Stochastic Algorithms for Maximizing Molecular Diversity

Dimitris K. Agrafiotis[†]

3-Dimensional Pharmaceuticals, Inc., 665 Stockton Drive, Suite 104, Exton, Pennsylvania 19341

Received May 15, 1997[⊗]

A common problem in the emerging field of combinatorial drug design is the selection of an appropriate subset of compounds for chemical synthesis and biological evaluation. In this paper, we introduce a new family of selection algorithms that combine a stochastic search engine with a user-defined objective function that encodes any desirable selection criterion. The method is applied to the problem of maximizing molecular diversity, and the results are visualized using Sammon's nonlinear mapping algorithm. By separating the search method from the performance metric, the method can be easily extended to perform complex multi-objective selections in advanced decision-support systems.

INTRODUCTION

In recent years, combinatorial chemistry and high throughput screening have revolutionized the way in which new drug candidates are being discovered. As it is practiced today, combinatorial chemistry is used merely as a source of compounds for mass screening. While this approach is very powerful, it still does not address the key, rate-limiting step in drug discovery which is the elaboration of sufficient SAR around a lead compound and the refinement of its pharmacological profile. Recently,[1,2] we presented a blueprint of an integrated system that permits the automatic chemical synthesis, refinement, and elaboration of bioactive compounds through the tight integration of high-speed parallel synthesis, structure-based design, and chemi-informatics. This system, known as DirectedDiversity, is an iterative optimization process that explores combinatorial space through successive rounds of selection, synthesis, and testing. Unlike traditional combinatorial approaches where the entire library is made and tested in a single conceptual step, DirectedDiversity physically synthesizes, characterizes, and tests only a portion of that library at a time. The selection of compounds is carried out by computational search engines that combine optimal exploration of molecular diversity with a directed search based on SAR information accumulated from previous iterations of the integrated machinery.

A central task of DirectedDiversity, and indeed any library design system, is to select an appropriate set of compounds (or building blocks) for physical synthesis and biological evaluation. In the absence of any structural or SAR information, a common strategy has been to identify a subset of compounds from some "virtual" collection that represents the molecular diversity that is present in the larger population. A number of different strategies have been used in the past to achieve this goal. A common approach is to define a set of clusters by some kind of clustering methodology[3−5] and then extract a diverse set by selecting a representative from each cluster.[6] Lajiness et al.[7,8] and Polinsky et al.[9] have employed a recursive procedure known as maximin, in which one starts with a randomly chosen compound and gradually builds up the diversity list by examining the remaining compounds and selecting the one that is most different from the already selected members. A major disadvantage of this algorithm is that it scales to the square of the compounds being considered, although improvements were recently reported by Holliday and Willett based on the cosine coefficient of similarity.[10,11] When the dimensionality is sufficiently low, the data space can be partitioned into a set of evenly distributed hypercubes, and a representative can be selected from each hypercube to make up the diversity list. This approach, which offers significant computational advantages, has been employed by Cummins[12] and Pearlman[13] who relied on factor analysis and B-Cut values, respectively, for constructing a low-dimensional representation of chemical space. Martin and co-workers[14] used principal component analysis and multidimensional scaling to derive a compact multivariate representation and then employed a d-optimal experimental design procedure to perform the diversity selection. Lin[15] has also reported a diversity metric based on information theory, but we later demonstrated[16] (using the algorithms described herein) that this approach had a strong and general tendency to oversample remote areas of the feature space and produce unbalanced designs. For more extensive reviews on diversity profiling, the reader is referred to Martin,[17] Martin,[18] and Agrafiotis.[19]

From a computational perspective, this problem consists of two parts. First, we need a *measure* to quantify the diversity of any conceivable subset of compounds from a given collection, and, second, we need efficient *search algorithms* for identifying the optimal (*i.e.*, most diverse) set from among the vast number of possibilities. In this paper, we introduce two different measures for quantifying diversity (one already in widespread use, and the other of our own device) and present a new solution to the selection problem based on simulated annealing. The method is tested using four artificial data sets and a collection of three combinatorial libraries characterized by means of spatial autocorrelation vectors, generously provided to us by Prof. J. Gasteiger and Dr. M. Wagener. To aid the analysis and confirm the performance of our algorithm, the results are visualized using Sammon's nonlinear mapping algorithm. Our goal is to demonstrate that the method is robust, general, and extensible and can be readily adapted to perform complex multiobjective selections in advanced experimental design systems. A more systematic discussion including a

[†] Tel: (610) 458-6045. Fax: (610) 458-8249. E-mail: dimitris@3dp.com.

[⊗] Abstract published in *Advance ACS Abstracts,* September 1, 1997.

S0095-2338(97)00033-4 CCC: $14.00

comparison of a large number of “diversity” functions along with alternative search algorithms such as evolutionary programming and genetic algorithms will be presented elsewhere.[20] This paper is of algorithmic nature; no attempt is made to validate these algorithms against their ability to separate biologically active from inactive compounds. For a discussion on the choice and validation of molecular descriptors for diversity profiling, the reader is referred to Brown,[21] Patterson,[22] and Delaney[23] and references therein.

METHODS

Sammon Projections. To visualize the results of our selection algorithm and assess its performance in higher-dimensional spaces, we used a multidimensional scaling algorithm developed by Sammon.[24] This technique has been employed by Barlow and Richards to display protein folding topologies in two dimensions[25] and most recently by the author for visualizing protein sequence relationships.[26]

Sammon mapping approximates local geometric relationships of vectorial samples in a two- or three-dimensional plot. In particular, given a finite set of n-dimensional samples $\{\mathbf{x}_i, i = 1, 2, ..., k; \mathbf{x}_\mathrm{i} \in R^n\}$, a distance function $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ between $\mathbf{x}_i$ and $\mathbf{x}_j$, and a set of images of $\mathbf{x}_i$ on a k-dimensional display plane $\{\mathbf{r}_i, i = 1, 2, ..., k; \mathbf{r}_i \in R^k\}$, the objective is to place $\mathbf{r}_i$ onto the display plane in such a way that their Euclidean distances $||\mathbf{r}_i - \mathbf{r}_j||$ approximate as closely as possible the corresponding values d_{ij}. This projection, which can only be made approximately, is carried out in an iterative fashion by minimizing an error function, $E(m)$, which measures the difference between the distance matrices of the original and projected vector sets

$$E(m) = \frac{\displaystyle\sum_{i<j}^{k} \frac{[d_{ij}* - d_{ij}(m)]^2}{d_{ij}*}}{\displaystyle\sum_{i<j}^{k} d_{ij}*} \qquad (1)$$

where m is the iteration number. $E(m)$ is minimized using instantaneous gradients and a steepest-descent algorithm. Details of this algorithm can be found elsewhere.[24-26]

Sammon mapping is ideally suited for both metric and nonmetric scaling. The latter is particularly useful when the (dis)similarity measure is not a true metric, i.e., it does not obey the distance postulates and, in particular, the triangle inequality (such as the Tanimoto coefficient[3]). Although an “exact” projection is only possible when the distance matrix is positive definite, meaningful projections can still be obtained even when this criterion is not satisfied.[27,28] In this case, the quality of the projection is determined by a sum-of-squares error function such as eq 1 or Kruskal’s “stress”:[27,28]

$$S = \sqrt{\frac{\displaystyle\sum_{i<j}^{k} (d*_{ij} - d_{ij})^2}{\displaystyle\sum_{i<j}^{k} d_{ij}^{\,2}}} \qquad (2)$$

The usefulness of multidimensional scaling stems from the fact that data in R^d are almost never d-dimensional. Although scaling becomes more problematic as the *true* dimensionality of the space increases, the presence of structure in the data is very frequently reflected on the resulting map. This was clearly demonstrated in our recent analysis of 390 protein kinase domains,[26] using as a property vector the 339 amino acids that comprised each of the aligned sequences. Indeed, we found that the Sammon maps (which represented the compression of a 339-dimensional space) were able to capture the essential features of the distance matrix and revealed clusters that were consistent with the known substrate specificities of these proteins. Of course, one can easily conceive of situations where Sammon mapping is not effective, particularly when the data is random. Fortunately, these situations rarely arise in practice, as some form of structure is always present in the data (see examples below).

