Diversity and Coverage of Structural Sublibraries Selected Using the SAGE and SCA Algorithms

Charles H. Reynolds,*,† Alexander Tropsha,‡ Lori B. Pfahler,§,|| Ross Druker,§ Subhas Chakravorty,§,|| G. Ethiraj,‡ and Weifan Zheng§,#

The R. W. Johnson Pharmaceutical Research Institute, Welsh and McKean Roads, Spring House, Pennsylvania 19477, The Laboratory for Molecular Modeling, School of Pharmacy, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, and Rohm and Haas Company, 727 Norristown Road, Spring House, Pennsylvania 19477

Received April 23, 2001

It is often impractical to synthesize and test all compounds in a large exhaustive chemical library. Herein, we discuss rational approaches to selecting representative subsets of virtual libraries that help direct experimental synthetic efforts for diverse library design. We compare the performance of two stochastic sampling algorithms, Simulating Annealing Guided Evaluation (SAGE; Zheng, W.; Cho, S. J.; Waller, C. L.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 738−746.) and Stochastic Cluster Analysis (SCA; Reynolds, C. H.; Druker, R.; Pfahler, L. B. Lead Discovery Using Stochastic Cluster Analysis (SCA): A New Method for Clustering Structurally Similar Compounds *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 305−312.) for their ability to select both diverse and representative subsets of the entire chemical library space. The SAGE and SCA algorithms were compared using u- and s-optimal metrics as an independent assessment of diversity and coverage. This comparison showed that both algorithms were capable of generating sublibraries in descriptor space that are diverse and give reasonable coverage (i.e. are representative) of the original full library. Tests were carried out using simulated two-dimensional data sets and a 27 000 compound proprietary structural library as represented by computed Molconn-Z descriptors. One of the key observations from this work is that the algorithmically simple SCA method is capable of selecting subsets that are comparable to the more computationally intensive SAGE method.

INTRODUCTION

The ability to synthesize and screen large numbers of compounds has created a revolution in the search for biologically active compounds.[1−3] However, despite increases in high throughput screening and new parallel and combinatorial synthesis methods it is still necessary to make some choices with respect to synthesis targets. It is simply impossible to make and screen every compound that might conceivably be made. This has led to the recent interest in methods for designing structurally diverse lead generation libraries.[4−21] These methods for selecting diverse subsets of molecules make use of a wide variety of structural descriptors and search algorithms. Some are based on standard statistical techniques such as d-optimal design[22] or cluster analysis.[23] Other methods have been proposed that are less deterministic including simulated annealing, genetic algorithms, and various Monte Carlo search algorithms.[10,13]

Two quite different diversity selection methods have been developed independently in our laboratories. One method, SAGE,[24] uses simulated annealing to optimize a diversity function. The second, SCA,[25,26] uses a single random pass through the structural library to select a subset of compounds that fall outside a predefined similarity or distance cutoff with respect to one another. While the methods differ, the objective is basically the same: to select a diverse subset of compounds that are representative of the larger structural library.

The objective of this study was to compare these methods using both simulated and "real" structural data sets.[27] The comparison was made using the u- and s-optimal metrics[28] as independent assessments of the diversity (s-optimal) or coverage (u-optimal) of the selected subsets. This comparison allowed us to assess how well the simple, but very fast, SCA algorithm compares to the more compute intensive, but presumably more robust, simulated annealing based sampling method, SAGE.

DESCRIPTORS

Molecular topology descriptors[29−31] were generated using the Molconn-Z program.[32] All atom dependent descriptors or descriptors that are dependent on the specific atom number were omitted from consideration. All remaining descriptors were range scaled. The descriptors in the simulated data sets are simply arbitrary coordinates in a two-dimensional space. The u-optimal and s-optimal scores were calculated using algorithms available in the SAS PROC OPTEX (version 8) program.[28]

SAGE ALGORITHM

The idea behind the SAGE algorithm[21] is that optimizing the diversity value (see above) over different subsets of *M*

* Corresponding author phone: (215)628-5675; e-mail: Creynol1@prius.jnj.com.
† The R. W. Johnson Pharmaceutical Research Institute.
‡ The University of North Carolina at Chapel Hill.
§ Rohm and Haas Company.
|| Present address: Merck, West Point, PA 19486.
Present address: Glaxo-Smithkline, King of Prussia, PA.

10.1021/ci010041u CCC: $20.00
Published on Web 10/12/2001

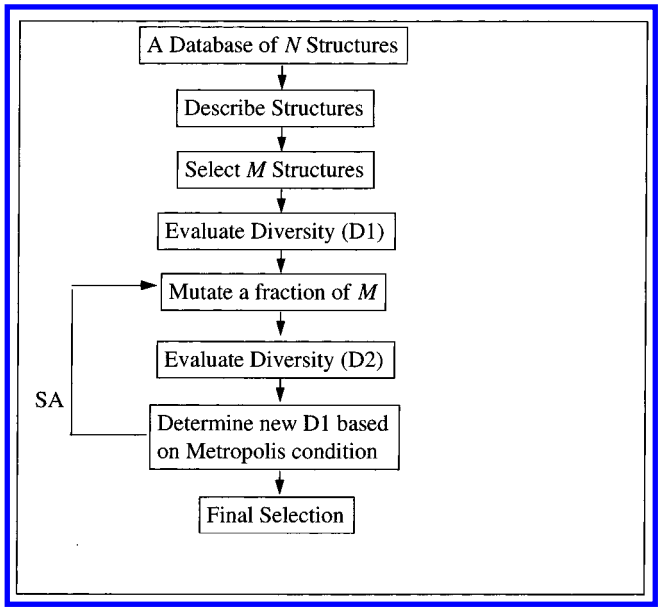

Figure 1. Outline of the SAGE diversity algorithm.

points (molecules), one should obtain the subset of *M* points (molecules) that are optimally diverse and representative in the descriptor space. Conceptually, one should compare the diversity values for all possible subsets of *M* compounds in order to get the most diverse subset. The diversity function was formulated as follows (eq 1)[21] where d_{ij} is the distance between any two points of the subset in the descriptor space.

$$D = \frac{1}{\sum_{i}^{m-1}\sum_{j>i}^{m}\frac{1}{\mathrm{dij}}_{a}} \tag{1}$$

The summation is over all pairwise distances between the *M* selected points (molecules). The reciprocal of the summation simply makes it true that a larger value represents a more diverse and representative sampling. Power *a* was set to 1 in all calculations.

