Rational Combinatorial Library Design. 3. Simulated Annealing Guided Evaluation (SAGE) of Molecular Diversity: A Novel Computational Tool for Universal Library Design and Database Mining

Weifan Zheng,[†] Sung Jin Cho,[†,§] Chris L. Waller,[‡] and Alexander Tropsha*[,†]

Laboratory for Molecular Modeling, Division of Medicinal Chemistry, School of Pharmacy, University of North Carolina at Chapel Hill, North Carolina 27599-7360, and OSI Pharmaceuticals, Inc., 4727 University Drive, Suite 400, Durham, North Carolina 27707

Received May 18, 1998

We have developed a novel method for molecular diversity sampling called SAGE (simulated annealing guided evaluation of molecular diversity). Compounds in chemical databases or virtual combinatorial libraries are conventionally represented as points in multidimensional descriptor space. The SAGE algorithm selects a desired number of optimally diverse points (compounds) from a database. The diversity of a subset of points is measured by a specially designed diversity function, and the most diverse subset is selected using Simulated Annealing (SA) as the optimization tool. Application of SAGE to two simulated data sets of randomly distributed points in two-dimensional space afforded diverse and representative selection as judged by visual inspection. SAGE was also applied, in comparison with random sampling, to two other simulated data sets with points distributed among many clusters. We found that SAGE sampling covered significantly more clusters than the random sampling. By defining a fraction of data points as active, we also compared SAGE with random sampling in terms of hit rates. We showed that when the percentage of active points was low, the hit rates obtained by SAGE were always higher than those obtained by random sampling. When the percentage of active points was high, the performance of SAGE, in terms of individual hit rates, depended upon the data structure. However, in all cases, SAGE performed better than random sampling when cluster hit rates were used as the criterion.

INTRODUCTION

Combinatorial chemical synthesis and high throughput screening have significantly increased the speed of the drug discovery process.[1−3] However, it is still impossible to synthesize all of the library compounds in a reasonably short period of time. As many as 3000^3 (2.7×10^{10}) compounds can be synthesized from a molecular scaffold with three different substitution positions when each of the positions has 3000 different substituents. If a chemist could synthesize 1000 compounds per week, 27 million weeks (~0.5 million years) would be required to synthesize all these compounds. Furthermore, many of these compounds can be structurally similar to each other, and chemical information contained in the library can be redundant. Thus, there is a need for rational library design (i.e., rational selection of a subset of building blocks for combinatorial chemical synthesis) so that a maximum amount of information can be obtained while a minimum number of compounds are synthesized and tested. Similarly, there is a closely related task in computational database mining, i.e., rational sampling of a subset of compounds from commercially available or proprietary databases for biological testing.

There are two types of experimental combinatorial chemistry and high throughput screening research directions, namely, targeted screening and broad screening.[1,2] The former approach involves the design and synthesis of chemical libraries with compounds that either are analogues of some active compounds or can specifically interact with the biological target under study. This is desired when a lead optimization (or evolution) program is pursued. On the other hand, a broad screening project involves the design and synthesis of a large array of maximally diverse chemical compounds, leading to diverse (or universal) libraries that are then tested against a variety of biological targets. This design strategy is suited for lead identification programs. Thus, two categories of computational tools should be developed and validated to meet the needs of the two different projects.

In a targeted screening project, computational library design involves the selection of a subset of chemical building blocks from an available pool of chemical structures. This subset of selected building blocks affords a limited virtual library with a high content of compounds similar to a lead molecule. Molecular similarity is quantified using a chosen set of molecular descriptors and similarity metrics.[4,5] Building blocks can also be chosen such that the resulting virtual library could have a high percentage of compounds that are predicted to be active from a preconstructed QSAR model.[6] In cases where the structure of the biological target is known, one can select building blocks so that the resulting library compounds are stereochemically complementary to the binding site structure of the underlying target.[7,8] Other

* To whom correspondence should be addressed.
[†] University of North Carolina at Chapel Hill.
[‡] OSI Pharmaceuticals, Inc.
[§] Current address: Combinatorial Drug Discovery, Brystol-Myers Squibb Company, Wallingford, CT 06492-7660.

10.1021/ci980103p CCC: $18.00 © 1999 American Chemical Society
Published on Web 06/09/1999

approaches to targeted library design using different criteria have also been considered recently.[9] Similar approaches have long been used in targeted database mining, which were based on the principle of either molecular similarity[10−18] or structure-based drug design.[19−22]

In a broad screening project, computational library design or database mining involves the selection of a subset of compounds that are optimally diverse and representative of available classes of compounds, leading to a nonredundant chemical library or a set of nonredundant compounds for biological testing. Reported methods can be generally classified into several categories: (1) cluster sampling methods, which first identify a set of compound clusters, followed by the selection of several compounds from each cluster;[23−30] (2) grid-based sampling, which places all the compounds into a low dimensional descriptor space divided into many cells, and then chooses a few compounds from each cell;[31] and (3) direct sampling methods, which try to obtain a subset of optimally diverse compounds from an available pool by directly analyzing the diversity of the selected molecules.[32−35] Recently, many reports have been published addressing various aspects of diversity analysis in the context of chemical library design and database mining.[36−43]

For both types of library design, appropriate molecular descriptors with good neighborhood behavior should be developed and validated. This means that in an ideal case most of the close neighbors of an active compound in the descriptor space will also be active while those surrounding an inactive compound will be inactive. Several recent papers have addressed this issue.[30,44−46]

Recently, we have initiated the development of algorithms and software for diversity analysis (DA project) in the context of database mining and the design of both targeted and diverse chemical libraries. Our approaches to targeted library design have been discussed earlier.[5,6] In this paper, we discuss our new method for the optimal diversity selection that belongs to the category of direct sampling. This method, which we call SAGE (simulated annealing guided evaluation) of molecular diversity, selects a desired number of optimally diverse compounds from an available virtual (potential combinatorial library) or actual database.

