Automated Pharmacophore Identification for Large Chemical Data Sets[1]

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

Chemoinformatics Group, Research Information Systems, Glaxo Wellcome Inc.,
Research Triangle Park, North Carolina 27709, Laboratory for Molecular Modeling, School of Pharmacy,
University of North Carolina, Chapel Hill, North Carolina 27599, and Medicinal Chemistry,
Alcon Laboratories Inc., Fort Worth, Texas 76134

Received May 25, 1999

The identification of three-dimensional pharmacophores from large, heterogeneous data sets is still an unsolved problem. We developed a novel program, SCAMPI (statistical classification of activities of molecules for pharmacophore identification), for this purpose by combining a fast conformation search with recursive partitioning, a data-mining technique, which can easily handle large data sets. The pharmacophore identification process is designed to run recursively, and the conformation spaces are resampled under the constraints of the evolving pharmacophore model. This program is capable of deriving pharmacophores from a data set of 1000−2000 compounds, with thousands of conformations generated for each compound and in less than 1 day of computational time. For two test data sets, the identified pharmacophores are consistent with the known results from the literature.

INTRODUCTION

The recent progress of combinatorial chemistry[2,3] and high throughput screening (HTS) techniques[4,5] is providing a revolution for drug discovery 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, leading to the tremendous increase of the quantity of data in the drug discovery cycle. However, how to timely analyze this large amount of data and convert it into utilizable information remains a big challenge for chemoinformaticians. Among many unsolved problems is the automated pharmacophore identification process for large chemical data sets.

The concept of a pharmacophore, which can be described as the key chemical *features* and the spatial relationships among them (*configurations*) in the determination of the biological activities of chemical compounds,[6] is one of the most important concepts in medicinal chemistry and has played an important role in drug discovery. It can help medicinal chemists gain insight about the key interactions between ligand and receptor, especially when the 3D ligand−receptor complex structure has not been determined. It can be used as the search query for a 3D database search, which is called a pharmacophore search, and has been demonstrated as a very productive way for lead compound discovery.[7] It can also be used for 3D QSAR analysis, grouping together the compounds that follow the same binding mode and indicating the possible 3D alignment rules.[8]

However, pharmacophore identification was traditionally a process heavily dependent on the experience, skill, and intuition of medicinal chemists. Although often successful, humans work under some limitations. For example, humans cannot handle large amounts of data, cannot keep working for a long time, etc. During the past decade, we have seen many efforts made to involve the contribution of computational methods into this process.[9] These efforts are usually called *automated pharmacophore identification (or pharmacophore mapping or pharmacophore recognition)*. To the best of our knowledge, the approaches and programs that were specifically developed for this purpose include AAA (active analogue approach),[10−12] ensemble distance geometry,[13] DISCO,[14] Catalyst/Hypo,[15,16] Catalyst/HipHop,[17,18] Chem-X/pharmacophore,[19] Apex-3D,[20] DANTE,[21,22] etc., and use ILP (inductive logic programming) system Progol to do 3D pharmacophore identification.[23] Furthermore, many QSAR programs also have a certain pharmacophore identification function, like Compass,[24] etc. However, all these methods have the following limitations: (a) None of them was originally designed for large, heterogeneous chemical data sets. They are inherently limited to small data sets, which contain less than 100 compounds according to the published results. (b) Except for Catalyst/Hypo and Apex-3D, most of them consider only the structural information provided by a small number of the most active compounds. However, this is usually not a good strategy for analyzing a large chemical data set. (c) Except for Progol and DANTE, most of them presume that all the compounds follow the same mechanism and, therefore, cannot handle the situation of multiple binding modes, which are expected in a large chemical data set. All these limitations make the existing pharmacophore identification methods unsuitable for analyzing large chemical data sets such as HTS data sets.

Reported here is a novel computational program, SCAMPI (statistical classification of activities of molecules for pharmacophore identification), that was developed specifically for identifying pharmacophores from large, heterogeneous chemical data sets by jointly utilizing data-mining and

* Corresponding author. E-mail: ssy0487@glaxowellcome.com. Telephone: 919-483-8456. Fax: 919-483-2494.

[†] Glaxo Wellcome Inc.

[‡] University of North Carolina.

[§] Alcon Laboratories Inc.

10.1021/ci990327n CCC: $18.00 © 1999 American Chemical Society
Published on Web 08/19/1999

computational chemistry techniques. The methodology and implementation details of this program are described, followed by application to two large, public chemical data sets. Finally, some general considerations about the automated pharmacophore identification for large chemical data sets are discussed.

METHODOLOGY

Although the simplest way to analyze a large chemical data set is to convert it into a small data set by sorting the activities of compounds and then considering only a small number of the most active compounds, this strategy obviously ignores much of the data and has the following disadvantages. (a) Some weakly active compounds may follow a mechanism different from the most active compounds and point to novel and promising lead compounds in the next round of biological assay. The chance to discover these lead compounds will thus be lost if we consider only the most active compounds. (b) Inactive and weakly active compounds also contain valuable information about the structure−activity relationships, e.g., by indicating the structural features that significantly decrease the activities. (c) Selecting the cutoff value of activity for the compounds to be analyzed is usually highly arbitrary. Different choices may significantly influence the final results. Therefore, SCAMPI is designed to directly derive pharmacophores from a large number of compounds, by integrating recursive partitioning and fast conformation search methods.

Theoretically, a pharmacophore identification process needs to search two spaces:[17] *the conformational space* that represents all the reasonable 3D conformational structures for each individual compound and *the correspondence space* that indicates all the possible correspondences of chemical features and configurations among different compounds. The pharmacophore identification problem can be better understood by posing the following questions: (a) *How do you search the conformational space?* (b) *How do you search the correspondence space?* (c) *How do you combine these two searches?* We will describe the SCAMPI methodology by answering these three questions in order.

Conformational Search. Most of the existing pharmacophore identification methods try to use a small number (usually tens) of conformers to completely cover the whole conformational space of a compound. These representative conformers are generated either by some conformational search method followed by clustering analysis[25] or by some sophisticated method that is specifically developed for generating diverse conformational structures, like the "poling" method.[26] Energy minimization is often used to restrict the generated conformations within the low-energy regions. However, this kind of strategy has the following drawbacks: (a) Tens of conformers may not be enough to completely cover the whole conformational space of a highly flexible compound, which often exists in a large chemical data set.[27] (b) Receptor-bound conformations may not be located in the low-energy regions determined by the force-field calculation.[28] (c) The heavy computational burden of this strategy prohibits its application to large chemical data sets.

