mapping methods like DISCO. However, DISCO has no way to find two distinctively different pharmacophore models, like those in Figures 3 and 4, simultaneously, because it was not designed to handle mixtures of mechanisms of action and thousands of compounds at the same time.

SAR Tree Derived from the Multiple-Conformation Database. Following the same protocol, the SAR tree was derived from the multiple-conformation database and shown in Figure 5. An F-test for this tree gives a value of 127.61, slightly better than the F-value of single-conformation SAR tree in Figure 1.

Two active nodes, N10101 and N011101, are found in this SAR tree. The former contains 35 compounds and the latter contains 26 compounds. Furthermore, the key features that lead to these two active nodes also look like those in the single-conformation SAR tree. DISCO was used again to derive the pharmacophore models for these two active nodes. The same procedure as above was followed. First, a total of 362 977 conformational structures were dropped down the SAR tree using Pachinko to find those individual conformations located in the active nodes. Among these conformations, the most active compounds (potency = 3) were chosen as the input structures for the DISCO pharmacophore mapping. For the 441 conformations of 35 most active compounds located in the node N10101, DISCO generated the identical pharmacophore model as shown in Figure 3. The pharmacophore distances are 5.90 ± 0.5 Å between the hydrophobic center and the triple bond center, 3.01 ± 0.5 Å between the triple bond center and the H-bond acceptor center, and 3.87 ± 0.5 Å between the H-bond acceptor center and the hydrophobic center. However, only six conformational structures of three compounds were dropped into node N011101. A more careful examination revealed that seven other atom pair features coexisted with the one shown at the last split point on the path to node N011101. These eight atom pair features all have the same smallest *p*-value. Although they coexisted in the descriptor

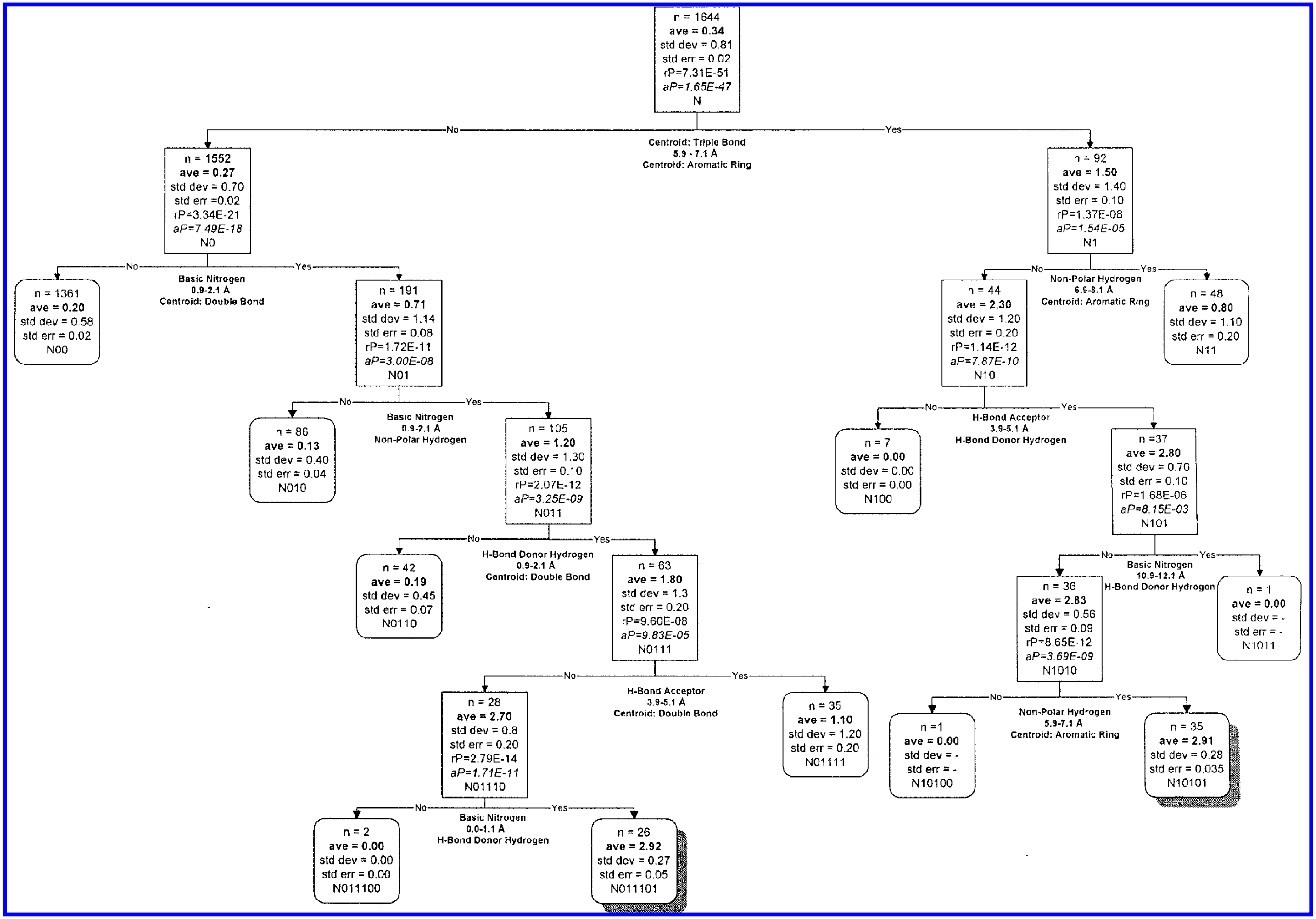

Figure 5. The SAR tree of MAO data set generated by SCAM/RP using multiple-conformation representation.

