Lead Discovery Using Stochastic Cluster Analysis (SCA): A New Method for Clustering Structurally Similar Compounds

Charles H. Reynolds,* Ross Druker, and Lori B. Pfahler

Rohm and Haas Company, 727 Norristown Road, Spring House, Pennsylvania 19477

Received July 24, 1997

We have developed an algorithm that clusters structural databases using topological similarity. The first step in this procedure is to identify a set of probe structures that all fall outside a defined similarity score cutoff with respect to one another. This list of probes is then used to bin the remaining compounds in the database. In the last step, some housekeeping is performed to ensure that each compound in the dataset is either a probe or is contained in one and only one bin. We have applied this clustering method to a database of ∼27 000 compounds for which we have screening level biological data. Analysis of the resulting clusters shows that clusters defined by an active probe are much more likely to contain other active compounds than clusters defined by an inactive probe. Indeed, the incidence of active compounds in bins with active probes is anywhere from 6 to 10 times greater than the incidence of active compounds in the database as a whole. This results demonstrates the power of simple two-dimensional topological descriptors, and serves to validate our clustering algorithm.

INTRODUCTION

The application of molecular topology has taken on an increasingly significant role in the discovery of biologically active molecules.[1−4] For example, topological similarity has been exploited to mine structural databases for new analogues based on a lead structure,[5,6] and to improve the efficiency of lead identification screening programs employing high-throughput screening (HTS).[7] Topological similarity, or dissimilarity,[8] is also an important concept in combinatorial synthesis programs where one wants to design diverse or biased libraries for synthesis.[9−11]

The use of molecular similarity in the design of new biologicals, or any other new product, is based on the assumption that similar molecular structures impart similar molecular properties. Therefore, if one can classify a set of compounds by their degree of structural similarity, one might also be classifying them with respect to their molecular properties. Of course, our success at doing this depends on a number of factors. These factors include: the degree to which a particular property is actually influenced by structure, the ability of molecular descriptors to fully describe molecular structures, and the scoring algorithm for assessing similarity. Many topological descriptors[5,6,12−17] and scoring algorithms have been proposed. All of these have advantages and disadvantages, but in the end, similarity tends to be a subjective concept with no widely accepted definition.

Molecular similarity is also highly dependent on the particular property being studied. This dependency can be illustrated with a simple example. If one compares pyridine to benzene using atom pairs (AP) and topological torsions (TT), two very different scores result (Table 1). In the case of AP descriptors, the similarity score is 0.67. By comparison, the TT descriptors give only a similarity of 0.33 between benzene and pyridine. Is one descriptor demonstrably better than the other? It depends on the property. If the property is size, then both the AP and TT scores are much too small because in reality benzene and pyridine have almost identical sizes, as demonstrated by their very similar molar refractivity (MR) values. If, however, one is interested in electron withdrawing/donating ability (σ_p), then the similarity scores seem, if anything, too high. Benzene has an essentially zero σ_p value, meaning that it is neither strongly electron donating or withdrawing. In contrast, the σ_p for pyridine is large and positive, indicating a group that is strongly electron withdrawing. Likewise, there is a big difference in the solubilities of benzene and pyridine in water, as evidenced by the significant difference in their experimental log *P* values. Thus, descriptors that emphasize the similarities of benzene and pyridine are better for properties related to size and shape, whereas descriptors that emphasize the differences are better for σ_p and log *P*.

Table 1. Comparison of Similarity Scores and Molecular Properties for Benzene and Pyridine

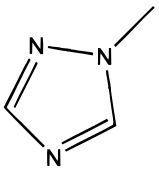
property/ descriptor	benzene	pyridine[a]	Δ	score[b]
AP				0.67
TT				0.33
MR[c]	2.54	2.30	0.24	
log *P*[c]	2.13	0.65	1.48	
σ_p[c]	−0.01	0.44	0.45	

[a] 4-Pyridyl. [b] Scored using the Carhart algorithm. [c] Corwin Hansch, Albert Leo and David Hoekman, *Exploring QSAR: Hydrophobic, Electronic and Steric Constants;* American Chemical Society: Washington, D.C., 1995.

In this paper we describe a method for clustering two-dimensional (2D) structures based on molecular topology. The relative efficiency of clustering algorithms and similarity

* Author to whom correspondence should be sent. E-mail: creynolds@rohmhaas.com.

S0095-2338(97)00056-5 CCC: $15.00
Published on Web 02/10/1998



Topological Torsions	Atom Pairs
C3(1)-N3(3)-C2(2)-N2(2)	C3(1)-2-C2(2)
C3(1)-N3(3)-N2(2)-C2(2)	C3(1)-3-C2(2)
C2(2)-N3(3)-N2(2)-C2(2)	C3(1)-1-N3(3)
C2(2)-N2(2)-C2(2)-N3(3)	C3(1)-2-N2(2)
C2(2)-N2(2)-C2(2)-N2(2)	C3(1)-3-N2(2)
N3(3)-N2(2)-C2(2)-N2(2)	C2(2)-2-C2(2)
N2(2)-C2(2)-N3(3)-N2(2)	C2(2)-1-N3(3)
	C2(2)-2-N3(3)
	C2(2)-1-N2(2)
	C2(2)-2-N2(2)
	N3(3)-1-N2(2)
	N3(3)-2-N2(2)
	N2(2)-2-N2(2)

Figure 1. Topological torsion (TT) and atom pair (AP) descriptors for substituted triazole. For example, the first TT describes four connected atoms with the following atom types: (sp^3 *C* with one nonhydrogen attachment) − (sp^3 *N* with three nonhydrogen attachments) − (sp^2 *C* with two nonhydrogen attachments) − (sp^2 *N* with two nonhydrogen attachments). The atom pairs follow the same convention except that the middle number denotes the number of bonds between atoms.

descriptors have been the subject of several recent articles.[12,18,19] We have tested our method against a large database of compounds for which we have screening level biological data, which allows us to test our clustering algorithm and molecular descriptors in an objective and quantitative manner.