Dammkoehler, Marshall, and their colleagues followed a different strategy by developing a novel systematic conformational search method, called constrained search, for their pharmacophore identification approach (AAA).[29,30] This method is based on an equation that expresses the variable distance between two nonbonded atoms as a function of a single rotational variable:[31,32]

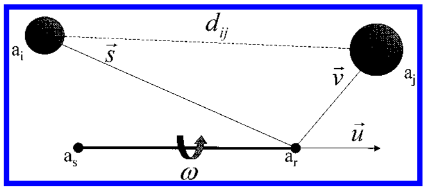

Figure 1. Illustrating eq 1. During the rotation of a rotatable bond that is defined by two anchor atoms *s* and *r*, the positions of atoms *i*, *s*, and *r* remain constant, while atom *j* moves around the axis of rotation. Atom *j* can be covalently linked to *r* or connected to *r* by a sequence of nonrotatable bonds. $\vec{s}$ and $\vec{v}$ are the vectors from atom *r* to atom *i* and to atom *j*, respectively, while *u* is the unit directional vector following the axis of rotation.

$$d_{ij}^{\ 2}(\omega) = k_1 + k_2 \cos \omega + k_3 \sin \omega \qquad (1)$$

As illustrated in Figure 1, d_{ij} is the distance between two atoms *i* and *j*; ω is the rotational angle around the axis of a rotatable bond, which is define by two anchor atoms *r* and *s*; and k_1, k_2, and k_3 are all constants and can be calculated from the vectors $\vec{s}$, $\vec{\nu}$, and $\vec{u}$[29] (cf. Figure 1). If an acceptable conformational structure is defined as a structure that satisfies bump checking and does not contain any bad van der Waals contacts among atoms, the distance d_{ij} will be larger than the sum of the van der Waals radii of atoms *i* and *j*, c_{ij} leading to

$$k_1 + k_2 \cos \omega + k_3 \sin \omega > c_{ij}^{\ 2} \qquad (2)$$

This inequality can be viewed as a distance constraint in the conformational search. After some mathematical transformations, inequality 2 can be converted into a quadratic form so that the ω range that corresponds to all the sterically allowed conformations can be analytically determined.[30] This ω range can be called the acceptable rotational range. The intersection of the acceptable rotational ranges for all the nonbonded atom pairs located at both ends of a rotatable bond produces the final acceptable rotational ranges.[30] With the rotational angle sampled within these common ranges, the generated conformation will definitely satisfy van der Waals bump checking. Hence, this algorithm is actually a look-ahead algorithm. It can determine the acceptable rotational ranges before really rotating a rotatable bond, while a traditional trial-and-error algorithm usually rotates a rotatable bond blindly, followed by either bump checking to decide if the generated conformation is acceptable or energy minimization to release the bad contacts. This difference makes this constrained search algorithm much faster than most of the traditional systematic search algorithms by 3 or 4 orders of magnitude, according to the original reports.[29,30,33] Furthermore, this algorithm can be easily extended to other distance constraints. For example, if *i* and *j* (cf. Figure 1) now represent pharmacophoric points and we want to do a conformational search with the distance between them constrained within a certain distance range, the ω ranges that satisfy this pharmacophoric distance constraint can be determined analytically in a similar way.

Due to its speed and adaptation to the additional pharmacophoric distance-constrained search, this algorithm is adopted in SCAMPI for fast conformational searching. Several modifications have been introduced to make it run even faster or to satisfy some special requirements of SCAMPI. Some of these modifications are described as follows.

(a) Random Search Instead of Systematic Search Used. Once identified, the acceptable rotational ranges are randomly sampled in SCAMPI, rather than in the systematic way as in the original constrained search method. The major consideration here is that systematic search may be severely hindered or virtually stopped by highly flexible compounds, which often exist in a large chemical data set. Therefore, the random search strategy is chosen to get a broad sampling of the whole conformational space, with a cutoff value (default 1000) preset for the maximum number of conformers that are allowed for each compound, to limit the whole computation time within an acceptable range. Clustering analysis is not followed to try to derive diverse, representative conformations, because we use the recursive partitioning method to identify pharmacophores from a large group of conformations (described later). In essence, we let the biological activity data select the appropriate conformations.

(b) Minimum Accessible Distance Calculation Used To Further Increase the Computational Speed. According to the original constrained search method, if there are M atoms at one end of a rotatable bond and N atoms at the other end, the calculation based on inequality 2 needs be repeated $M \times N$ times before obtaining the final acceptable rotational ranges by intersection. However, we noticed that in most cases the final acceptable rotational ranges are only determined by a small number of atom pairs (usually far less than $M \times N$). The reason is that the minimum accessible distance between two atoms in most of the atom pairs is larger than the sum of their van der Waals radii so that there is no bad van der Waals contact between them no matter how the rotatable bond is rotated. Since the calculation of a minimum accessible distance[34] is much easier and faster than that of an acceptable rotational range, we introduce this calculation as the first step to examine all the atom pairs. Only those pairs of atoms whose minimum accessible distances are smaller than the sums of their van der Waals radii will be further considered for the calculation of acceptable rotational ranges. A similar consideration is also applied to the case of the pharmacophoric distance constraint. Only those pharmacophoric pairs whose minimum accessible distances are smaller than the upper boundaries of their distance constraint ranges and maximum-accessible-distances are larger than the lower boundaries will be further considered for the calculation of acceptable rotational ranges. Our experience has indicated that the inclusion of such a calculation as a filter can significantly increase the computational speed.

(c) Conformational Search Implemented in Both Cartesian and Internal Coordinates. In the original constrained search method, a conformational search of a flexible ring is transformed to a chain conformational search by breaking one of the ring bonds and adding a tight distance constraint between the two terminal atoms of the broken bond. This kind of tight distance constraint will cost a considerable portion of computational time for this algorithm to find a successful ring closure. However, the conformational search can be implemented in two different systems: a Cartesian coordinate system where the Cartesian coordinates of each atom are directly perturbed, and an internal coordinate system where the torsional angle of each rotatable bond is directly modified.[35] Obviously, Cartesian coordinates are more suitable for the flexible ring conformational search, while internal coordinates are a natural choice for the chain conformational search. Therefore, the conformational search in SCAMPI is implemented in both systems. A flowchart of the conformational search procedure in SCAMPI is illustrated in Figure 2.

Correspondence Search. The earliest pharmacophore identification methods, like the active analogue approach and ensemble distance geometry, simplify the correspondence search by requiring the user to indicate the correspondence of pharmacophoric features among different compounds. Most of the new pharmacophore identification methods depend on some pairwise comparison algorithm[36] to detect the chemical features and configurations that can be shared by all the active compounds. However, this kind of strategy is usually sensitive to the choice of training compounds and computationally expensive, therefore inherently limited to small data sets.

The correspondence search in SCAMPI uses the recursive partitioning algorithm, as implemented in the FIRM and SCAM programs,[37−40] to identify the chemical features and configurations that are statistically most significantly correlated with the biological activities. An easily interpreted dendrogram or tree diagram is generated in which the statistically best molecular descriptors are used to split a large data set into multiple smaller and more homogeneous subsets. The advantages of recursive partitioning include the following: (a) It is inherently fast when compared with many other methods for grouping compounds.[41] (b) It overcomes the difficulties of handling nonlinear relationships and strong interactions in large data sets.[42] (c) It can detect the multiple mechanisms by dropping the compounds with different mechanisms into the different terminal nodes of a dendrogram.