bit-string of a compound, they could not exist in a reasonable conformational structure simultaneously. Thus, this split is actually an erroneous split. When Pachinko used these eight features to split the flow of conformations, very few conformations would be led to the right because most conformational structures could not possess them simultaneously. As we have mentioned, this kind of error comes from the way we generate the compound descriptor in the case of multiple-conformation database. After finding this split is a misleading one, we went back to check the conformational structures passing through node N01110. Thirty-nine conformational structures of 20 most active compounds were found and entered into DISCO and the exactly same pharmacophore model as Figure 4 was generated, because of the rigidity of this model.

An interesting result is that compounds AL21572 and AL21573 (cf. Figure 2), which locate in a separated node in the single-conformation tree (see node N001 in Figure 1), are found to be together with other active compounds in the active node N10101 in the multiple-conformation tree (cf. Figure 5). They all satisfy the pharmacophore model illustrated in Figure 3. Obviously, these two compounds have distorted their side chains to attain this less extended conformation. This is a good example that a multiple-conformation representation can help find a correct classification for some flexible compounds.

CONCLUSIONS

From the results of this work, it is demonstrated again that RP analyses can find reasonable SAR rules in a very large data set, even in the huge 3D space, hundreds of thousands of conformations and thousands of different descriptors in this work. To our knowledge, this is the first time 3D-SARs are successfully derived from such a large, heterogeneous data set. Hundreds of thousands of 3D structures were considered when we used multiple conformations to represent a compound; the successful retrieval of the key features from such a huge pool of data seems impressive. It is much like finding the proverbial “needle in the haystack”.

Compared with the previous 2D RP work,[10−13] 3D RP analyses provides us more comprehensible SAR rules, which can directly lead to one or several pharmacophore models when used with some other pharmacophore mapping programs such as DISCO. Often, the SAR rules derived by 2D RP analyses may be difficult to rationalize in 3D. This work also gives us a good example of a strategy to derive 3D pharmacophore model from a large, diverse structure−activity data set. We can use RP to classify a large data set by grouping the active compounds probably having the same binding mode together and obtaining the useful key structural features as well. Then, some pharmacophore mapping program like DISCO, which was originally designed to handle small data sets, can be used to find the pharmacophore model for each group of active compounds, by the use of the key structural features provided by the RP analysis.

Although the multiple-conformation representation of the 3D structures of compounds was designed to hopefully include direct information on active conformations into the descriptor and provide the chance to obtain a better SAR model, the present results indicate that its performance is not appreciably better than the single-conformation representation. Mixing the information of many conformations into a single molecular descriptor not only makes it more difficult for RP to pick up the essential features but also can induce occasional erroneous rule generation as illustrated above. The single-conformation representation provided comparable SAR trees with the multiple-conformation representation in this work, but its reliability and predictability is still limited by the quality of the generated structure—whether it is enough close to the real bound conformation.

In future studies, we need to find a better way to guide the RP to separate the bound conformations from other irrelevant conformations. A possible way is to closely combine the conformation search process with the RP analysis process. Another hurdle before the 3D SAR analysis of large data sets is the efficient search of the conformational space. At present, it is still the rate-limiting step in the whole process and costs over 95% of the computational efforts. Furthermore, the excluded volume and external pharmacophore feature points located at the binding site were not directly considered in this work, although some relative information can sometimes be derived from the SAR tree. (For example, see the N11 node in Figure 5. The negative rule seems to say a certain feature is related to the excluded volume.) More directly introducing these concepts into molecular descriptors would provide more information about the interaction between the compounds and their receptor and hopefully lead to a better result of the 3D SAR analysis.

REFERENCES AND NOTES

(1) Patents pending. We have patents pending on these methods.

(2) Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Bodor, S. P. A.; Gordon, E. M. Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries. *J. Med. Chem.* **1994**, *37*, 1233−1251.

(3) Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. Applications of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies, and Future Directions. *J. Med. Chem.* **1994**, *37*, 1385−1401.

(4) Kubunyi, H. *QSAR: Hansch Analysis and Related Approaches*; VCH: 1993.

(5) Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; American Chemical Society; 1995.