In principle, there are two approaches to characterize the behavior of any new computational technique. First, through retrospective analysis of a chemical database with known structures and activities and, second, through thorough analysis of simulated data sets with known data distribution. The latter approach allows separating the issues of descriptor development, selection, and validation, which are unavoidable when dealing with actual data sets, from those of the sampling efficiency and accuracy. Therefore, we have concentrated on the second approach, which allows exploring the advantages and caveats of the sampling strategies. Since simulated data sets were used to replace real chemical databases, compound selection became choosing a subset of points from the collection of points. We have created several simulated data sets with different data structure and analyzed them using SAGE vs random sampling in terms of diversity coverage and simulated hit rates. Three types of computational experiments were performed: (1) visualization of the rationally selected points in two-dimensional (2D) space; (2) quantification of the information content in terms of the percent coverage of the clusters; and (3) computer simulation of the hit rates in terms of the percentage of active points (molecules) obtained by SAGE vs random sampling. We found that SAGE affords very diverse selection of points as judged both visually and by percent coverage. However, the performance of SAGE with respect to the hit rates depended on the data structure and did not generally exceed that of the random sampling when the percentage of active compounds in the simulated data sets was relatively high. This result identifies theoretical boundaries of the expected efficiency of computational methods for sampling chemical diversity even when theoretically ideal descriptors are used to characterize compounds.

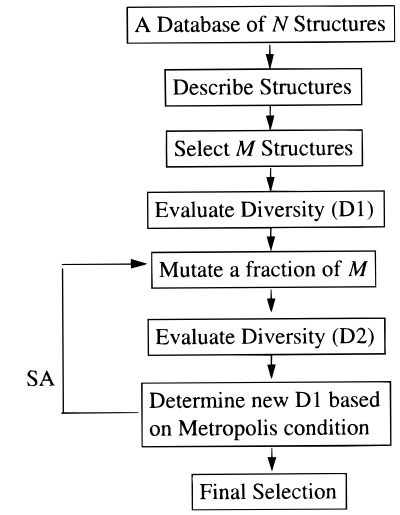

Figure 1. Flow chart of the SAGE algorithm.

COMPUTATIONAL DETAILS

General Design of SAGE. The general goal of the diverse sampling is to select a relatively small subset of *M* molecules from an available pool of much bigger size *N* so that the selected subset represents as many chemical classes of compounds as possible. A special diversity function (see below) has been designed to measure the diversity of selected compounds. Compounds are conventionally characterized by descriptors, such as molecular connectivity indices,[47−49] atom pairs,[50] or other available molecular descriptors.[51] Therefore, each compound is ultimately represented as a point (or a vector) in a multidimensional descriptor space. A stochastic search algorithm using simulated annealing has been implemented to identify the most diverse set of M points. Due to the use of simulated annealing as an optimization protocol, we called this method simulated annealing guided evaluation of chemical diversity, or SAGE. The general flow chart of the SAGE method is shown in Figure 1, and the algorithm is described in more detail below.

Diversity Function. One of the key aspects of diversity sampling is to design a mathematical function that adequately measures the diversity of a subset of *M* selected molecules. Since each molecule is represented by a point in multidimensional descriptor space, the distance between two points, such as Euclidean distance, Tanimoto coefficient, or Mahalanobis distance,[52] measures the *dissimilarity* between the two molecules. Thus, the overall diversity function should be based on all pairwise distances between molecules in the

selected subset. The most diverse subset of points should be characterized by the maximum value of the diversity function. This property of the diversity function is reminiscent of the energy function for a physical system of limited size consisting of M positive point charges, which will be distributed as far away from each other as possible after the system reaches the minimum energy state. Thus, the reciprocal of the energy function of such a system has been modified and used as the diversity function D (eq 1)

$$D = \frac{1}{\sum_{i}^{M-1} \sum_{j>i}^{M} \frac{1}{d_{ij}^{a}}} \tag{1}$$

where d_{ij} is the distance between any two points of the subset in descriptor space and the summation is over all pairwise distances between the M selected points (molecules). Clearly, a larger value of D represents a more diverse and representative sampling. The power a was set to 1 in all of the experiments reported below; however, it could be set to any other value.

The SAGE Algorithm. The major objective of SAGE is to obtain a subset of M points (molecules) that are optimally diverse and representative in the descriptor space. Conceptually, we should compare the diversity values for all possible subsets of M compounds in order to obtain the most diverse subset. However, according to simple combinatorics, the number of all combinations of M objects selected from an available pool of N objects becomes very large for large values of N as indicated by eq 2.

$$C_N^M = \frac{N!}{M!(N-M)!} \tag{2}$$

Thus, an exhaustive evaluation of all the combinations of M points (molecules) is computationally prohibitive. In fact, Kuo[53] and Ghosh[54] have proved independently that the maximum diversity problem, including both MAXISUM (meaning that the sum of all pairwise distances is maximal) and MAXIMIN (meaning maximizing the minimal pairwise distance) formulations, is NP-hard. Therefore, more advanced optimization techniques are needed to ensure the effectiveness of the diversity sampling. In this paper, we have adapted simulated annealing (SA)[55,56] as an efficient stochastic optimization technique. The SA algorithm was implemented in SAGE as follows.

1. For real chemical database mining or library design, each of the N compounds in a database is represented by a vector of molecular descriptors and geometrically mapped to one point in a multidimensional space. In this paper a simulated data set of N points in 2D or 100D space was used in place of a database of chemical structures.

2. A subset of M points is randomly selected from N available points.

3. The diversity value ($D1$) for this subset of M points is evaluated according to eq 1.

4. A new subset of M points is obtained by replacing a fraction (e.g., one-third) of M points in the current subset by other points that are randomly selected from the available pool.

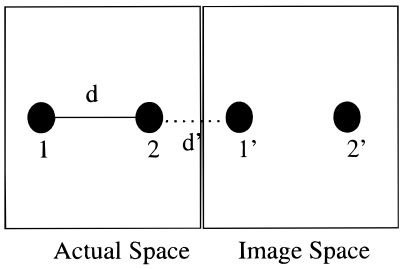

Figure 2. Definition of the edgeless distance using a periodic boundary condition. An "image space" in 2D is obtained by translating the actual data space in eight directions (toward four edges and four corners of the actual space) thus creating eight images of the original data space. If $d < d'$, then d is used as the distance between points 1 and 2; otherwise, d' is defined as the distance between points 1 and 2.

```
cal_dist_PBC(obj1, obj2){
        dist = 0.0;
        for(ii = 0; ii < dim; ii++){
                diff = fabs(dataset[obj1][ii] - dataset[obj2][ii]);
                if(diff > fabs(dim_size[ii] - diff)) diff = fabs(dim_size[ii]-diff);
                dist += diff*diff;
        }
        dist = sqrt(dist);
        return(dist);
}
```

Figure 3. Pseudocode for the calculation of edgeless distance in multidimensional space.