The Student's t-test is used as the split criterion in SCAMPI to recursively partition a whole data set into multiple subsets, until each subset cannot be split any further. For each Student's t-test, a binary molecular descriptor matrix is generated in which each row represents an individual compound, each column represents a different molecular descriptor, and each position has a binary value, 1/0, indicating the presence or absence of a molecular descriptor in a compound.[43] Then, each one of the molecular descriptors is sequentially checked, and the data set is split into two subsets according to whether that molecular descriptor exists in a compound or not. The Student's t-value is computed according to the following formula:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{M} + \frac{1}{N}}\sqrt{\frac{\mathrm{SSX} + \mathrm{SSY}}{M + N - 2}}} \qquad (3)$$

where

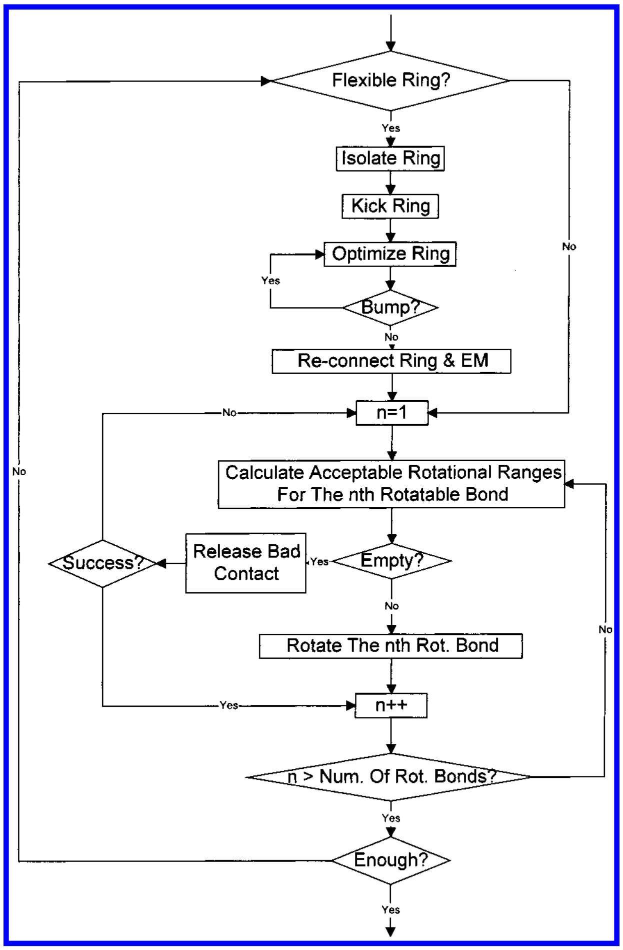

Figure 2. Flowchart of the conformational search procedure in SCAMPI. (1) If there are any flexible rings in the compound, each flexible ring is isolated by cutting off its side chain(s) but keeping its nearest side-chain neighbors. (2) Each flexible corner of the isolated rings is randomly kicked.[56] (3) Several steps (20−100 steps) of energy minimization are used to optimize each ring's structure to an acceptable conformation; if it is failed after 100 steps, the flexible ring will be randomly kicked again from the initial structure. (4) All the side chains are reconnected to the flexible rings to form the complete compound. (5) Several steps (20−100 steps) of energy minimization are used again to optimize the whole compound's structure to release the possible bad contacts between the side chains and rings; if it is failed after 100 steps, the flexible rings will be randomly kicked again. (6) With the flexible rings treated as rigid groups, the constrained search algorithm, as described in text, is used to calculate the acceptable rotational ranges for the first rotatable bond. If the acceptable torsional ranges are not empty, a rotational angle will be randomly picked from them and the first rotatable bond will be correspondingly rotated. Then, the algorithm moves on to the next rotatable bond. (7) If the acceptable torsional ranges are empty, a "release procedure" will be used to relieve the existing bad contacts.[57,58] If this release attempt succeeds, the constrained search algorithm moves on to the next rotatable bond; otherwise, the whole conformational build-up process will be started again from the first rotatable bond.

$$\mathrm{SSX} = \sum_{i=1}^{M}(X_i - \bar{X})^2 \qquad \mathrm{SSY} = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

$$\bar{X} = \sum_{i=1}^{M} X_i/M \qquad \bar{Y} = \sum_{i=1}^{N} Y_i/N$$

X_1, X_2, ..., X_M are the activities of the compounds in the first subset, and Y_1, Y_2, ..., Y_M are the activities of the compounds in the second subset. M and N are the numbers of the compounds in these two subsets, respectively. The molecular descriptor that gives the largest t value is chosen as the descriptor for the split, if its corresponding Bonferroni-adjusted p value is smaller than some cutoff value (default 0.01), which indicates the acceptable significance. The Bonferroni adjustment multiplies the raw Student's t-test p value by the number of variables under consideration, thereby taking into account the number of statistical tests in order to avoid the increased probability of a false positive split due to multiple testing.[44]

SCAMPI follows a pharmacophore build-up procedure similar to those in Catalyst/HipHop[17] and DANTE:[21] a two-point pharmacophore is searched at first, and then a new pharmacophoric point is searched and added if it is found; this process continues until no more pharmacophoric points can be found. Consequently, SCAMPI generates all the possible two-point molecular descriptors for each compound, at first. These two-point descriptors are composed of two chemical features and the "binned" distance[45] between them. Once the Student's t-test determines the most significant one from these two-point descriptors, it is treated as the first two pharmacophoric features. Then, all the possible three-point molecular descriptors for each compound are generated. These three-point descriptors retain the feature of the most significant two-point descriptor that has been identified and, therefore, are only composed of a third chemical feature and its two binned distances to each one of the two pharmacophoric features that have been identified. From these three-point descriptors, the Student's t-test is used again to determine the third pharmacophoric feature. Following this procedure, the four-point, five-point, etc., descriptors are generated sequentially with the evolving pharmacophore model, until no new pharmacophoric points can be found.

There are two kinds of splits at each split point: positive and negative. A positive split is one in which the subnode that contains the split descriptor is the more active node; i.e., the subset of compounds containing the split descriptor is more active on average than the subset of compounds not containing the split descriptor; otherwise, it is a negative split. Negative split descriptors are sometimes difficult to explain in terms of classic pharmacophore modeling, but they may also provide valuable information, like the excluded volume. Therefore, SCAMPI allows the user to select the split methods, although the default choice is positive split only.

Combining the Conformational Search and Correspondence Search. All the existing pharmacophore identification methods treat the conformational search and correspondence search as two separate steps. A set of representative conformers is generated first to hopefully cover the whole conformational space of each compound, and then pharmacophores are identified from, and only from, these representative conformations; no new conformers are generated in the pharmacophore identification process.