Finally, we must point out that Sammon mapping (and multidimensional scaling in general) is not limited to two or three dimensions. Martin[14] has employed classical multidimensional scaling techniques to compress 1133 2048-bit fingerprints into seven-dimensional vectors of continuous variables that reproduced all 642 000 pairwise similarities with a relative standard deviation of only 10%. Experiments with fingerprints and other forms of binary data carried out in our laboratories have confirmed these findings.[29]

Selection. The selection problem can be formulated as follows: given an n-membered virtual library and a number k, find the k most diverse compounds in that population. The number of different k-membered subsets of an n-membered set is given by the binomial

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \qquad (3)$$

In effect, the selection of the most diverse set of compounds involves a search over the k-membered subset space of the n-membered set. As is immediately apparent from eq 3, the problem is NP-complete, and the cardinality of that space is enormous even for the most conservative cases encountered in real combinatorial drug design. This means that the solution cannot be determined by enumerating every possible k-membered subset but must be found using an alternative technique. In the system described here, this search is carried out using a stochastic search algorithm based on simulated annealing. Although in the general case k itself is also subject to optimization, in the DirectedDiversity system[1] it is known a priori and represents the number of compounds that can be made in parallel in a single iteration of the automated synthetic machinery (typically an integral multiple of 96 compounds per run).

Diversity Measure. In the system presented here, each compound is represented by a d-dimensional vector of independent, orthogonal, continuous variables, computed either directly in the form of molecular descriptors or indirectly through multidimensional scaling, principal component preprocessing, or some other equivalent technique. Although the nature, scaling, information content, and covariance of those descriptors is of paramount importance if the solution is to be chemically and statistically relevant, this issue does not affect the operation of the algorithm and will not be addressed here. As will become apparent below, orthogonality and normalization are particularly important, as they affect the size of the unit hypercube and, thus, the operation of the algorithm.

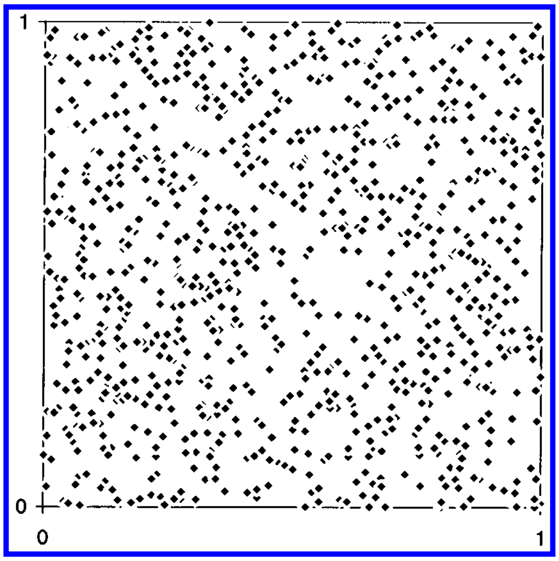

Figure 1. Data set 1.

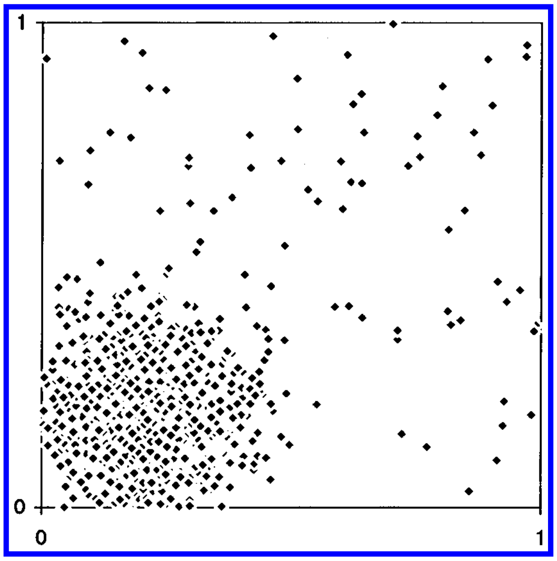

Figure 2. Data set 2.

Almost every selection algorithm reported to date employs some concept of chemical distance, which measures the similarity or dissimilarity between two compounds.[3-12] Implicit in most of these approaches is the fact that each selected compound is used as a representative of some neighborhood or domain in feature space. Indeed, if one could define and associate a neighborhood of chemical space to each compound in the virtual collection, the most diverse set would clearly be the one that exhibited the least amount of overlap between the neighborhoods associated with each of its elements.

In more formal terms, we define the total diversity volume, V_{tot}, as the d-dimensional volume spanned by the input variables in the n-membered virtual library. If k is the number of compounds to be selected, the neighborhood of each compound is defined by dividing the total volume into k hypercube partitions whose volume is given by

$$V_0 = \frac{V_{\mathrm{tot}}}{k} \tag{4}$$

For purposes of scaling, each feature is usually normalized from 0 to 1, in which case $V_{\mathrm{tot}} = 1$ and $V_0 = 1/k$. Each compound in the n-membered library is then associated with a hypercube of size V_0, centered at that point. The diversity $f(S)$ of any given k-membered subset, S, of the virtual library can be expressed simply as the fraction of the total volume that is occupied by the hypercubes centered at each point in S:

$$f(S) = \frac{V_S}{V_{\mathrm{tot}}} \tag{5}$$

It is evident that the most diverse k-membered subset can be identified by maximizing the function $f(S)$ over the domain of S. Alternatively, we can minimize the function $g(S) = 1 - f(S)$, which measures the degree of overlap between the neighborhoods associated with the elements of S. In theory, volume computations should be confined within the boundaries of the feature space defined by the virtual library. An edge effect would need to be applied, so that if a point is close to the boundaries of the confining hypercube, the portion of its neighborhood that extends beyond those boundaries is not taken into consideration. In practice, however, this is not usually necessary for reasons that will become apparent below.

While a straightforward application of this principle works well in two or three dimensions, it becomes more problematic as the dimensionality of the space increases. The following examples illustrate why. Consider, for instance, the fraction of the volume of a d-dimensional hypercube contained within the inscribed hypersphere:[30]

$$f_d = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \tag{6}$$

For $d = 1, 2, 3, 4, 5, 6$, and 7, f_d is 1, 0.785, 0.524, 0.308, 0.164, 0.081, and 0.037, respectively. It is clear that as d increases, the center of the hypercube becomes insignificant and its volume is concentrated in its corners. This apparent paradox has also been demonstrated by Wegman[31] by considering the hypervolume of a thin shell, i.e., the volume contained within two concentric hyperspheres, one with radius r and the other with a slightly smaller radius, $r - \epsilon$. The fraction of the volume of the larger sphere contained within the two spheres is given by

$$f_s = \frac{V_d(r) - V_d(r - \epsilon)}{V_d(r)} = 1 - \left(1 - \frac{\epsilon}{r}\right)^d \xrightarrow[d\to\infty]{} 1 \tag{7}$$

Thus, for higher dimensions, the volume of the hypersphere is concentrated on its surface. These two simple examples illustrate that the concept of “neighborhoods” in higher dimensions is somewhat distorted: if the neighborhoods are “local”, they are virtually empty; if they are not empty, then they are not “local”. This has important consequences in many diversity profiling applications.