5. The diversity value ($D2$) for this new subset of M points is evaluated based on eq 1.

6. If $D2 \geq D1$, the new subset of M points (molecules) is accepted and used as the current selection. If $D2 < D1$, the new subset is accepted as the current selection only if the following Metropolis condition[56] is satisfied, i.e.,

$$\text{rnd} < e^{-(D1-D2)/T} \tag{3}$$

where rnd is a random number uniformly distributed between 0 and 1, and T is a parameter analogous to the temperature in the Boltzmann distribution law.

7. Steps 4−6 are repeated until the termination condition is satisfied. The temperature-lowering scheme and the termination condition used in this work have been adapted from Sun et al.[57] Thus, every time a new subset is accepted or when a preset number of successive iterations of selection does not lead to a better solution in terms of the diversity value, the temperature is lowered by 10% (the default initial temperature is 1000). The calculations are terminated when either the current temperature is lowered to the value of $T = 10^{-6}$ or the ratio between the current temperature and the temperature corresponding to the best subset found (i.e., with the highest diversity value so far) is equal to or less than 10^{-6}.

A modified version of SAGE was also implemented in which the distance between two points (molecules) was defined using periodic boundary conditions (PBC) as in edgeless Kohonen neural network,[58] so that the *edge effect* (see Results and Discussion) can be eliminated. The definition of edgeless distance is given in Figure 2 and a C-like pseudocode that implements PBC is given in Figure 3. This version of SAGE is referred to as SAGE/PBC.

Random Sampling of a Subset of *M* Points from a Data Set. In order to compare the results of SAGE with random sampling, M points have been selected randomly as follows.

A random number r between 1 and N was generated and the rth point of the simulated data set was selected. This process was repeated for M times until M different points (molecules) have been selected. These M points were then considered as a random sampling of the data set.

Simulated Data Sets. 1. *Simu1.* One thousand points were randomly generated which were uniformly distributed in 2D space. The ranges of both coordinates were [−3.0, 3.0]. The above ranges were chosen based on the consideration that most of the coordinate values would fall within [−3.0, 3.0] if autoscaling were used[59] in real chemical database mining or library design. The ranges can also be chosen arbitrarily for the purpose of the simulation. For example, they can lie within [0, 1] if range scaling is simulated.

2. *Simu2.* This data set was generated as follows. First, nine cluster centers were defined at different locations in 2D space with coordinate values within [−3.0, 3.0]. Second, a random number (between 1 and 100) of points for each cluster were generated randomly around each cluster center within a radius of 0.5. Finally, additional points were generated which were randomly distributed in the 2D space so that a total of 1000 points were obtained. This data set simulates the situation where clusters of molecules exist in their descriptor space and the number of members for each cluster is different, i.e., some regions are more densely populated than others are.

3. *Cluster1* and *Cluster2* Data Sets. A certain number of cluster centers were generated in a 2D or 100D geometrical space, which were away from each other by more than a preset distance (3 times larger than the cutoff radii for each cluster, see below). The ranges of all the coordinates were within [−3.0, 3.0]. Ninety-nine cluster centers were generated in 2D space for *Cluster1*, and 95 cluster centers were generated for *Cluster2* in 100D space. Then, a random number (between 1 and 100) of points for each cluster were generated randomly around each cluster center within a cutoff distance of 0.5, so that no members from two different clusters overlapped. This led to two simulated data sets, namely, *Cluster1* with 951 points distributed among 99 clusters in 2D space, and *Cluster2* with 950 points distributed among 95 clusters in 100D space. These data sets simulated the situations where many classes of compounds exist as separate clusters in the descriptor space. *Cluster2* simulated a more realistic situation in rational library design or database mining, where molecules are mapped onto multidimensional descriptor space.

Computer Simulation of Hit Rates for SAGE and Random Sampling. The major goal of the rational design of a diverse library is to increase the hit rates of lead compounds vs random sampling. Thus, we have applied SAGE and random sampling to a simulated data set (*Simu2*, see above) with 1000 nonuniformly distributed points in 2D space and compared the hit rates obtained by both methods. The definition of *active points* (i.e., active molecules) and *active clusters* as well as the definition of *individual hits* and *cluster hits* are described below followed by the discussion of the experimental design of the hit rate simulations.

Definition of Active Clusters. C active clusters with size R were defined by randomly placing C nonoverlapping circles of radius R into the 2D space of the *Simu2* data set. Points that happened to lie within each of the C circles were defined as *active points* whereas points outside the circles were defined inactive. Each cluster with active points is referred to as an *active cluster*. This design simulated an ideal data set where all active molecules are clustered together and away from inactive molecules.

Definition of an Individual Hit and a Cluster Hit. If a sampled point happened to be an active point, it was counted as an *individual hit*. Thus, the number of active points sampled by a particular method was defined as the *individual hit rate* obtained by that method. On the other hand, in order to characterize the representivity of a sampling, we defined a *cluster hit* as follows. If one or more points were sampled from an active cluster, this whole cluster was counted as a cluster hit. Therefore, the number of active clusters sampled by a particular method was defined as the *cluster hit rate* obtained by that method.

Experimental Design. In order to simulate different scenarios in real chemical database mining and chemical library design, several important factors, which could influence the hit rates, were examined. These variable factors include (1) the geometrical size of each active cluster (R), (2) the number of active clusters (C) in a data set, and (3) the locations of active clusters in the descriptor space. Thus, the design of our computational experiments was as follows.