According to simple combinatorics, the number of combinations for choosing *M* objects from an available pool of *N* objects (eq 2) will become huge when *N* becomes big.

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

Therefore, an exhaustive search of all the combinations of *M* points (molecules) is computationally expensive. In fact, Kuo and Ghosh have proved independently that the maximum diversity problem, including both MAXISUM and MAXIMIN formulations,[33,34] is NP-hard. Thus, more advanced optimization techniques are needed to ensure the effectiveness of the diversity sampling. SAGE uses Simulated Annealing (SA), which is one of the most efficient stochastic optimization techniques. The general flowchart of the algorithm is given in Figure 1, and a detailed description of the implementation of SA in SAGE is 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, simulated data sets of *N* points in 2D space were also used in place of a database of chemical structures.

2. A subset of *M* points (molecules) is randomly selected from *N* available points (molecules).

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

4. A new subset of *M* points (molecules) is obtained by replacing a fraction (e.g. 1/3) of the *M* points (molecules) in the current subset by other points (molecules) that are randomly selected from the available pool.

5. The diversity value (*D2*) for this new subset of *M* points (molecules) 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 is satisfied, i.e.

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

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

7. Steps 4−6 are repeated until the termination condition is satisfied. Thus, every time when a new subset is accepted or when a preset number of successive iterations of selection do 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}.

SCA ALGORITHM

The stochastic cluster analysis method (SCA) has been described elsewhere.[22] This algorithm does not entail any standard optimization such as simulated annealing or genetic algorithm. Instead SCA uses a randomized single pass through the data set combined with a user defined similarity cutoff to quickly extract a diverse sample of compounds. The first step in this process is to define the similarity cutoff (e.g. 0.75) that will be used as the maximum similarity score allowed between any two probe (selected) structures. A random structure is then selected from the database as a starting point in the diversity search. Beginning with the randomly chosen structure, the algorithm evaluates structures sequentially to determine if they fit the required condition of having a similarity score with all previously selected probes less than the cutoff. When a new probe is found that satisfies this condition, it is added to the list of probes and a new random starting point is selected. This process is repeated until there are no compounds left in the data set that can satisfy the selection criterion (i.e. no similarity score relative to any previously selected structure greater than the cutoff). The selected probe structures constitute a diverse subset, as defined by the SCA procedure.

If the objective is only to select a diverse subset, the program can be terminated after this step. If one wants to assign the remaining unselected compounds to clusters, a second step is initiated in which the unselected compounds are assigned to clusters based on similarity (using the same

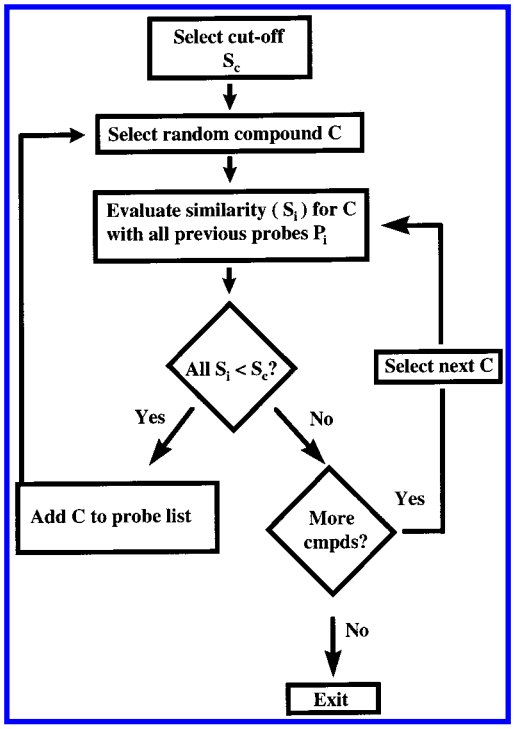

Figure 2. Outline of the SCA diversity algorithm.

similarity cutoff value). For our purposes here it was unnecessary to generate clusters. The flowchart for the diversity step in SCA is given in Figure 2. The essence of SCA is a random walk through the data set with radii of exclusion surrounding each selected compound (probe). This algorithm has the advantage of scaling linearly with the size of the data set. Thus, while it might not be expected to be as effective as more complex methods that use sophisticated optimization algorithms, it is very efficient computationally.

DATA SETS

To assess SAGE and SCA independently from any topology descriptors three simulated 2-D data sets were created for testing. These data sets contain 800−1000 points with distributions that range from relatively uniform to highly clustered. These data sets are convenient testbeds for comparing SAGE and SCA because they are small, have a well defined structure, and contain only two simulated descriptors. The latter property allows for simple visualization of the subset selections. The simulated data sets are denoted as DS1−DS3. Data set DS1 contains 1000 datapoints that are randomly (and fairly isotropically) distributed in the 2-D descriptor space. The distribution of data points in DS3 is less homogeneous than DS1 with many obvious clusters. Data set DS2 is intermediate between DS1 and DS3 in terms of the distribution of points in descriptor space. It is also slightly smaller with 812 datapoints. These three simulated data sets were all subjected to SCA and SAGE in order to determine if the distribution of datapoints in the overall descriptor space has any effect on either or both selection algorithms.

