Selecting Combinatorial Libraries to Optimize Diversity and Physical Properties

Valerie J. Gillet,*,† Peter Willett,† John Bradshaw,‡ and Darren V. S. Green‡

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom, and GlaxoWellcome Research and Development Limited, Gunnels Wood Road, Stevenage, SG1 2NY, United Kingdom

Received July 23, 1998

The program SELECT is presented for the design of combinatorial libraries. SELECT is based on a genetic algorithm with a multi-objective fitness function. Any number of objectives can be included, provided that they can be readily calculated. Typically, the objectives would be to maximize structural diversity while ensuring that the compounds in the library have "drug-like" properties. In the examples given, structural diversity is measured using Daylight fingerprints as descriptors and either the normalized sum of pairwise dissimilarities, calculated with the cosine coefficient, or the average nearest neighbor distance, calculated with the Tanimoto coefficient, as the measure of diversity. The objectives are specified at run time. Combinatorial libraries are selected by analyzing product space, which gives significant advantages over methods that are based on analyzing reactant space. SELECT can also be used to choose an optimal configuration for a multicomponent library. The performance of SELECT is demonstrated by its application to the design of a two-component amide library and to the design of a three-component thiazoline-2-imine library.

INTRODUCTION

Compound selection is currently a topic of great importance in the drug discovery process.[1−3] The need to select compounds arises from the fact that the number of compounds that is available for testing far exceeds the capacity of high throughput screening (HTS).[4] This disparity is often true for corporate databases and is increasingly the case with the development of combinatorial chemistry experiments, where large numbers of compounds are synthesized simultaneously. Using combinatorial chemistry it is easy to plan synthetic schemes that could generate potentially massive numbers of compounds. There is, hence, a need to be able to select compounds that are both diverse yet also representative of some larger collection. In lead generation experiments the selection strategy is usually to choose as diverse a subset as possible so that all the different types of biological activity within a larger collection are sampled using as few compounds as possible. Consequently, much effort has gone into devising different measures of structural diversity to assist in the design of combinatorial libraries,[5] where we use the phrase *combinatorial library* to refer to a library that is enumerated from *subsets* of reactants.

In most approaches to compound selection, the methods are applied at the reactant level[6] on the assumption that a diverse set of reactants will result in a diverse set of products. In a previous study,[7] we have shown that greater diversity can be achieved when the selection method involves analyzing product space rather than reactant space. In that work we gave a brief description of how a genetic algorithm could be used to select diverse reactants by an analysis of a fully enumerated virtual library, where we use the phrase *virtual library* to represent a library that is enumerated from all available reactants and that would normally only exist in a computational representation. Although diversity is an important criterion in the design of combinatorial libraries, other criteria are also of importance; for example, the compounds within a library should have "drug-like" characteristics.[8] In this paper, we describe the program SELECT, which has been developed to select combinatorial libraries that are optimized for diversity and also for user-defined physical properties. SELECT is based on the previous genetic algorithm in that the selection criteria are applied to the fully enumerated virtual library; however, the original algorithm has been extended so that the selection criteria can include a number of factors, such as the diversity and the physical properties of the library. The effectiveness of SELECT is demonstrated through examples.

GENETIC ALGORITHMS FOR LIBRARY SELECTION

Compound selection can be a computationally intensive task. Selection of the maximally diverse subset is computationally unfeasible because it requires evaluation of

$$\frac{N!}{n!(N-n)!}$$

subsets, where a subset of n compounds is selected from a library containing N compounds. In the previous paper[7] we showed how it is possible to conceptualize a two-component virtual library of size N_1N_2 as a two-dimensional (2D) matrix. A combinatorial library of size n_1n_2 can then be selected by intersecting n_1 rows with n_2 columns of the matrix. Finding an optimal library is then equivalent to exploring all permutations of rows and columns of the matrix to identify that

* Author to whom correspondence should be sent. E-mail: v.gillet@sheffield.ac.uk.
† Krebs Institute for Biomolecular Research.
‡ GlaxoWellcome Research and Development Limited.

10.1021/ci980332b CCC: $18.00
Published on Web 12/17/1998

library with the largest value of some user-defined criterion of "goodness". This method represents an enormous search space even for libraries of moderate size. Genetic algorithms (GAs)[9,10] have been developed as effective methods for exploring large search spaces, such as those that are characteristic of virtual combinatorial libraries.

A GA is the computer equivalent of Darwinian evolution. In a GA, the problem space is represented by a population of chromosomes, and genetic operators, such as crossover and mutation, are applied to evolve new potential solutions to the problem. A fitness function is used to judge the value of each potential solution.

GAs have already been applied to the problem of compound selection although in contexts different from that which is described here. Sheridan and Kearsley[11] developed a GA for the design of a library with the maximal chance of containing actives in a particular biological assay. A chromosome of the GA encodes a single library product that is constructed from fragments extracted from fragment pools. Hence, the GA optimizes a population of individual products, and fragments that occur frequently in the final products can be identified and used in a combinatorial synthesis. Weber et al.[12] developed a GA that optimizes the actual biological response for the compounds within a combinatorial library. Each chromosome in the GA represents a single product compound of the reaction. The fitness function involves actually performing the corresponding reaction and testing the product. The method was able to find compounds with micromolar activity after synthesizing only 400 out of a possible 160 000 molecules. A similar procedure has been described by Singh et al.[13]

Closer in spirit to the work described here is the program GALOPED, which was developed by Brown and Martin[14] for the design of combinatorial mixtures. In GALOPED, each chromosome represents a library of compounds and the algorithm was designed to handle the specific problem of deconvolution that occurs with mixtures. Thus, the selection strategy is based on both diversity and optimizing the molecules for deconvolution. Diversity is assessed by clustering at the 2D or three-dimensional (3D) level. The precursors that are available at each site are clustered using 2D MACCS structural keys and Ward's agglomerative clustering. The 3D families or partitions are identified by examining all potential pharmacophore points in each cluster. The greater the number of clusters or partitions covered by a library, the greater is its diversity. Measuring diversity in product space requires that the full library is enumerated and clustered. Thus, the procedure is too costly to allow clustering at the 3D level and the 2D clustering is limited to libraries of around 200 000 structures. SELECT differs from GALOPED in the way in which the libraries are encoded in the GA, the criteria that are to be optimized, and the methods used to analyze diversity. Moreover, SELECT has been designed to handle libraries that are the result of the parallel synthesis of descretes, and deconvolution is hence not a consideration. The diversity measures implemented in SELECT are based on calculating intermolecular similarities.