The SCAMPI strategy is to combine the conformational search and correspondence search together and let them depend on each other. The sampling completeness of descriptors is used as the stop criterion for the conformational search, since a further conformational search will not add more information for the following statistical test once the possible descriptors have been found. Furthermore, the identified pharmacophoric features and configurations are imposed as additional constraints in the next-round confor-

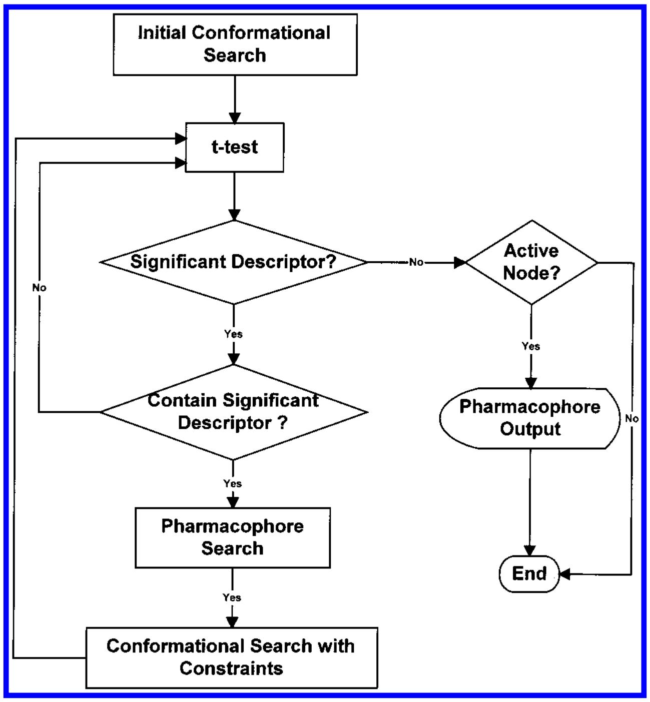

Figure 3. Flowchart for the general design of SCAMPI.

mational search so that the conformational subspace that satisfies the already identified part of a pharmacophore model can be sampled more thoroughly. This strategy is illustrated as the flowchart in Figure 3 and is described as follows.

SCAMPI starts the pharmacophore build-up process with the conformational search for each compound, during which the presence of “extended” two-point descriptors is recorded. The extended two-point descriptors have the form of “chemical feature ID”−“binned distance”−“chemical feature ID”, indicating two chemical feature points and the binned distance between them. If for continuous *n* times (default 100) no new such descriptors are found or the maximum number of conformations that are allowed for a compound (default 1000) has been reached, the conformational search process is stopped. All the generated conformations are temporarily saved in memory space for future use.

Next, all the extended two-point descriptors are converted to the corresponding “standard” two-point descriptors, which have the form of “chemical feature type”−“binned distance”−“chemical feature type” as has been described. Next, the Student’s *t*-test is used to check all these standard two-point descriptors to find the most significant one to split the data set. If found, it will be treated as the first two-point pharmacophore and the data set will be correspondingly divided into two subsets: the compounds containing this two-point pharmacophore and the compounds not containing it.

For all the compounds containing this two-point pharmacophore, a pharmacophore search using the Ullman algorithm[46,47] is done for all the saved conformations of these compounds to identify where these two pharmacophoric points are so that the pharmacophoric distance constraint can be added between them in the next round of the conformational search.

During this round of the conformational search, the presence of extended three-point descriptors is recorded. These extended three-point descriptors have the form of “chemical feature ID”−“binned distance”−“binned distance”, indicating a new chemical feature point and its binned distances to the two pharmacophoric points that have been identified. The same stop criterion as in the first-round conformational search is also used here to decide when to stop this round of conformational search. Then, all the extended three-point descriptors are converted to the corresponding standard three-point descriptors, which have the form of “chemical feature type”−“binned distance”−“binned distance” and will be checked by the Student’s *t*-test again to find the third pharmacophoric point.

For all the compounds not containing the first two-point pharmacophore, the Student’s *t*-test will be used again to analyze all of their standard two-point descriptors, to find the most significant one to split them. If found, it will be treated as the second two-point pharmacophore.

The above iterations are repeated, as shown in Figure 3, until no more significant splits can be found or the default maximal number of pharmacophoric points (default 5) has been reached. Finally, SCAMPI will output the pharmacophore models indicated by the active terminal nodes in the final dendrogram.

Thus, the pharmacophore identification strategy that SCAMPI follows is actually the adaptive sampling strategy in which the sampling procedure is dependent on the values of some variables of interest that are observed in the sampling process.[48] The choice of this strategy is based on our understanding that both the conformational space and the correspondence space are huge and are not easily sampled completely, especially for large chemical data sets. Therefore, we have to choose some more rational sampling strategies, like the adaptive sampling strategy, to increase the sampling efficiency and limit the whole computational time within an acceptable range.

IMPLEMENTATION

Much attention has been paid to the programming of SCAMPI, which currently contains about 12 000 lines of C code. Some implementation details are described as follows.

SCAMPI reads multiple mol2 files containing compound structures and a data file containing biological activities as input. The output is linked to a graphic interface program so that the diagram tree can be directly displayed. SCAMPI also outputs a mol2 file for each active terminal node. Each mol2 file contains the active conformational structures for all the compounds dropped into that terminal node, aligned with each other according to the pharmacophore model indicated by that node. The active conformation for each compound is chosen as the one of all the conformations ever generated, which not only satisfies the identified pharmacophore model but also has the lowest internal energy. The user can then use SYBYL to view these possibly active conformations and the related pharmacophore model.

SCAMPI follows the work of Greene et al.[49] to define chemical features. The chemical feature types that are already implemented in SCAMPI are listed in Table 1. The user is allowed to choose any of them when starting the program. To detect these chemical features, two search systems have been coded: a substructure search using functionality fragments as search queries to detect the chemical features defined by a group of atoms, like guanidine, etc., and a rule-based search to detect the chemical features defined by a

Table 1. Chemical Feature Types Implemented in SCAMPI

types	description
negative charge center	carboxylic group, sulfinic group, phosphinic group, etc.[49]
positive charge center	nitrogen in primary, secondary, and tertiary amines, etc.[49]
hydrogen bond acceptor	nitrogen, oxygen, and sulfur with at least 1 available lone pair electron
hydrogen bond donor	nitrogen, oxygen, sulfur, or the terminal of a triple bond linked with hydrogen
aromatic ring center	center of aromatic 5- or 6-membered ring
hydrophobic center	center of fragment containing more than 3 hydrophobic atoms[49]
nitrogen	nitrogen atom
oxygen	oxygen atom
sulfur	sulfur atom
phosphor	phosphorus atom
fluorine	fluorine atom
halogen	chlorine atom, bromine atom, and iodine atom

single atom and its linked neighbors, like tertiary nitrogen, etc. A rule-based search system has the advantage of speed, while a substructure search system can detect complicated functionality groups and also allow the user to include the definitions of new chemical features.

A modified Tripos force field[50] is implemented in SCAMPI for all the molecular mechanics calculations, including the terms of bond, angle, torsion, out-of-plane, and van der Waals interactions. A distance constraint term will also be added between pharmacophoric points during the constrained conformational search. An electrostatic interaction term is not included in order to simplify the calculations.

Memory and speed optimizations are also important for practical use. Thus, a sparse matrix technique is utilized to conserve the memory space, and a hash-table search is used to accelerate the computation.