A fourth "real" database was also used to assess SAGE and SCA. This database consists of approximately 27 000 compounds from the Rohm and Haas structure archive. Each of these compounds has been evaluated for biological activity in a range of low-level screens. Based on previous analysis of these data set,[26] we believe it is relatively diverse in

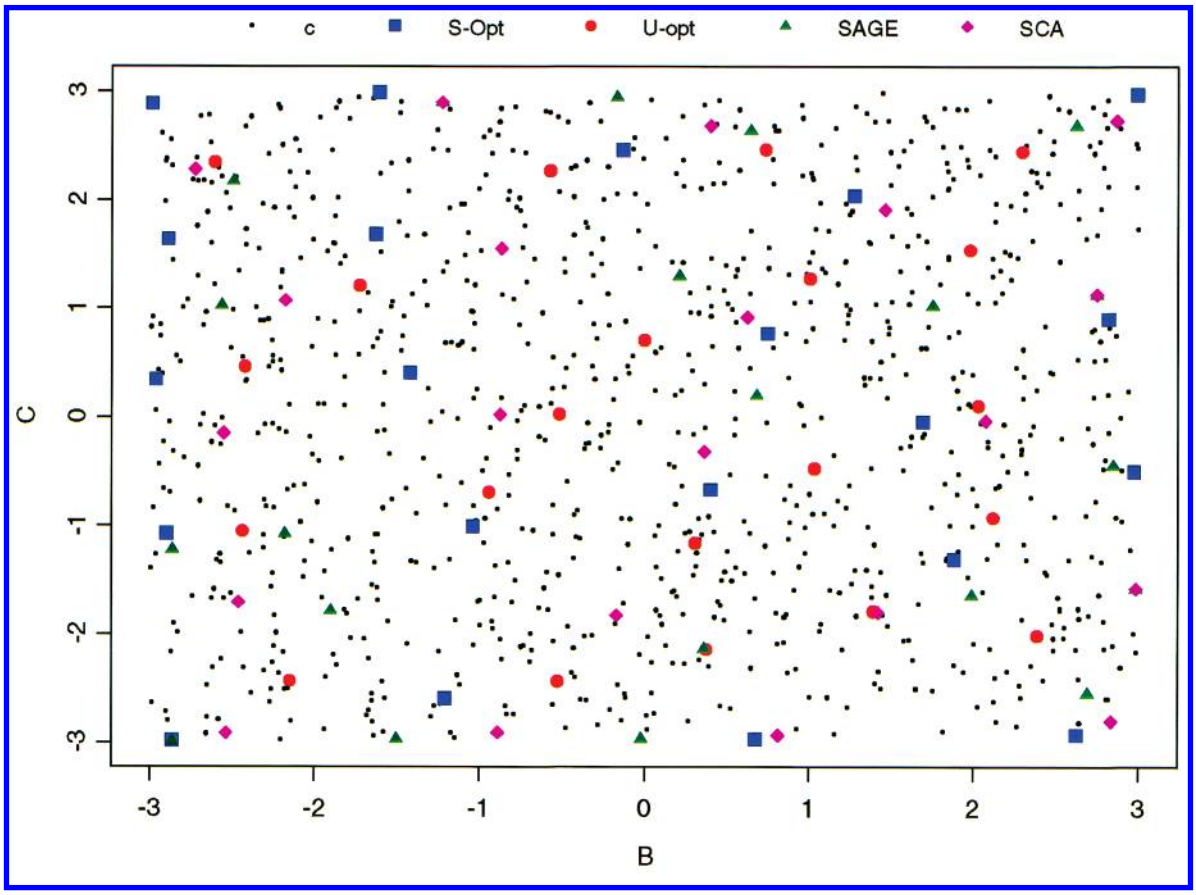

Figure 3. Subsets selected by SAGE, SCA, and the u- and s-optimal routines in SAS for DS1.

chemistry and contains both very unique structures (singletons) and large collections of very similar analogues. As such it should provide a reasonable test case for SAGE and SCA.

RESULTS AND DISCUSSION

Diverse subsets were generated for simulated data set DS1 using both SAGE and SCA. The selection criteria (number of compounds and distance cutoff for SAGE and SCA, respectively) were set in order to give the same number of selections (21) for both methods. The results of the SAGE and SCA runs for DS1 are shown graphically in Figure 3. Visual inspection indicates that both methods result in a fairly diverse subset. SCA appears to give better coverage, i.e., some regions appear to be under represented by SAGE.

While informative, simple visual inspection of a distribution of points on a graph is not an objective method for comparing diversity algorithms. Therefore, we wanted a method for evaluating the quality of the subsets that is objective, unrelated to our algorithms, and generally accepted. The problem of selecting a diverse or representative subset of datapoints from a larger set is not at all unique to chemistry and has led to a number of statistical techniques for making and assessing subset selections. We have chosen to use two statistical design metrics, namely s-optimal and u-optimal scores, to evaluate the diversity and coverage provided by the SAGE and SCA selected subsets. These scoring algorithms are given in eqs 4 and 5 below.[28]

$$U - opt = \sum_{x \in C} d(x,D) \tag{4}$$

$$S - opt = \frac{N_D}{\sum_{y \in D} \frac{1}{d(y,(D-y))}} \tag{5}$$

In eqs 4 and 5: C is the set of all possible compounds (the full library), D is the selected set, N_D is the number of compounds in the selected set, $d(y, D-y)$ is the minimum distance of a selected compound y to the selected set (D) excluding y, and $d(x,D)$ is the minimum distance of a selected compound x to the selected set (D). Thus u-opt seeks to minimize the sum of the distance from each compound in D to the entire library in C, while s-opt seeks to maximize the harmonic mean distance from each compound in D to all the other compounds in D (maximum spread).

The u-optimal and s-optimal scores for the SAGE and SCA subsets are given in Table 1 along with the SAS PROC OPTEX[28] optimized u- and s-optimal based selections. In addition, the selected datapoints from optimized u- or s-optimal selection are displayed graphically in Figure 3. Additional runs were carried out using SCA and SAGE to select larger subsets from DS1. These results are also given in Table 1. When comparing the results in Table 1 it is important to remember that smaller u-optimal scores are better (better coverage), while larger s-optimal scores are better (greater diversity).

