A New Approach to Finding Natural Chemical Structure Classes

Jun Xu[†]

Boehringer Ingelheim Pharmaceuticals, Inc., 900 Ridgebury Road, Ridgefield, Connecticut 06877-0368

Received November 8, 2001

In modern drug discovery, large compound libraries need to be compared and the diversity of compound libraries needs to be analyzed. Classification algorithms are important tools for accomplishing these tasks. In this paper, a chemical structural scaffold based classification approach is reported. The goals of the approach are to find natural structure families from a large (millions of entries) compound library within a feasible time period and to view the library in two-dimensional data space using chemically meaningful methods.

Introduction

To design general and focused combinatorial libraries for the purpose of identifying new lead scaffolds, one needs to analyze the chemical diversity of large, real, or virtual libraries of compounds (typically hundreds of thousands or millions). This is a difficult and controversial task. To evaluate a chemical compound library, the following questions are raised: (1) How diverse is the library? (2) How are the structures distributed in the chemical space? (3) What are the structural differences between a compound library and a recognized drug? (4) What are the structural differences between a combinatorial library and the libraries in the compound inventory? (5) Which compounds are structurally closest to existing drug candidates or leads?

To answer these questions, one can either classify a library to analyze the number of compound classes and their distributions or map the compounds of a library to two- or three-dimensional space in order to visually review how the compounds are grouped.

Compound Classification. The term cluster analysis (CA) was first used by Tryon in 1939.[1] The term actually encompasses a number of different classification algorithms. Since its introduction, many CA algorithms have been reported. They belong to two categories: hierarchical clustering and partitional (nonhierarchical) clustering. Hierarchical clustering rearranges objects in a tree structure. A nonhierarchical cluster algorithm, the Javis−Patrick algorithm (also known as nearest-neighbor cluster algorithm), is commonly used to cluster chemical structures.[2] Conventionally, approaches using topology-based compound classifications involve the following steps: (1) computing descriptors from connection tables, (2) selecting principal components from the descriptors by means of principal component analysis (PCA)[3] or factor analysis (FA),[4] (3) normalizing the principal descriptors so that they are comparable in value, (4) selecting a similarity or distance measurement to compute the similarity or distance between two structures, and (5) choosing a cluster algorithm to group structures. Willett has published a very interesting review of this area.[5] Taraviras, Ivanciuc, and Cabrol-Bass have recently reported a comparison of the most recently available structure clustering methods.[6]

Without the correct input, clustering algorithms cannot correctly locate natural structural families. The clustering results depend on many parameters, such as descriptor selection, data normalization, similarity metrics, etc. Hierarchical clustering algorithms do not reveal the number of structural families and how they are distributed until a hierarchical threshold is chosen. However, the selection process for choosing the threshold is not rationally determined. Nonhierarchical clustering methods, such as *K*-mean and *K*-nearest-neighbor algorithms, produce even more arbitrary results. *K*-mean clustering algorithms assume the user knows the number of the clusters before clustering. This obviously is not the case. Since the computing complexity of a *K*-mean algorithm is factorial to the number of data points, the number of computing iterations has to be limited to make the computation viable. However, the condition for potentially unlimited iterations is inherent in the fact that the user must specify the number of clusters before clustering. *K*-nearest-neighbor algorithms may produce better results in comparison with *K*-mean algorithms, but it still asks the user to choose the number of nearest neighbors. This is typically unknown to the user. Therefore, the results of *K*-nearest-neighbor algorithms are still rather arbitrary.

Mapping a Compound Library. Using the concept that a picture is worth a thousand words, we will approach the understanding of library diversity by representing the library as a picture. To map a compound library to a two-dimensional graph, multidimensional scaling (MDS)[7] and neural network (NN)[8] approaches can be used. A structure can be represented as an array of structural descriptors (numbers). If there are n ($n > 3$) descriptors, then we say the structure is represented as a point of n-dimensional space. MDS is an approach for projecting a point of n-dimensional space to two-dimensional space, which we can then view.

MDS is not so much an exact procedure as rather a way to “rearrange” objects in an efficient manner. It allows one to arrive at a configuration that best approximates the observed distances. It actually moves objects around in the space defined by the requested

[†] Current address: Discovery Partners International, Inc., 9640 Towne Centre Drive, San Diego, CA 92121. E-mail: junxu@DiscoveryPartners.com. Phone: 858-455-8600. Fax: 858-546-3081.

10.1021/jm010520k CCC: $22.00
Published on Web 10/23/2002

number of dimensions and checks how well the distances between objects can be reproduced by the new configuration. In other words, MDS uses a function minimization algorithm that evaluates different configurations with the goal of maximizing the goodness of fit (or minimizing “lack of fit”).[9]

As an NN mapping method, the self-organizing map (SOM) is effectively a vector algorithm that has been quantized. It creates reference vectors in a high-dimensional input space and uses them to approximate the input patterns in an image space in an ordered fashion. It does this by defining local order relations between the reference vectors so that they are made to depend on each other as though their neighboring values would lie along a hypothetical “elastic surface”.[10,11] The SOM, by preserving local features, is therefore able to approximate the point density function $p(x)$ of a complex, high-dimensional input space, so it is presented as two dimensions.

As a simple method, we can use PCA to represent a library with a two-dimensional graph without losing too much information. We can select two principal descriptors from n ($n > 3$) descriptors. This requires that the two principal components can explain at least 85% of the data. It is normally hard to reach such a criterion.

These dimension reduction approaches do not always work well. To validate the dimension reduction results, we need a technology to permit us to map a graphed point to its structure drawing. This process involves chemical structure related data visualization technology, which is offered commercially.[12] However, even if we use this technology and can make some progress in solving the problems of the dimension reduction approach, that approach is still lacking. The mapped dimensions have no chemical meaning. Therefore, bench chemists cannot understand the new dimensions.

In recent years, new techniques for chemical structure diversity analysis continue to be reported.[13−15] It is expected that more novel and powerful approaches will emerge in coming years. In this paper, we propose a scaffold-based classification approach (SCA) in order to classify compound libraries. It is based on compound topological scaffolds. In addition, we propose new ways to graph the chemical diversity of compound libraries in a chemically understandable manner. By means of SCA, it is possible to find natural chemical structural families. The relationships between identified structural families can be viewed in two- or three-dimensional space, and each dimension can have a chemical explanation. For example, on the basis of the result of SCA, a two-dimensional chemical space can be defined by “structural complexity” as the X axis and by “cyclicity” as the Y axis. Compounds that have larger numbers of atoms and bonds will have greater values of structural complexity. In the same structural class, compounds that have a larger number of ring bonds will have greater values of cyclicity.