To avoid these complications and in the interest of performance, we approximate the overlap between a pair of points by a simple function of their distance. The hypercubes given in eq 4 are now replaced by hyperspheres, whose radius is approximated by eq 8

$$r_0 = d\sqrt{\frac{1}{k}} \tag{8}$$

where d is the dimensionality of the (normalized) feature

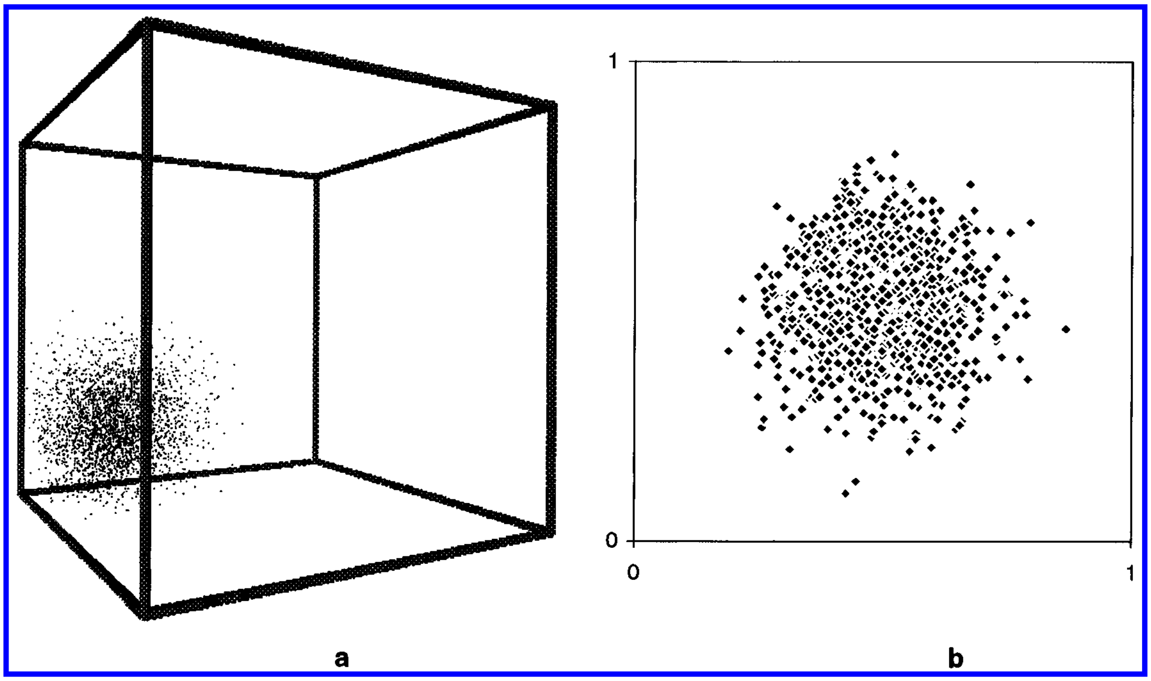

Figure 3. Data set 3: a, 3D scatter plot and b, Sammon projection.

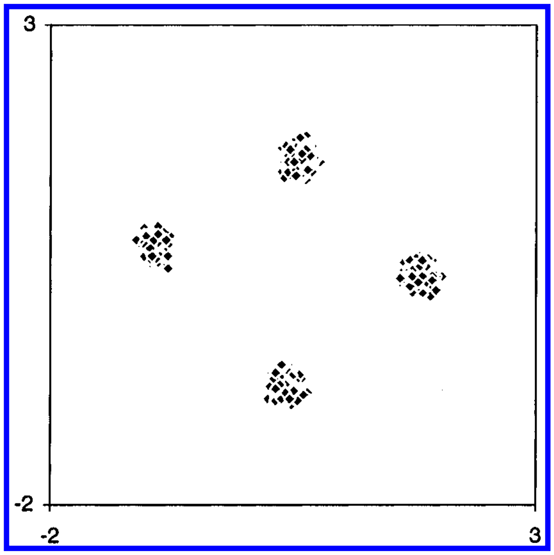

Figure 4. Sammon projection of data set 4.

space. The actual penalty function is given by eq 9

$$g(S) = \sum_{i<j}^{k} O(i,j) + \sum_{i}^{k} E(i) \tag{9}$$

where i, j are used to index the elements of S, k is the cardinality of S, and $O(i,j)$ and $E(i)$ are given by eq 10 and 11, respectively

$$O(i,j) = \begin{cases} \min(100, 2r_0/r_{ij} - 1), & r_{ij} \le 2r_0 \\ 0, & r_{ij} > 2r_0 \end{cases} \tag{10}$$

and

$$E(i) = \lambda(1 - \sum_{j=1}^{d} e_{ij}/dr_0) \tag{11}$$

where r_{ij} is the distance between the ith and jth points, e_{ij} is the distance of the ith point from the nearest boundary along the jth coordinate truncated at r_0, and λ is a scaling factor.

The penalty function in eq 9 consists of two terms: an overlap penalty, $O(i,j)$, and an edge penalty, $E(i)$. The overlap term in eq 10 is essentially a reciprocal repulsion function, whose nonlinear functional form serves to over-penalize close contacts and keep points above some minimum separation. This term vanishes beyond contact, i.e., if the distance between two points is larger than the sum of the radii of their hyperspheres, the overlap penalty is zero regardless of how far these points are from each other.

The value of λ is used to scale the edge and overlap terms with respect to each other and represents the maximum edge penalty applied to a point located at one of the corners of the feature space. For $\lambda = 1$, the maximum edge penalty is equal to the overlap penalty at separation r_0. In nonuniform or sparse distributions where a suboptimal choice must be made, this scheme ensures that the edge effect by itself will not cause the selection algorithm to choose points below some minimum separation and thus lead to close contacts. In this study, λ is (empirically) set to $(d - 1)/d$, where d is again the dimensionality of the input space. In practice, and for the reasons outlined above, the edge effect is of secondary importance and can be neglected.

The diversity measure described above is used here merely as an example; it works well in practice but is limited in the sense that it requires a multivariate representation of chemical space. As pointed out by one of the reviewers, there are representations (such as binary fingerprints) that are not readily amenable to this kind of analysis. To illustrate that additional diversity functions can be easily accommodated by the algorithm, we have also tested the well-known maximin function

$$f(S) = \max_{i}(\min_{j \neq i}(d_{ij})) \tag{12}$$

or its variant

$$f(S) = \sum_{i} (\min_{j \neq 1}(d_{ij})) \tag{13}$$

where S is any given k-membered subset of the n-membered virtual library, and i, j are used to index the elements of S.

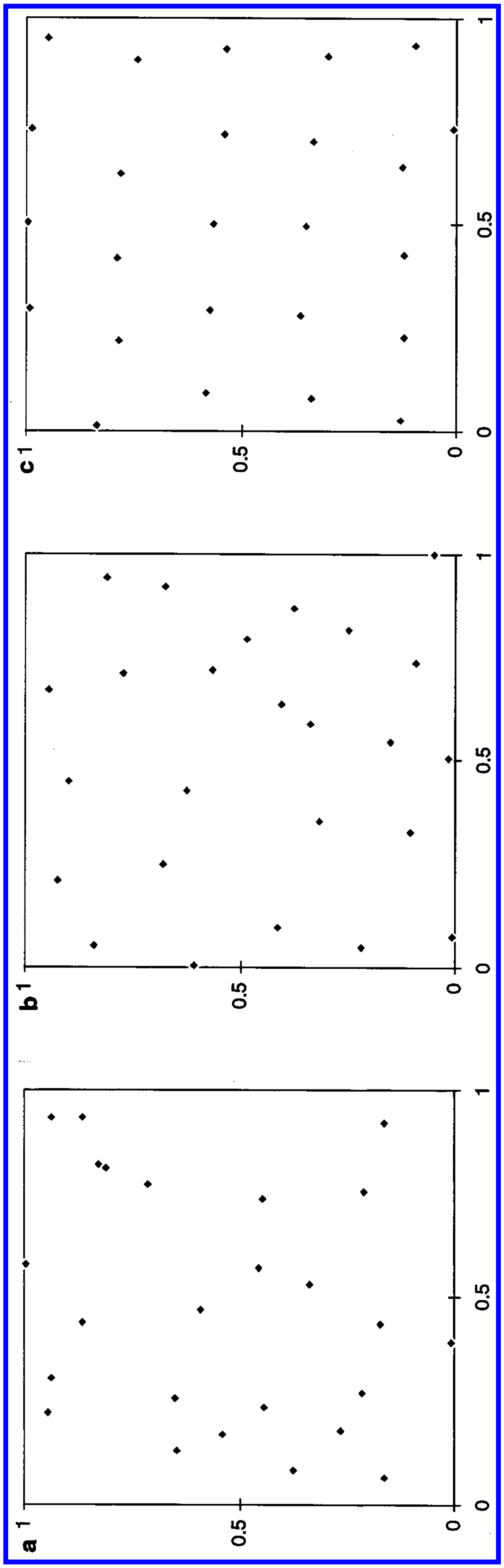

Figure 5. Twenty-five point selection from data set 1 (2D): a, initial state; b, best state found after first cycle; and c, best state found after 31st cycle.