C active clusters (with $C = 5$, 10, 15, ..., 30) of three different geometrical sizes ($R = 0.1$, 0.2, and 0.3) were defined for *Simu2* data set, and their locations in 2D space were defined randomly. For instance, for $R = 0.1$, we first defined $C = 5$ active clusters, 10 active clusters, etc. For each value of C, the locations of the active clusters in geometrical space could be different in real chemical data sets. To simulate this effect, the process of defining C active clusters was repeated 100 times using different random seeds, leading to 100 different distributions of active clusters in the 2D space. Then, for each of the 100 cases, both SAGE and random sampling were applied to sample M points (e.g., $M = 40$), and the individual hit rate and cluster hit rate (see definitions above) were determined for both sampling methods. Therefore, for both SAGE and random sampling, an average individual hit rate as well as an average cluster hit rate was obtained from 100 individual hit rates and 100 cluster hit rates, respectively. The same procedure was repeated for $M = 80$, 100, etc. Finally, the average individual and cluster hit rates for both methods were calculated for different values of M, different number of clusters $C = 5$, 10, 15, ..., 30 and different sizes of clusters, $R = 0.1$, 0.2, and 0.3.

RESULTS AND DISCUSSIONS

Visual Evaluation of SAGE Sampling. The purpose of this experiment was to visually evaluate the sampling results obtained by SAGE and SAGE/PBC. More specifically, we wanted to examine whether the points sampled by SAGE and SAGE/PBC were representative, i.e., distributed evenly in the geometrical space defined by all points. We have applied SAGE and SAGE/PBC to *Simu1 and Simu2* simulated data sets. In both cases, 100 points were sampled from an available pool of 1000 points. The results obtained with *Simu1*, where 1000 points are uniformly distributed in 2D space, are given in parts a and b of Figure 4 for SAGE and

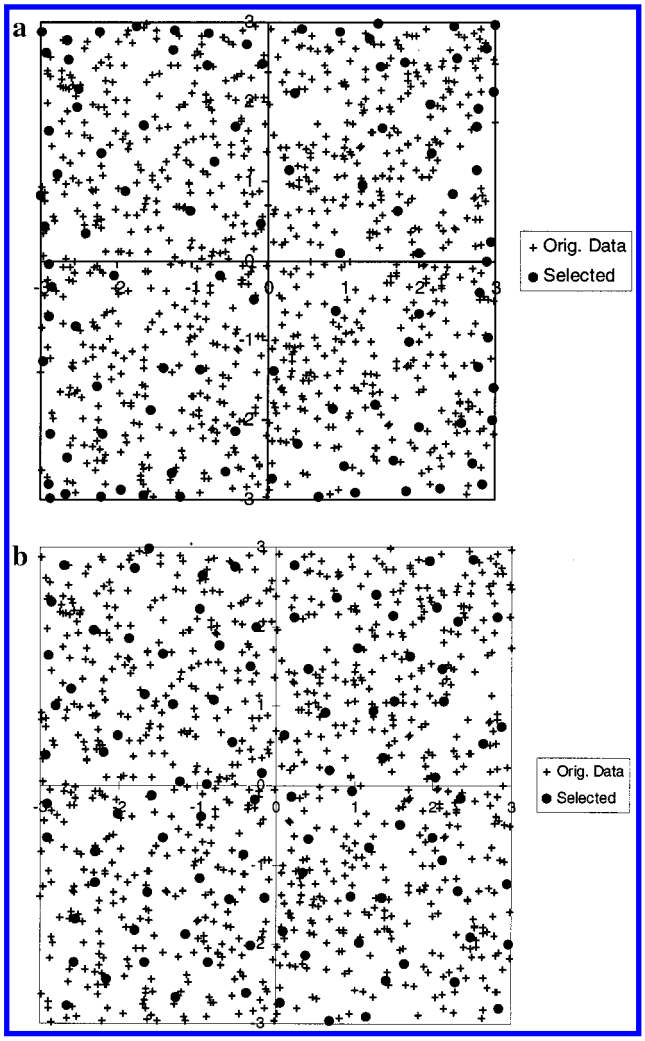

Figure 4. Selection of 100 points from data set *Simu1* using SAGE (a) and SAGE/PBC (b).

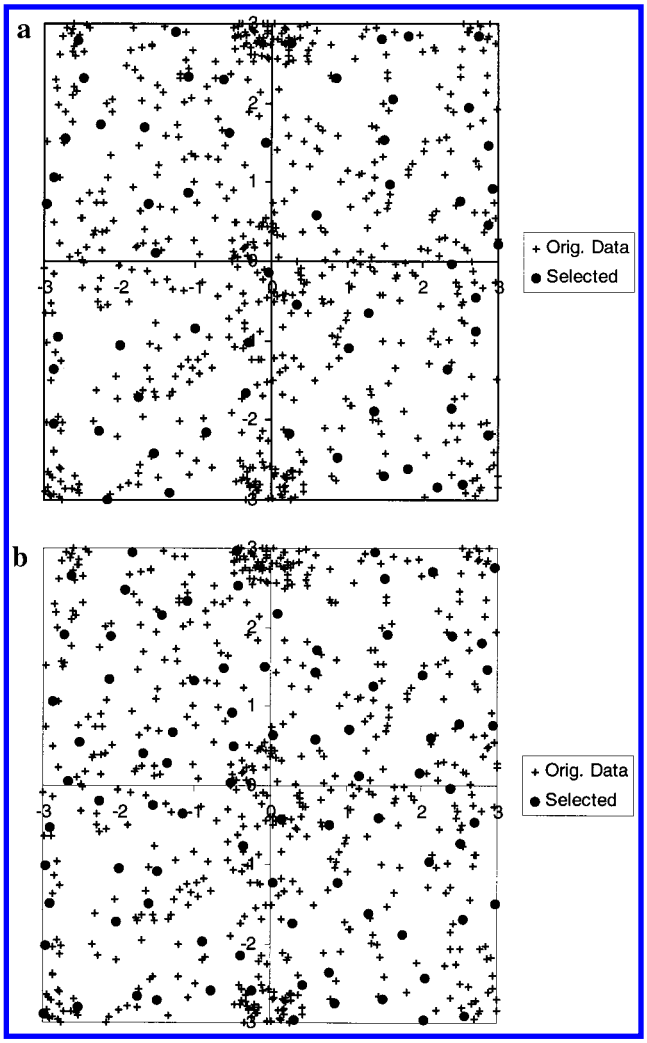

Figure 5. Selection of 100 points from data set *Simu2* using SAGE (a) and SAGEIPBC (b).