Good and Lewis[15] describe a library design tool called HARPick that is based on simulated annealing for optimizing reagent selection by an analysis of product space. Their method includes a flexible scoring function that allows diversity to be optimized alongside other properties such as molecular shape profiles. Diversity calculations are based on pharmacophore profiling and require that pharmacophore keys are generated for all potential products. The processing requirements for performing these calculations and the memory required to store all pharmacophores found for each potential product limit the approach to libraries of $<100\,000$ compounds. Molecular properties are optimized by forcing the library to adopt an even distribution of shapes. SELECT is similar in concept to HARPick but utilizes a GA as its optimization method, rather than simulated annealing. The diversity measure used in SELECT is much less time consuming to calculate and the memory requirements are substantially less, thus allowing the method to be applied to larger virtual libraries; for example, SELECT is currently being used at GlaxoWellcome[16] to design libraries of 12 000 compounds selected from a virtual library of 800 000 compounds. In SELECT, physical property profiles of a library can be optimized to be similar to profiles found in a reference collection, such as the World Drugs Index, in an effort to produce diverse libraries that have "drug-like" characteristics.

SELECT was designed for multicomponent combinatorial libraries. Initially, the full combinatorial library is enumerated using all the reactants from all the reactant pools. For example, a three-component library that has 10 possible reactants in each pool will result in a fully enumerated library of 1000 products. Daylight fingerprints[17] and the physical properties that are to be optimized are calculated for each product. SELECT then examines the characteristics of combinatorial libraries selected from the fully enumerated library and identifies libraries that best fit the user-specified design criteria for the final library.

SELECT. Each chromosome in the GA represents a combinatorial library selected from the fully enumerated library. A chromosome is partitioned, and the number of partitions is equal to the number of components (or reactant pools) in the library. The number of genes in each partition corresponds to the number of reactants to be selected from the respective reactant pool. The genes are integers, with each gene representing a reactant number in the respective pool. The standard genetic operators of mutation and crossover are applied with the extra constraint that each partition of the chromosome contains unique integers, corresponding to unique reactants from the respective reactant pool. Crossover occurs within a single partition with the other partitions remaining unchanged; for example, if the crossover point occurs in the second partition of a three-partition chromosome (representing a three-component reaction), then the first and third partitions are unchanged in the child chromosomes, and mixing occurs only within the second partition. Crossover and mutation are illustrated in Figures 1(a) and 1(b), respectively. The rate of mutation versus crossover is determined by a user-defined parameter. Chromosomes are selected for crossover or mutation using roulette wheel selection so that the probability of a chromosome being selected is proportional to its fitness. The GA is a steady-state GA with no duplicates. Decoding of a chromosome is equivalent to enumerating the combinatorial library it represents. The fitness function, which is described in the next section, is then applied to the combinatorial library.

A niching procedure was implemented in the GA.[18] The niching technique was originally developed to enable GAs

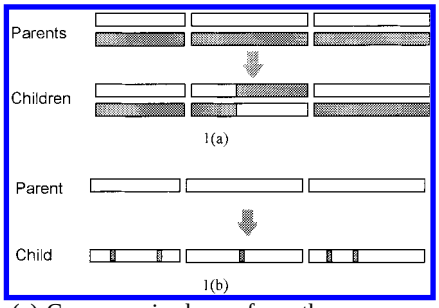

Figure 1. (a) Crossover is shown for a three-component combinatorial library consisting of three partitions. The crossover point, chosen at random, occurs in the second partition. The effect of crossover is shown in the child chromosomes where mixing has occurred in the second partition only. (b) Mutation can occur at any position in the chromosome. The number of genes that are mutated in one operation is determined by a user-defined parameter.

to provide solutions to multimodal problems. In SELECT, niching is used in an attempt to prevent the GA from converging on a suboptimal solution. Niching operates by first allowing the GA to converge on a solution as normal. This solution then forms the center of a niche and an area around it is removed from the search space. For example, all libraries with a user-specified number of reactants in common with the library at the center of the niche are included within the niche and are hence excluded from the search space. The GA is then rerun using the reduced search space and a second solution is found. In total, the GA is run N times to create N niches and the fittest solution in the N niches is then identified as the overall best solution. In SELECT, the size of the niche around a solution is determined by the maximum number of reactants selected from a pool that can be shared with any new chromosome. If this number is exceeded by the library represented by the chromosome, then the chromosome represents a solution within a niche that has already been found and hence it is removed from the search. The number of niches, N, and the number of reactants that define the size of a niche are user-specified parameters.

One advantage of niching in combinatorial library design is that it allows "good" solutions to be found in different regions of the search space. In general, libraries from different regions of the search space will be composed of different reactants, which allows the user to browse through different solutions and can be useful when additional criteria, such as cost or availability, are important selection criteria.