APPLICATIONS

Monoamine Oxidase (MAO) Inhibitors. Of the 1650 MAO inhibitors provided by Abbott Laboratories,[51] CONCORD[52] successfully converted 1644 compounds from 2D to 3D structures. The structures and activities of these 1644 compounds were used as the input for SCAMPI.

With a single run of SCAMPI, a recursive partition tree was generated, as shown in Figure 4. The computational time was about 5.1 CPU hours on a SGI R10000 machine, and a total of 943 304 conformations were generated. Chemical features of the triple-bond center and carbon were included, and negative splits were allowed. From Figure 4, we can see two major active nodes, N101 and N01, indicated as shaded nodes. The pharmacophore corresponding to node N101 is composed of an aromatic ring center, a triple-bond center, and a positive charge center located on nitrogen, as illustrated in Figure 5. This model is consistent with the experimental evidence that propargylamines (e.g., compound AL19120 in Figure 5) are suicide inhibitors that can irreversibly inhibit MAOs through covalent attachment to its flavin cofactor.[53] Furthermore, there is a negative split on the pathway to node N101, implying that the existence of a side chain at the vicinity of the triple-bond terminus may significantly decrease activity. This is also consistent

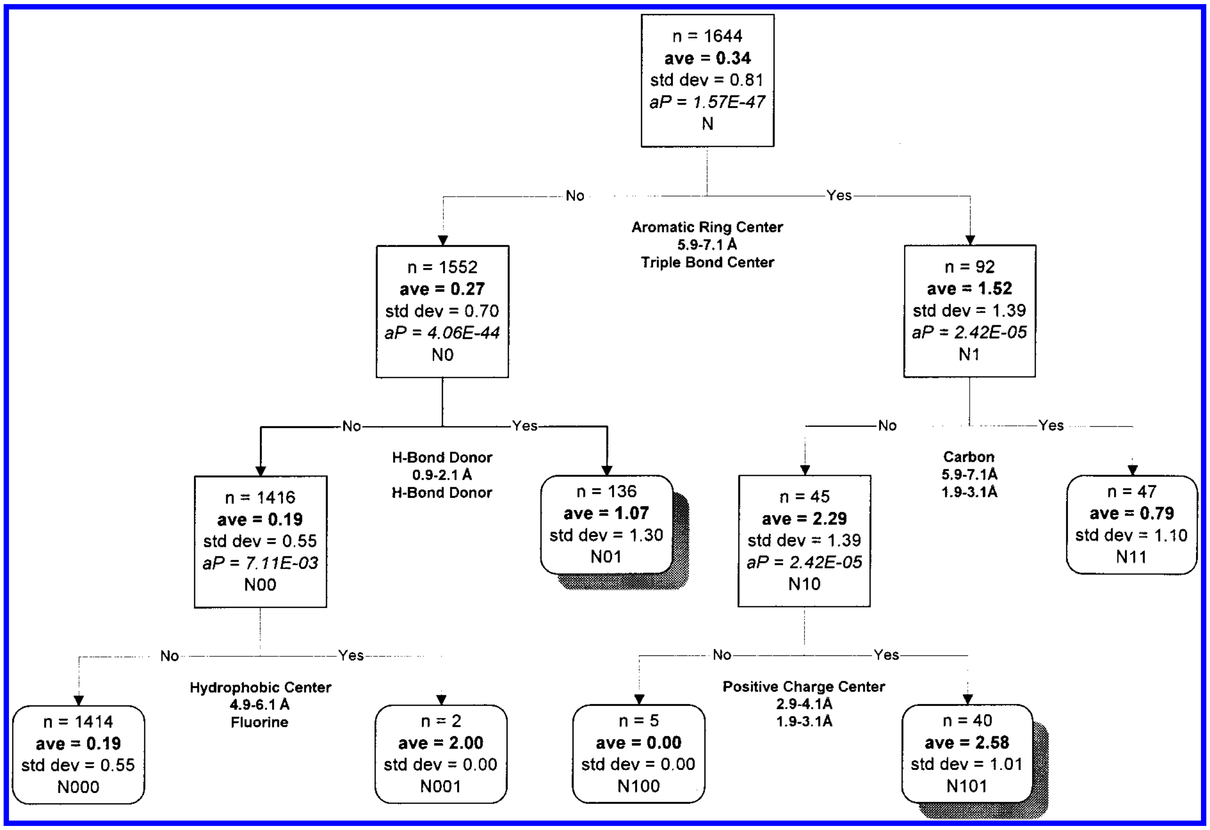

Figure 4. SCAMPI tree for the MAO data set. For each node, "*n*" is the number of compounds, "ave" represents the average potency of the compounds, "std dev" represents the standard deviation of the potencies of the compounds, and "aP" is the Bofferroni-adjusted *p* value for the split. The active nodes are indicated as the shaded nodes. Two-point descriptors are indicated as two chemical features and the distance range between them, three-point descriptors are indicated as a chemical feature and its two distance ranges to the first two pharmacophoric points, and so on.

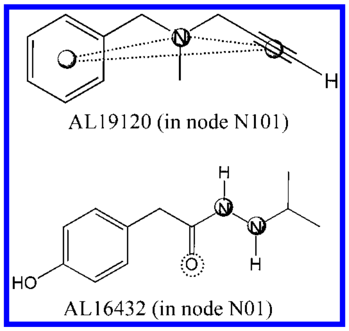

Figure 5. Pharmacophores identified by SCAMPI for the MAO data set.

with the inhibition mechanism that propargylamines use their triple-bond terminus to attack flavin.[53] The pharmacophore corresponding to node N01 contains two hydrogen bond donors located on nitrogen, as illustrated in Figure 5. After checking the compounds in this node, we found another highly correlated carbonyl group at the adjacent position. So it is actually a hydrazide feature. It is known that these hydrazides (e.g., compound AL16432 in Figure 5) can be acetylated at first and then hydrolyzed to hydrazines, which can act as nonselective, irreversible inhibitors to various macromolecules including MAOs.[54]

Therefore, SCAMPI successfully found two different pharmacophores from this particular data set, a clear demonstration that SCAMPI has the capability to detect multiple mechanisms of action or binding modes coexisting in a large chemical data set.

Angiotensin-Converting Enzyme (ACE) Inhibitors. This data set originally contained 114 ACE inhibitors provided by Tripos Inc.[55] One thousand new compounds were randomly picked up from the WDI (World Drug Index) database, and CONCORD[52] successfully converted 932 of them from 2D to 3D structures. These 932 compounds were then added into the original ACE data set, acting as negative compounds. The activities of the 114 ACE inhibitors are expressed as continuous pIC_{50} values, and the WDI compounds are arbitrarily assigned a pIC_{50} value of 0. The structures and activities of these 1046 compounds were used as the input for SCAMPI.