SAGE/PBC, respectively. In general, the points selected by SAGE are distributed evenly in 2D space (Figure 4a). However, a careful inspection of this distribution indicated that more points were sampled along the edges and fewer were sampled around the inner region of the space. We reasoned that this *edge effect* was due to the nature of our diversity function (cf. eq 1), which causes more points to occupy edges to increase the diversity value. The implementation of periodic boundary conditions in SAGE/PBC (cf. Computational Details) eliminated this effect: indeed, the points selected by this algorithm were distributed more uniformly upon visual inspection than those obtained by SAGE (cf. parts a and b of Figure 4).

Both SAGE and SAGE/PBC were also applied to the *Simu2* data set where 100 points were sampled from 1000 points, nonuniformly distributed in 2D space. The results are given in parts a and b of Figure 5 for SAGE and SAGE/PBC, respectively. Both methods appeared to obtain comparable results based on the visualization of the sampled points in 2D space, although SAGE/PBC (Figure 5b) seemed to perform slightly better than SAGE (Figure 5a). These results indicated that the difference between SAGE and SAGE/PBC might not be as significant in cases where original data points are not uniformly distributed in space.

Quantitative Analysis of the Space Coverage Obtained by SAGE and SAGE/PBC. One of the major goals in the design of diverse combinatorial libraries is to ensure that the sampled molecules (or points) represent as many classes of compounds (or points) as possible. Thus, the aim of this experiment was to determine how many clusters of points could be covered by SAGE and SAGE/PBC in comparison with random sampling, when a given number of points were sampled. Thus, we have applied SAGE, SAGE/PBC, and random sampling to two simulated data sets, *Cluster1 and Cluster2*. Each sampling method was applied 3 times using different random seeds. The results are given in Tables 1 and 2. When the number of sampled points was much smaller than that of clusters in the data set, there was virtually no difference between the three methods (cf. first row of Tables 1 and 2). This implies that when the number of sampled compounds was very small compared to that of the natural clusters in the data set, rational sampling could not provide any advantages over random sampling. As the number of

Table 1. Comparison of the Number of Clusters Covered by SAGE, SAGE/PBC, and Random Selection for the *Cluster1* Simulated Data Set

	SAGE			SAGE/PBC			random		
M	run 1	run 2	run 3	run 1	run 2	run 3	run 1	run 2	run 3
9	9	9	9	9	9	9	9	9	9
50	50	50	50	50	50	50	38	39	37
99	78	84	88	86	83	80	59	58	61

Table 2. Comparison of the Number of Clusters Covered by SAGE, SAGE/PBC, and Random Selection for the *Cluster2* Simulated Data Set

	SAGE			SAGE/PBC			random		
M	run 1	run 2	run 3	run 1	run 2	run 3	run 1	run 2	run 3
9	9	9	9	9	9	9	9	8	9
50	50	50	50	50	50	50	40	36	37
99	89	89	90	86	89	89	63	58	60

sampled points increased, both SAGE and SAGE/PBC covered more clusters than corresponding random sampling. For instance, when 50 points were sampled, both SAGE and SAGE/PBC covered 50 different clusters (100% of maximal diversity coverage) for both data sets, while random sampling covered less than 40 clusters (i.e., <80% of maximal coverage). For the first data set (Table 1), when 99 points were sampled, both SAGE and SAGE/PBC covered, on average, ca. 83 clusters (i.e., >83% of maximal coverage). However, for random sampling, less than 60 clusters were covered, amounting to <60% of maximal diversity coverage. For the second data set (in 100D space), similar results were obtained. For example, when 95 points were sampled, both SAGE and SAGE/PBC covered 86−90 clusters (i.e., 91−95% of the maximal coverage) while random sampling covered only 58−63 clusters (i.e., 61−66% of maximal coverage). Thus, we concluded that both SAGE and SAGE/PBC obtained a better diversity coverage than random sampling, since they could explore more clusters of points (corresponding to classes of compounds in real database mining) when the same amount of points (i.e., compounds) were sampled.

As in the previous experiment, it was also noted that for these two simulated data sets (i.e., *Cluster1* and *Cluster2*) both SAGE and SAGE/PBC performed equally well. Therefore, we concluded that the edge effect is not as important in both nonuniformly distributed and clustered data sets.

Computer Simulation of the Hit Rates: SAGE vs Random Sampling. The *Simu2* data set with active clusters of three different sizes ($R = 0.1, 0.2, 0.3$) was sampled, and the results are given in Figures 6−8, respectively. Each figure displays the average hit rates (either individual hit rates or cluster hit rates, according to the definitions above) obtained by SAGE and random sampling for different number of active clusters C in the data set. For instance, Figure 6a compares the average individual hit rates obtained by SAGE and random sampling for $C = 5$, 10, 15, 20, 25, and 30. For each C, both SAGE and random sampling were applied for different number of points M sampled, with $M = 40$, 80, etc. For instance, the bars corresponding to 5s and 5r compare the hit rates obtained by SAGE and random sampling for different number of M when $C = 5$. Similarly, the bars corresponding to 10s vs 10r compare the hit rates obtained by SAGE and random sampling for different number of M when $C = 10$.

When $R = 0.1$, the percentage of active points (active compounds) in the data set was in the range of 0.65−3.96%. In all cases of different number of C, the individual hit rates obtained by SAGE were higher than those obtained by corresponding random sampling (Figure 6a). An even better performance of SAGE vs random sampling was observed when cluster hit rates were considered. For all different numbers of C, the cluster hit rates obtained by SAGE were no less than 100% higher than those obtained by random sampling (Figure 6b).