Comparison of the SAGE and SCA scores in Table 1 for data set DS1 shows that SCA actually gives better results than SAGE for this data set. Since both methods are stochastic in nature, the results may differ slightly from run to run. Indeed the SCA s- and u-optimal scores are sandwiched between the SAGE scores and the scores obtained by optimizing either the s- or u-optimal criteria using SAS.

Table 1. Comparison of the s- and u-Optimal Scores for the Simulated Data Sets, DS1−DS3

data set	similarity distance cutoff	criterion	score u-opt	score s-opt
DS1 (random)	0.200	SCA	0.092	0.218
	21 probes	SAGE	0.108	0.167
		u-optimal	0.083	0.173
		s-optimal	0.102	0.237
		poor selection	0.630	0.014
		random[a] (av)	0.115	0.084
	0.100	SCA	0.048	0.110
	66 probes	SAGE	0.056	0.073
		u-optimal	0.044	0.091
		s-optimal	0.048	0.117
		random[a] (av)	0.061	0.042
	0.0500	SCA	0.022	0.057
	209 probes	SAGE	0.029	0.026
		u-optimal	0.019	0.048
		s-optimal	0.023	0.060
		random[a] (av)	0.028	0.022
DS2 (intermediate)	0.200	SCA	0.088	0.210
	18 probes	SAGE	0.103	0.161
		u-optimal	0.074	0.148
		s-optimal	0.090	0.247
		poor selection	0.365	0.007
		random[a] (av)	0.112	0.068
	0.100	SCA	0.046	0.144
	54 probes	SAGE	0.051	0.071
		u-optimal	0.038	0.082
		s-optimal	0.045	0.124
		random[a] (av)	0.057	0.036
	0.0500	SCA	0.021	0.059
	157 probes	SAGE	0.024	0.030
		u-optimal	0.018	0.044
		s-optimal	0.022	0.062
		random[a] (av)	0.028	0.018
DS3 (clustered)	0.200	SCA	0.082	0.210
	18 probes	SAGE	0.105	0.122
		u-optimal	0.069	0.136
		s-optimal	0.096	0.231
		poor selection	0.411	0.008
		random[a] (av)	0.111	0.072
	0.100	SCA	0.046	0.112
	46 probes	SAGE	0.061	0.049
		u-optimal	0.041	0.071
		s-optimal	0.051	0.129
		random[a] (av)	0.059	0.039
	0.0500	SCA	0.023	0.058
	135 probes	SAGE	0.026	0.031
		u-optimal	0.020	0.044
		s-optimal	0.023	0.060
		random[a] (av)	0.029	0.021

[a] The average over 100 separate random selections should provide a good representation of the full data set, even though in many cases the individual random selections are very poor with regard to coverage and diversity. See Figures 9 and 10.

The same procedure was carried out for data sets DS2 and DS3. SCA and SAGE were used to select 18 and 21 datapoints from these data sets. The results are given graphically in Figures 4 and 5. The s- and u- optimal scores for these selections are also given in Table 1. As was true for DS1, both SAGE and SCA appear to give reasonably representative, and diverse, subsets of DS2 and DS3. For example, both methods select at least one representative of every cluster in DS3, with the exception of one cluster that is not sampled by SAGE. In any event, both selections look