Fitness Function. The GA is designed to optimize both the diversity and the physical properties of a combinatorial library and, hence, the fitness function can consist of a number of terms. This function has the general form

$$f(n) = w_D(1 - D) + w_C(1 - C) + w_{f_1}\Delta f_1 + ...$$

with its exact form being specified at run time. The first two terms represent diversity terms and are described next. These are followed by any number of terms that are designed to optimize the physical properties of the library. D is the diversity of the library, which can be measured in one of two different ways: the normalized sum of pairwise dissimilarities calculated using the cosine coefficient (D_{SUM});[19] and the average nearest neighbor distance calculated using the Tanimoto coefficient (D_{NN}).[20] In each case, the molecules are represented by Daylight fingerprints of size 1024. Calculating diversity as the normalized sum of pairwise dissimilarities using the cosine coefficient is the most efficient measure because the centroid algorithm[21] can be used, resulting in the calculation having time complexity O(N), where N is the number of molecules in the library. For example, the diversity of library A is given by

$$D_{SUM}(A) = 1 - \frac{\mathrm{DOTPROD}(A_C,A_C)}{N(A)^2}$$

where A_C is the centroid of A, DOTPROD(A_C,A_C) is the dot product of the centroid of A with itself, and there are $N(A)$ compounds in library A. Measuring diversity as the average nearest neighbor distance has time complexity of O(N^2) and is thus limited to the selection of small libraries. The GA is designed to minimize the fitness function and, hence, the diversity term is included as $(1 - D)$.

The second term in the fitness function $(1 - C)$ is designed to force the library to be maximally different from some reference collection. For example, it may be desirable to design a library that is diverse with respect to a corporate collection, or having designed one library for a given combinatorial experiment it may be desirable to design a second that is diverse with respect to the first. A combinatorial library can be forced to complement an existing reference collection by maximizing the diversity that would result if the two collections were combined together. The combined diversity that would result if the selected library (called library A) was added to the reference collection (called library X) can be calculated efficiently when diversity is measured as the sum of pairwise dissimilarities using the cosine coefficient. The calculation requires the centroid A_C of library A, and the centroid X_C of the reference collection X. From Turner et al.,[21,22] the diversity of the combined collection AX is given by

$$D_{SUM}(AX) = 1 - (\{\mathrm{DOTPROD}(A_C,A_C) + \mathrm{DOTPROD}(X_C,X_C) + 2 \times \mathrm{DOTPROD}(A_C,X_C)\}/\{(N(A) + N(X))^2\})$$

In practice, the centroid of the reference collection, X_C, is precalculated and stored. The centroid of the combinatorial library, A_C, is calculated from the fingerprints of the molecules encoded in the chromosome, and C is calculated as $D_{SUM}(AX)$. The combined diversity term is included in the fitness function as $(1 - C)$ because the GA is designed to minimize the fitness function.

The remaining terms in the fitness function relate to the physical properties of the libraries. Appropriate physical properties will be those that are thought to influence the ability of a compound to act as a drug (e.g., molecular weight, rotatable bond profile, ClogP, etc.). A physical property of the library is optimized by comparing the distribution of its values in the library with the distribution of values of the same property in some reference collection. SELECT can optimize any property that can be calculated for the molecules in the combinatorial library and for which a reference distribution is supplied. The distribution of a property is binned and the number of bins and the ranges of values covered by each bin can be varied according to the property being measured (the distribution of the reference

collection is precalculated and stored). The term Δf_1 is then calculated as the scaled difference between the profile root mean square (RMS) of the library and the reference collection, using the following equation

$$\Delta f_1 = \sqrt{\frac{\sum_{b=1}^{N}(A_b - X_b)^2}{N}}/F_1$$

where N is the number of bins used to represent the distribution of the feature; A_b is the percentage of the library A that occupies bin number b; X_b is the percentage of the reference collection that occupies bin number b; and F_1 is a scaling factor that is specified by the user at run time so the RMS difference is scaled to be of the same order of magnitudes as the diversity index.

Implementation Details. The fitness function of the GA requires that the library represented by a chromosome be enumerated so that the fingerprint and physical properties can be calculated for each product molecule in the library. The fitness function is applied very many times during a run of the GA, and performing these calculations each time is very costly. Hence, prior to running the GA, the full library is enumerated from the reactant pools, and the fingerprints and any physical properties that are to be optimized are calculated for each molecule in the fully enumerated library and stored in memory. This procedure requires 152 bytes plus 4 bytes per physical property per molecule so that a virtual library of 10 000 molecules requires 1.56 Mbytes of memory to store sufficient information to allow the selection of a library optimized on diversity and rotatable bond profile. During the running of the GA, the appropriate values for a combinatorial library are retrieved by lookup. The reference centroid and reference distributions of physical properties are pre-calculated, the weights and terms to be included in the fitness function are defined, the population size and crossover versus mutation rate are defined, and the numbers of reactants required from each pool are specified.

RESULTS AND DISCUSSION

The effectiveness of the GA is investigated here by using it to select combinatorial libraries for two different combinatorial syntheses. The libraries were used to test the effectiveness of the GA and were extracted from the literature: they are an amide library and a thiazoline-2-imine library.[23]

Amide Library. The first library was used to examine the parameters of the GA. The amide library is a two-component library where amines in one pool are reacted with carboxylic acids from another to form amides. A virtual library was built using 100 amines and 100 carboxylic acids, the reactant pools each being formed by extracting structures at random from SPRESI.[24] The full library of 10 000 amides was enumerated, and the Daylight fingerprint and the number of rotatable bonds were calculated for each molecule within the library. In each of the runs described later, the GA was configured to select 20 amines and 20 carboxylic acids (i.e., a 20 × 20 combinatorial library from the full 100 × 100 virtual library). The chromosomes of the GA, therefore, contain two partitions, each of size 20. The genes in each partition can have integer values between 1 and 100 and they represent reactants in the corresponding reactant pool. The chromosomes were initialized to random integers in the range 1−100, with the constraint that all of the genes within a partition are unique.