This function has the advantage that it can be used with any conceivable dissimilarity index and does not require a metric space. In practice, we have found eq 13 to be smoother and thus much easier to optimize in a Monte-Carlo environment. Since the electronic release of this paper, more diversity functions have been tested, some with surprising results.[16] A more systematic comparison of alternative diversity functions and search algorithms will appear elsewhere.[20]

Simulated Annealing. The task of searching for the optimal (most diverse) set involves a search over the entire set of k-membered subsets of the n-membered virtual library, for the one that minimizes the penalty function $g(S)$. As we explained before, this is an NP-hard problem, that defies enumeration.

Simulated annealing is a global, multivariate optimization technique based on the Metropolis Monte-Carlo search algorithm. The method starts from an initial random state and walks through the state space associated with the problem of interest by generating a series of small, stochastic steps. An objective function maps each state into a value in R that measures its energy or fitness. In the problem at hand, a state is a unique k-membered subset of compounds from the n-membered virtual library, its energy is the diversity penalty associated with that set as measured by eqs 9, 12, or 13, and the step is a small change in the composition of that set (usually of the order of 1−10% of the points comprising the set). While downhill transitions are always accepted, uphill transitions are accepted with a probability that is inversely proportional to the energy difference between the two states. This probability is computed using Metropolis', $p = e^{-\Delta E/K_BT}$, or Felsenstein's, $p = 1/(1 + e^{\Delta E/K_BT})$ acceptance criterion. The later ensures that the transition probability never exceeds 0.5 regardless of the energy difference between the two states and thus prohibits the system from performing random walks. Boltzmann's constant, K_B, is used for scaling purposes, and T is an artificial temperature factor used to control the ability of the system to overcome energy barriers. To circumvent the difficulty of selecting an appropriate value for K_B, in our implementation this is not a true constant but is adjusted based on an estimate of the mean transition energy. In particular, at the end of each transition, the mean transition energy is updated, and the value of K_B is adjusted so that the acceptance probability for a mean uphill transition at the final temperature is 0.1%. The temperature is reduced using an exponential cooling schedule with a half-width of 5−10 deviation units (Figure 6). Other cooling schedules, such as linear, Gaussian, and Lorentzian, have also been tested; in general, schedules that involve more extensive sampling at lower temperatures seem to perform best, although it is also important that sufficient time must be spent at higher temperatures so that the algorithm does not get trapped into local minima.

We must point out that simulated annealing is just one variant of a stochastic search method that is ideally suited for this optimization problem. Other methods such as evolutionary programming and genetic algorithms have also been tested in this context and will be presented in a subsequent publication.[20]

Software. The software was implemented using an object-oriented approach and the C++ programming language. It is based on 3-Dimensional Pharmaceuticals' generic C++ simulation classes and the Mt Toolkit[32] and the RogueWave foundation libraries (Tools.h++ and the Standard C++ Library, Math.h++ and Lapack.h++).[33] The calculations

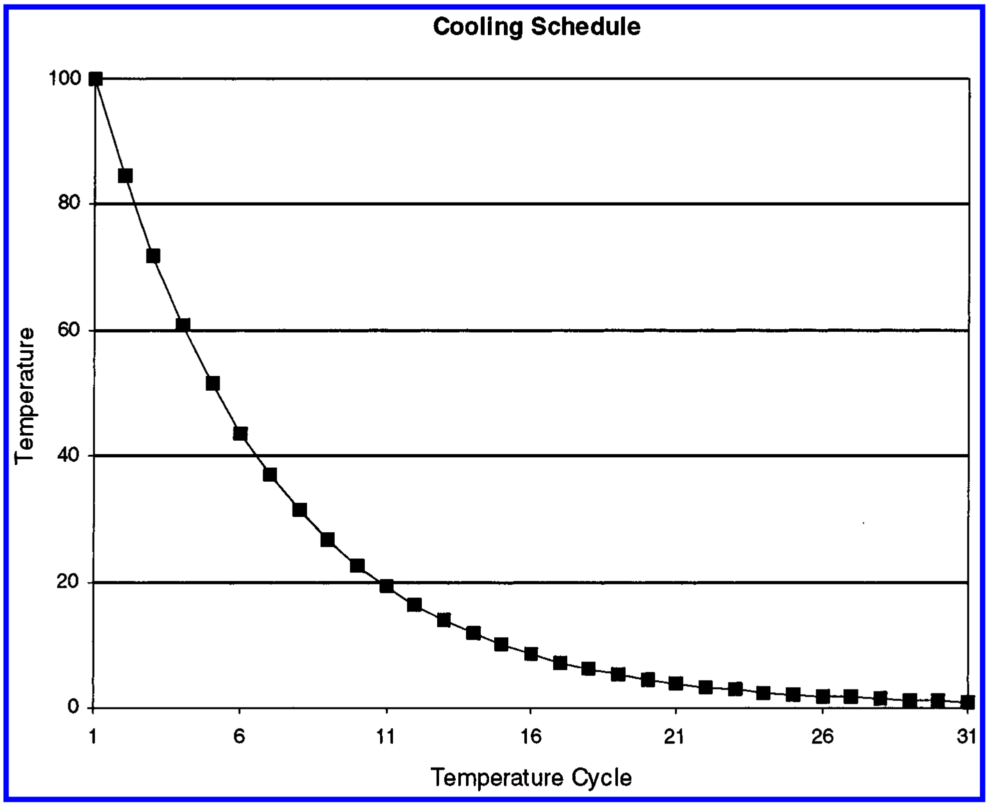

Figure 6. Exponential cooling schedule.

Table 1. Data Sets Used in the Analysis

set no.	no. dimensions	no. points	no. clusters
1	2	10 000	0
2	2	10 000	1
3	3	10 000	1
4	30	10 000	4

were performed on a Silicon Graphics Challenge workstation running Irix 6.2.

RESULTS AND DISCUSSION

The performance of our algorithm was initially studied using four data sets constructed artificially and designed to explore the effects of dimensionality and clustering in the data. The parameters used to derive them are shown in Table 1. Each data set consists of 10 000 data points whose coordinates were determined in the manner described below.

The first data set consists of two-dimensional uniformly distributed random vectors (Figure 1). The property space is densely populated, does not exhibit any significant clustering, and is designed to reveal the structure of the "true" minimum. Conversely, the remaining data sets exhibit strong clustering and were designed to illustrate the ability of our algorithm to escape local areas of high density and sample the feature space in a thorough and unbiased way. Data sets 2 and 3 were generated using two- and three-dimensional random vectors, 90% of which were distributed normally around a single (randomly chosen) cluster center, and the remaining 10% were uniformly distributed in the unit square (cube). The corresponding density functions are shown in Figures 2 and 3, respectively. Figure 3 shows two views of the data: the full three-dimensional representation (Figure 3a) and a Sammon projection in two dimensions (Figure 3b). The fourth data set consists of 30-dimensional random vectors distributed around four randomly chosen Gaussian clusters (again, 10% of these vectors were drawn from a uniform distribution). The Sammon projection of this data is shown in Figure 4. For clarity, only 1000 points are shown in this figure; these points were chosen at random and provide a sufficiently accurate visual representation of the underlying density function. In all cases, clustered points were generated using one-dimensional normal random deviates with mean zero and standard deviation one using Marsaglia's algorithm,[34] followed by centering around a randomly chosen cluster center, and rejection if the point exceeded the boundaries of the unit (hyper)cube.