Scaffold-Based Classification Approach (SCA)

Most conventional approaches for clustering chemical compounds are based on structural descriptors. The SCA does not use structural descriptors to classify compounds. It groups compounds into the same class if they share the same topological scaffold or class center. But what is a topological scaffold? Before defining a topological scaffold, we introduce the concepts of “ring bond” and “linker bond” with the following definitions: (1) A ring bond is a bond in which the two atoms in the bond are both in the same ring. (2) A linker bond is a bond that is not itself a ring bond. However, both of its two atoms, directly or indirectly, connect to a ring or rings. (3) A chain bond is a bond that is neither ring bond nor linker bond. On the basis of these definitions, a topological scaffold is defined as follows: (4) A topological scaffold is a structure that has ring bonds and linker bonds but has no chain bonds.

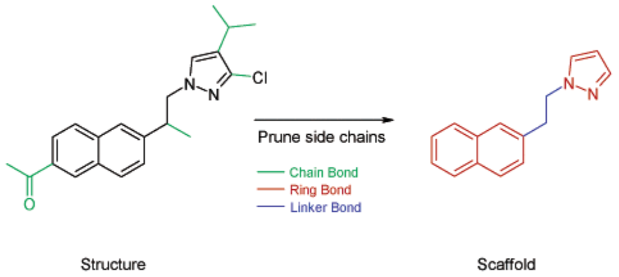


Figure 1. Deriving a structural scaffold.

Figure 1 illustrates the definition of a topological scaffold.

Definition 4 permits a computer to consistently derive scaffolds from chemical structures. However, it should be noted that not all chemists would agree with the specified definition.

The above definitions cover all cyclic compounds. However, the SCA must treat acyclic compounds in special ways as follows: (1) For saturated acyclic compounds, their scaffold is Q−Q (where Q is any heteroatom). If there are no heteroatoms, then their scaffold is C−C. (2) For unsaturated acyclic compounds, their double bonds and triple bonds are considered as ring bonds. Table 1 lists the examples of these special treatments.

Unsaturated bonds (normally double bonds) that directly connect to a ring system are recognized as parts of that ring system, since they change the chemical behavior of the ring system. An example is shown in Figure 2.

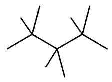
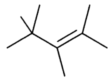
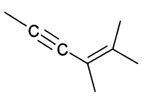
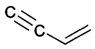
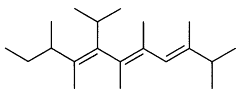
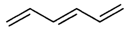
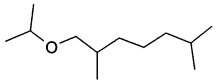
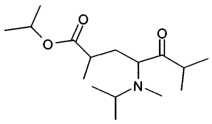
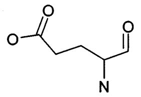
Classification Processes. The SCA classifies compound libraries in the following steps.

Step 1. It finds all nonredundant scaffolds. The scaffold for each structure is derived by pruning all existing side chains. Hydrogen atoms are not recognized as side chains. If the resulting scaffold is not in the scaffold list, then it is appended to the scaffold listing.

Step 2. It sorts the scaffold list. The scaffolds are sorted in ascending order of structural complexity. The structural complexity is computed as follows:

(1) A reference vector $\mathbf{V}_v$ is found in order to calculate the structural complexity for every scaffold. $\mathbf{V}_v$ is the virtual scaffold vector. It consists of four structural descriptors: (1) the maximum number of the smallest set of smallest rings (sssrs), (2) the maximum number of heavy atoms, (3) the maximum number of bonds, where covalent bonds between hydrogen atoms and other atoms are excluded and, (4) the maximum sum of heavy atomic numbers.

Table 1. Acyclic Compound and Their Sacffolds

No.	Acyclic Compound	Scaffold
1		C—C
2		C=C
3		
4		
5		O—C
6		

(2) These four structural descriptors are computed for every structure "*i*" and stored in vector $\mathbf{V}_i$.

(3) The complexity for scaffold $\mathbf{S}_i$ is calculated as

$$\text{complexity}(\mathbf{S}_i) = \frac{||\mathbf{V}_i + \mathbf{V}_\text{v}|| - ||\mathbf{V}_i - \mathbf{V}_\text{v}||}{||\mathbf{V}_i + \mathbf{V}_\text{v}||} \qquad (1)$$

(4) The scaffolds are then sorted in ascending order by complexity. After the order position of a scaffold in the scaffold list is sorted, the position is specified as its class ID. A scaffold represents a topological class center. Therefore, a scaffold ID is also a class ID (CID) that is associated with its complexity. Scaffold $i + 1$ is structurally more complicated than scaffold i. Complexity differences will thus also show structural differences.

Step 3. It classifies structures. Every structure will have its class ID, which is its scaffold (class center) number or position in the scaffold list. The similarity between the structure and its scaffold is the class membership of the structure. Since the differences between a structure and its scaffold are due to side chains, then those structures with fewer side chains have higher membership values. Therefore, membership is defined as "cyclicity".

Membership is based on the following structural descriptors: the sum of heavy atomic numbers (a), the number of rotating bonds (r), the number of 1° nodes (d1), the number of double bonds (db), the number of triple bonds (tb), the number of 2° nodes (d1). If $\mathbf{S}_\text{c} = \{a, r, \text{d1}, \text{db}, \text{tb}, \text{dl}\}$ represents a scaffold, and $\mathbf{S}_i = \{a', r', \text{d1}', \text{db}', \text{tb}', \text{dl}'\}$ represents the structure belonging to that scaffold, then membership of the structure is calculated as follows:

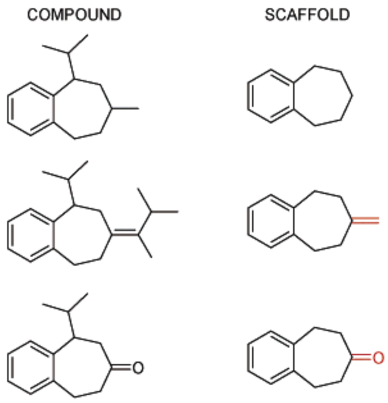


Figure 2. Compound and scaffold. When a double bond is directly attached to a ring, it is recognized as the part of the ring and it is not recognized as a chain bond.

$$\text{membership}(\mathbf{S}_{i \to c}) = \frac{||\mathbf{S}_i + \mathbf{S}_\text{c}|| - ||\mathbf{S}_i - \mathbf{S}_\text{c}||}{||\mathbf{S}_i + \mathbf{S}_\text{c}||} \qquad (2)$$

Note that steps 1 and 3 require a structure match algorithm. There are four options to match structures: (1) topological match, where no atom types and bond types are recognized; (2) bond-topological match, where only bond types are recognized; (3) atom-topological match, where only atom types are recognized; (4) chromatic match, where both atom and bond types are recognized.