SIMSEARCH

We have developed a program at Rohm and Haas called SimSearch that allows us to rapidly screen 2D databases for topologically similar compounds. Using this program it is possible to score the similarity of any compound with respect to any other compound. SimSearch is based on the approach pioneered at Lederle wherein molecular structures are represented as sets of fragment-based descriptors.[15] The molecular descriptors available in SimSearch include the topological-torsion (TT) and atom-pair (AP) descriptors outlined by Sheridan and Venkataraghavan,[15] as well as related descriptors developed in our lab. In each case, the descriptor is divided into two parts: an atom type and topological path. In our implementation, the atom type can take a variety of forms. Following the TT and AP examples, the atom type is typically the element, hybridization and the number of connected nonhydrogen bonds (Figure 1). We have also experimented with other atom types, including generic atom types, such as hydrogen bond donor/acceptor, size, and lipophilicity.

The second part of the molecular descriptor contains the topology information and is defined either as the number of bonds in the shortest path between atoms (AP) or as the number of adjacent atoms to include in the descriptor. In the case of TTs, four adjacent atoms are used to define a descriptor. We have also experimented with different connected path lengths. We refer to these as diads, triads or tetrads (equivalent to TTs) depending on the number of atoms in the descriptor. Finally, SimSearch allows us to use multiple path lengths concurrently. For example we have employed the diads, triads, and tetrads concurrently as molecular descriptors.[20] Similar schemes involving multiple or blended descriptors have been suggested by other groups.[6,7] Using multiple path lengths has the advantage of increasing the total number of descriptors and tends to produce smoother changes in similarity scores, particularly when comparing small molecules.

Similarity is scored using eq 1 following the example of Carhart et al.[13] The Carhart scoring algorithm is a slight departure from the more common Tanimoto coefficient, but has the same behavior in the limiting cases. If two molecules are identical, the score is 1.0, and if they have no common descriptors, the score is 0.0. Intermediate cases give scores that fall between these two extremes.

$$S_{AB} = 2(N_C)/(N_A + N_B) \quad (1)$$

In eq 1, S_{AB} is the similarity score for comparison of structures A and B, N_A is the number of descriptors in A, N_B is the number of descriptors in B, and N_C is the number of descriptors common to A and B. In practice, each descriptor is assigned a unique integer number and each entry into the database is stored as an array of integers. Comparison of any probe molecule against the database simply entails matching descriptors in the probe against descriptors for each structure in the database.

A simple example of an analogue search can be illustrated with the commercial antifungal Systhane[21] (**1**) in Figure 2. If **1** is used as a probe to search our internal database using AP descriptors, SimSearch will match **1** against every structure in the database and produce a similarity score (S). The results are ordered from highest similarity score to lowest, and the user can select any arbitrary cutoff. In this case, the highest scoring compounds include a number of interesting hits (Figure 2). All of these compounds[22−24] (**2**−**7**) have obvious structural features in common with **1** and also exhibit at least low-level antifungal activity.

Even though these compounds are similar, it would be difficult to construct a substructure query that would identify them all without being so generic as to include large numbers of uninteresting compounds. For example, although **2** represents a minor substitution of methyl for cyano, structures **3** and **4** contain different heterocycles. Indeed, the only substructure common to all six compounds is very simple (**8**) and poorly represents this class of compounds.

DIVERSITY AND CLUSTERING

Given the capability to search a database for similar compounds, it follows directly that one could also use the same criteria to search for diverse compounds. Interest in assembling large sets of structurally diverse compounds has grown rapidly as pharmaceutical and agrochemical companies have begun implementing high throughput screens. These screens have created a huge appetite for compounds to feed them, but it is generally recognized that these screens can be used most productively in lead generation mode if the compounds put through the screens are diverse.[9,25] Similarly, combinatorial chemistry has made it possible to synthesize hundreds to tens of thousands of compounds rapidly. Whereas early work in combinatorial chemistry[26] simply concentrated on synthesis of large numbers of compounds, more recently it has been recognized that a pure numbers game is inefficient. This realizaiton has led to schemes for ensuring that combinatorial libraries for lead generation have a high degree of diversity.