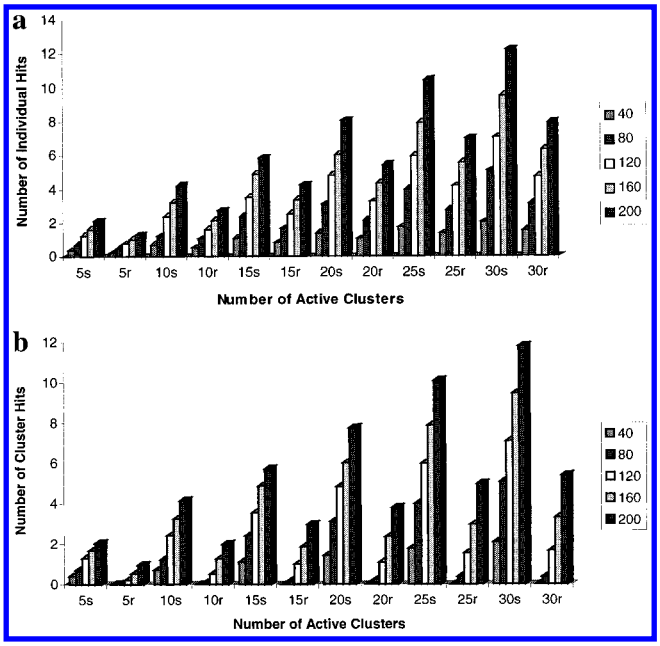

Figure 6. Comparison of individual (a) and cluster (b) hit rates obtained by SAGE (s) and random (r) sampling (cluster size $R = 0.1$) for different number of active clusters C in the *Simu2* data set (see text for further discussion).

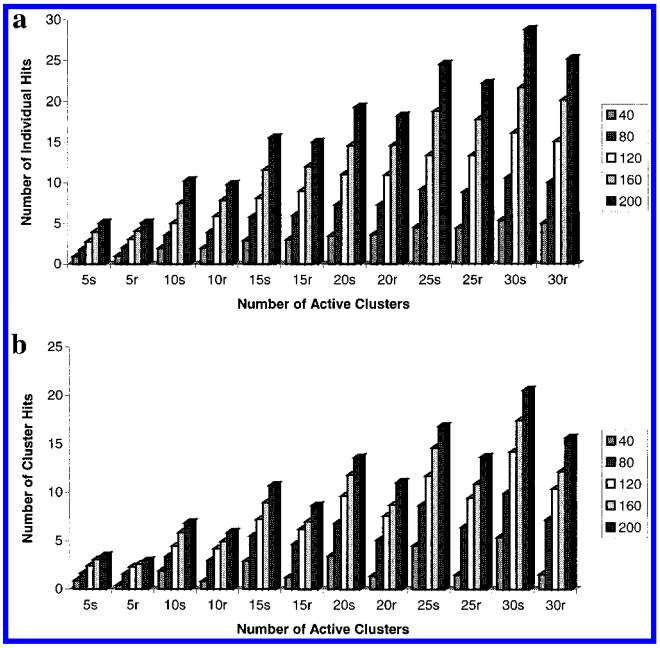

Figure 7. Comparison of individual (a) and cluster (b) hit rates obtained by SAGE (s) and random (r) sampling (cluster size $R = 0.2$) for different number of active clusters C in the *Simu2* data set (see text for further discussion).

With $R = 0.2$, the percentage of active points in the data set was in the range of 2.56−12.6%. When individual hit rates were considered, SAGE performed about the same as or slightly better than random sampling in all cases of different number of C (Figure 7a). However, when cluster hit rates were considered, SAGE performed clearly better than corresponding random sampling, especially when the number of C increased (Figure 7b).

When R increased to 0.3, the percentage of active points in the data set increased to 5−16.8%. In this case, SAGE

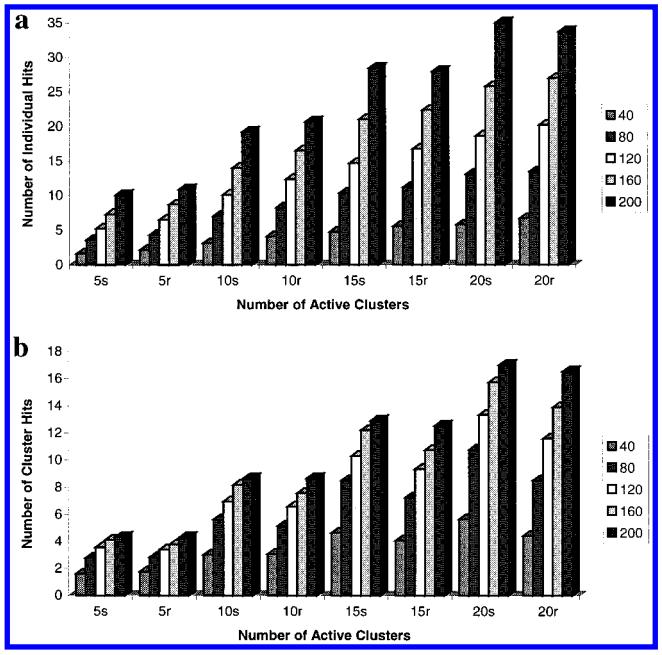

Figure 8. Comparison of individual (a) and cluster (b) hit rates obtained by SAGE (s) and random (r) sampling (cluster size $R = 0.3$) for different number of active clusters C in the *Simu2* data set (see text for further discussion).

performed about the same as or even slightly worse than the corresponding random sampling in terms of individual hit rates (Figure 8a). Furthermore, when cluster hit rates were considered, SAGE performed the same as or slightly better than random sampling, especially when the number of active clusters increased (Figure 8b).

The above observations imply that when the percentage of active compounds in the library or database is low (in the range of 0.65−4%), SAGE performs significantly better than random sampling. This is an encouraging result, since in most of the combinatorial chemical synthesis/high throughput screening projects the percentage of active compounds is as small as in this simulated case. However, if the percentage of active compounds in the data set increases, the number of active compounds (hits) obtained by random sampling increases proportionally. This suggests that when the percentage of active compounds in the library is relatively high, SAGE (and, probably, any other cluster sampling method) performs no better than random sampling in terms of *individual hit rate*. There is a common view that the poor performance of cluster sampling strategies is due to the use of nonideal descriptors. However, our simulations indicate that this result is due to the structure of the data set and is controlled by the content of active compounds (hits) regardless of the nature of the descriptors, since we have used simulated data sets in which descriptors were assumed *ideal*. Nevertheless, when *cluster hit rate* was considered as the criterion, SAGE performed better than or as well as random sampling in all tested cases, which indicated that the information content obtained by SAGE was always higher than that obtained from random sampling.