Table 1. Amide Libraries Optimized for Diversity Alone[a]

pop[b]	no. Its[c]	t (s)[d]	D_{SUM}
10	1110 (291)	162 (43)	0.593 (0.004)
20	1020 (224)	224 (49)	0.594 (0.002)
50	940 (163)	412 (71)	0.595 (0.001)

[a] SELECT was run to choose amide libraries that are optimized on diversity alone, with diversity measured as the sum of pairwise dissimilarities (in the fourth column). [b] The effect of varying the population size (pop) is shown. In each case, the results are averaged over 10 runs, with standard deviations in brackets. [c] The second column shows the number of iterations required for the GA to converge. [d] The third columns shows the execution time in s (SG R10000 195 MHz processor).

The first set of runs was designed to investigate different population sizes. SELECT was parameterized to measure diversity alone (i.e., w_D was set to 1.0 and no other terms were included in the fitness function). The diversity measure optimized D_{SUM} so that the best solution was, therefore, the library with the highest diversity. SELECT was run 10 times for each population size and the number of niches was set to one. In each run, iterations of the GA were continued until there was no change in the fitness function over 250 iterations, at which point SELECT was said to have converged. The average diversity achieved and its standard deviation (in brackets) are shown in Table 1. The best solution found was a combinatorial library with diversity 0.597. For comparison, the diversities of combinatorial libraries selected at random averaged over 100 libraries were calculated as 0.508 with standard deviation 0.015, and the least diverse library found by running SELECT with w_D set to −1.0 has diversity 0.314. Hence, SELECT is able to find libraries that are significantly more diverse than libraries selected at random. The average run time was 7.2 min on a Silicon Graphics R10000 processor running at 195 MHz, and the memory required to store the virtual library was 1.56 Mbytes.

As the population size increases from 10 to 50, the average diversity over 10 runs increases slightly and less variation in values is seen over the 10 runs. When the GA is run with a small population of chromosomes, it is more likely to get trapped in a local optimum and, hence, the larger variation in values seen with the smaller populations. In all the subsequent experiments, the GA was run with a population of 50 chromosomes. The larger population sizes require longer running times; however, in general, fewer runs are required to find a minimum value for the fitness function because there is little variation over several runs (see the second column of Table 1).

The next set of experiments was designed to investigate the effect of optimizing both diversity and the rotatable bond profiles of the selected libraries. The distribution of rotatable bonds in a subset of WDI[25] was calculated and used as a reference profile. The subset of WDI consisted of 15 441 structures and was created by removing the following structures: any structures that contained elements other than C, N, O, F, P, S, Cl, Br, and I; structures with no activity

Table 2. Amide Libraries Chosen with Different Characteristics[a]

w_D	w_{RB}	D_{SUM}	Δ_{RB}
1.0	0.0	0.595 (0.001)	0.381 (0.025)
0.0	1.0	0.528 (0.007)	0.169 (0.004)
1.0	1.0	0.574 (0.004)	0.170 (0.002)

[a] SELECT was run to choose amide libraries with different characteristics. Each row shows the average results and standard deviation (in brackets) over 10 runs. The population size was 50. The first row shows the results of running SELECT to optimize diversity alone. In the second row the library is designed to have a "drug-like" rotatable bond profile. In the third row the library is designed to be "drug-like" and diverse.

class assigned; structures that are labelled as "trial-prep"; and structures belonging to the activity classes pesticides and plant hormones (except for fungicides), zootoxins, toxins, surfactants, diagnostics, chelators, and absorbents. Rotatable bonds were defined using the following SMARTS[17] definition:

```
[!$([NH]!@C(=O))&!D1&!$(*#*)]-&!@[!$([NH]!@
                    C(=O))&!D1&!$(*#*)]
```

The distribution was represented by 20 bins: the first bin recording the number of structures containing exactly 0 rotatable bonds; the second bin recording the number of structures containing exactly 1 rotatable bond; and so on, with the 20th bin recording the number of structures containing ≥ 19 rotatable bonds. Nearly 5% of the WDI subset consists of structures with ≥ 19 rotatable bonds, which is shown as a peak at the end of the distribution. Highly flexible compounds such as these are not generally considered to be "drug-like" and, hence, the compounds were removed, resulting in a smoother distribution that tends to zero for large numbers of rotatable bonds.

Table 2 shows the results in terms of diversity and rotatable bond profile when different terms and weights are used in the fitness function optimized by SELECT. In each case, the population size is 50 and the number of niches is set at one. The average results and standard deviations over 10 runs are shown. The first row shows the results for the previous experiment; that is, when SELECT is run to optimize diversity alone and when the population size is 50. The distributions of rotatable bonds in the combinatorial libraries found in these runs were calculated and compared with the reference profile. The RMS difference was scaled by dividing by 10. The difference in rotatable bond profile between the combinatorial libraries and the reference profile, Δ_{RB}, is shown in the last column. A range of different values of Δ_{RB} were found (from 0.329 to 0.408) so that libraries with similar diversity values can have significantly different rotatable bond profiles. The average number of rotatable bonds per molecule in the most diverse library found is 8.9 as compared with 6.2 for WDI. Inspection of this library shows that it contains a number of highly flexible molecules that are not obvious candidates for lead compounds.