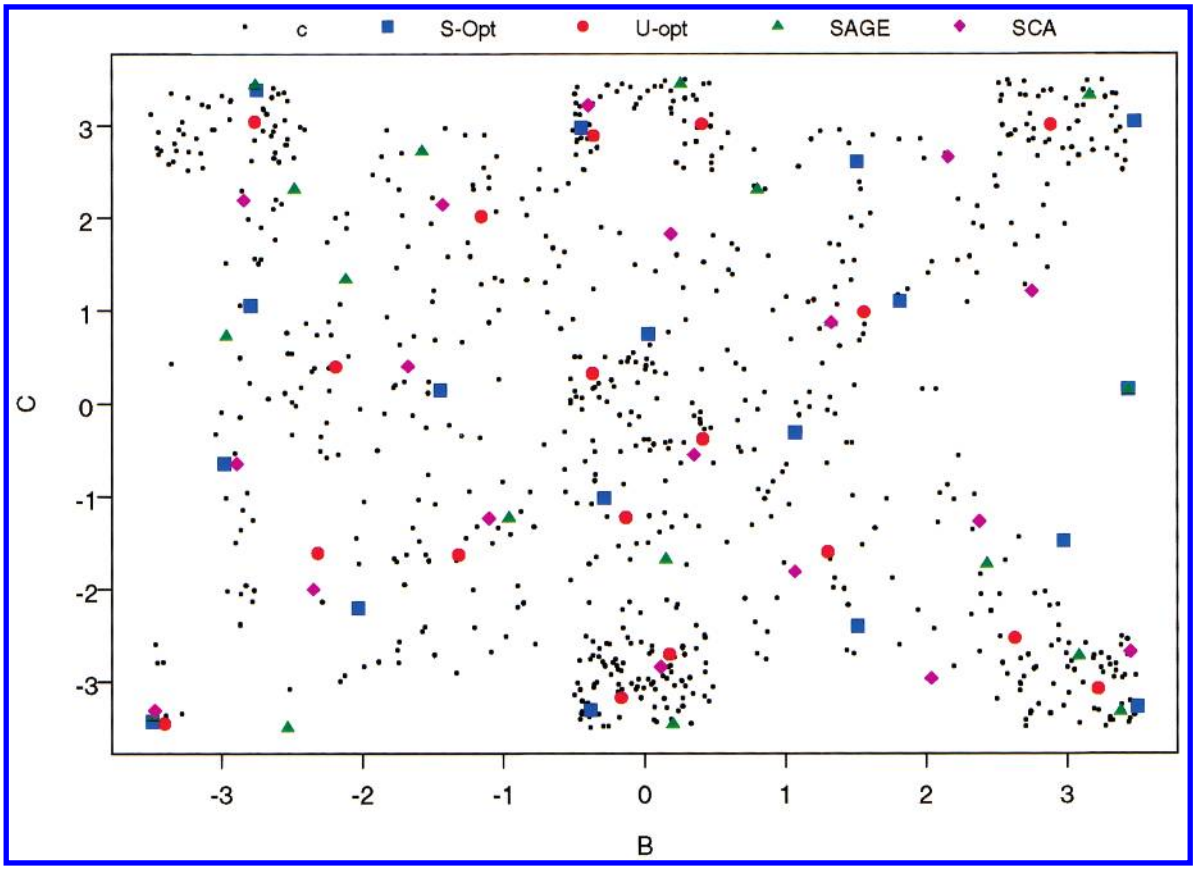

Figure 4. Subsets selected by SAGE, SCA, and the u- and s-optimal routines in SAS for DS2.

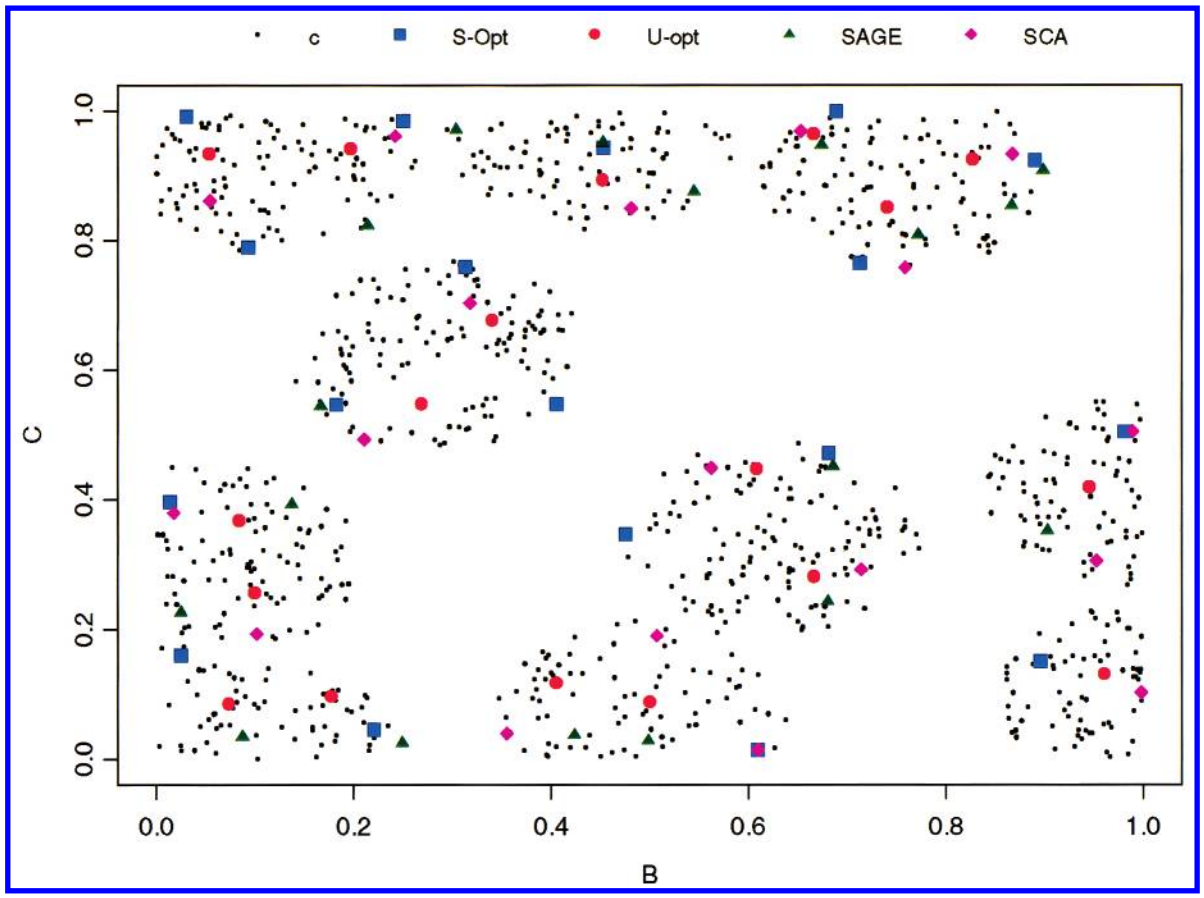

Figure 5. Subsets selected by SAGE, SCA, and the u- and s-optimal routines in SAS for DS3.

very reasonable visually. Further, the u- or s-optimal results (obtained by optimization within SAS) while better quantitatively do not look qualitatively different when viewed graphically (Figures 4 and 5).

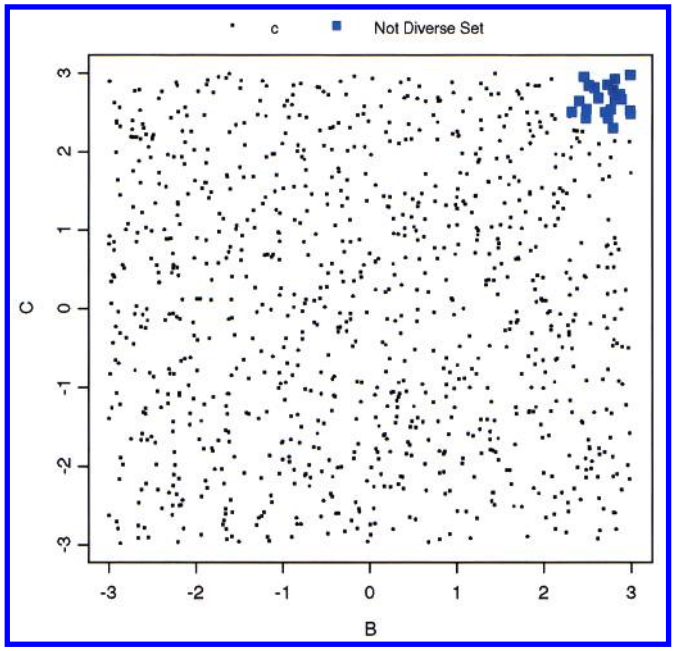

Figure 6. Intentionally bad subset selected from data set DS1.