The 30-dimensional example was designed specifically to enable visual inspection of the results. The clusters were intentionally tight, so that a "clean" Sammon projection would be possible. The reader may notice that from a statistical perspective the fourth data set is not truly 30-dimensional. Indeed, principal component analysis reveals that the first three PCs account for 81.8% of the total variance, while 21 PCs are needed to explain 90% of the variance. This, however, has little bearing on the performance of the algorithm, which still operates on a true 30-dimensional space.

For purposes of visualization, selections were limited to 25 compounds. The simulations were carried out in 30 temperature cycles, using 1000 sampling steps per cycle, an exponential cooling schedule (Figure 6) the Metropolis acceptance criterion, and the penalty function in eq 9. Transitions represented single-point mutations in the composition of the current set, i.e., a new state was derived by changing only one of the 25 points comprising the old state (in general, however, for optimal performance the step size should gradually decrease during the course of the simulation).

Figure 5 shows three snapshots of the selection process for data set 1. As expected, the initial state of the system (Figure 5a), which represents a random choice of 25 compounds, exhibits a nonuniform distribution. Some areas

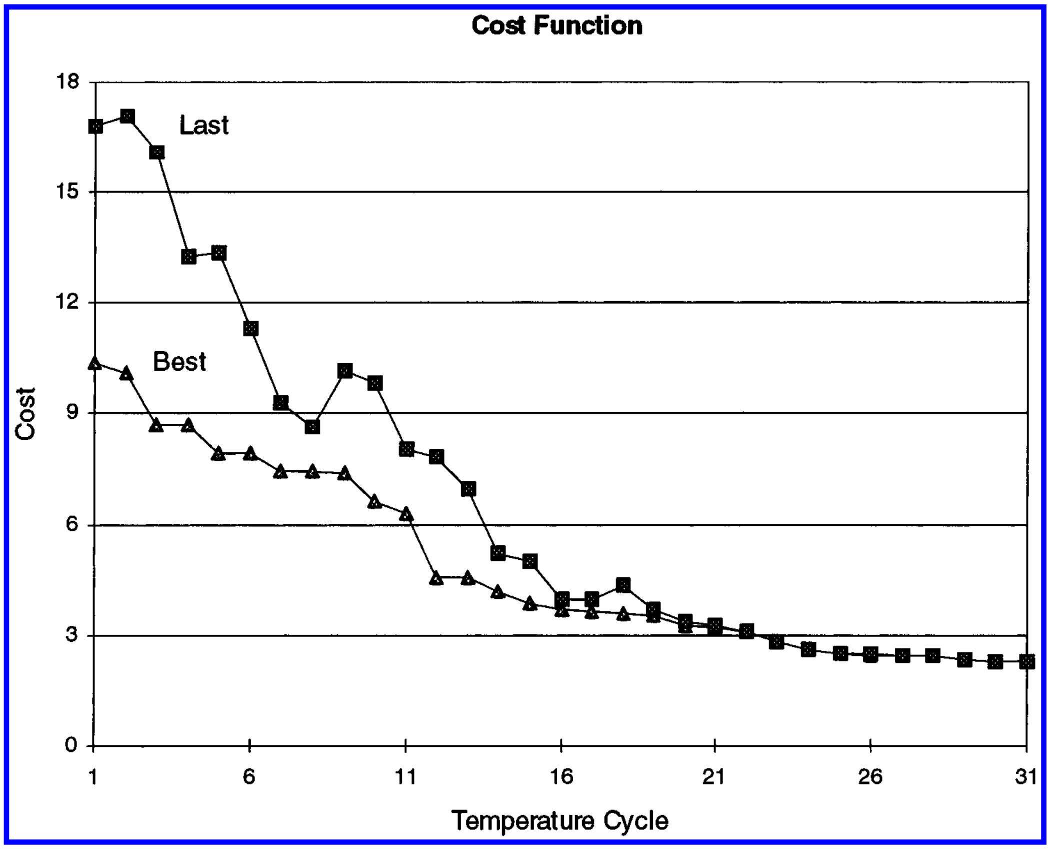

Figure 7. Energy (cost) as a function of time. "Best" and "Last" represent the best and last state, respectively, found at the end of each temperature cycle. Since simulated annealing allows uphill transitions, these states are not always the same.

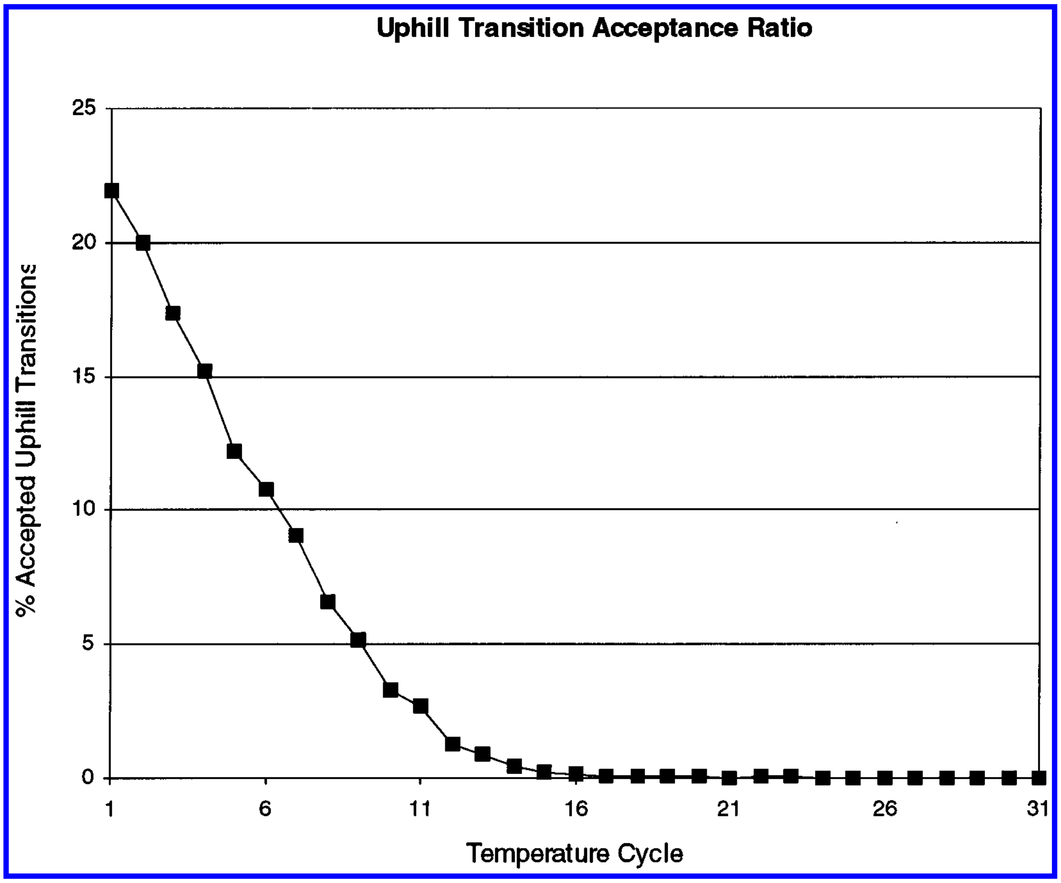

Figure 8. Percent of accepted uphill transitions as a function of time.

of the property space are oversampled (a number of close contacts are observed), while others are not sampled at all. However, as the simulation progresses (Figure 5b), the spread improves significantly, and eventually the selected points assume an optimal distribution (Figure 5c). We found that the algorithm is very robust, and in every case that we studied multiple runs converged to the same or nearly the same solution.