These options control the classification results. Option 1 will yield the smallest number of classes because it ignores atom types and bond types. Option 1 will group benzene, hexane, and pyridine in the same class. Only option 4 can recognize these as different scaffolds.

Results

The SCA has been implemented in C/C++ on Windows 95, Windows NT, and UNIX platforms. The executable program is callable by ISIS/Base through ISIS/PL. Using the SCA, we have classified compounds selected from the four following databases: ACD (250 468 structures); NCI (126 554, MDL 1994); CMC (only 4591 oral drugs have been taken into account); and MDDR (only 6347 launched or preclinical compounds have been taken into account). Note that many compounds that are collected in the CMC and MDDR databases are excluded because they are nonoral drugs. These include compounds such as radiopaque agents, imaging agents, dental resins, veterinary compounds, sweeteners, and peptides or proteins. It requires 1 h 42 min to classify all 387 960 structures on an NT laptop (Compaq, Armada E700).

A total of 57 186 classes were found from ACD, NCI, CMC, and MDDR databases. The class centers for computing scaffold complexities are as follows: (1) maximum number of smallest set of smallest rings is 33; (2) maximum number of non-hydrogen atoms is 183; (3) maximum number of non-hydrogen-involved bonds is 206; (4) maximum sum of atomic order numbers is 1247. The classification results are depicted in Figure 3. They are based on using Spotfire V5.0[20]

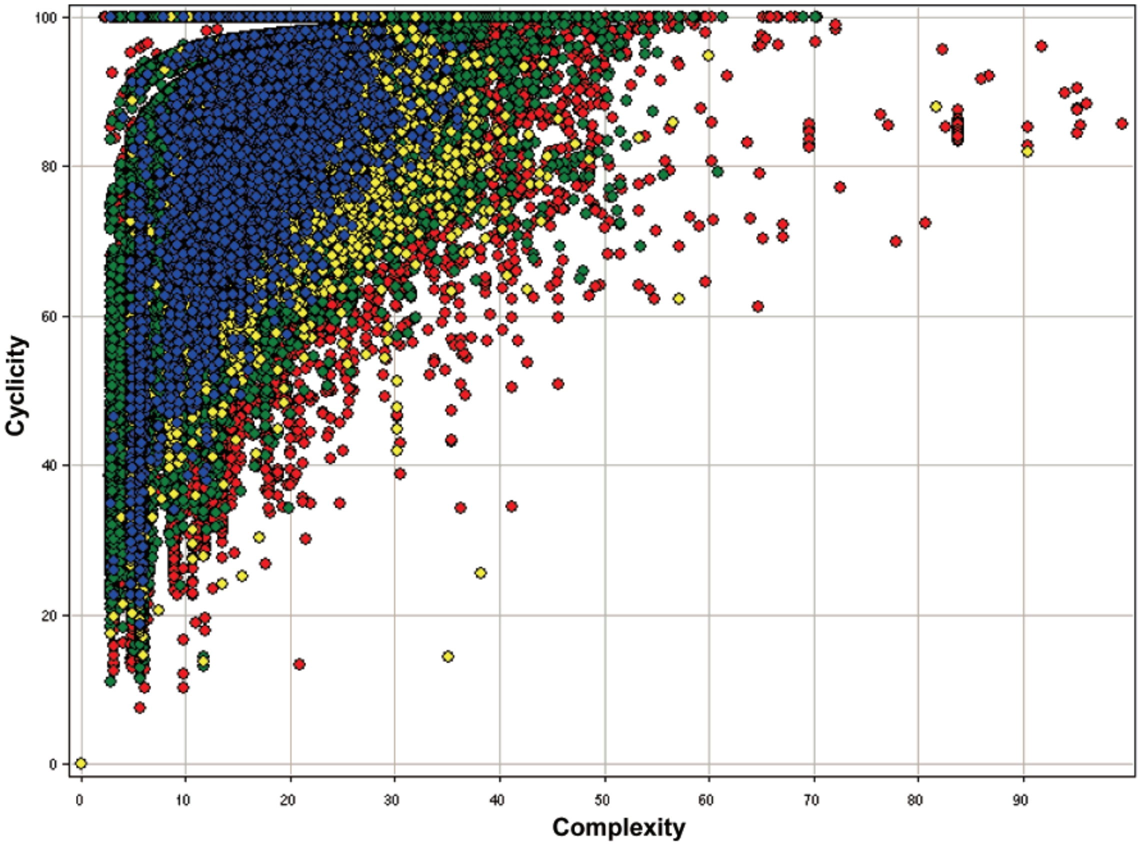


Figure 3. SCA diversity map. The SCA diversity map can be used to compare ACD (red), NCI (green), MDDR (yellow), and CMC (blue) databases. It is easy to see that orally active drugs are distributed in a narrower region compared with other compound libraries.