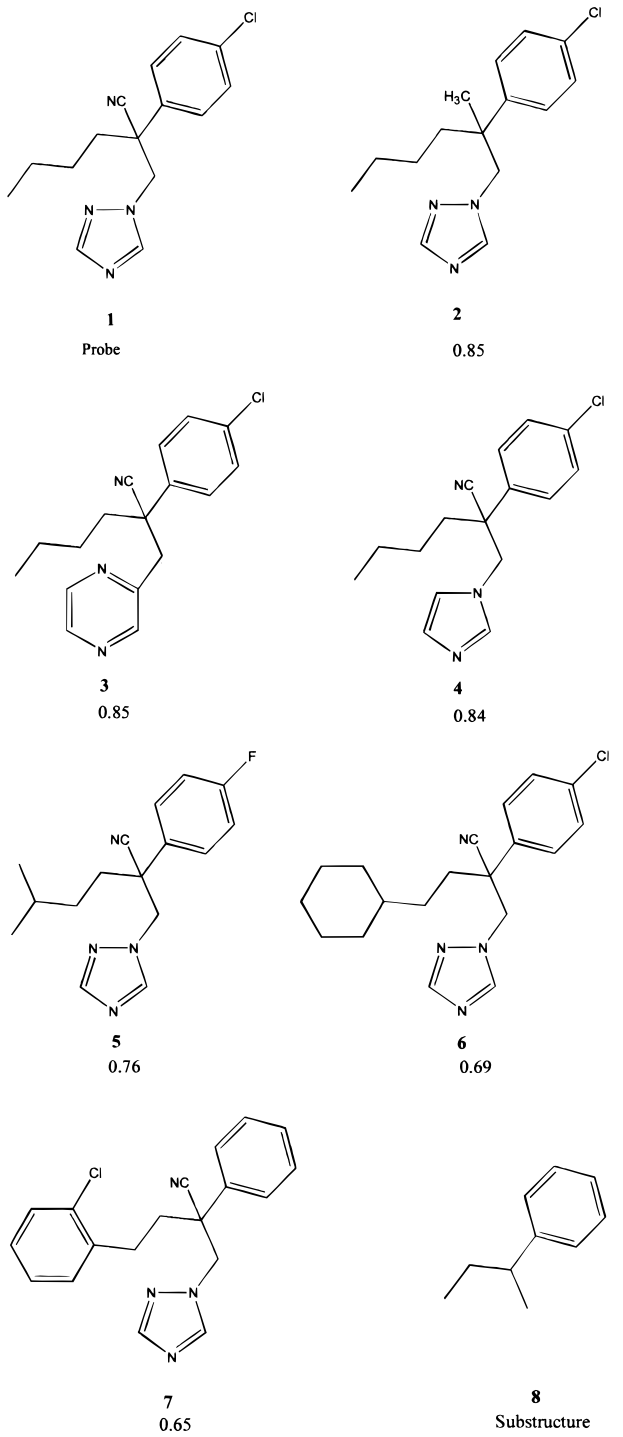


Figure 2. Structures and similarity scores for antifungal compounds extracted from the Rohm and Haas database using *Systhane* **(1)** as the probe molecule. The maximum common substructure for all seven structures is also given **(8)**.

One approach to the diversity problem is to cluster a structural database or virtual library based on some kind of structural criteria. In our case we have chosen to use 2D topological descriptors as already outlined. Standard approaches for clustering[12,27] can be broken into two broad categories: hierarchical and nonhierarchical. The hierarchical approaches can be further categorized as agglomerative or divisive. In these approaches, the database is either divided successively until a predetermined number of clusters has been created, or members are successively grouped together until the predetermined number of clusters has been assembled. In either case, a dendrigram is created that maps N members in 1 cluster to N members in N clusters. In a nonhierarchical approach, a nearest neighbor list is created and used to assemble members into related clusters. An example of this is the Jarvis-Patrick[28] clustering algorithm, which has been widely used to cluster structural databases.

There are many reasons why one might want to cluster a database of molecular structures.[19,29,30] Two of the most practical reasons are to identify representative compounds from a structural database or virtual compound library for screening or synthesis. In addition, we were interested in using a clustering algorithm to validate our similarity methods and descriptors. If it is possible to cluster databases where we have some biological data in a way that groups compounds with like activity, that will serve to validate the methods used to assign similarity. Further, it is sometimes useful just to be able to determine if a database offering is very diverse or if most of the structures fall into a small number of homologous structural classes.[31]

We had three objectives in mind when designing a clustering algorithm. First, we wanted a method that would divide a database into an "appropriate" number of clusters based on the structures and their relative similarity rather than some predefined number. Having to specify the number of clusters is a significant shortcoming of most clustering algorithms that create a defined number of clusters without regard to the fact that this sometimes requires grouping very unlike structures together. Second, we wanted a method that would allow us to cluster additional structures without starting from scratch. This objective requires an algorithm that can begin with a set of clusters and add future structures to existing clusters or create new clusters as their structural topology dictates. Third, of course, any method has to be computationally tenable for very large structural databases. Speed is one of the most significant problems with hierarchical methods, but even the more efficient nonhierarchical approaches scale formally as N^2. None of the standard clustering procedures we were able to identify in the literature met all three requirements.

STOCHASTIC CLUSTER ANALYSIS (SCA)

We have developed an algorithm[20] that clusters structural databases using topological similarity in two steps (Figures 3 and 4). The first step is analogous to "dissimilarity" searches proposed by other workers. Dissimilarity searches[32] involve picking a compound at random and then looking for a dissimilar, often maximally dissimilar, compound. This process is repeated for a given number of steps until a subset of highly diverse compounds is assembled. In our algorithm we also use dissimilarity in the first (diversity) step. But rather than searching for maximally dissimilar compounds, we define a similarity cutoff value (S_c). Any pair of compounds with $S < S_c$ are accepted as diverse.