SELECT was then run to minimize the difference in rotatable bond profiles between the library being selected and the reference profile. The results are shown in the second row of Table 2. The average result in this case has a rotatable bond profile that is much more similar to that of WDI (as

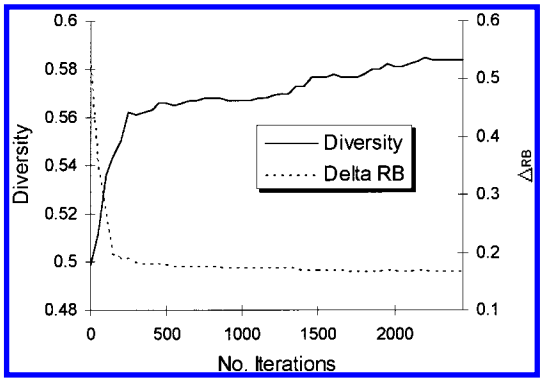

Figure 2. The evolution of an amide library optimized for diversity and rotatable bond profile. The diversity of the library and the RMS difference in rotatable bond profile of the library compared with the rotatable bond distribution in WDI are shown.

reflected in the markedly lower value for Δ_{RB}); however, this is achieved at a considerable loss in the diversity of the library (D_{SUM} averaged over 10 runs is 0.528). The remaining runs attempted to optimize both diversity and the rotatable bond profile for the library. It can be seen that a compromise between diversity and the optimum distribution of rotatable bonds is achieved when each term is weighted equally. In this case, the best library over the 10 runs has an average of 6.3 rotatable bonds per molecule and diversity of 0.581. Figure 2 shows how diversity D_{SUM} and Δ_{RB} vary over a run (when they are equally weighted). Figure 3a shows the distribution of rotatable bonds for a library optimized on diversity alone (Lib 1) superimposed on the distribution of rotatable bonds in WDI, and Figure 3b shows the rotatable bond profile of a library optimized for diversity and rotatable bonds (Lib 2) superimposed on the distribution in WDI.

Finally the effect of niching was investigated. The runs in Table 2 were repeated but in each case SELECT was run using five niches with a niche size of 0.8; that is, all libraries with 80% or more reactants in common with the library at the center of a niche were included within the niche and hence were excluded from the search space. The overall best solution over all the niches was then presented as the best library. In each case, SELECT was run 10 times and the overall best solutions were averaged. It can be seen from Table 3 that a slightly improved result is achieved by the use of niching; it hence appears that the niching process forces the GA to search more of the search space and it is more likely that the optimum solution will be found. As already mentioned, niching is also a useful way of indicating "good" solutions that occupy different regions of structure space.

Thiazoline-2-Imine Library. SELECT was then used to investigate a three-component library. The library is based on a thiazoline-2-imine template[23] and the reaction is shown in Figure 4. The R1 reactants are isothiocyanates, the R2 reactants are amines, and the R3 reactants are haloketones. Reactants for each pool were extracted at random from SPRESI. The pools consisted of 12 isothiocyanates, 99 amines, and 54 halo-ketones, representing a fully enumerated virtual library of 70 092 thiazoline-2-imines. The library was enumerated, Daylight fingerprints were calculated and stored, and rotatable bonds and Andrews' binding energies[26] were calculated for each of the molecules. The Andrews' binding energy is an example of a readily calculable property that

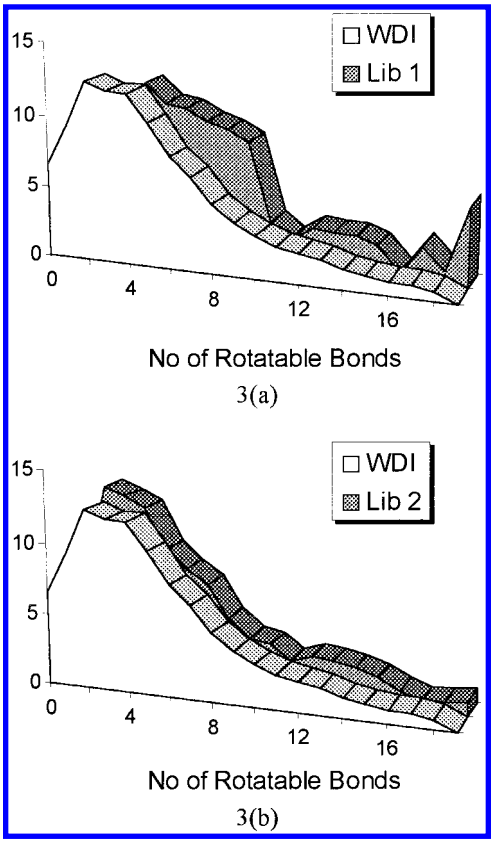

Figure 3. (a) The area labeled WDI shows the distribution of rotatable bonds in WDI. Lib 1 shows the distribution of rotatable bonds in an amide library selected for optimum diversity. (b) Lib 2 shows the distribution of rotatable bonds in an amide library selected for optimum diversity and rotatable bond profile. The average number of rotatable bonds per molecule are 6.2 for WDI, 8.9 for Lib 1, and 6.3 for Lib 2.

Table 3. Amide Libraries Chosen by Niching[a]

w_D	w_{RB}	D_{SUM}	Δ_{RB}
1.0	0.0	0.597 (0.001)	0.362 (0.026)
0.0	1.0	0.524 (0.006)	0.167 (0.001)
1.0	1.0	0.582 (0.002)	0.170 (0.001)

[a] The runs in Table 2 were repeated but using niching. For each set of weights, the results over 10 runs were averaged, and in each run, the best solution (the solution with the lowest value of the fitness function) found over five niches of size 0.8 was used to calculate the average result and standard deviation.

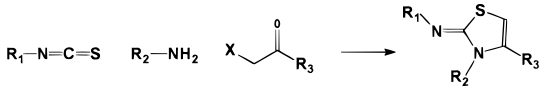

Figure 4. The thiazoline-2-imine library.

can rapidly identify molecules that are large, flexible, and lipophilic. The presence of such undesirable molecules was a problem encountered with some early combinatorial libraries. SELECT was parameterized to select eight isothiocyanates, 40 amines, and 20 haloketones, representing a combinatorial library of 6400 molecules.