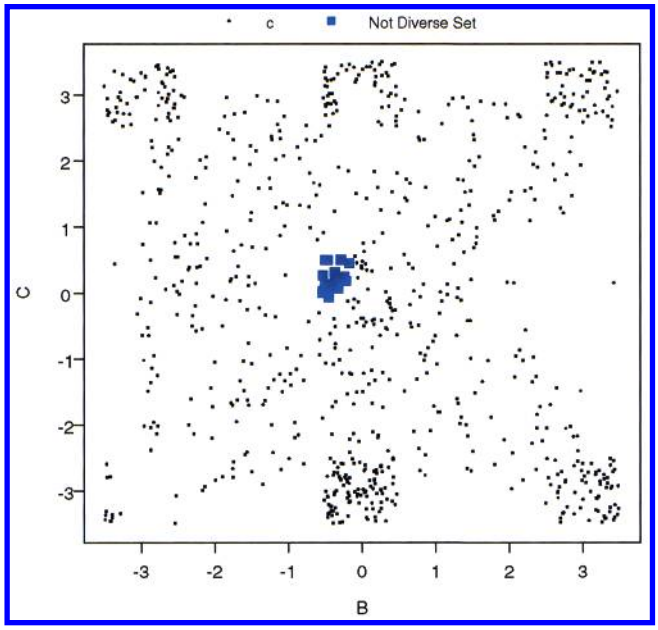

Figure 7. Intentionally bad subset selected from data set DS2.

To put the SCA and SAGE results into perspective subsets for DS1−DS3 were created where the selection was made deliberately poor. These can be seen in Figures 6−8. The u- and s-optimal scores for these obviously very poor subsets are also given in Table 1 for comparison. For example, the u-optimal score for this intentionally poor subset from data set DS1 is more than six times larger than for any other subset. Similarly, the s-optimal score for this "bad" subset is eight times smaller than for any other subset of DS1. (Again, one should keep in mind that smaller u-optimal and larger s-optimal scores are favored.) As a final assessment of SAGE and SCA, we computed u- and s-optimal scores for 100 random selections from data set DS3. The distribution of u-optimal and s-optimal scores for these 100 random selections are given in Figures 9 and 10. Bars are overlaid on the histograms showing where the scores fall for SAGE, SCA, u-optimal (optimized), and s-optimal (optimized)

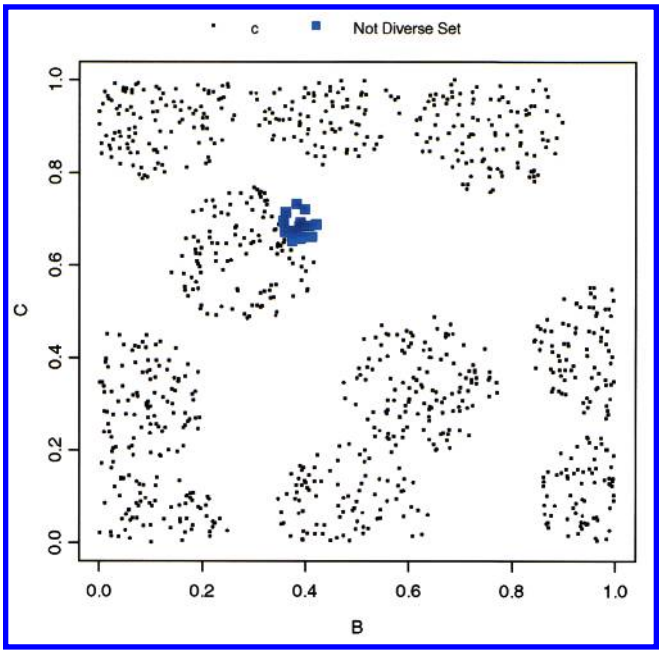

Figure 8. Intentionally bad subset selected from data set DS3.

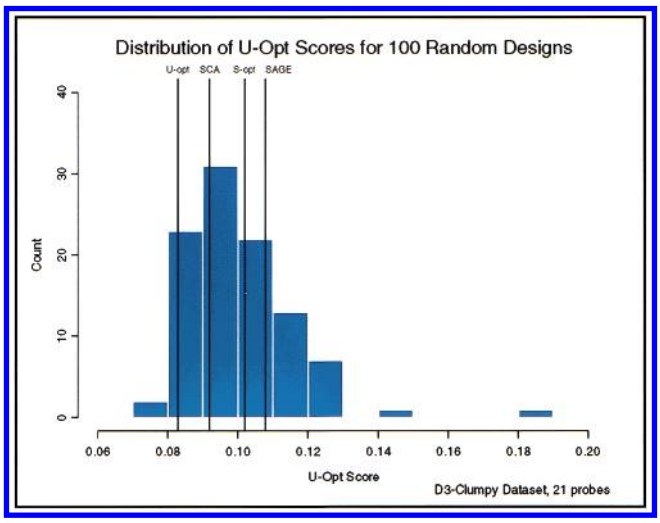

Figure 9. Histogram of 100 random selections from data set DS3. Vertical bars show u-optimal scores for u-optimal design (SAS), SCA, and s-optimal design (SAS) and SAGE, respectively.

selections. For the u-optimal scores (Figure 9) the maximum in the distribution of u-optimal scores for the random selections falls below the s-optimal (optimized) and SAGE scores. SCA and, of course, the u-optimal (optimized) subsets are better with only approximately 25% of the random selections being as good as SCA and only 2% better than u-optimal (optimized). Thus, while some random selections provide good coverage, many random selections are very poor relative to any of the systematic algorithms. This graphic illustrates the risks inherent in random selection as a picking strategy. Comparison of the s-optimal scores is even more revealing (Figure 10). All of the random selections are much poorer than any of the systematic selections (i.e. SAGE, SCA, u-optimal, or s-optimal). These graphs provide compelling evidence that both SCA and SAGE are able to give consistently superior subsets relative to random selection,[35] particularly with regard to diversity.