The diversity step to select diverse probes of a structural database has the following basic procedure (Figure 3): (1) A similarity score is chosen as S_c (e.g., 0.65). This score is the maximum similarity allowed between probe structures. (2) A random compound is selected from the database. At the start of the program this compound becomes the first probe, P_1. In subsequent steps this compound is just a starting point for the search. (3) Beginning at the randomly

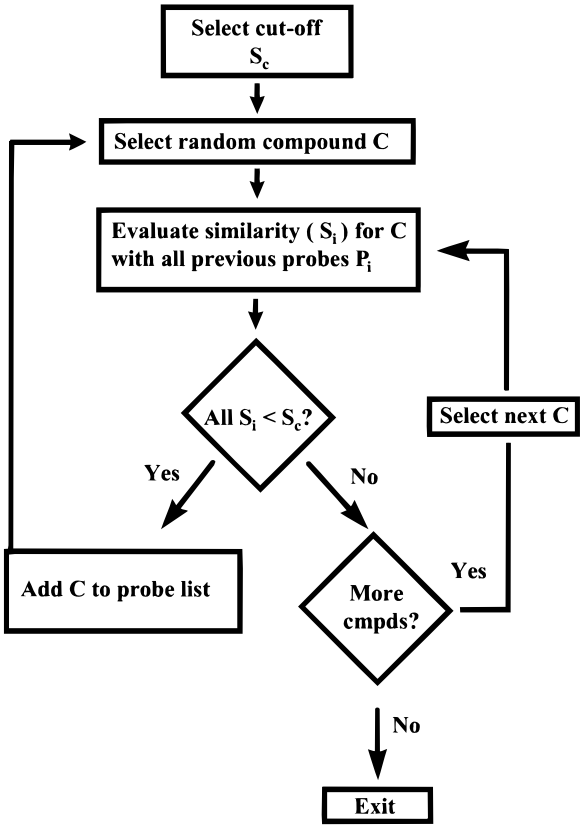


Figure 3. Procedure for identifying dissimilar probes to define the clusters: S_c is the similarity score cutoff, *C* represents a compound from the database, S_i is the similarity score for a randomly selected compound with one member of the probe list, and P_i are members of the current probe list.

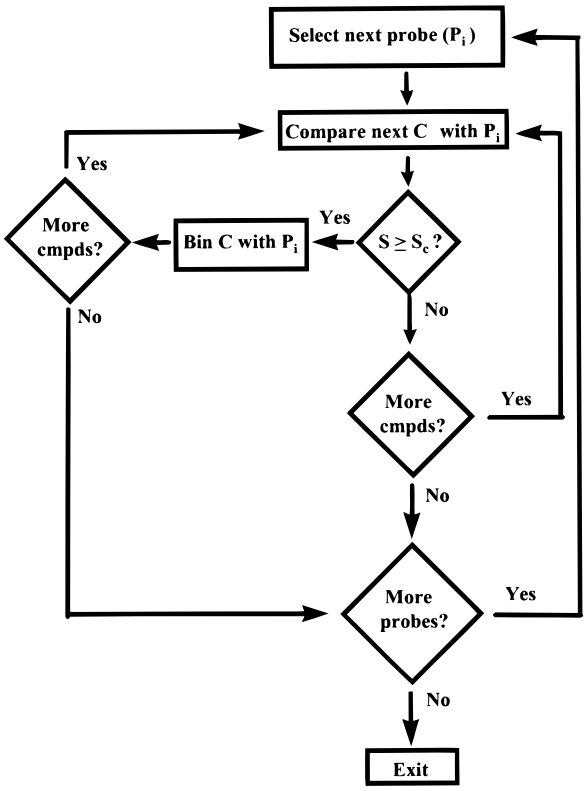


Figure 4. Procedure for assigning remaining structures (*C*) to clusters defined by the probe molecules (P_i).

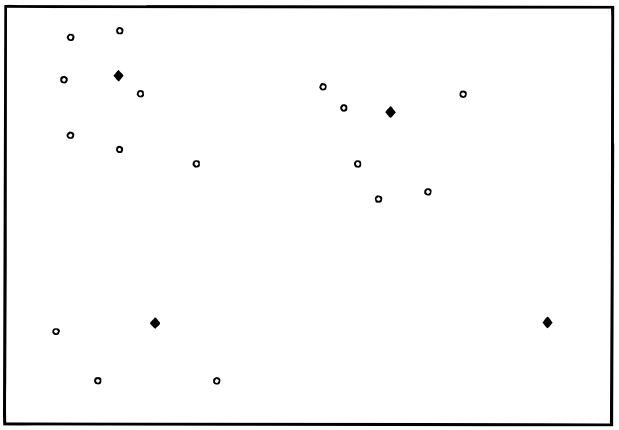


Figure 5. An idealized 2D representation of the multidimensional descriptor space. The circles represent structures in descriptor space and the diamonds represent structures that have been selected as probes.

selected starting point, evaluate the structure to determine its similarity score (*S*) with respect to all previous probes. If $S < S_c$ with respect to all previous probes it becomes a new probe and is added to the probe list. Go to step 2. If $S > S_c$ similar to any previous probe it is skipped and the next structure in the database is evaluated. Go to step 4. (4) Select the next structure and determine its similarity score (*S*) with respect to all previous probes. If $S < S_c$ with respect to all previous probes it becomes a new probe and is added to the probe list. Go to step 2. If $S > S_c$ similar to any previous probe it is skipped and the next structure in the database is evaluated. If there are no more compounds in the database, the diversity search terminates.

At this point a series of structural probes has been defined. Each probe has the property that it is less than S_c similar to any other probe. These data are represented as filled diamonds in Figure 5.

After the probes have been identified, a similarity search of the whole database is run using each probe structure. The similarity step uses the following procedure (Figure 4): (1) Select a probe (P_i) molecule from the set of diverse probes. If there are no more probes in probe list, the similarity function terminates. (2) Select a compound in the database that is not a probe. Calculate the similarity score (*S*) with respect to P_i. (3) If $S \geq S_c$ with respect to P_i, then the compound is added to cluster list for P_i. (4) If all (nonprobe) compounds have been evaluated, go to step 1, otherwise go to step 2.

For P_i, all structures in the database that are $> S_c$, similar to P_i, are assigned to a cluster defined by P_i (Figure 6). In some cases, a particular probe will have many members in the cluster it defines, in others cases it will be the only member (singleton). It is possible that some compounds will fall within the cutoff (S_c) with respect to more than one probe (Figure 7).