Another interesting observation was that with the increase of the number of active clusters C, the performance of SAGE also increased, both in terms of individual and cluster hit rates (Figures 6−8). This is, in fact, a desired feature in the design of "a universal library", which, by definition, will be tested against *many different* biological targets. In this case, if the data set contains active compounds for any target, these compounds should be distributed among different clusters in the descriptor space. It also implies that the most discriminating set of descriptors should be used to ensure that active compounds are grouped into more clusters, if only a higher individual hit rate is being sought.

SUMMARY

We have developed a novel computational tool (SAGE) for the diversity sampling of chemical databases. The two most crucial aspects of every diversity selection methodology are the use of representative molecular descriptors and the sampling algorithm. In this paper, we have separated these two aspects by applying our sampling method (SAGE) to various simulated data sets of geometrical points in 2- and 95-dimensional spaces. Furthermore, we have specifically addressed the issue of hit rates (i.e., the percentage of "active" compounds selected by a diversity sampling algorithm vs random selection). For this purpose, we designed perhaps practically impossible but theoretically important extreme cases of simulated chemical databases where all active compounds (points) belong to the same "active" cluster. Such extreme cases imply situations when ideal chemical descriptors are used to characterize chemical compounds so that all bioactive compounds appear similar to each other and dissimilar from inactive compounds in a data set.

The SAGE method selects a subset of optimally diverse compounds (represented as points in multidimensional descriptor space) from an available pool of existing compounds (points). One of the key components of this method is the diversity function (eq 1) that evaluates the global diversity of any selected subset of compounds (points). The most diverse subset is obtained by optimizing (maximizing) this function via sampling of different subsets of compounds (points) using simulated annealing as the stochastic optimization tool. There are several distance base diversity functions that can be considered, e.g., MAXISUM and MAXIMIN mentioned earlier. After some experimenting we chose the function in eq 1 due to its computational efficiency and its ability to provide the most even distribution of selected points. (For instance, in our preliminary experiments MAXISUM was biased toward selecting points in the corners of the data space.)

Initial application of SAGE to two simulated data sets showed that the subsets of selected points were both diverse and representative (based on a visual analysis) in 2D space. When SAGE was applied to two other simulated data sets with points distributed in many clusters, we found that SAGE afforded a higher coverage of clusters than corresponding random sampling. We also showed, by a computer simulation, that hit rates obtained by SAGE were always higher than those obtained by random sampling, when the percentage of active points (compounds) was low (0.65−4%). On the contrary, when the percentage of active compounds (points) was high (4−15%), the performance of SAGE, in terms of *individual hit rates*, was dependent upon how many active clusters existed in the descriptor space. Nevertheless, in all cases, SAGE performed either better than or as well as random sampling, when *cluster hit rate* was used as the criterion, suggesting that SAGE always obtained a higher information content than random sampling. This indicates

that SAGE is capable of selecting a nonredundant subset of compounds that adequately represent the whole volume of chemical diversity space. Another important implication is that SAGE works the best when active compounds are distributed in many different clusters in the descriptor space, which is the desired feature of "universal" libraries. We continue to investigate whether conclusions of this paper remain valid when SAGE is applied to real chemical databases with known structures and activities.

ACKNOWLEDGMENT

This work was supported in part by PHS grant MH 40537 and Center grants HD03310 and MH33127. W.Z. acknowledges the 1996 Award from the Chemical Structure Association Trust and the graduate assistantship from EPA/UNC Toxicology Research Program, Training Agreement #T901915, with the Curriculum in Toxicology, UNC-CH. A.T. acknowledges the research support from Rohm and Haas, Inc.

REFERENCES AND NOTES

(1) Gordon, E. M.; Barret, 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.

(2) Gallop, M. A.; Barret, R. W.; Dower, W. J.; Fodor, 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) Warr, W. A. Combinatorial Chemistry and Molecular Diversity. An Overview. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 134−140.

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

(5) Zheng, W.; Cho, S. J.; Tropsha, A. Rational Combinatorial Library Design. 1. Focus-2D: A New Approach to the Design of Targeted Combinatorial Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 251−258.

(6) Cho, S. J.; Zheng, W.; Tropsha, A. Rational Combinatorial Library Design. 2. Rational Design of Targeted Combinatorial Peptide Libraries Using Chemical Similarity Probe and the Inverse QSAR Approaches. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 259−268.

(7) Zheng, Q.; Kyle, D. J. Computational Screening of Combinatorial Libraries. *Bioorg. Med. Chem.* **1996**, *4* (5), 631−638.

(8) Kick, E. K.; Roe, D. C.; Skillman, A. G.; Liu, G.; Ewing, T. J.; Sun, Y.; Kuntz, I. D.; Ellman, J. A. Structure-Based Design and Combinatorial Chemistry Yield Low Nanomolar Inhibitors of Cathepsin D. *Chem. Biol.* **1997**, *4* (4), 297−307.

(9) Brown, R. D.; Martin, Y. C. Designing Combinatorial Library Mixtures Using A Genetic Algorithm. *J. Med. Chem.* **1997**, *40* (15), 2304−2313.

(10) Willett, P. Algorithms for the Calculation of Similarity in Chemical Structure Databases. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley & Sons, Inc.: New York, **1990**; pp 43−63.

(11) Fisanick, W.; Lipkus, A. H.; Rusinko, A., III. Similarity Searching on CAS Registry Substances. 2. 2D Structual Similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (1), 130.

(12) Fisanick, W.; Cross, K. P.; Rusinko, A., III. Similarity Searching on CAS Registry Substances. 1. Global Molecular Property and Generic Atom Triangle Geometric Searching. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (6), 664.

(13) Kearsley, S. K.; Sallamack, S.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (1), 118−127.

(14) Sheridan, R. P.; Miller, M. D.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (1), 128−136.

(15) Downs, G. M.; Willett, P. Similarity Searching in Databases of Chemical Structures. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers, Inc.: New York, 1996; Vol. 7, pp 1−65.