The details of the annealing simulation are illustrated in Figures 7 and 8. Figure 7 shows the value of the cost function at the end of each temperature cycle for a single annealing run. Red points indicate the cost of the last accepted state, and blue points represent the cost of the best state discovered at the end of each cooling step. Within the first 10−15 cycles, the algorithm is able to extract the gross features of the global minimum and recover most of the diversity in the system. The final cycles are spent refining that global minimum with relatively minimal improvements in the fitness function. The asymptotic convergence of the two curves manifests the decreasing ability of the Metropolis search algorithm to perform high energy uphill transitions. Indeed, at higher temperatures the algorithm is able to overcome substantial energy barriers, but this ability is diminished at lower temperatures, and the system is eventually frozen around the 20th cycle. This effect is also illustrated in Figure 8 which plots the percentage of accepted uphill transitions during each temperature cycle as a function of time.

As we mentioned earlier, data set 1 allows us to look at the structure of the "true" minimum, at least in the two-dimensional case. This is due to the uniform distribution of the data points and dense population of the feature space. The remaining data sets, on the other hand, allow us to assess the performance of our algorithm in sparsely populated and highly clustered landscapes, which are typically encountered in combinatorial molecular property distributions.

Figure 9 shows three snapshots of a 25-point selection from data set 2 using the same simulation parameters that were used for set 1. Here, the effects of a random choice are much more clearly visible. Since the data is highly clustered, a random sample is an extremely poor choice for sampling the diversity of the system; the initial state is concentrated around the cluster center in the lower left quadrant of the map. Again, within the first iteration the situation improves dramatically, but there is still some visible tendency to oversample the region surrounding the cluster center. By the end of the simulation, however, these defects are removed, and the final selection samples every corner of the property space in an unbiased and uniform way. This is also true for the three-dimensional case (data set 3), as shown in Figures 10 and 11.

Figure 11 is the first example of the use of Sammon mapping for visualizing molecular diversity in higher-dimensional spaces. Sammon mapping is a nonlinear mapping technique that allows a set of multidimensional data samples to be projected onto a display plane in such a way that the pairwise distances of the images match as closely as possible the corresponding distances in the original space. Although the projection is only approximate, it preserves the topology of the original space and allows for qualitative conclusions to be drawn regarding the distribution and density of the data samples. In the problem at hand, the projection illustrates the tendency of the algorithm to move away from clearly discernible local clusters and sample larger domains of the property space. In this case, one should not expect an optimal distribution of the images on the display plane but rather a qualitative tendency to increase the spread and avoid any cluster formations which are indicative of oversampling. This is indeed the case for both the 3- (Figure 11) and 30-dimensional (Figure 12) spaces used in this study.

As mentioned above, eq 9 is only one of many different diversity functions that can be used for compound selection.

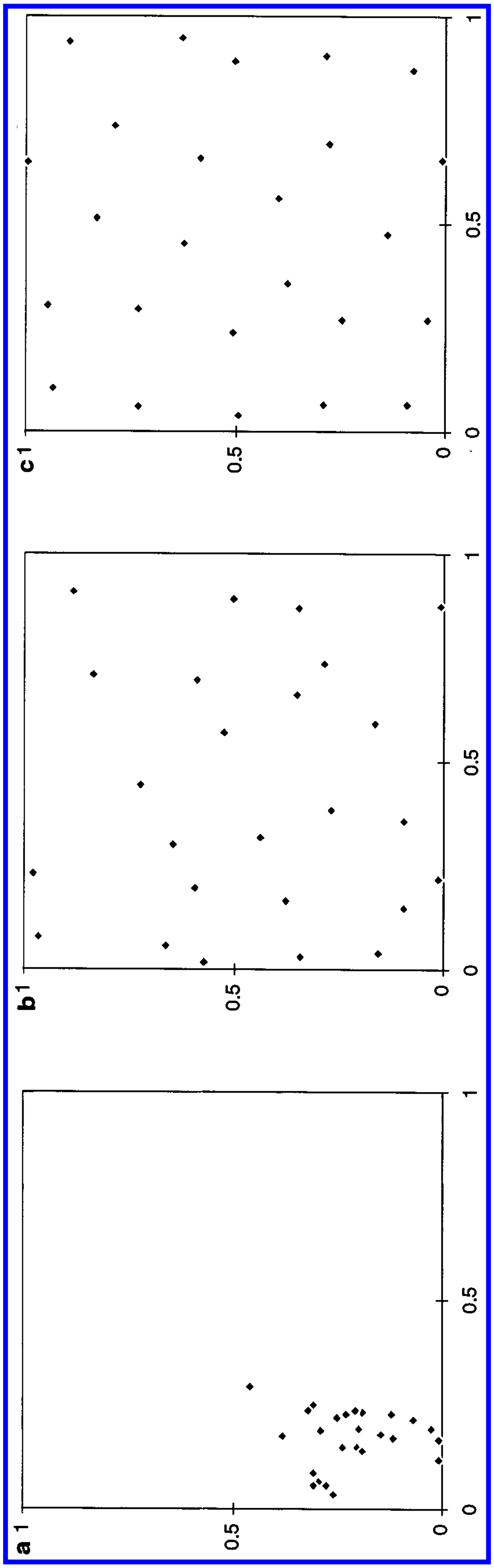

Figure 9. Twenty-five point selection from data set 2 (2D): a, initial state; b, best state found after first cycle; c, best state found after 31st cycle.

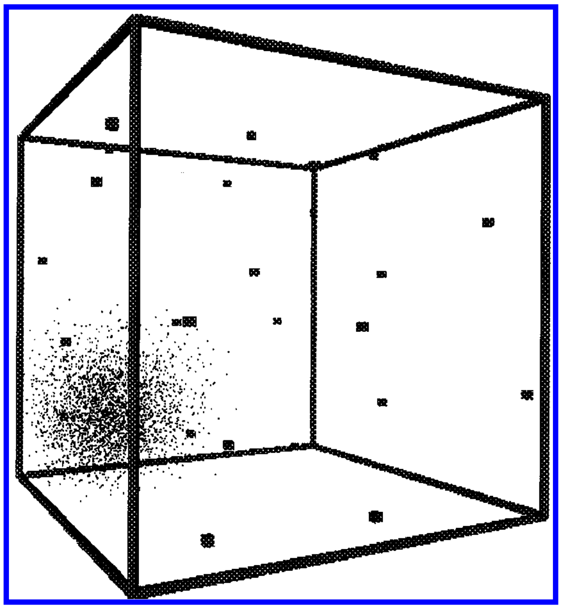

Figure 10. Twenty-five point selection from data set 3.

This function requires a vectorial representation of chemical space and is not readily amenable to some other commonly used similarity coefficients. For this reason, we have tested our algorithm using a variant of the maximin function (eq 13). The results for the selection of 25 compounds from the second data set are shown in Figure 13 and indicate a similar behavior. In fact, this was the case with every data set used in this study, and, in the interest of brevity, the results are not shown here.