The first set of runs was designed to optimize diversity alone. The results averaged over 10 runs for a population of size 50 and one niche are shown in Table 4. The most diverse library had D_{SUM} of 0.429. For comparison, the diversity of libraries selected at random averaged over 100 libraries is 0.384, with standard deviation 0.007, and the least diverse

Table 4. Optimal Diversity of the Thiazoline-2-imine Library[a]

pop	no. Its	t (s)	D_{SUM}
50	720 (118)	5132 (840)	0.427 (0.001)

[a] SELECT was run to optimize the diversity of the thiazoline-2-imine library. The results over 10 runs are averaged, with standard deviations shown in brackets. For comparison, the diversity of 100 libraries selected at random is 0.384 (0.007) and the diversity of the most similar library found by running select with $w_D = -1.0$ is 0.322.

Table 5. Thiazoline-2-imine Libraries with Different Characteristics[a]

w_D	w_{RB}	w_{BE}	no. Its	D_{SUM}	Δ_{RB}	Δ_{BE}
1.0	0.0	0.0	720 (118)	0.427 (0.001)	0.701 (0.018)	0.535 (0.019)
0.0	1.0	0.0	770 (181)	0.385 (0.004)	0.505 (0.002)	0.656 (0.029)
0.0	0.0	1.0	1175 (329)	0.398 (0.004)	0.699 (0.027)	0.363 (0.005)
1.0	1.0	0.0	1025 (274)	0.404 (0.003)	0.507 (0.002)	0.598 (0.026)
1.0	0.0	1.0	1065 (330)	0.413 (0.003)	0.707 (0.022)	0.366 (0.004)
1.0	1.0	1.0	1470 (352)	0.400 (0.002)	0.544 (0.008)	0.385 (0.005)

[a] SELECT was run to choose thiazoline-2-imine library libraries with different characteristics. The results are shown for optimizing each of the criteria diversity, rotatable bond profile, and binding energy profile alone and in combination. The parameters w_D, w_{RB}, and w_{BE} are the weights assigned to diversity, rotatable bonds, and Andrews' binding energy, respectively.

library (found by running SELECT with w_D set to -1.0) has diversity of 0.322. The average run time was 1.4 h on a Silicon Graphics R10000 processor running at 195 MHz, and the memory requirement for the virtual library was 11.2 Mbytes.

Next, the effect of optimizing diversity alongside rotatable bond profiles and binding energies was investigated. The rotatable bond profile for WDI was used as a reference profile, as described previously. The Andrews' binding energies (in kcal/mol) were calculated for the 15 441 subset of WDI used in the previous experiments, and the distribution represented by 20 bins. Each bin covered a range of 4 units, with the first bin representing the number of structures in the WDI subset having Andrews' binding energies of <-6.0, the second bin representing the number of structures in the range -6.0 to -1.9, and the 20th bin representing the number of structures with binding energy ≥ 7.0 kcal/mol. The distribution for WDI again showed a peak in the 20th bin, with a relatively large number of molecules having very large binding energy. These structures were removed from the distribution and so the final bin was again set to 0. When the libraries are optimized for rotatable bond profile or binding energy profile, the appropriate distributions in the combinatorial libraries were calculated and compared with the appropriate reference profiles. The RMS difference was scaled by dividing by 10 for both features.

The results for various weighting schemes used in the fitness function are shown in Table 5. The evolution of two thiazoline-2-imine libraries resulting from running SELECT with different fitness functions is shown in Figure 5. The library Lib 3 was optimized for diversity alone, whereas library Lib 4 was optimized for diversity and rotatable bond profile. Figure 5(a) shows the evolution of diversities of the two libraries. Figure 5(b) shows the evolution of the rotatable bond profiles. Figure 5(c) shows the simultaneous evolution of diversity and rotatable bond profile for Lib 4. Figures 5(a) and 5(b) taken together indicate that a better rotatable bond

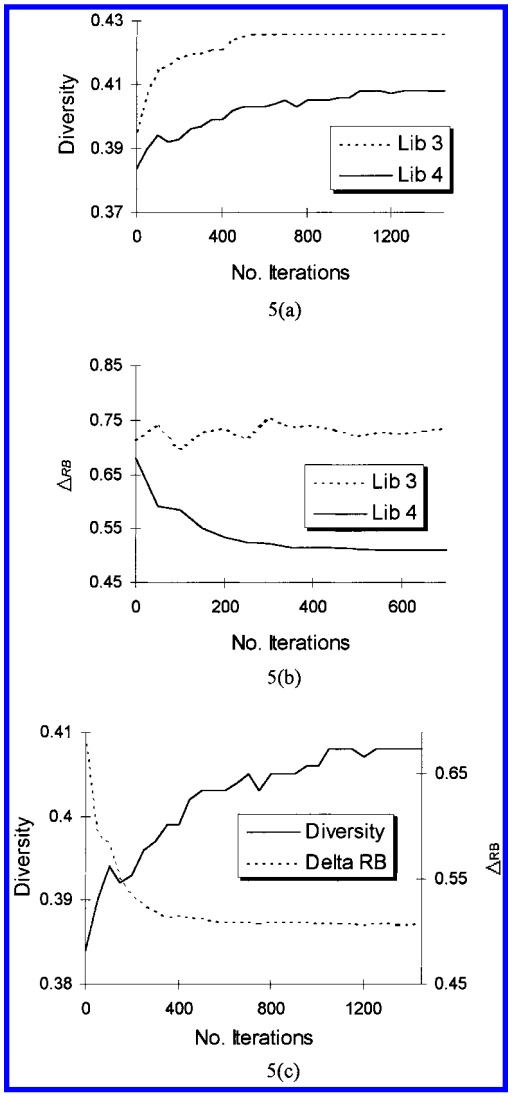

Figure 5. The evolution of two different thiazoline-2-imine libraries is shown. Lib 3 was optimized for diversity alone, whereas Lib 4 was optimized for diversity and rotatable bond profile. (a) The evolution of diversities of the two libraries. (b) The evolution of the rotatable bond profiles. (c) The simultaneous evolution of diversity and rotatable bond profile for Lib 4, where D_{SUM} is 0.408 and Δ_{RB} is 0.507. Figures 5a and 5b show that a better rotatable bond profile is achieved at the expense of some diversity in the library.