The final step in SCA is to look for compounds that appear in more than one cluster, which can happen because a compound falls within the similarity cutoff criterion for more than one probe. Compounds are only allowed to remain in one cluster. Therefore, in the last step, all compounds that appear in multiple clusters are evaluated, and placed in the cluster represented by the probe with which they have the greatest similarity. The entire process is illustrated in Figures 5-7. At the end of this process, all of the compounds in the database are either a probe or belong to one or more clusters.

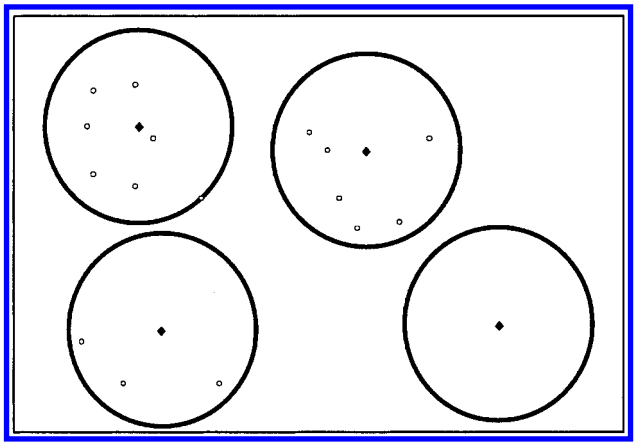


Figure 6. The large circles represent a radius of similarity in descriptor space. Compounds (small circles) within a large circle have similarity scores greater than the cutoff with respect to the central probe (diamonds)

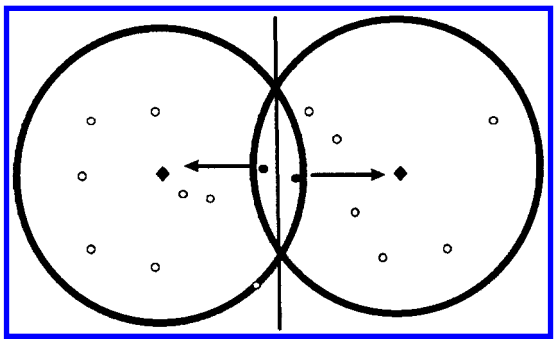


Figure 7. Compounds that initially fall within two overlapping clusters (filled circles) are moved into the cluster where they share the highest similarity with the probe (diamonds).

It should be emphasized that no attempt is made to eliminate singletons. If a structure is found to be unique, it is allowed to stand as a cluster of one. This overall approach is analogous to a method developed independently in Pearlman's group at Texas for validation of structural metrics.[33,34]

Although it may appear convoluted at first pass, the SCA method just outlined has many advantages over traditional clustering schemes. First, the division of structures into distinct clusters is driven exclusively by S_c, not by a predetermined number of clusters. This procedure means that the number of clusters and the number of members in any cluster are dictated solely by the similarity between compounds. Thus, it is possible to characterize the diversity of a database by the number of clusters generated at a specific S (e.g., $S = 0.80$). All of the probes are guaranteed to be different by some consistent similarity measure, and there should be no occurrences of compounds being "forced" into a cluster with which it has low similarity to accommodate other clustering criteria. By contrast, other nonhierarchical methods that rely on nearest neighbor lists (e.g., Jarvis-Patrick) are very susceptible to placing diverse structures in the same cluster, which is probably the reason that Jarvis-Patrick[35] clustering fared so poorly in the comparitive study of Brown and Martin.[12]

Another advantage of SCA is its extensibility. Because the probes and clusters are strictly determined by the similarity cutoff rather than any external criteria, it is always possible to add new structural data to a cluster analysis without need to recluster structures that have already been analyzed. One can simply cluster additional structures as if they had been part of the previous database. The only exception is the final step where redundant clusters are eliminated. This feature helps reduce the computational cost for projects that are ongoing.

Table 2. Percent Active Compounds Using TT Descriptors

similarity cutoff	0.50	0.65	0.80
number of bins	4296	8836	16 029
average activity			
database as whole	0.07	0.07	0.07
all bins >5	0.07	0.08	0.10
active bin[a]	0.46	0.55	0.64
inactive bin[b]	0.04	0.04	0.05
ratio active/random	6.6	7.8	9.1

[a] Probe compound is active. [b] Probe compound is inactive.

Table 3. Percent Active Compounds Using AP Descriptors

similarity cutoff	0.50	0.65	0.80
number of bins	3655	9562	17 737
average activity			
database as whole	0.07	0.07	0.07
all bins >5	0.07	0.08	0.09
active bin[a]	0.43	0.56	0.70
inactive bin[b]	0.04	0.04	0.03
ratio active/random	6.1	8.0	10.0

[a] Probe compound is active. [b] Probe compound is inactive.

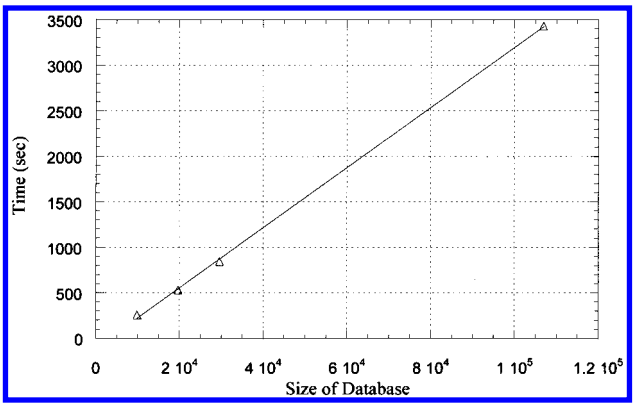


Figure 8. Comparison of run times for clustering databases ranging in size from 9898 to 106 909 structures. The computation time increases linearly with database size.