In general, the u-optimal metric measures coverage (i.e. representative subset selection), and the s-optimal metric

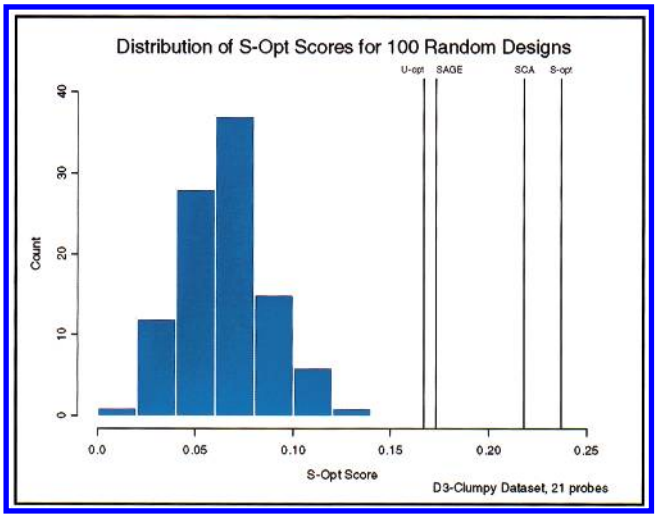

Figure 10. Histogram of 100 random selections from data set DS3. Vertical bars show s-optimal scores for u-optimal design (SAS), SAGE, SCA, and s-optimal design (SAS), respectively.

measures diversity (among members of the selected subset). With this in mind, it is logical that random selection, while inferior to the systematic methods, at least has a better chance of being representative than being diverse. This is especially true when one considers clumpy data sets, such as DS3. During random selection any compound has an equal chance of being picked whether it is in a cluster or not. On average, this will lead to multiple selections from large clusters and lower diversity. An algorithm that is designed to optimize diversity will be biased against multiple selections from a cluster preventing selections that degrade diversity. It is, therefore, important to keep in mind what the objective is when extracting subsets. Is the goal to represent the original data set as effectively as possible, with higher representation going to larger clusters, or is the goal a diverse subset of the original data set with as little redundancy as possible. The best selection criterion/criteria may be quite different for these two goals.

The 27 000 compound data set extracted from the Rohm and Haas corporate collection represents a more realistic test for the diversity sampling methods; it also allows us to verify some conclusions about diversity selection made on the basis of simulated data sets. In both cases the structural library was represented using 312 molecular topology descriptors calculated by Molconn-Z (see Supporting Information). Even though we have observed many collinearities between different Molconn-Z descriptors, the entire set of range-scaled values were used rather than a reduced set or the most significant principal components. Our primary interest in this study was not to validate any particular set of descriptors but to assess our diversity algorithms.

The SAGE and SCA results for the Rohm and Haas library are summarized in Table 2. Again three different cutoff values (in the case of SCA) or requested subset sizes (in the case of SAGE) were used to select subsets ranging from 1345 to 3875 compounds. In this case the size of the data set precludes calculating *optimized* u- or s-optimal scores from SAS. However, it is possible to calculate u- and s-optimal scores for the selections from SCA and SAGE. Comparison of the u- and s-optimal scores for SCA and SAGE in Table 2 shows some interesting differences between the two methods. The u-opt scores for SAGE are better than the u-opt scores for SCA for all three subsets. However, the s-optimal scores for SCA tend to be better than the s-optimal scores for SAGE. The practical interpretation of these results is that SCA appears to provide a subset that has poorer coverage than SAGE (u-optimal score) but greater diversity than SAGE (s-optimal score).

Table 2. Comparison of s- and u-Optimal Scores for the Rohm and Haas Library

similarity distance cutoff	criterion	u-opt score	s-opt score
0.900	SCA	0.648	1.010
1345 probes	SAGE	0.514	0.423
0.800	SCA	0.563	0.901
1982 probes	SAGE	0.481	0.363
0.650	SCA	0.698	0.736
3875 probes	SAGE	0.647	0.294

These results are very informative with respect to diversity analysis. Based on this comparison it appears that the very computationally fast (i.e. cpu-minutes on an R10000) SCA method gives results that are comparable to the more computationally intensive (i.e. cpu-days) simulated annealing based SAGE method. Thus SCA can be expected to provide a reasonably diverse subset for large data sets where more accurate, but more complex, methods such as SAGE are too computationally expensive. Further, these results argue for using a fast method, such as SCA, as a starting point for more robust, optimization techniques, such as SAGE.

Finally, we would like to point out that SCA has another advantage, particularly relative to typical clustering algorithms, when selecting representative samples from very large data sets (e.g. corporate structural archives). This advantage is that it is not necessary to begin from scratch each time one adds new compounds to the database. SCA can be used to determine if new compounds fall into unique clusters or clusters that have already been defined without recomputing the entire database. This is a significant factor when the goal is to cluster a large corporate database that is growing rapidly.

CONCLUSIONS

SAGE and SCA have been compared using a variety of simulated and real structural data sets We employed the s-optimal and u-optimal scores as reasonable and unbiased reference values for comparing results of diversity sampling with respect to optimal chemical space representation. This comparison shows that SAGE might have an advantage in terms of selecting compounds that are representative (coverage) of the complete data set, while SCA may give greater structural variety (diversity) between selected datapoints (compounds). Thus it is important to consider the objective when selecting a subset. Depending on the circumstance, the goal might be to represent the original data set as effectively as possible, with higher representation going to larger clusters, or the goal might be to select a maximally diverse subset of the original data set with as little redundancy as possible. The best selection algorithm and criteria are likely to be quite different for these two different objectives.