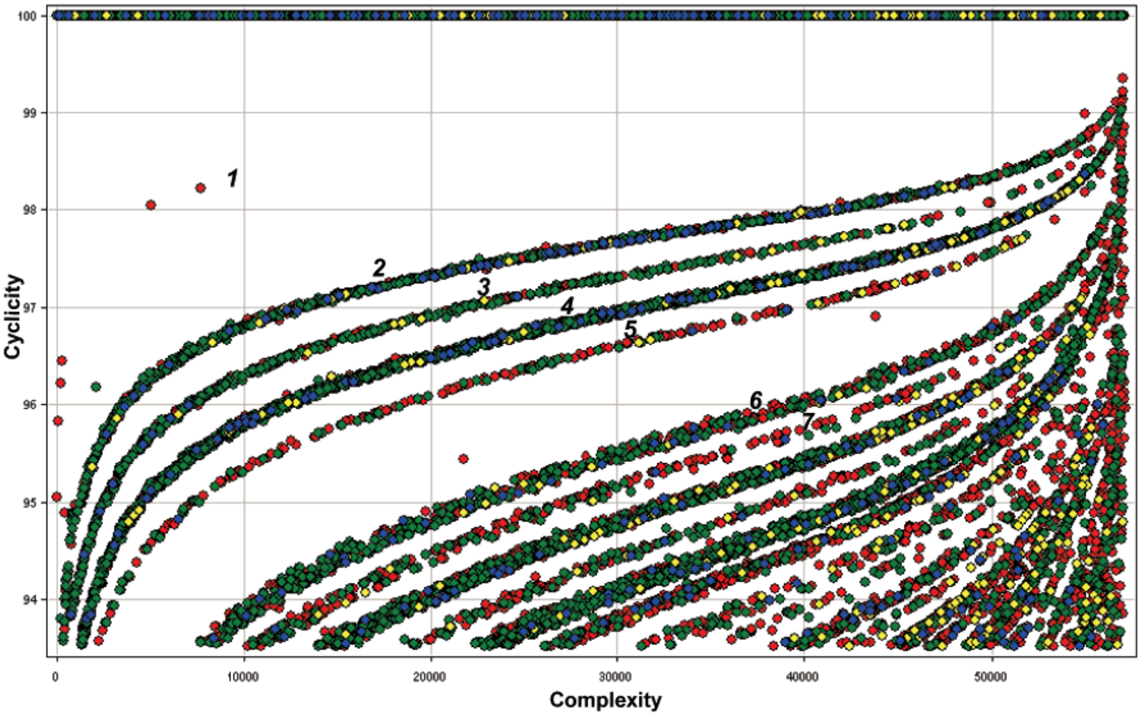


Figure 4. Chemical diversity patterns in the SCA map.

In Figure 3, the *X* axis represents the complexity and the *Y* axis represents the class membership. Because the greater values of a membership correspond to all higher ring bond rates, we assign the *Y* axis the label as cyclicity. Both the *X* and *Y* axes are normalized to 100%. Blue dots represent CMC compounds, yellow dots represent MDDR compounds, green dots represent NCI compounds, and red dots represent ACD compounds. It is clear that chemical diversity increases among the databases in the order CMC < MDDR < NCI < ACD.

In Figure 3, the most complicated compound is an ACD compound that is located in the upper-right corner as circled. It is actually a DNA molecule.

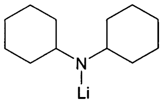
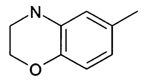
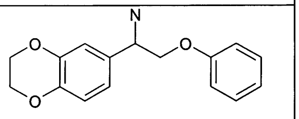
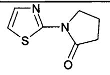
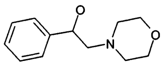
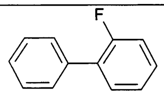
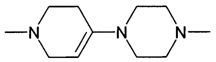
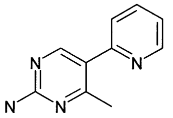
The SCA class map has four corner regions that represent four classes of compounds as follows: (1) the A (upper left) corner represents the compounds with simpler rings and fewer side chains; (2) the B (bottom left) corner represents the compounds with simpler rings but longer and more complicated side chains; (3) the C (upper right) corner represents the compounds with more complicated ring systems but fewer complicated side chains; (4) the D (bottom right) corner represents the compounds with more complicated ring systems and more complicated side chains.

The D corner is almost always empty because it represents compounds with the most complicated ring systems and many side chains. These are very hard to synthesize. This region is probably not very interesting. Most "chainlike" compounds are located around the B corner.

As might be expected, drugs are compounds with moderately complicated rings and side chains. At the top of Figure 3, the compounds represented by a straight line (cyclicity of 100%) are pure scaffolds. These compounds have no side chains. There are a number of drugs reported in the CMC database that are pure scaffolds.

Since the scaffolds are sorted by ascending complexity, the compound class ID (CID) that is also associated with complexity is in ascending order. In other words, the greater a CID is, the greater the scaffold complexity will be. When the complexity is replaced with CID on the *X* axis and the upper half of Figure 3 is closely examined, a number of interesting curves emerge, as shown in Figure 4. These curves are designated as chemical diversity patterns. To understand why these curves are formed, seven of the curves of Figure 4, as marked, are examined closely. The results are presented in Table 2.

Table 2. Some Chemical Diversity Patterns (Examples Are from ACD Database)

Pattern #	Meaning	Example
1	Single Li substituted compounds	
2	Single methyl substituted compounds, some times ethylene-substituted compounds	
3	Single primary amine compounds	
4	Single carbonyl group or alcohol group compounds	 
5	Single fluoride substituted compounds	
6	Double methyl substituted or single ethyl substituted compounds	
7	Single primary amine and single methyl substituted compounds	

It is clear that each curve represents a set of compounds with a common side chain or side chains but different scaffolds. By reviewing the chemical diversity pattern map, we can conclude the following.

(1) Chemical diversity space is discrete, not continuous. Therefore, some "diversity holes" will never be filled and should not cause concern.

(2) Some diversity patterns, such as curve 1 in Figure 4, may not be of interest to medicinal chemists because they are not "druglike".

(3) Some diversity patterns, such as curves 2−4 in Figure 4, are druglike because many CMC database drugs are aligned on these curves (colored by blue). For scaffolds with missing compounds on these curves, medicinal chemists are advised to add corresponding side chains or scaffolds to reconnect the broken diversity patterns. A greater opportunity to discover new drugs can thus be provided.

(4) The diversity patterns become broader at lower cyclicity values because there are more ways to add side chains to a scaffold. For example, if we have two carbon atoms and one oxygen atom to add to a scaffold, we will have eight different ways of implementation: ({∼C,∼C,∼O}, {∼C−C,∼O}, {∼C−O,∼C}, {∼O−C,∼C}, {∼C−C−O}, {∼C(C)−O}, {∼O−C−C}, {∼C−O−C}). To simplify this example, we have not considered the positions on a scaffold and the bonding types. As we see, however, fuzzier diversity patterns mean that the chemistry in that region is getting more complicated.

For a smaller compound library (less than a few thousand compounds), the diversity patterns are not presented clearly because the number of scaffolds is too small to provide continuous curves. However, if the diversity patterns of a large library (over 10 000 compounds) are not apparent, the library is either very biased or not diverse.