SCA is relatively efficient compared with other nonhierarchical methods because, in the early part of a clustering run, it is usually only necessary to sample a small number of structures to find one that has a value of S less than the cutoff. Furthermore, once a compound has been visited once in the dissimilarity step, it need not be considered again. The number of evaluations needed in the second similarity step can vary widely, but in most practical cases, the number of probes that must be checked against the database is a small fraction of the total. For the examples in Tables 2 and 3 the number of probes ranged from 15 to 58% of the database. All of these factors lead to a method that scales linearly with database size. A plot of SCA timings versus database size for databases ranging from 10 000 to 106 000 compounds is given in Figure 8. It takes <1 h to cluster >100 000 compounds on an IBM RS6000−530 class workstation.

One potential criticism of this method is that the probes are not required to be in the center of the clusters in descriptor

space (Figure 6), which is a consequence of selecting the probes solely on the basis of their similarity with other probes. Thus one could argue that, in some cases, another compound in the cluster would be a better representative of the cluster.

TESTING THE SCA METHOD

To test the SCA method, a subset of the Rohm and Haas compound archive was created. The test database (DBT) contains ~27 000 compounds for which first level herbicidal screening data is available. This screen is a simple one-dose, yes/no test to identify compounds for more advanced testing. The biological data in DBT is precisely the type of data one is likely to rely on for screening large numbers of compounds for new leads, a primary focus for our SCA method. On average, 7% of the compounds in DBT are scored as "active" structures. In terms of chemical structure, the database is a mixture of compounds resulting from directed synthesis programs and a random selection of diverse compounds acquired from a variety of independent sources. If our similarity descriptors are relevant to this assay and the clustering algorithm is valid, then structures should be clustered so as to group active compounds together.

To test our clustering algorithm, we clustered DBT using TT and AP with S_c values of 0.50, 0.65, and 0.80. We then analyzed the resulting clusters to determine the percentage of active compounds in clusters defined by an active or inactive probe. If all is working as we hope, the clusters represented by an active probe should be significantly enriched with active compounds relative to clusters represented by inactive probes.

The average percentage of active compounds across all bins, active bins (active probe) and inactive bins (inactive probe), are given in Table 2. To prevent small bins or singletons from unduly skewing the results, only bins that have five or greater members have been used to determine the averages. It is clear from Table 2 that the probes are indeed predictive with respect to the biological activity of the members of the cluster. For the TT-based clusters, the active bins have hit rates ranging from 46 to 64% depending on the S_c. As one would expect, a higher S_c leads to a better hit rate in the clusters with active probes because the degree of similarity required for compounds to be clustered together is more stringent. Enhancement in the hit rate for compounds in active clusters over the random hit rate ranges from a factor of 6.1 to 10. Clearly, compounds that are more similar to one another are being clustered together, and this structural similarity is reflected in the biological activity. Similar results are seen if the AP descriptor is used (Table 3). These results are a tremendous improvement over random sampling and are in stark contrast to the poor results reported by Taylor[36] for clustering a small test set of agrochemicals. We have also used SCA to cluster a small database constructed using 1000 random compounds from the Available Chemicals Directory (ACD)[37] and 27 compounds that represent four classes of structures with known biological activity. The results of this exercise are consistent with the large database just described, and are reported as Supporting Information.

Another approach is to conduct a theoretical exercise in which the objective is to identify as many active compounds from a database as possible in the fewest number of tests.

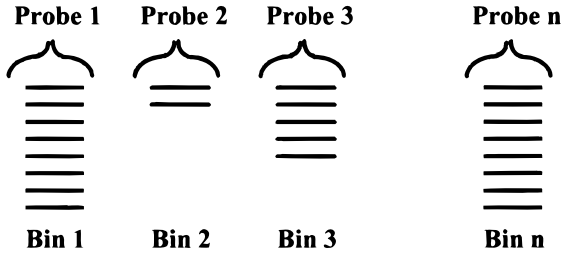


Figure 9. Each probe acts as a structural representative for the compounds in its cluster.

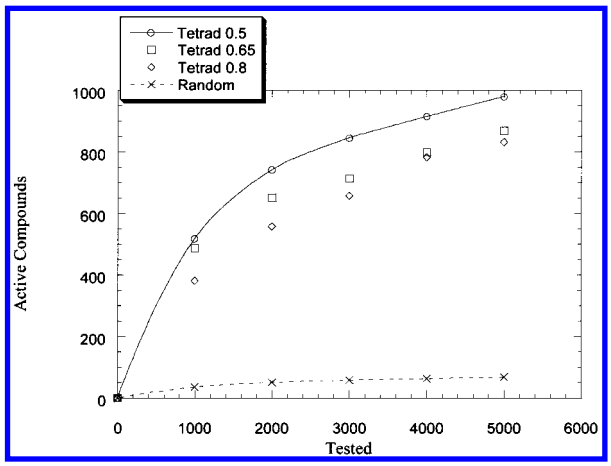


Figure 10. Plot of the number of active compounds found versus the number tested. The solid line is for a search using the TT descriptors at a similarity score of 0.50. The dashed line is for a random search.

This would simulate clustering in lead identification mode. The procedure is to test probe molecules (dissimilar samples of the database) until an active probe is found, and then test all the compounds in the cluster of that probe. This procedure amounts to skimming across the database sampling diverse structures until activity is found, followed by thorough exploration of similarity space about active probes (Figure 9). This method is somewhat analogous to the exercise proposed by Kearsley et al.,[5] except their searches were based only on similarity.