profile is achieved at the expense of some diversity in the library. The rotatable bond profiles for the combinatorial libraries Lib 3 and Lib 4 are shown in Figure 6 superimposed on the rotatable bond profile for WDI. The average number of rotatable bonds per molecule in library Lib 3 is 11.4, whereas this value has fallen to 9.4 for Lib 4 (compared with the mean value of 6.2 for WDI). Lib 4, therefore, represents a more "drug-like" library than Lib 3.

Figure 7 shows the evolution of a library that is optimized for diversity and binding energy profile. The binding energy profile of this library (Lib 5) is shown in Figure 8 superimposed on the binding energy profile of WDI. For comparison, the binding energy profile of a library optimized for diversity alone, Lib 6, is also shown. Finally, Figure 9 shows the evolution of a thiazoline-2-imine library optimized for diversity, rotatable bond profile, and binding energy profile. In this case, the diversity of the library (D_{SUM}) is 0.401.

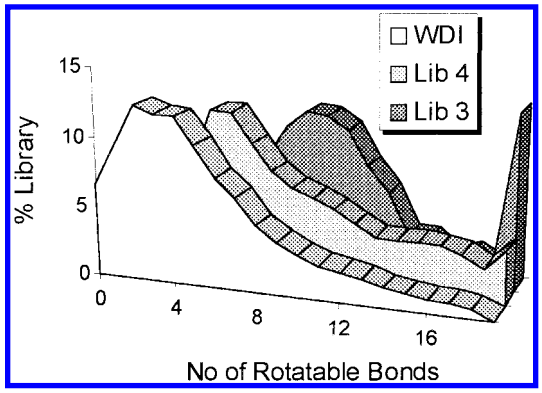

Figure 6. The area labeled WDI shows the distribution of rotatable bonds in WDI. Lib 3 shows the distribution of rotatable bonds in a thiazoline-2-imine library selected for optimum diversity. Lib 4 shows the distribution of rotatable bonds in a thiazoline-2-imine library selected for optimum diversity and rotatable bond profile. The average numbers of rotatable bonds per molecule are 6.2 for WDI, 11.4 for Lib 3, and 9.4 for Lib 4.

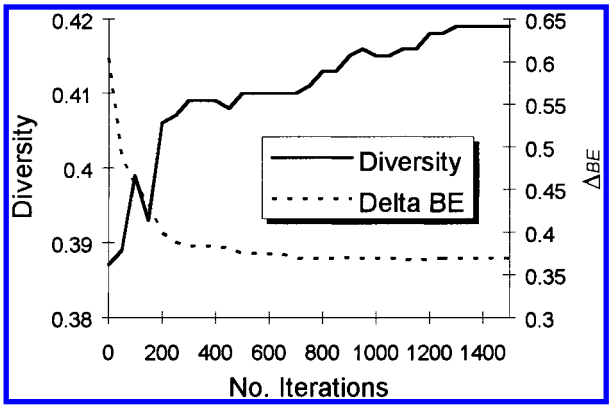

Figure 7. The evolution of a thiazoline-2-imine library (Lib 5) optimized for diversity and binding energy profile.

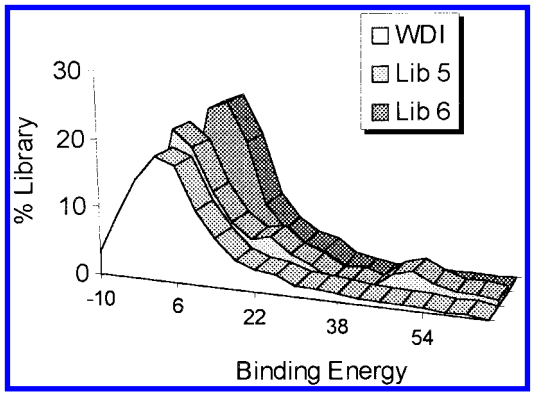

Figure 8. The binding energy profile of Lib 5 is shown together with the binding energy profile of a library optimized for diversity alone (Lib 6) superimposed on the binding energy profile for WDI.

The effect of niching is shown in Table 6. Some of the runs in Table 5 were repeated, but in each case, SELECT was run using five niches with a niche size of 0.8 as for the amide library. The first four rows of Table 6 show the average results and standard deviations for the best solution in each run, where the best solution is the library with the lowest scoring fitness function. The last row of Table 6 shows the results for the same runs as shown in row four; however, in this case, the best library in each run is chosen as the library with greatest diversity. Again the slightly improved results over Table 5 suggest that the niching process forces

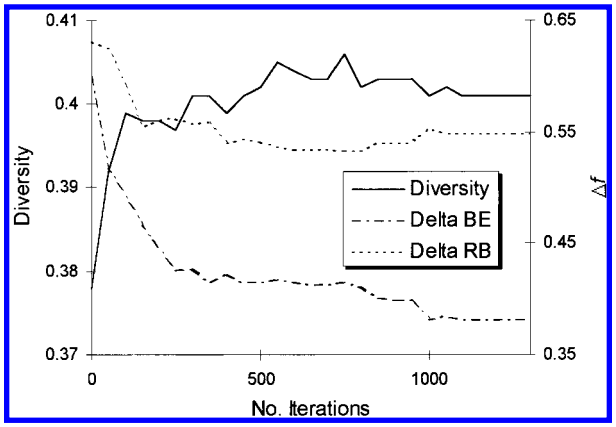

Figure 9. The evolution of a thiazoline-2-imine library optimized for diversity, rotatable bond profile, and binding energy profile.