The diversity pattern maps can also show structural family distribution. In the CMC database, one of the largest classes is the benzene-scaffold cluster. This cluster has 459 compounds; most of them are anesthetic, antiadrenergic (β-receptor), bronchodilator, anorexic, antineoplastic, adrenergic, analgesic, and antihypertensive drugs. Another large class contains steroid compounds (see Figure 5).

By examination of the diversity patterns (see Figure 6), it can be seen that many oral drugs in the CMC database have a single hydroxyl, carbonyl, or methyl group while the oral drugs that have a single primary amine or single fluoride group are relatively rare.

A compound library can be mapped in many different ways depending on the focus of the scientists. Synthetic chemists typically prefer scaffold-based mapping because they are considering how to make compounds. Scientists who are more involved in drug design may want to see a map that is based on other structural

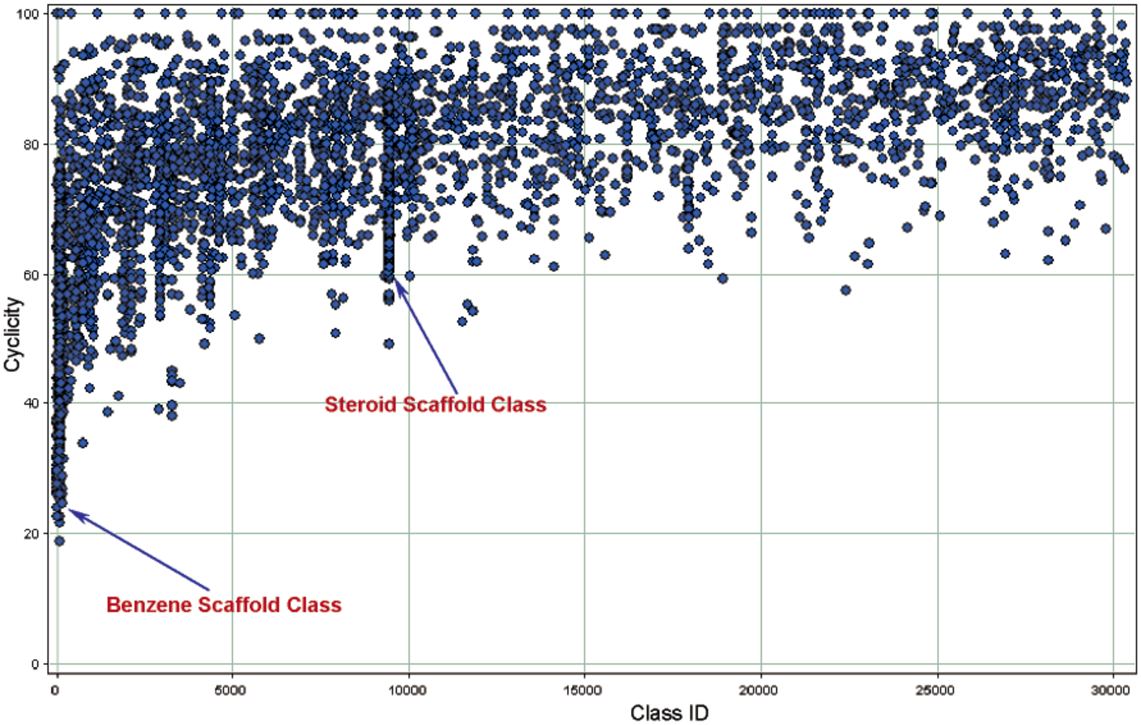


Figure 5. Larger scaffold families in CMC database. Larger scaffold families in CMC database are benzene family and steroid families.

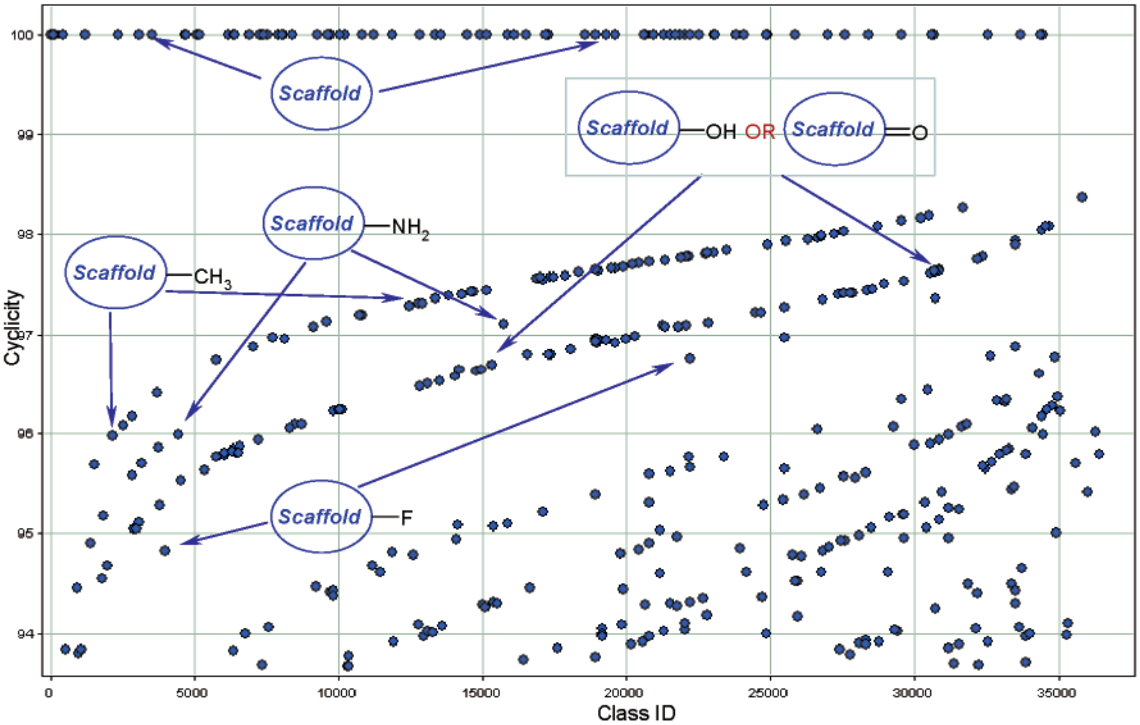


Figure 6. Diversity patterns of CMC oral drugs.