The procedure just described was carried out for the same database of 27 000 compounds clustered at similarities of 0.50, 0.65, and 0.80. Probe molecules were "tested" in order of bin size (large bins first) for activity. When an active probe was found, the bin was also searched for active compounds. A running total of active compounds found was tabulated against the total number of compounds examined. The results of this tabulation are given in Figure 10 as a plot. The *x*-axis gives the number of compounds evaluated for activity, and the *y*-axis gives the total number of active compounds found. If the search were truly random, this comparison should produce a line with a slope equal to the average activity in the database, in this case 7%.

Examination of Figure 10 shows that active compounds are found at a much better than random rate when one uses the cluster probes as an indicator of activity for the entire bin. It is interesting to note that although the probe is obviously more representative of the bin when a high S_c, such as 0.80, is used (Tables 2 and 3), this strategy is actually more successful when a lower S_c, 0.50, is used. At 3000 compounds tested the 0.50, 0.65, and 0.80 values of S_c lead to 846, 714, and 657 active compounds, respectively.

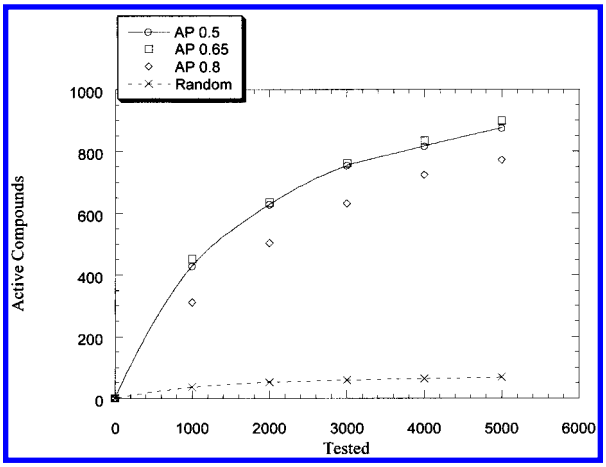


Figure 11. Plot of the number of active compounds found versus the number tested. The solid line is for a search using the AP descriptors at a similarity score of 0.50. The dashed line is for a random search.

Although at first glance this might seem counterintuitive, it is simply a function of the size of the bins at different similarity values. If a low S_c is used (e.g., 0.50), the number of bins is smaller and each bin is correspondingly larger. This means that larger portions of the database are represented by each probe. Thus, although the probe is a poorer representative, it allows more efficient differentiation between large groups of active and inactive compounds. By contrast, clustering with a high S_c (e.g., 0.80) leads to more and smaller bins. So, although the probe is more representative of the bin, the decision to test or discard a bin provides less leverage. For comparison, random sampling of the database would be expected to produce only 210 hits for 3000 compounds tested.

Obviously, there is a limit at which the probe becomes such a poor predictor of activity in the bin that the benefit of larger bins is erased by the inaccuracy introduced by judging large numbers of structures based on a relatively dissimilar probe. Close examination of the AP plot (Figure 11) illustrates this effect. The 0.65 curve is significantly higher than the 0.80 curve, as just seen. However, when we drop further to an S_c of 0.50, the curve does not improve at all. Indeed, it is slightly poorer. Again, comparing the number of active compounds found for 3000 tested, the number of active compounds found using S_c values of 0.50, 0.65, and 0.80 are 754, 761, and 631, respectively.

It is interesting to note that even though SCA is a stochastic method, the list of probes is not necessarily completely different every time it is run against a particular database. Many probes are found repeatedly when searching the same database multiple times using the same S_c value. For example, we have clustered DBT 10 times using an S_c value of 0.65. Almost half of the probes for 10 runs show up more than five times. A histogram of the percentage of probes found 1−10 times is given in Figure 12. The largest probabilities are for probes that appear only once and probes that appear in all 10 probe lists.

CONCLUSIONS

We have developed a simple clustering method (SCA) that lends itself to classifying chemical structures base on simple 2D topology descriptors. But, this approach should be applicable to a wide variety of topology descriptors. We have validated the use of SCA and descriptors, such as TT and AP, using a subset of our internal structural database. This clustering exercise clearly demonstrates that compounds with high similarity scores also tend to exhibit similar biological activity in this assay. The average activities in bins with an active probe were anywhere from 6 to 10 times greater than the average of the entire database. This enrichment in activity is certainly too great to be attributable to chance and serves to validate the use of similarity, the AP and TT descriptors, and the SCA clustering algorithm for dividing structural databases into chemically similar bins.

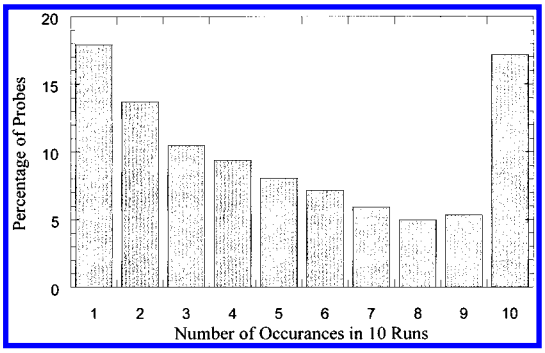


Figure 12. Histogram showing the percentage of probes that appear 1−10 times in 10 separate SCA runs for the test database.

Supporting Information Available: Structures of 27 compounds seeded into a sample database and a table of the resulting clusters containing those compounds (6 pages). See any current masthead page for ordering and Web access instructions.