(16) Perry, N. C.; van Geerestein, V. J. Database Searching on the Basis of Three-Dimensional Molecular Similarity Using the SPERM Program. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (6), 607.

(17) Hagadone, T. R. Molecular Substructure Similarity Searching: Efficient Retrieval in Two-Dimensional Structure Databases. *J. Chem. Inf. Comput. Sci.* **1992**, *3* (5), 515.

(18) Judson, P. N. Structural Similarity Searching Using Descriptors Developed for Structure−Activity Relationship Studies. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (4), 1028.

(19) Gschwend, D. A.; Good, A. C.; Kuntz, I. D. Molecular Docking towards Drug Discovery. *J. Mol. Recognit.* **1996**, *9* (2), 175−186.

(20) Clark, D. E.; Westhead, D. R.; Sykes, R. A.; Murray, C. W. Active-Site-Directed 3D Database Searching: Pharmacophore Extraction and Validation of Hits. *J. Comput.-Aided Mol. Des.* **1996**, *10* (5), 397−416.

(21) Mizutani, M. Y.; Tomioka, N.; Itai, A. Rational Automatic Search Method for Stable Docking Models of Protein and Ligand. *J. Mol. Biol.* **1994**, *243* (2), 310−326.

(22) Sobolev, V.; Wade, R. C.; Vriend, G.; Edelman, M. Molecular Docking Using Surface Complementarity. *Proteins* **1996**, *25* (1), 120−129.

(23) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nonhierarchic Cluster Analysis Methods in Chemical Information Systems: Selection of Compounds for Bilogical Testing and Clustering of Substructures Search Output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109−118.

(24) Lawson, R. G.; Jurs, P. C. Cluster Analysis of Acrylates to Guide Sampling for Toxicity Testing. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 137−144.

(25) Hodes, L. Clustering a Large Number of Compounds. 1. Establishing the Method on an Initial Sample. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 66−71.

(26) Whaley, R.; Hodes, L. Clustering a Large Number of Compounds. 2. Using the Connection Machine. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 345−347.

(27) Shemetulskis, N. E.; Dunbar, J. B.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the Diversity of a Corporate Database Using Chemical Database Clustering and Analysis. *J. Comput.-Aided Mol. Des.* **1995**, *9* (5), 407−416.

(28) 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* (6), 644.

(29) Downs, G. M.; Willett, P.; Fisanick, W. Similarity Searching and Clustering of Chemical-Structure Databases Using Molecular Property Data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094.

(30) 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* (3), 572−584.

(31) Pearlman, R. S. Novel Software Tools for Addressing Chemical Diversity. *Network Sci.* **1996**, http://www.awod.com/netsci/Science/Combichem/feature08.html.

(32) 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* (9), 1431−1436.

(33) Agrafiotis, D. K. Stochastic Algorithms for Maximizing Molecular Diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 841−851.

(34) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and Visualization of Molecular Diversity and Combinatorial Libraries. *Mol. Diversity* **1996**, *2*, 64−74.

(35) Zheng, W.; Cho, S. J.; Waller, C. L.; Tropsha, A. Simulated Annealing Guided Evaluation (SAGE) of Diversity: A Novel Computational Tool for Diverse Chemical Library Design and Database Mining. *Book of Abstracts*, 213th National Meeting of the American Chemical Society, San Francisco, CA, 1997; American Chemical Society: Washington, DC, 1997; CINF-015.

(36) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating Biologically Active Compounds in Medium-Sized Heterogeneous Data Sets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (6), 1205−1213.

(37) Polinski, A.; Feinstein, R. D.; Shi, S.; Kuki, A. LiBrain: Software for Automated Design of Exploratory and Targeted Combinatorial Libraries. In *Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery*; Chaiken, I. M., Janda, K. D., Eds., ACS Conference Proceeding Series, **1996**; pp 219−232.

(38) Turner, D. B.; Tyrell, S. M.; Willett, P. Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18−22.

(39) Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular Diversity in Chemical Databases: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750−763.

(40) Pickett, S.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36* (6), 1214−1223.

(41) Myers, P. L.; Greene, J. W.; Saunders, J.; Teig, S. L. Rapid, Reliable Drug Discovery. *Today's Chemist at Work.* **1997**, *July/August*, 46−53.
(42) Good, A. C.; Lewis, R. New Methodology for profiling Combinatorial Libraries and Screening Sets: Cleaning Up the Design Process with HARPick. *J. Med. Chem.* **1997**, *40*, 3926−3936.
(43) Gillet, V. J., Willett, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf Comput. Sci.* **1997**, *37*, 731−740.
(44) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: a Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39* (16), 3049−3059.
(45) Matter H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40* (8), 1219−1229.
(46) 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), 1−9.
(47) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(48) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Research Studies Press: Chichester, England, 1986.
(49) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. In *Reviews in Computational Chemistry II*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1991; pp 367−422.
(50) 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.
(51) Downs, G. M.; Willett, P. Similarity Searching in Databases of Chemical Structures. In *Reviews in Computational Chemistry, Volume 7*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers, Inc., New York, 1996; pp 1−65.
(52) Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michotte, Y.; Kaufman, L. *Chemometrics: A Textbook*. Elsevier: Amsterdam, 1988.
(53) Kuo, C.-C.; Glover, F.; Dhir, K. S. Analyzing and Modeling the Maximum Diversity Problem by Zero-One Programming. *Decision Sci.* **1993**, *24* (6), 1171−1185.
(54) Ghosh, J. B. Computational Aspects of the Maximum Diversity Problem. *Oper. Res. Lett.* **1996**, *19*, 175−181.
(55) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.
(56) Kirkpatrick, S.; Gelatt, C. D., Jr.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671−680.
(57) Sun, L.; Xie, Y.; Song, X.; Wang, J.; Yu, R. Cluster Analysis By Simulated Annealing. *Comput. Chem.* **1994**, *18*, 103−108.
(58) Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem.* **1993**, *32* (4), 503.
(59) Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michotte, Y.; Kaufman, L. *Chemometrics: A Textbook*. Elsevier: Amsterdam, 1988.

CI980103P