descriptors. For this reason, the SCA also outputs the following structural descriptor values: (1) AE, average electronegativity; (2) HD, number of H bond donors; (3) HA, number of H bond acceptors; (4) AB, number of aromatic bonds; (5) ATMS, number of non-H atoms; (6) BNDS, number of non-H-involved bonds; (7) SSSRS, number of smallest set of smallest rings; (8) AZ, average atomic numbers; (9) RB, number of rotating bonds

Normally, we keep cyclicity as the *Y* axis and switch descriptors for the *X*axis in order to view the different types of diversity patterns. By comparing a proposed library against druglike libraries, such as the CMC and MDDR oral drug libraries, we are able to visually filter out non-drug-like compounds that fall outside the druglike range for a specific descriptor. For example, AE in most CMC and MDDR oral drugs ranges from 2.6 to 3.0 while AE in NCI compounds ranges between 2.41 and 3.46. However, the majority of NCI compounds have an AE range between 2.55 and 3.02. Therefore, we understand that the compounds with AE values out of

Table 3. ACD Compounds with Minimum and Maximum AE Values

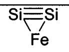
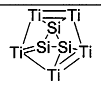
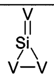
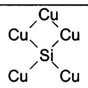
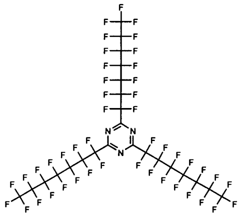
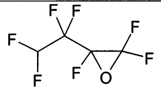
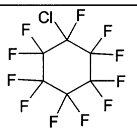
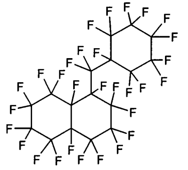
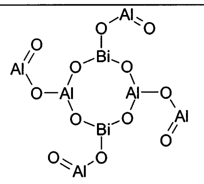
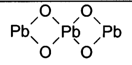
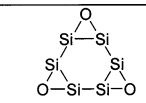
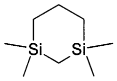
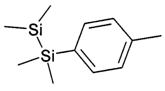
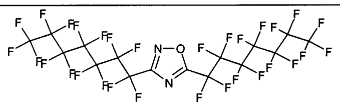
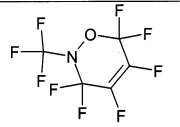
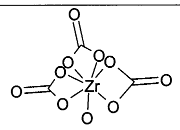
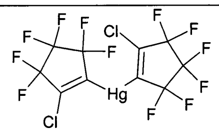
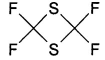
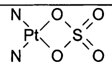
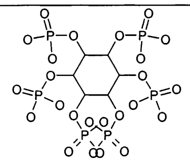
AE Value	Compound
1.9	
1.9	
1.9	
1.9	
3.46	
3.46	
3.46	
3.46	
3.44	
3.44	

Table 4. NCI Compounds with AE Values beyond Range of 2.55−3.02

AE Value	Compound
2.41	
2.42	
2.45	
3.46	
3.44	
3.23	
3.35	
3.27	
3.20	
3.08	

this range are non-drug-like compounds. To validate this observation, we examined the compounds that are out of the druglike AE range and list the examples in Tables 3 and 4. Obviously, the AE descriptor does have the capacity to discriminate drug-like compounds from non-drug-like compounds.

Other descriptors that we use to filter out non-drug-like compounds through the diversity map are the number of hydrogen bond donors and the number of hydrogen bond acceptors. Examples are shown in Figures 7 and 8. As expected, most CMC database compounds (blue) have less than five hydrogen donors. However, many compounds in the ACD database have more than five hydrogen donors. These ACD compounds are sugar-like, nucleic-acid-like, or peptide-like compounds. Again, the number of hydrogen bond acceptors of the ACD compounds ranges from 0 to 84 while the number of hydrogen acceptors of the CMC drugs ranges from 0 to 8.

It is also interesting to examine the number of smallest set of smallest rings (sssrs). As shown in Figure 16, the most complicated polycyclic compound in the ACD database is fullerene C60. The power of the sssrs descriptor permits it to easily identify non-drug-like compounds because they possess too many rings. It has been reported that the best druglike ranges for sssrs is from 0 to 4. Most marketed drugs have less than six rings.[16] The sssrs values, however, in the ACD compound database, range from 0 to 31.

Finally, we found that the average atomic number, excluding H atoms (AZ), has a capacity for distinguish-

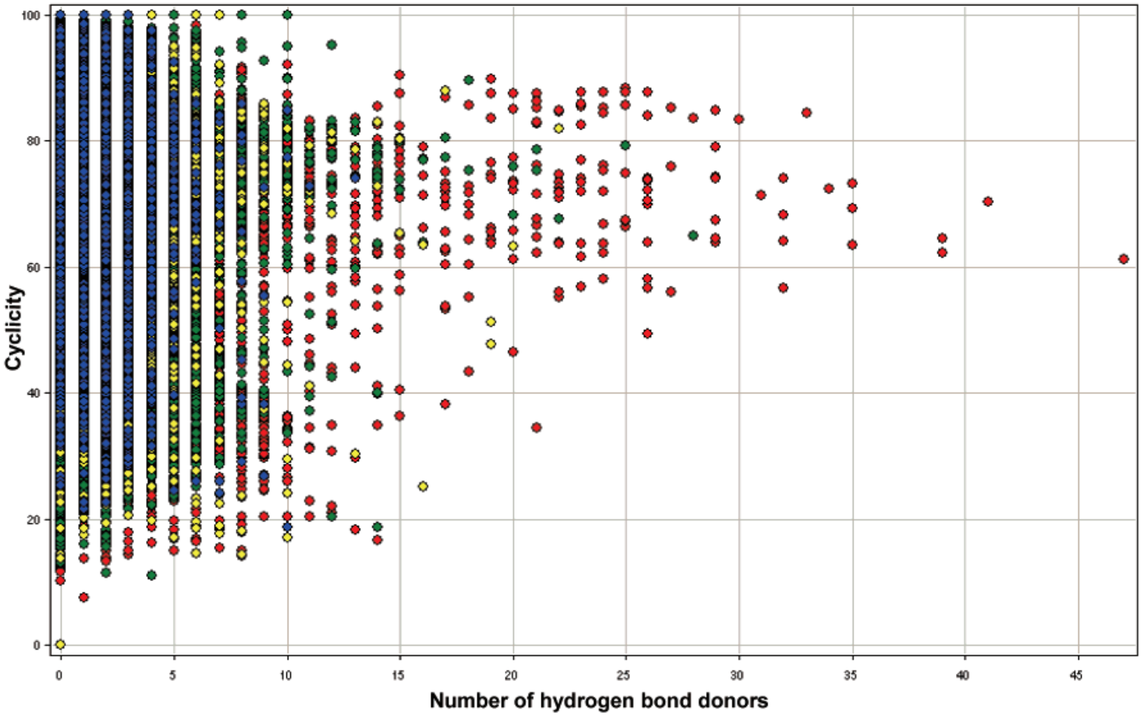


Figure 7. Cyclicity–HD (H bond donors) map. The cyclicity–HD (map is used to compare ACD (red), NCI (green), MDDR (yellow), and CMC (blue) databases.