With a single run of SCAMPI, a recursive partition tree was generated, as shown in Figure 6. The computational time was about 8.1 CPU hours on a SGI R10000 machine, and a total of 573 798 conformations were generated. From Figure 6, we can see two major active nodes, also indicated as shaded nodes. The two corresponding pharmacophores are illustrated in Figure 7. The first one contains a negative charge center located on a carboxylate group, a carbonyl oxygen atom, another negative charge center located on another carboxylate group, and a nitrogen atom. The second pharmacophore contains a negative charge center located on a carboxylate group, a carbonyl oxygen atom, and a sulfur atom located on a thiolate group.

A comparison of the two pharmacophores in Figure 7 clearly shows the similarity between their first three points. They share the same geometry and two of the three pharmacophoric features. A literature search indicates that

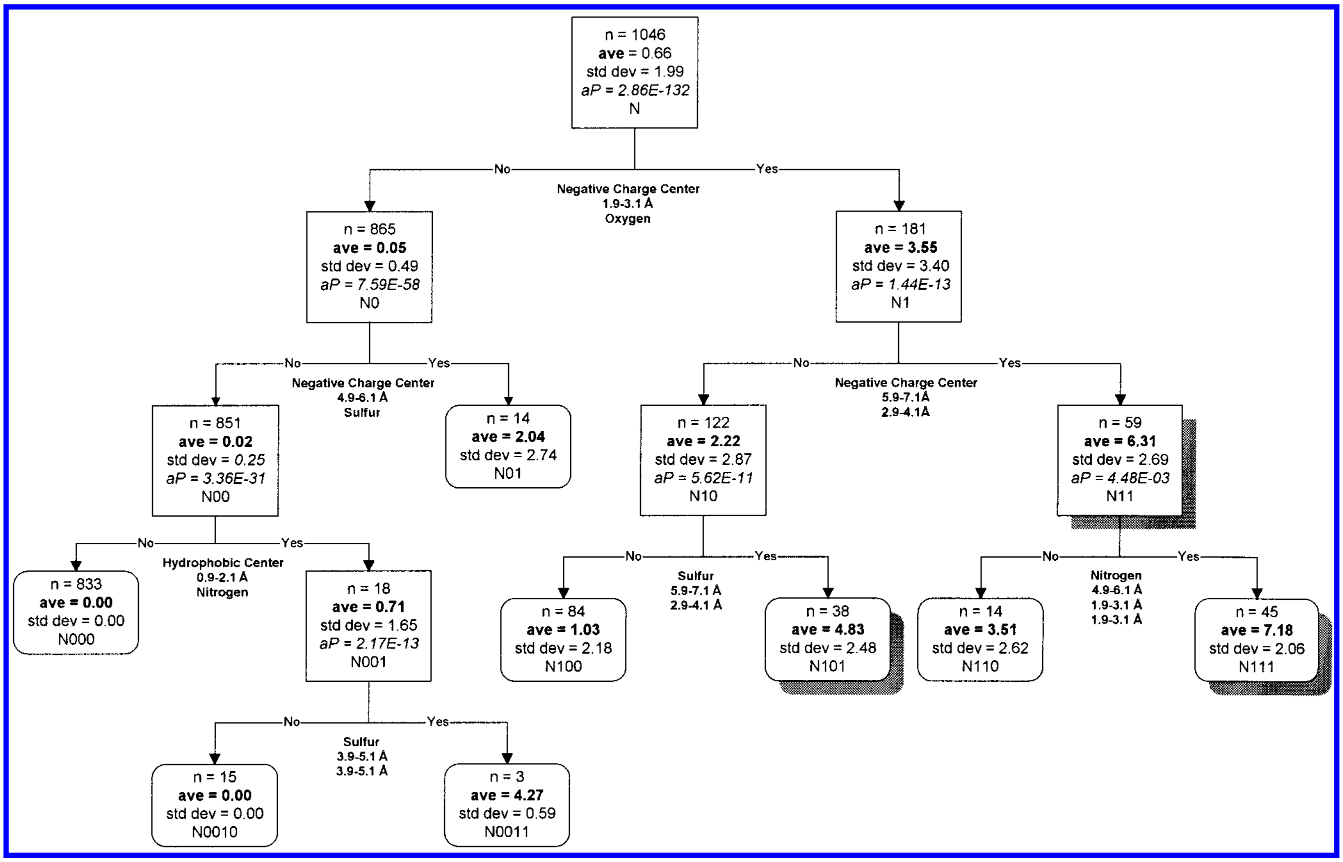

Figure 6. SCAMPI tree for the ACE data set. For each node, “*n*” is the number of compounds, “ave” represents the average potency of the compounds, “std dev” represents standard deviation of the potencies of the compounds, and “aP” is the Bofferroni-adjusted *p* value for the split. The active nodes are indicated as the shaded nodes. Two-point descriptors are indicated as two chemical features and the distance range between them, three-point descriptors are indicated as a chemical feature and its two distance ranges to the first two pharmacophoric points, and so on.

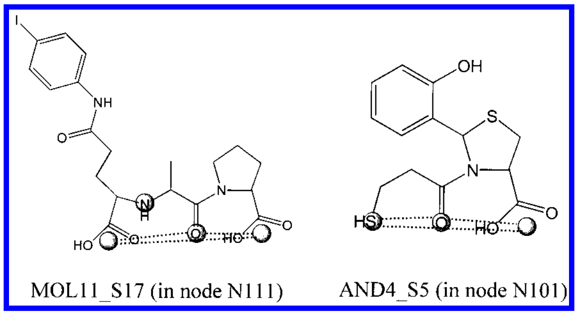

Figure 7. Pharmacophores identified by SCAMPI for the ACE data set.

they actually follow the same binding mode, where the carboxylate and the thiolate groups bind to the same zinc atom in ACE.[8] The commonly acceptable pharmacophore is composed of a negative charge center, a carbonyl oxygen as the hydrogen-bond acceptor, and a zinc binding site. Since we did not define a special chemical feature for zinc binding, SCAMPI has no way to find this feature. Instead, it splits the compounds following this binding mode into two different terminal nodes. The fourth point, a nitrogen atom, in the first pharmacophore is statistically significant, but we have not found any literature supporting its reality.

This data set demonstrates again that SCAMPI can quickly suggest the pharmacophore model is consistent with the known result. It also exemplifies that a careful study of the SCAMPI output is sometimes necessary for capturing the real pharmacophore. In the terminology of data mining, this is called the involvement of domain knowledge.

DISCUSSION

Automated pharmacophore identification is a computationally intensive process where one needs to search both conformational and correspondence spaces. For the compounds that are highly flexible and contain multiple chemical functional groups, a complete search of each space is computationally daunting and their combination essentially impossible. The attempt to exhaust all the possibilities in both the conformational search and the correspondence search for a large chemical data set, which not only contains a large number of compounds but also may include some highly flexible compounds, is almost an impractical endeavor with the current or foreseeable computational capability. Hence, we choose to use less exhaustive but more feasible strategies in the conformational search and pharmacophore buildup to detect pharmacophores within an acceptable computational time. SCAMPI is not designed to find all the possible, high-quality pharmacophore models in a single run of the program; it is designed to quickly examine large, heterogeneous data sets. Some arguments about the SCAMPI methodology need to be discussed further.