(6) King, R. D.; Muggleton, S.; Lewis, R. A.; Sternberg, M. J. Drug Design by Machine Learning: the Use of Inductive Logic Programming to Model the Structure−Activity Relationships of Trimethoprim Analogues Binding to Dihydrofolate Reductase. *PNAS* **1992**, *89*, 11322−11326.

(7) King, R. D.; Muggleton, S.; Srinivasan, A.; Sternberg, M. J. Structure−Activity Relationships Derived by Machine Learning: the Use of Atoms and Their Bond Connectivities to Predict Mutagenicity by Inductive Logic Programming. *PNAS* **1996**, *93*, 438−442.

(8) Klopman, G. Artificial Intelligence Approach to Structure−Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7321.

(9) Klopman, G. MULTICASE. 1. A Hierarchical Computer Automated Structure Evaluation Program. *Quant. Struct.-Act. Relat.* **1992**, *11*, 176−184.

(10) Young, S. S.; Hawkins, D. M. Analysis of a 2^9 Full Factorial Chemical Library. J. *Med. Chem.* **1995**, *38*, 2784−2788.

(11) Young, S. S.; Hawkins, D. M. Using Recursive Partitioning to Analyze a Large SAR Data Set. *SAR QSAR Env. Res.* **1998**, *8*, 183−193.

(12) Hawkins, D. M.; Young, S. S.; Rusinko, A. Analysis of a Large Structure−Activity Data Set Using Recursive Partitioning. *Quant. Struct.-Act. Relat.* **1997**, *16*, 1−7.

(13) Rusinko, A.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. SCAM: Statistical Classification of Activities of Molecules Using Recursive Partitioning. *J. Am. Chem. Soc.* To be submitted.

(14) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.

(15) Barnard, J. M.; Downs, G. A. Clustering of Chemical Structures on the Basis of 2-D Similarity Measures. J. *Chem. Inf. Comput. Sci.* **1992**, *32*, 644−649.

(16) Downs, G. M.; Wilett, P. In *Advanced Computer-Assisted Techniques in Drug Discovery*; van de Waterbeemd, H., Ed.; VCH: Weinheim; 1994; Vol. 3.

(17) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(18) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: a New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.

(19) Nilikantan, R.; Bauman, N.; Venkataraghavan, R. New Method for Rapid Characterization of Molecular Shapes: Application in Drug Design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79−85.

(20) Miller, M.; Schneider, J.; Sathyanarayana, B. K.; Toth, M. V.; Marshal, G. R.; Clawson, L.; Selk, L.; Kent, S. B. H.; Wlodawer, A. Structure of a Complex of Synthetic HIV-1 Protease with a Substrate-Based Inhibitor at 2.3 Å Resolution. *Science* **1989**, *246*, 1149−1152.

(21) Kubinyi, H. *3D QSAR in Drug Design: Theory, Methods, and Applications*; ESCOM Science Publishers: Leiden, 1993.

(22) Cramer, R. D.; Patterson, D. E.; Bruce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *10*, 5959−5967.

(23) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of Common Functional Configurations Among Molecules. *J Chem. Inf. Comput. Sci.* **1996**, *36*, 563−571.

(24) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J Chem. Inf. Comp. Sci* **1996**, *36*, 572−584.

(25) *CONCORD. A Program for the Rapid Generation of High Quality Approximate 3-Dimensional Molecular Structures*; The University of Texas at Austin and Tripos Associates: St. Louis, MO.

(26) *MACCS-II Manual*; Molecular Design Ltd.: San Leandro, CA.

(27) *SYBYL Manual*; Tripos Associate Inc.: St. Louis, MO.

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

(29) Clark, M.; Cramer, R. D.; Opdenbosch, N. V. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *8*, 982−1012.

(30) Hawkins, D. M.; Kass, G. V. In *Topics in Applied Multivariate Analysis*; Hawkins, D. H., Ed.; Cambridge University Press: 1982; p 269.

(31) Miller, R. G. *Simultaneous Statistical Inference*; Springer-Verlag: New York, 1981.

(32) *SAS Manual*; SAS Institute Inc.: Cary, NC.

(33) Nelson, S. D.; Mitchell, J. R.; Timbrell, J. A.; Snodgrass, W. R.; Corcoran, G. B., III. Soniazid and Iproniazid: Activation of Metabolites to Toxic Intermediates in Man and Rat. *Science* **1976**, *193*, 901−903.

(34) Maycock, A. L.; Abeles, R. H.; Salach, J. I.; Singer, T. P. The Structure of the Covalent Adduct Formed by the Interaction of 3-dimethylamino-1-propyne and the Flavine of Mitochondrial Amine Oxidase. *Biochemistry* **1976**, *15*, 114−125.

(35) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. A Fast New Approach to Pharmacophore Mapping and Its Application to Dopaminergiic and Benzodiazepine Agonists. *J. Comp.-Aided Mol. Des.* **1993**, *7*, 83−102.

(36) Johnson, C. L. Quantitative Structure−Activity Studies on Monoamine Oxidase Inhibitors. *J Med. Chem.* **1976**, *19*, 600−605.

CI980089G