REFERENCES AND NOTES

(1) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley and Sons: New York, 1990.

(2) Downs, G. M.; Willett, P. Similarity Searching in Databases of Chemical Structures In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Ed.; VCH: New York, 1996; Vol. 7, pp 1−66.

(3) Fisanick, W.; Lipkus, A. H.; Rusinko, III, A. Similarity Searching on CAS Registry Substances. 2. 2D Structural Similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 130−140.

(4) Stanton, D. T.; Murray, W. J.; Jurs, P. C. Comparison of QSAR and Molecular Similarity Approaches for a Structure−Activity Relationship Study of DHFR Inhibitors. *Quant. Struct.-Act. Relat.* **1993**, *12*, 239−245.

(5) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.

(6) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128−136.

(7) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(8) Holliday, J. D.; Willett, P. Definitions of Dissimilarity for Dissimilarity-Based Compound Selection. *J. Biomol. Screening* **1996**, *1*, 145−151.

(9) Ferguson, A. M.; Patterson, D. E.; Garr, C. D.; Underiner, T. L. Designing Chemical Libraries for Lead Discovery. *J. Biomol. Screening* **1996**, *1*, 65−73.

(10) Sheridan, R. P.; Kearsley, S. K. Using a Genetic Algorithm To Suggest Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 310−320.

(11) Tropsha, A.; Zheng, W.; Cho, S. J. Application of Topological Indices in Rational Design of Combinatorial Chemical Libraries. *Book of Abstracts*; 211th ACS National Meeting, New Orleans, LA, March 24−28; American Chemical Society: Washington, DC, 1996, CINF-068.

(12) 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.
(13) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(14) Pogliani, L. Modeling with Special Descriptors Derived from a Medium-Sized Set of Connectivity Indices. *J. Phys. Chem.* **1996**, *100*, 18065−18077.
(15) Sheridan, R. P.; Venkataraghavan, R. New Methods in Computer-Aided Drug Design. *Acc. Chem. Res* **1987**, *20*, 322−329.
(16) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure−Property Modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1991; Vol. 2, pp 367−422.
(17) Hall, L. H.; Kier, L. B.; Brown, B. B. Molecular Similarity Based on Novel Atom-Type Electrotopological State Indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1074−1080.
(18) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand−receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(19) Matter, H. Selecting Optimally Diverse Compounds from Structural Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(20) Druker, R.; Pfahler, L. B.; Reynolds, C. H. Finding a Needle in a Haystack: Using Topological Similarity to Identify Biologically Active Leads, *Book of Abstracts*; 21st ACS National Meeting, New Orleans, LA, March 24−28; American Chemical Society: Washington, D.C., 1996, COMP-047.
(21) Fujimoto, T. T. U.S. Patent 4,920,139, 1990.
(22) Reynolds, C. H.; Shaber, S. H. Rational Design of Novel Ergosterol Biosynthesis Inhibitor Fungicides. In *Computer-Aided Molecular Design: Applications in Agrochemicals, Materials, and Pharmaceuticals*; Reynolds, C. H., Holloway, M. K., Cox, H. K., Eds.; American Chemical Society: Washington, D.C., 1995; Vol. 589, pp 171−182, 2 plates − pp 230a-230b.
(23) Shaber, S. H.; Quinn, J. A.; Fujimoto, T. T. 2-Cyanoarylethyltriazoles as Agricultural Fungicides. Discovery of Fenbuconazole. In *Synthesis and Chemistry of Agrochemicals IV*; American Chemical Society: Washington, D.C., 1995; Vol. 584, pp 406−419.
(24) Fuimoto, T. T.; Quinn, J. A.; Egan, A. R.; Shaber, S. H.; Ross, R. R. *Pestic. Biochem. Physiol.* **1988**, *30*, 199−213.
(25) Johnson, M.; Lajiness, M.; Maggiora, G. Molecular Similarity: A Basis for Designing Drug Screening Programs. *Prog. Clin. Biol. Res.* **1989**, *291*, 167−171.
(26) Warr, W. Combinatorial Chemistry and Molecular Diversity. An Overview. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 134−140.
(27) Willett, P. Algorithms for the Calculation of Similarity in Chemical Structure Databases. In *Concepts in Applied Molecular Similarity*; Wiley: New York, 1990; pp 43−65.
(28) Jarvis, R. A.; Patrick, E. A. Clustering Using a Similarity Measure Based on Shared Nearest Neighbors. *IEEE Trans. Comput.* **1973**, *C-22*, 1025−1034.
(29) McGregor, M. J.; Pallai, P. V. Clustering of Large Databases of Compounds: Using the MDL "Keys" as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.
(30) Doman, T. N.; Cibulskis, J. M.; Cibulskis, M. J.; McCray, P. D.; Spangler, D. P. Algorithm 5: A Technique for Fuzzy Similarity Clustering of Chemical Inventories. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1195−1204.
(31) Turner, D. B.; Tyrrell, S. M.; Willett, P. Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18−22.
(32) Bawden, D. Molecular Dissimilarity in Chemical Information Systems. In *Chemical Structures 2. The International Language of Chemistry*; Warr, W. A., Ed.; Springer-Verlag: Heidelberg, 1993; pp 383−388.
(33) Pearlman, R., personal communication.
(34) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. In *3D QSAR and Drug Design: Recent Advances*; Kubinyi, H., Martin, Y., Folkers, G., Ed.; ESCOM: Amsterdam, in press.
(35) Barnard, J. M.; Downs, G. M. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644−649.
(36) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59−67.
(37) Available Chemicals Directory (ACD), distributed by Molecular Design Ltd., San Leandro California.

CI970056L