First, some active conformations may be missed, since a random search rather than a systematic search has been chosen as the conformational search strategy in SCAMPI. Nevertheless, we believe that the subsequent statistical test can tolerate this degree of missing if it is not too serious, because by using a statistical test we focus on the characteristics of the whole data set, not the individual compounds. This robustness will prevail as long as not too many active compounds have missed their active conformations. If we can sample most of the region in the conformational space of a compound and the possibility of missing a particular conformation is small, then the possibility of simultaneously missing the conformations relevant to a particular pharmacophore by multiple compounds will be even smaller. Using the default set of parameters in SCAMPI, usually hundreds to thousands of conformations will be generated for each compound, depending on its flexibility and complexity; we believe that in most cases this is enough for sampling the conformational space of a drug-size compound.

Second, bump checking may be argued as being too rough a criterion to generate structures of high quality, and therefore, some of the generated conformational structures may have very high internal energies and may not be attainable when binding to the receptor. However, extensive force-field calculation is computationally too expensive for the millions of conformations of thousands of compounds. Thus, we choose to use a coarse but fast conformational search method, so as to limit the computational time within an acceptable range. Furthermore, we believe that these high-energy conformations can also be tolerated by the following statistical test because the activity data will help it pick up the correct active conformations. However, including many high-energy conformations does increase the possibility of a false positive, so we suggest the user check the active conformations suggested by SCAMPI to see if their corresponding pharmacophores are energetically favorable.

Third, it has been advocated that a pharmacophore model should include other kinds of constraints, such as angular constraint, torsional constraint, and excluded volume; presently we consider only distance constraints in SCAMPI, once again for the sake of improving speed. Obviously, distance constraints alone may not be enough to model some complicated pharmacophores (for example, reflecting chirality centers), so we suggest the user look for other constraints after the compounds have been classified into individual terminal nodes.

Fourth, suppose all the compounds in a data set have a common scaffold. It has been argued that a method that looks for structural differences to explain the difference in activities will ignore the common scaffold in assigning the pharmacophore.[20] We consider this to be a limitation of the data set itself, not the descriptors or analysis methods. Recursive partitioning overcomes this problem by using a large, structurally heterogeneous chemical data set. Such a data set can come from a company collection or combinatorial synthesis with a wide range of structures. Therefore, at least at the first split, we have enough diversity of structures and activities to detect the significant pharmacophoric features and configurations. When the data set has been partitioned into small subsets, the structures and activities of compounds in an individual subset become more and more homogeneous and a statistical test may not be able to detect the additional pharmacophoric features and configurations. To avoid this defect, we suggest using the traditional pharmacophore identification methods, for example, DISCO,[14] to analyze the compounds in each active terminal node or just visually checking the structures of those compounds, since the number of compounds in each terminal node is typically small.

Increasing the speed of the conformational search is definitely needed. Presently, it is the rate-limiting step in SCAMPI and costs over 95% of the whole computation time. This improvement can come from methodology development, software programming, etc. Another direction of improvement is multiple splitting or multiple tree generation. SCAMPI uses the greedy search strategy to generate a single tree; only the statistically most significant descriptor is chosen to split the data set. This strategy simplifies the search process but may miss the global optimal tree, i.e., the best pharmacophore model in this case. Therefore, multiple splitting should be tried to examine more possibilities in the pharmacophore search. Furthermore, the detection and interpretation of correlated descriptors also presents an interpretation problem: if there are several perfectly correlated descriptors for a split, only one of them will be picked as the split descriptor. For now, we do not expect that an algorithm can completely solve this problem, as the choice of descriptors and training compounds is also influential. A good graphic interface, specifically designed for the pharmacophore identification, may help domain experts interpret the important correlated features. Finally, an active conformation should not only satisfy the pharmacophore model but also possess the correct conformation of other parts of a compound, so some optimization methods need to be developed for the active conformations output by SCAMPI before they can be used for a traditional 3D QSAR analysis.

With the development of combinatorial chemistry, the availability of large commercial collections, and the speed-up of high-throughput screening, it is obvious that we will have more structure−activity data available, and consequently, data mining will become a more important step in the drug discovery process. Although data mining is not a totally new technique and has been widely applied in many fields, like finance, marketing, credit approval, etc., it is just beginning to be applied in the drug discovery process. Much effort still needs to be done before various data-mining methods can be successfully applied to drug discovery data. Apart from the logistical aspects of data collection, cleaning, and adjustment, how to combine data-mining and computational chemistry techniques is worth much attention. SCAMPI should be viewed as an effort in this direction.

CONCLUSION

According to the published work, SCAMPI is probably the first program that is specifically designed for the pharmacophore identification for large, heterogeneous chemical data sets. It uses the recursive partitioning method to detect the pharmacophore features and configurations and a fast conformational search algorithm to explore the conformational space. The application examples have demonstrated that it is able to identify multiple pharmacophores from a structurally heterogeneous data set containing about 1000−2000 compounds with a computational time of less than 1 day.

ACKNOWLEDGMENT

We acknowledge the contribution of some colleagues at Glaxo Wellcome to this work: Dr. Christophe Lambert gave valuable suggestions on the software development; Dr. James Bentley selected the WDI compounds; Dr. Leping Li provided some C codes for the conformational manipulation. We also thank Abbott Labs for generously supplying the MAO data set and Tripos Inc. for providing the ACE data set.

REFERENCES AND NOTES

(1) The methods reported in this paper are in the patent process.

(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) Sittampalam, G. S.; Kahl, S. D.; Janzen, W. P. High-throughput screening: advances in assay technologies. *Curr. Opin. Chem. Biol.* **1997**, *1*, 384−391.

(5) Silverman, L.; Campbell, R.; Broach, J. R. New assay technologies for high-throughput screening. *Curr. Opin. Chem. Biol.* **1998**, *2*, 397−403.

(6) Humblet, C.; Marshall, G. R. Pharmacophore identification and receptor mapping. *Annu. Rep. Med. Chem.* **1980**, *15*, 267−276.

(7) Wang, S.; Zaharevitz, D. W.; Sharma, R.; Marquez, V. E.; Lewin, N. E.; Du, L.; Blumberg, P. M.; Milne, G. W. A. The discovery of novel, structurally diverse protein kinase C agonists through computer 3D-database pharmacophore search. Molecular modeling studies. *J. Med. Chem.* **1994**, *37*, 4479−4489.

(8) DePriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 3D-QSAR of angiotensin-converting enzyme and thermolysin inhibitors: a comparison of CoMFA models based on deduced and experimentally determined active site geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372−5384.

(9) Good, A. C.; Mason, J. S. Three-dimensional structure database searches. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1996; Vol. 7, pp 67−117.

(10) Marshal, G. R.; Barry, C. D.; Bosshard, H. E.; Dammkoehler, R. A.; Dunn, D. A. In *Computer-Assisted Drug Design*; ACS Symposium Series 112; American Chemical Society: Washington, DC, 1979; pp 205−226.

(11) Motoc, I.; Dammkoehler, R. A.; Marshall, G. R. A three-dimensional structure−activity relatioships and biological receptor mapping. In *Mathematics and Computational Concepts in Chemistry*; Ellis Horwood: Chichester, 1985; pp 222−251.