Finally, to test the performance of our algorithm in a “real” setting, we decided to apply it on a collection of three combinatorial libraries characterized by means of spatial autocorrelation vectors, generously provided to us in electronic form by Prof. J. Gasteiger and Dr. Markus Wagener. These libraries were obtained by combining 19 L-amino acids with three rigid central scaffolds based on cubane, xanthene, and adamantane, respectively.[35] Each compound was described by a 12-dimensional autocorrelation vector, which represented the distribution of the electrostatic potential on the van der Waals surface of the molecule. These autocorrelation vectors were then used to train a Kohonen network, which very nicely separated the xanthene from the cubane and adamantane libraries, as one would expect based on the three-dimensional geometry of the scaffold and the relative disposition of the four amino acid R-groups. Details can be found in the original publication.[35]

We used these autocorrelation vectors to construct a three-dimensional Sammon map of these libraries, highlighted in blue (xanthene), green (cubane), and red (adamantane) in Figure 14. The map is sufficiently accurate, as manifested by a Sammon and Kruskal stress value of only 10 and 8%, respectively (see eqs 1 and 2). It is clear that the Sammon map is not only able to reproduce the sharp separation between the xanthene and cubane/adamantane libraries that was observed in the self-organized maps but also to reveal a more subtle distinction between the cubane and adamantane libraries, that was not captured by the Kohonen network. We believe this separation is consistent with the subtle geometrical differences between the two scaffolds and, in particular, the differences in the relative separations at the R-groups. A selection of 96 diverse compounds was then performed using the diversity function described above (eq

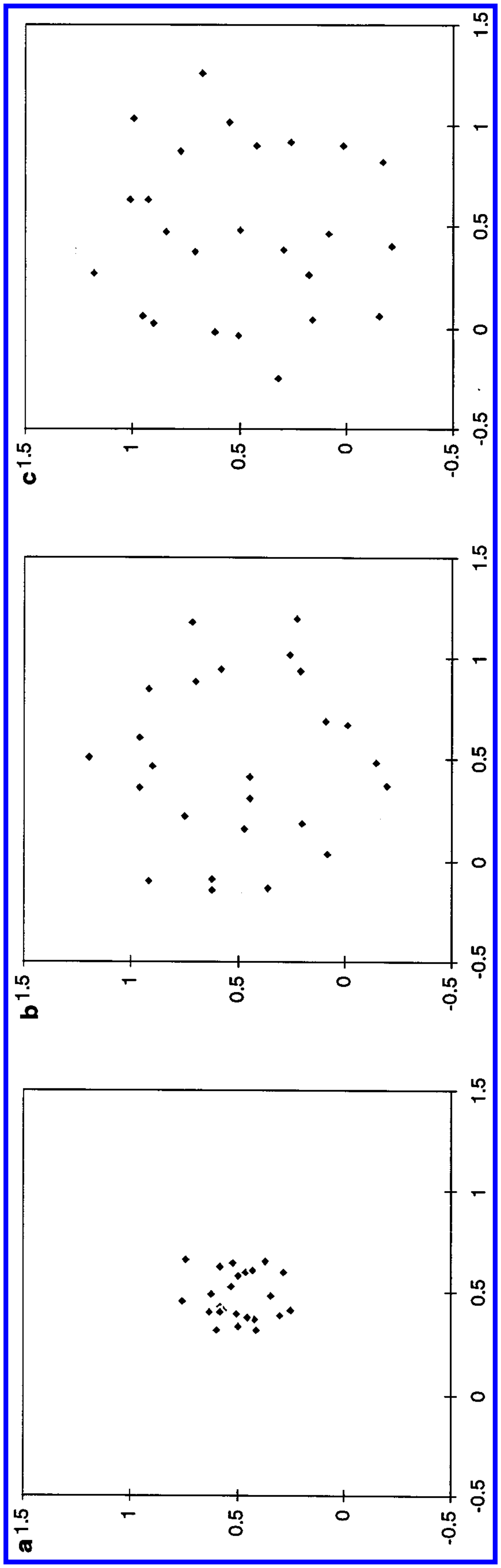

Figure 11. Twenty-five point selection from data set 3 (3D) using simulated annealing: a, initial state; b, best state found after first cycle; and c, best state found after 31st cycle.

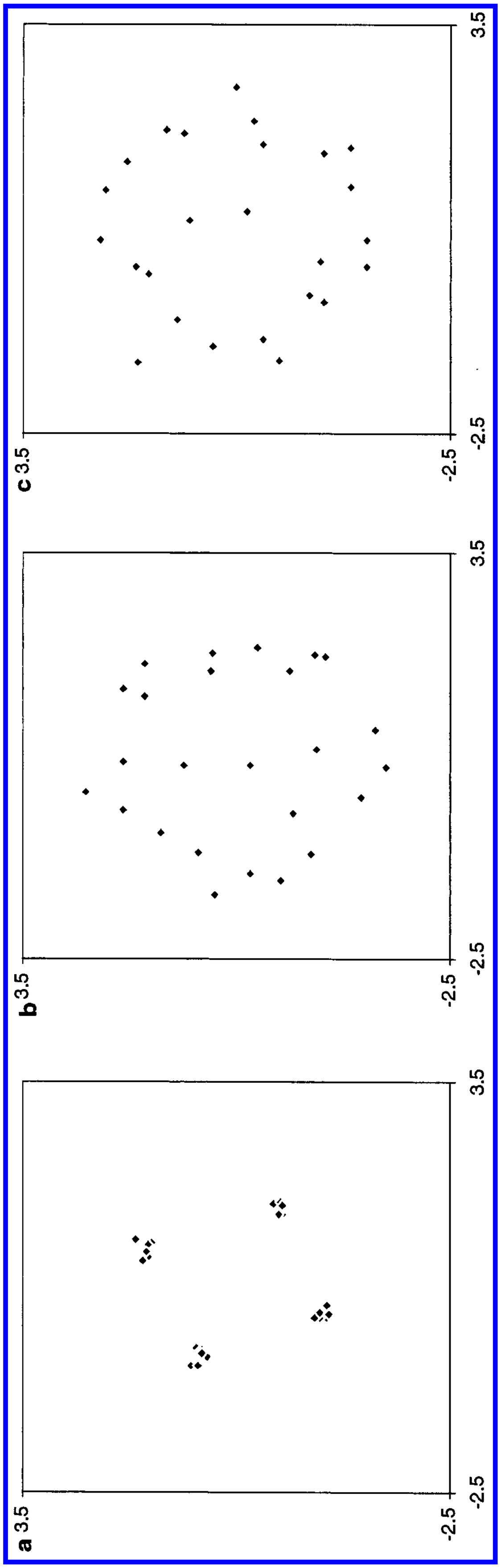

Figure 12. Twenty-five point selection from data set 4 (30D): a, initial state; b, best state found after first cycle; and c, best state found after 31st cycle.

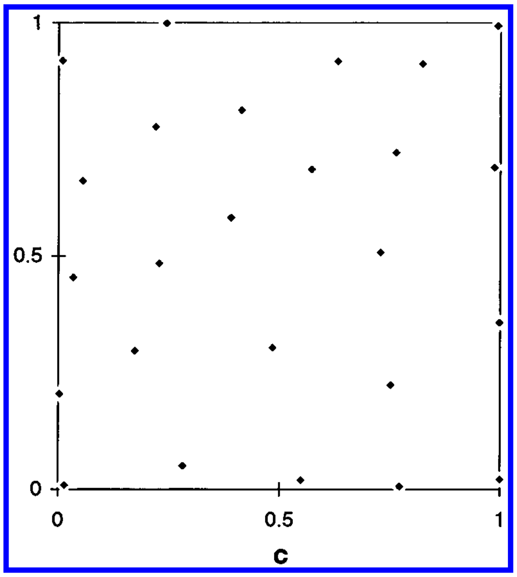

Figure 13. Twenty-five point selection from data set 2 using the maximin diversity function.

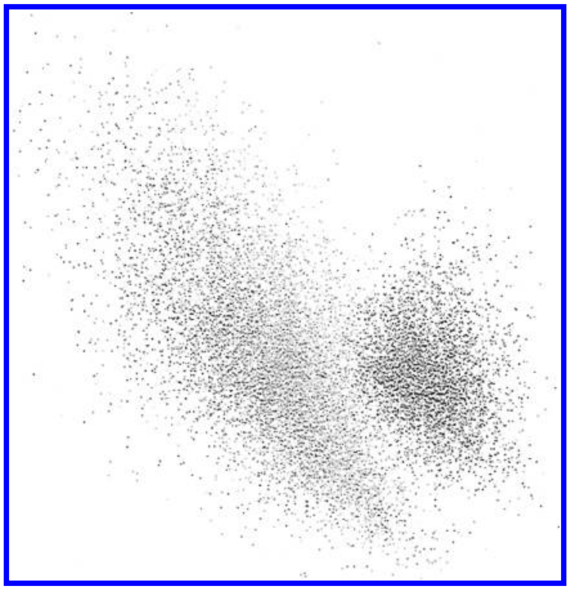

Figure 14.