Both SAGE and SCA are far superior to random selection, particularly with regard to diversity. One significant result is the finding that the computationally efficient SCA method provides reasonable subset selection with regard to coverage

and diversity. As such it can be used reasonably in cases where the amount of data is too great for an optimization method, such as SAGE, or may be used as a fast method for generating a very reasonable starting point for further optimization.

We also suggest that the s-optimal and u-optimal scores serve as unbiased reference values for comparing different diversity and clustering algorithms.

Supporting Information Available: A list of the 312 Molconn-Z descriptors. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Gordon, E. M.; Kerwin, J. F. *Combinatorial chemistry and molecular diversity in drug discovery*; John Wiley & Sons: New York, 1998.
(2) Warr, W. Combinatorial Chemistry and Molecular Diversity. An Overview. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 134−140.
(3) Nielsen, J. Combinatorial chemistry. *Chem. Ind. (London)* **1994**, 902−5.
(4) 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*, 3049−3059.
(5) 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*, 1431−6.
(6) Ramaswamy, N.; Bauman, N.; Haraki, K. S. Database diversity assessment: new ideas, concepts and tools. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 447−452.
(7) Bures, M. G.; Martin, Y. C. Computational methods in molecular diversity and combinatorial chemistry. *Curr. Opin. Chem. Biol.* **1998**, *2*, 376−380.
(8) Van Drie, J. H.; Lajiness, M. S. Approaches to virtual library design. *Drug Discovery Today* **1998**, *3*, 274−283.
(9) 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.
(10) Agrafiotis, D. K. Stochastic Algorithms for Maximizing Molecular Diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 841−851.
(11) Brown, R. D.; Martin, Y. C. Designing Combinatorial Library Mixtures Using a Genetic Algorithm. *J. Med. Chem.* **1997**, *40*, 2304−2313.
(12) Gillet, V. J.; Willett, P.; Bradshaw, J.; Green, D. V. S. Selecting Combinatorial Libraries to Optimize Diversity and Physical Properties. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 169−177.
(13) Clark, R. D. OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.
(14) Kubinyi, H.; Combinatorial, C. Combinatorial and computational approaches in structure-based drug design. *Curr. Opin. Drug Discovery Dev.* **1998**, *1*, 16−27.
(15) Martin, Y. C.; Brown, R. D.; Bures, M. G. Quantifying diversity. *Comb. Chem. Mol. Diversity Drug Discovery* **1998**, 369−385.
(16) 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*, 1219−1229.
(17) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(18) Schnur, D. Design and Diversity Analysis of Large Combinatorial Libraries Using Cell-Based Methods. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 36−45.
(19) Waldman, M.; Hassan, M. Novel metrics for the optimization of molecular diversity of combinatorial libraries. *Book of Abstracts*, 215th ACS National Meeting, Dallas, March 29-April *2* 1998, COMP-072.
(20) Zheng, W.; Hung, S. T.; Saunders, J. T.; Johnson, S. R.; Seibel, G. L. *A web-based computational tool for combinatorial library design that simultaneously optimizes multiple properties*; *Book of Abstracts, 220th ACS National Meeting*, Washington, D.C., 2000, CINF 53.
(21) Zheng, W.; Hung, S. T.; Saunders, J. T.; Seibel, G. L. *Piccolo: A tool for combinatorial library design via multicriterion optimization*; Pacific Symposium on Biocomputing, 2000; Vol. 5, pp 585−596.
(22) Martin, E. J.; Critchlow, R. E. Beyond Mere Diversity: Tailoring Combinatorial Libraries for Drug Discovery. *J. Comb. Chem.* **1999**, *1*, 32−45.
(23) 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*, 644−9.
(24) Zheng, W.; Cho, S. J.; Waller, C. L.; Tropsha, A. Rational Combinatorial Library Design. 3. Simulated Annealing Guided Evaluation (SAGE) of Molecular Diversity: A Novel Computational Tool for Universal Library Design and Database Mining. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 738−746.
(25) 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, 211th ACS National Meeting*, New Orleans, LA, March 24−28, 1996, COMP-047.
(26) Reynolds, C. H.; Druker, R.; Pfahler, L. B. Lead Discovery Using Stochastic Cluster Analysis (SCA): A New Method for Clustering Structurally Similar Compounds. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 305−312.
(27) Tropsha, A.; Reynolds, C. H.; Lappe, M.; Ethiraj, G.; Zheng, W.; Chakravorty, S.; Druker, R.; Pfahler, L. B. *Comparing the SCA and SAGE methods for assessing chemical diversity*; *Book of Abstracts, 216th ACS National Meeting*, Boston, August 23−27, 1998, COMP-081.
(28) *SAS/QC Software: Usage and Reference, Version 6*, 1st ed.; SAS Institute: Cary, NC, 1995; Vol. 1.
(29) Hall, L. H.; Kier, L. B. *The molecular connectivity chi indexes and kappa shape indexes in structure−property modeling*; *Reviews in Computational Chemistry*, Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1991; Vol. 2, pp 367−422.
(30) 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.
(31) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The electrotopological state*; Academic Press: San Diego, CA, 1999.
(32) *Molconn-Z*, 3.15 ed.; Edusoft: Ashland, VA.
(33) Ghosh, J. B. Computational aspects of the maximum diversity problem. *Operations Res. Lett.* **1996**, *19*, 171−181.
(34) Kuo, C.-C.; Glover, F.; Dhir, K. S. Analyzing and modeling the maximum diversity problem by zero-one programming. *Decision Sciences* **1993**, *24*, 1171−1185.
(35) It has been pointed out previously that the difference between diverse and random selections disappear as the sample size becomes very small relative to the full data set. Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graphics Modeling* **1998**, *15*, 372−385.

CI010041U