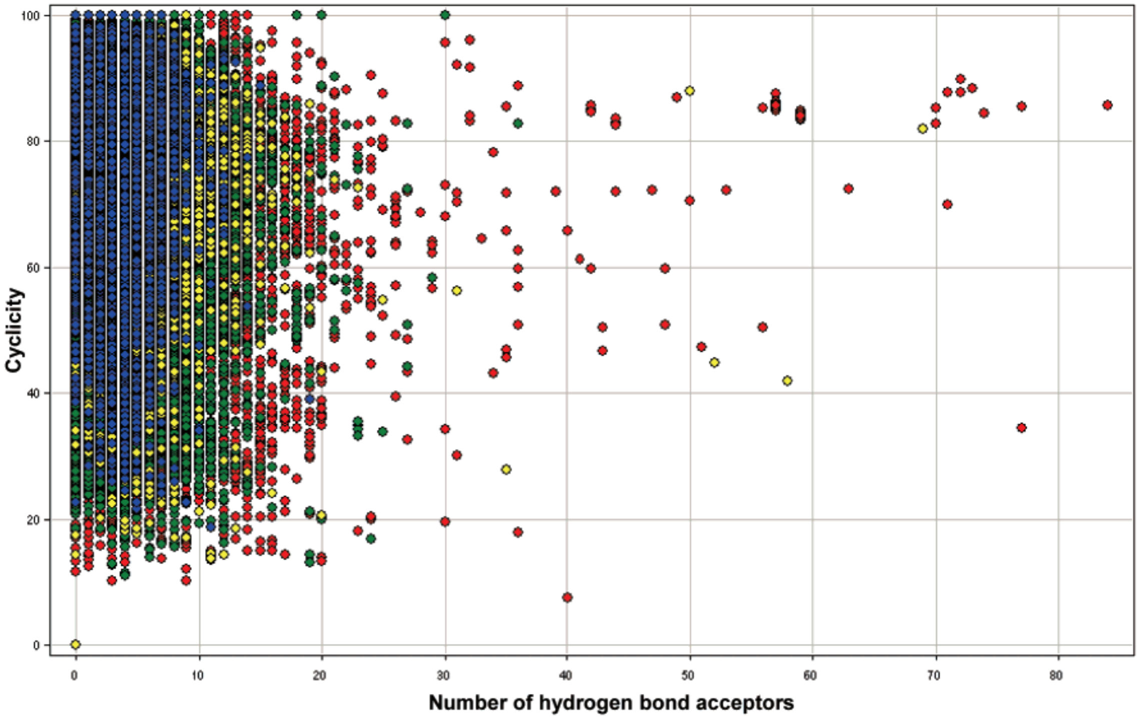


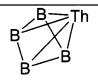
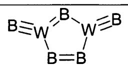
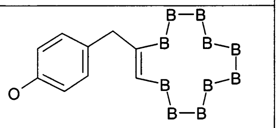
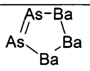
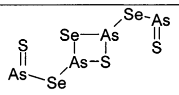
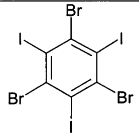
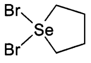
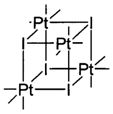
Figure 8. Comparison of structure databases of ACD (red), NCI (green), MDDR (yellow), and CMC (blue) databases on the distribution of number hydrogen bond acceptors.

ing the compounds that have a high content of heavier or lighter heteroatoms. Some interesting compound examples are listed in Table 5. We were immediately able to identify the small molecule with the longest chain in the cyclicity–rotating bonds map. This molecule is from the ACD database, and it has a benzene ring with a side chain as long as 180 single bonds.

Discussions

Most popular structure clustering algorithms are based on structural descriptors. However, medicinal chemists intuitively group their compounds based on scaffolds and functional groups. This difference in approach hinders the use of clustering results by

Table 5. ACD Structures with Minimum and Maximum AZ Values

AZ	Structure
5	
5	
5.6	
33	
28.2	
25	
	
18.29	
17.75	

medicinal chemists. The SCA was developed to eliminate this difference. Also, the SCA seeks to unify chemical library classification and mapping into one computation task. The purpose of this approach is to provide an efficient and intuitive tool for medicinal chemists to analyze and compare the drug-likeness and chemical diversity of large-scale libraries.

On the basis of our experience, the SCA is much faster than conventional approaches when classifying the same size library. Furthermore, the SCA can classify multiple libraries together for comparison purposes. These functions allow us to visually determine if two libraries are complimentary. The "diversity holes" can be identified from discontinuities in the diversity patterns in the SCA map (see Figure 4). In addition, a chemist can fill the "diversity holes" by adding scaffolds or placing functional groups on specific scaffolds by examining the peripheral environment in the SCA map.

Scientists from different disciplines may have very different understandings of the scaffold concept. In graph theory, a topological scaffold is a structure that has no side chain. To synthetic chemists, a scaffold might be a structure that can be made by a feasible chemical synthetic reaction. To biochemists, a scaffold might be a structure that can be a key component for a biologic target through assays. Therefore, for a given chemical structure, there can be many scaffold partitions. Figure 9 shows an example.[17] A biologic scaffold is associated with a biologic target. For a given compound, its biologic scaffold partition can be different when the compound is tested against a different target. Similarly, a compound may have a different synthetic scaffold partition when it is associated with a different synthetic strategy. If a topological scaffold covers the most common parts of biologic and chemical scaffold partitions, then the SCA works well. Otherwise, the SCA results may produce unsatisfactory results.