(12) Mayer, D.; Naylor, C. B.; Motoc, I.; Marshall, G. R. A unique geometry of the active site of angiotensin-converting-enzyme consistent with structure−activity studies. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 3−16.

(13) Sheridan, R. P.; Nilakantan, R.; Dixon, J. S.; Venkataraghavan, R. The ensemble approach to distance geometry: application to the nicotinic pharmacophore. *J. Med. Chem.* **1986**, *29*, 899−906.

(14) 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 dopaminergic and benzodiazepine agonists. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83−102.

(15) Catalyst/Hypo Tutorial, version 2.0; BioCAD Corp.: Moutain View, CA, 1993.

(16) Sprague, P. W. Automated chemical hypothesis generation and database searching with Catalyst. *Perspect. Drug Discov. Des.* **1995**, *3*, 1−20.

(17) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of common functional configurations among molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 563−571.

(18) HipHop Turorial, version 2.3; Molecular Simulation Inc.: Sunnyvale, CA, 1995.

(19) Davies, K.; Upton, R. 3D pharmacophore searching. *Net. Sci.* (http://www.netsci.org/Science/Cheminform/feature02.html).

(20) Golender, V.; Vesterman, B. APEX-3D expert system for drug design. *Net. Sci.* (http://www.awod.com/netsci/Science/Compchem/feature09.html).

(21) Van Drie, J. H. Strategies for the determination of pharmacophoric 3D database queries. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 39−52.

(22) Van Drie, J. H.; Nugent, R. A. Addressing the challenges posed by combinatorial chemistry: 3D databases, pharmacophore recognition and beyond. *SAR QSAR Environ. Res.* **1998**, *9*, 1−21.

(23) Finn, P.; Muggleton, S.; Page, D.; Srinivasan, A. Pharmacophore discovery using the inductive logic programming progol. In *Machine Learning, Special Issue on Applications and Knowledge Discovery*; Kluwer Academic Publishers: Boston, 1998; pp 1−33.

(24) Jain, A. N.; Dietterich, T. G.; Lathrop, R. H.; Chapman, D.; Critchlow, R. E., Jr.; Bauer, B. E.; Webster, T. A.; Lozano-Perez, T. Compass: a shape-based machine learning tool for drug design. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 635−652.

(25) Bravi, G.; Gancia, E.; Zaliani, A.; Pegna, M. SONHICA (Simple Optimized Non-Hierarchical Cluster Analysis): a new tool for analysis of molecular conformations. *J. Comput. Chem.* **1997**, *18*, 1295−1311.

(26) Smellie, A.; Teig, S. L.; Towbin, P. Poling: promoting conformational variation. *J. Comput. Chem.* **1994**, *16*, 171−187.

(27) Hurst, T. Flexible 3D searching: The directed tweak technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190−196.

(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) Dammkoehler, R. A.; Karasek, S. F.; Shands, E. F. B.; Marshall, G. R. Constrained search of conformational hyperspace. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 3−21.

(30) Dammkoehler, R. A.; Karasek, S. F.; Shands, E. F. B.; Marshall, G. R. Sampling conformational hyperspace: techniques for improving completeness. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 491−499.

(31) Motoc, I.; Dammkoehler, R. A.; Marshall, G. R. In *Mathematics and Computational Concepts in Chemistry*; Trinajstic, N., Ed.; Ellis Horwood: Chichester, 1986; pp 222−251.

(32) Motoc, I.; Dammkoehler, R. A.; Mayer, D.; Labanowski, J. *Quant. Struct.-Act. Relat.* **1986**, *5*, 99−105.

(33) Beusen, D. D.; Shands, E. F. B. Systematic search strategies in conformational analysis. *Drug Discov. Today* **1996**, *1*, 429−437.

(34) As illustrated in Figure 1, the minimum accessible distances between atoms *i* and *j* can be easily calculated as the distance between them when the atoms *i*, *s*, *r*, and *j* are in the same plane and form a syn configuration.

(35) Leach, A. R. A survey of methods for searching the conformational space of small and medium-sized molecules. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1991; Vol 2, pp 1−55.

(36) Brint, A. T.; Willett, P. Algorithms for the identification of three-dimensional maximal common substructures. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 152−158.

(37) Young S. S.; Hawkins D. M. Analysis of a 2^9 full factorial chemical library. *J. Med. Chem.* **1995**, *38*, 2784−2788.

(38) 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.

(39) Young, S. S.; Hawkins, D. M. Using recursive partitioning to analyze a large SAR data set. *SAR QSAR in Environ. Res.* **1998**, *8*, 183−193.

(40) Chen, X.; Rusinko, A.; Young, S. S. Recursive partitioning analysis of a large structure−activity data set using three-dimensional descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1054−1062.

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

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

(43) Since each compound may be represented by multiple conformations, we assign the absence of a particular molecular descriptor to a compound if none of its representative conformations contains it; otherwise, we assign the presence of that molecular descriptor to that compound.

(44) Miller, R. G. *Simultaneous Statistical Inference*; Springer-verlag, New York, 1981.

(45) The width of each distance bin is set as 1.0 Å as the default. The adjacent bins have 20% overlap with each other so that the actual width of each distance bin is 1.2 Å. Any distance located in the overlap region is assigned to both bins. All the distances longer than 20 Å are assigned to the last bin. Thus, 0.0−1.1 Å is bin no. 0, 0.9−2.1 Å is bin no. 1, 1.9−3.1 Å is bin no. 2, ... and 19.9−∞ Å is bin no. 20. The introduction of this "fuzzy distance" concept is to alleviate the unfavorable boundary effects of the distance bins.

(46) Ullmann, J. R. An algorithm for subgraph isomorphism. *J. Assoc. Comput. Mach.* **1976**, *23*, 31−42.

(47) Brint, A. T.; Willett, P. Pharmacophoric pattern matching in files of 3D chemical structures: comparison of geometric searching algorithms. *J. Mol. Graphics* **1987**, *5*, 49−56.

(48) Thompson, S. K. *Sampling*; Wiley: New York, 1992.

(49) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical function queries for 3D database search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.

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

(51) 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. Comput. Sci.* **1996**, *36*, 572−584.

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

(53) 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.

(54) 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.

(55) *Molecular diversity manager generates lead followup synthesis candidates*; Tripos Inc.: St. Louis, MO, 1995.

(56) Flexible ring corner is defined as the nonfused atom in a flexible ring. After a flexible ring is isolated, its average plane is determined. Then, the spatial position of each flexible corner is changed by randomly assigning a displacement with the length between 0 and 2 Å and the direction pointed to the opposite side of the average ring plane. The chiralty of each corner atom is recorded and monitored to make sure the correct chiralty is kept.

(57) This release strategy is similar to the 1D minimization strategy used by Smellie et al. (ref 58), except that we use the same constrained search algorithm as described in the text to do the bond rotation.

(58) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of conformational coverage. 1. validation and estimation of coverage. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 285−294.

CI990327N