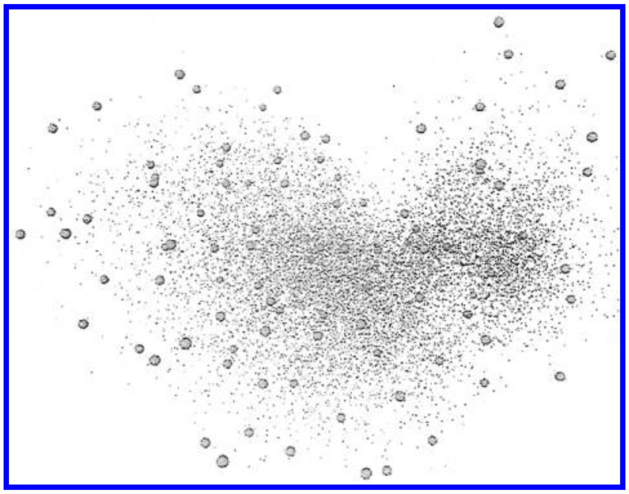

Figure 15.

9), and the results are illustrated in Figure 15, in a slightly different orientation. It is clear that the selection algorithm was able to sample the property space occupied by the three libraries in an unbiased and uniform way that is consistent with our expectations.

CONCLUSIONS

One of the most important applications of computational chemistry in combinatorial drug discovery is to assist in the design of libraries that maximize the diversity and information content of the resulting species. In this paper, we introduced a new family of general selection algorithms that combine a stochastic search engine with a user-defined objective function that encodes any desirable selection criterion. The method was applied to the problem of maximizing molecular diversity in multivariate descriptor spaces, and the results were visualized using Sammon's nonlinear mapping algorithm. We demonstrated that the method is capable of sampling remote, sparsely populated areas of property space and can easily escape areas of high local density, which a typical shortcoming of some clustering methodologies. Because the search engine and the performance metric are treated as independent entities, this technique can be readily extended to perform selections based on any suitably encoded, user-defined selection criterion. Indeed, many existing diversity algorithms such as maximin or d-optimal design can be recast in the form of an optimization problem by devising an appropriate objective function. Furthermore, other important criteria can be directly incorporated into the penalty function, such as the cost of starting materials and/or the compliance to existing SAR or pharmacophore models, and can be combined to perform complex multiobjective selections in advanced decision support systems.

ACKNOWLEDGMENT

The author would like to thank Drs. Raymond Salemme, Edward Jaeger, and John Spurlino of 3-Dimensional Pharmaceuticals and Prof. Micahel Levitt of Stanford University for their useful suggestions and comments.

REFERENCES AND NOTES

(1) Agrafiotis, D. K.; Bone, R. F.; Salemme, F. R.; Soll, R. M. System and method of automatically generating chemical compounds with desired properties. United States Patent 5,463,564, 1995.

(2) Graybill, T. L.; Agrafiotis, D. K.; Bone, R.; Illig, C. R.; Jaeger, E. P.; Locke, K. T.; Lu, T.; Salvino, J. M.; Soll, R. M.; Spurlino, J. C.; Subasinghe, N.; Tomczuk, B. E.; Salemme, F. R. Enhancing the drug discovery process by integration of high-throughput chemistry and structure-based drug design. In *Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery*; Chaiken, I. M., Janda, K. D., Eds.; ACS Conference Proceeding Series, 1996; pp 16−27.

(3) Willett, P. *Similarity and Clustering in Chemical Information Systems*; John Wiley & Sons: New York, 1987.

(4) Downs, G. M.; Willett, P. Clustering in chemical structure databases for compound selection. In *Chemometric Methods in Molecular Design*, H. van der Waterbeemd, Ed., VCH: Weinheim, 1994; pp 111−130.

(5) Downs, G. M.; Willett, P.; Fisanick, W. Similarity searching and clustering in large databases using molecular property data. *J. Chem. Info. Comput. Sci.* **1994**, *34*, 1094−1102.

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

(7) Lajiness, M. S.; Johnson, M. A.; Maggiora, G. M. Implementing drug screening programs using molecular similarity methods. In *QSAR: Quantitative Structure Activity Relationships in Drug Design*; Fauchere, J. L., Ed.; A. R. Liss: New York, 1989; pp 173−176.

(8) Johnson, M.; Lajiness, M. S.; Maggiora, G. M. Molecular similarity: a basis for designing drug screening programs. Fauchere, J. L., Ed.; A. R. Liss, New York, 1989; pp 167−171.

(9) Polinsky, 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.

(10) Holliday, J. D.; Willett, P. Definitions of dissimilarity for dissimilarity based compound selection. In press.

(11) 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.

(12) 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. Info. Comput. Sci.* **1996**, *36*, 750−763.

(13) Pearlman, R. S. Novel tools for addressing chemical diversity. *Network Science* **1996**, hhtp://www.awod.com/netsci/issues.

(14) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(15) Lin, S. K. Molecular diversity assessment: logarithmic relations of information and species diversity and logarithmic relations of entropy and indistinguishability after rejection of Gibbs paradox of entropy mixing. *Molecules* **1996**, *1*, 57−67.

(16) Agrafiotis, D. K. On the use of information theory for assessing molecular diversity. *J. Chem. Info. Comput. Sci.* In press.

(17) Martin, E. J.; Spellmeyer, D. C.; Critchlow, R. E.; Blaney, J. M. Does combinatorial chemistry obviate computer-aided drug design? In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1997; Vol. 10.

(18) Martin, Y. C.; Brown, R. D.; Bures, M. G. Quantifying diversity. In *Combinatorial Chemistry and Molecular Diversity*; Kerwin, J. F., Gordon, E. M., Eds.; John Wiley & Sons: New York. In press.

(19) Agrafiotis, D. K. Molecular Diversity. In *Encyclopedia of Computational Chemistry*; John Wiley & Sons: in press.

(20) Agrafiotis, D. K.; Jaeger, E. P. Manuscript in preparation.

(21) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors to ligand-receptor binding. *J. Chem. Info. Comput. Sci.* **1997**, *37*, 1−9.

(22) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validating molecular diversity descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(23) Delaney, J. S. Assessing the ability of chemical similarity measures to discriminate between active and inactive compounds. *Molecular Diversity* **1995**, *1*, 217−222.

(24) Sammon, J. W. A non-linear mapping for data structure analysis. *IEEE Trans. Comp.* **1969**, *C−18*, 401−409.

(25) Barlow, T. W., Richards, W. G. A novel representation of protein structure. *J. Mol. Graphics* **1995**, *13*, 373−376.

(26) Agrafiotis, D. K. A new method for analyzing protein sequence relationships based on Sammon maps. *Protein Science* **1997**, *6*, 287−293.

(27) Kruscal, J. B. Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Phychometrika* **1964**, *29*, 1−27.

(28) Kruskal, J. B. Non-metric multidimensional scaling: a numerical method. *Phychometrika* **1964**, *29*, 115−129.

(29) Agrafiotis, D. K.; Lobanov, V. S. Unpublished results.

(30) Scott, D. W. *Multivariate Density Estimation*; Wiley Interscience: New York, 1992.

(31) Wegman, E. J. Hyperdimensional data analysis using parallel coordinates. *J. Amer. Statist. Assoc.* **1990**, *85*, 664−675.

(32) Agrafiotis, D. K. The Mt toolkit: an object-oriented C++ class library for molecular simulations. Copyright 3-Dimensional Pharmaceuticals, Inc.: 1994−1997.

(33) RogueWave Software; Corvallis, OR.

(34) Box, G. E. P.; Muller, M. E.; Marsaglia, G. *Annals. Math. Stat.* **1958**, *28*, 610.

(35) Sadowski, J.; Wagener, M.; Gasteiger, J. Assessing similarity and diversity in combinatorial libraries by spatial autocorrelation functions and neural networks. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*(23), 2674−2677.

CI9700337