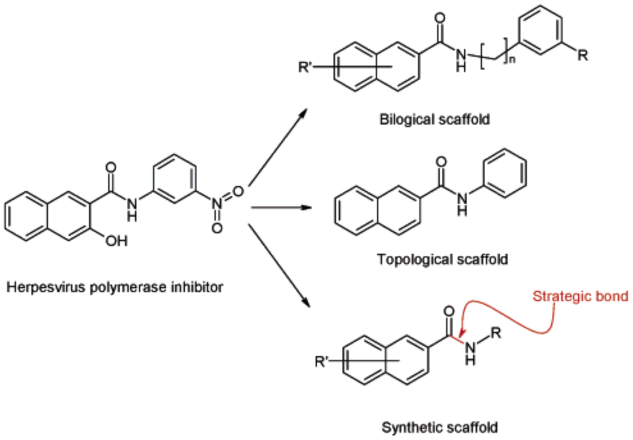


Figure 9. Biological scaffold, topological scaffold, and synthetic scaffold.

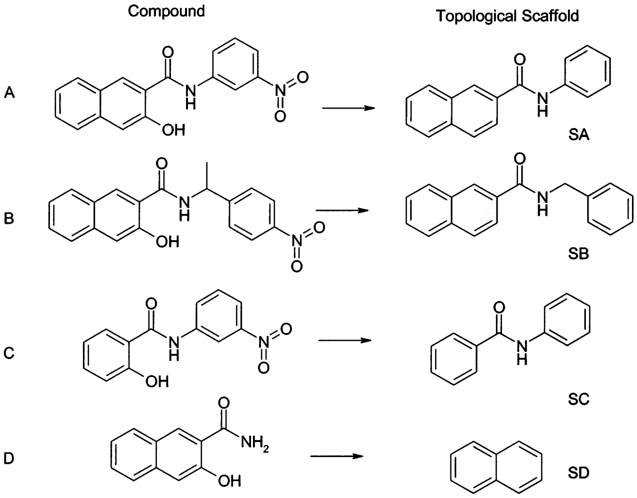


Figure 10. Scaffold discussions. Complicated cases are the following: scaffolds SC and SD are the substructures of scaffold SA; scaffolds SA and SB are similar, and they have small topological differences.

As shown in Figure 10, according to the SCA, compounds A–D have four different scaffolds (SA–SD). Note that scaffolds SC and SD are substructures of SA and that SA and SB are very similar (differing only in a CH_2 group). A chemist may suggest that we should group compound A together with compound B because SC is the substructure of SA. However, the further question is, would you as well group compound D with compound A? One may argue that SD is the substructure of SA, but it is not significant enough to group A and D together. For a given target, if the binding site has enough space to tolerate one more CH_2 group, then SA and SB should be considered as the same. Consequently, compounds A and B should belong to the same class, and the scaffold will be neither SA nor SB. The scaffold for compounds A and B is a nonspecific structure (the linker length is variable).

Owing to this type of problem, the current SCA can potentially produce too many classes and too many singletons. To try to correct this problem, we can group structures together. We can try this if their scaffolds are highly similar or if some scaffolds are the substructures of the other scaffolds, based on a similarity threshold or on substructure search rules. However, having tried this, we find that we can potentially group irrelevant compounds together and it is difficult to determine a scaffold for a class.

Another way to solve this type of problem is to use a set of predefined scaffolds. This strategy is similar to the method of LeadScope.[18] This approach makes chemical sense, but it requires human intervention, and it is biased on the basis of chemical experience. We believe that further improvements in the SCA will yield a solution to these problems.

Acknowledgment. The author thanks Drs. Curt Barthel, Hans Briem, John Proudfoot, and Richard Shaps for helpful discussion in testing the approach and in preparing this paper.

References

(1) Tryon, R. C. *J. Chronic Dis.* **1939**, *20*, 511−524.

(2) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press, Wiley: New York, 1987.

(3) Joliffe, I. T. *Principal Component Analysis*; Springer-Verlag: New York, 1986.

(4) Malinowski, E. H.; Howery, D. G. *Factor Analysis in Chemistry*; John Wiley & Sons: New York, 1980.

(5) Willett, P. Using Computational Tools To Analyze Molecular Diversity. In *A Practical Guide to Combinatorial Chemistry*; Czarnik, A. W., DeWitt, S. H., Eds.; American Chemical Society: Washington, DC, 1997; pp 17−48.

(6) Taraviras, S. L.; Ivanciuc, O.; Cabrol-Bass, D. Identification of Groupings of Graph Theoretical Molecular Descriptors Using a Hybrid Cluster Analysis Approach. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1128−1146.

(7) Cox, T. F.; Cox, M. A. A. *Multidimensional Scaling*; Chapman & Hall/CRC Press: Boca Rotan, FL, 2000 (ISBN 1-58488-094-5).

(8) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists: An Introduction*; VCH: Weinheim, Germany, 1993.

(9) Web address. http://www.statsoft.com/textbook/stmulsca.html#general.

(10) Kohonen, T.; Kangas, J.; Laaksonen, J. SOM_PAK, The Self-Organizing Map Program Package Available for Anonymous ftp User at Internet Site cochlea.hut.fi, version 1.2, November 1992.

(11) Bernard, P.; Golbraikh, A.; Kireev, D.; Chrétien, J. R.; Rozhkova, N. Comparison of chemical databases: Analysis of molecular diversity with self organising maps (SOM). *Analusis* **1998**, *26* (8) (October).

(12) Web address. www.spotfire.com.

(13) Agrafiotis, D. K.; Myslik, J. C.; Salemme, F. R. Advances in Diversity Profiling and Combinatorial Series Design. *Mol. Diversity* **1999**, *4*, 1−22.

(14) Hofmann, T.; Buhmann, J. M. Pairwise Data Clustering by Deterministic Annealing. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 1−14.

(15) Blatt, M.; Wiseman, S.; Domany, E. Superparamagnetic Clustering of Data. *Phys. Rev. Lett.* **1996**, *76*, 3251−3254.

(16) Xu, J.; Stevenson, J. Drug-Like Index: A New Approach To Measure Drug-Like Compounds and Their Diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177−1187.

(17) Oien, N. L.; et al. Broad-spectrum antiherpes activities of 4-hydroxyquinoline carboxamides, a novel class of herpes virus polymerase inhibitors. *Antimicrob. Agents Chemother.* **2002**, *46*, 724−730.

(18) Web address. www.leadscope.com.

JM010520K