Table 6. Thiazoline-2-imine Libraries Chosen by Niching[a]

w_D	w_{RB}	w_{BE}	D_{SUM}	Δ_{RB}	Δ_{BE}
1.0	0.0	0.0	0.427 (0.001)	0.704 (0.023)	0.528 (0.028)
1.0	1.0	0.0	0.405 (0.001)	0.507 (0.002)	0.599 (0.019)
1.0	0.0	1.0	0.415 (0.002)	0.729 (0.040)	0.366 (0.003)
1.0	1.0	1.0	0.397 (0.002)	0.549 (0.008)	0.379 (0.005)
1.0	1.0	1.0	0.405 (0.002)	0.559 (0.030)	0.412 (0.020)

[a] Some of the runs in Table 5 were repeated using niching. For each set of weights, the results over 10 runs were averaged. In the first four rows of the table, the averages and standard deviations were calculated by taking the best solution from each run, where the best solution is the one with the lowest value of the fitness function found over the five niches of size 0.8. In the last row of the table, the best solution for a run was taken to be the one with highest diversity found over the five niches of size 0.8.

the GA to search more of the search space, and it is more likely that the optimum solution will be found. When a multicomponent fitness function is used, the best solution in one niche may have a similar score as the best solution in a different niche; however, the relative values of the individual components of the fitness function may differ.

Average Nearest Neighbor Distance. In all of the experiments described so far, the diversity of a combinatorial library has been measured using D_{SUM} (i.e., the sum of pairwise dissimilarities over all molecules in the library). However, it has been shown that this diversity measure has a tendency to favor outliers, with the effect that libraries that contain molecules that have close near neighbors can score highly.[27] Hence, an alternative diversity measure has been implemented in SELECT; that is, the average nearest neighbor distance using the Tanimoto coefficient, D_{NN}. Indeed it would be possible to include any diversity index subject to the index being calculated sufficiently rapidly for the processing of large libraries. Using the D_{NN} index the amide runs were approximately 10 times slower than using the D_{SUM} index, taking on average 1.2 h on a Silicon Graphics R10000 processor running at 195 MHz. As already mentioned, the time complexity of the average nearest neighbor diversity measure is $O(N^2)$ because it involves the calculation of all the pairwise distances where the distances are measured using the Tanimoto coefficient. However, it is possible to reduce the number of similarity calculations performed for D_{NN} (by ordering the molecules in the set according to the number of bits set[27]), although the calculation remains much slower than using the D_{SUM} measure.

Table 7. Selecting 10 × 10 Libraries from the 100 × 100 Amide Library[a]

w_D	w_C	D_{SUM}	% in common
1.0	0.0	0.595 (0.002)	46.6 (12.3)
0.0	1.0	0.584 (0.003)	15.6 (3.1)

[a] The results are shown for selecting 10 × 10 libraries from the 100 × 100 amide library with different weights assigned to *D* and *C* in the fitness function. Initially, SELECT was run to generate a diverse library (Lib 7) with diversity measured to be 0.595. The first row represents 10 libraries that are also selected to be diverse. The final column gives the percentage of compounds in a library that are common to Lib 7, averaged over the 10 runs, with standard deviation in brackets. The second row shows the results for libraries that are selected to complement Lib 7 by including a combined diversity term *C* that is calculated using the centroid of Lib 7. The degree of overlap between each of the libraries represented in the second row with Lib 7 now averages 16%.

The average nearest neighbor measure can give useful results for small libraries, such as those used in the amide experiments, but its slowness prohibits its use for large libraries, such as the thiazoline-2-imine libraries, given our current computing resources.

Maximizing the Difference between Libraries. The amide library was used to examine the effect of forcing a library to be different from an existing collection. SELECT was run to choose a 10 × 10 library (Lib 7) optimized for diversity alone (i.e., w_D was set to 1.0 and all other weights were set to zero). The diversity of this library is 0.595. SELECT was then run 10 times to generate new diverse libraries, and the degree of overlap of each library with Lib 7 was calculated and averaged over the 10 runs. The results are shown in the first row of Table 7, where it can be seen that on average there is a large degree of overlap between each new library and Lib 7. The centroid of Lib 7 was then calculated and further runs were carried out to select libraries that are complementary to Lib 7. The results are shown in the second row of Table 7 where it can be seen that there is now a much smaller degree of overlap (approximately 16%) between the libraries and Lib 7.

Choosing an Optimum Configuration for a Library. It can be very useful to investigate the effects of changing the relative numbers of reactants used in a combinatorial synthesis, especially when some reactants are relatively rare or expensive to acquire. The diversity and physical property profiles of libraries that result from different configurations of reactants can be compared. For example, the thiazoline-2-imine reaction was used to investigate the effect on diversity of changing the configuration of potential libraries, where by configuration we mean the numbers of the different reactants required to produce a combinatorial library of a pre-defined size. Libraries consisting of 2880 compounds were selected according to the configurations shown in Table 8. The starting pools were the same as in the previous experiments. It can be seen that different configurations of the library can result in different maximum diversities and that varying the number of isothiocyanates is the most critical factor affecting the maximum diversity that is achievable, with the results indicating that 4−8 isothiocyanates are required to maximize diversity. Having restricted the number of isocyanates, there are then several different configurations available that result in libraries of very similar maximum diversities and the numbers of amines and haloketones can

Table 8. Effect on Diversity of Changing the Configuration of the Thiazoline-2-imine Library[a]

pool 1	pool 2	pool 3	D_{SUM}
1	72	40	0.386
2	72	20	0.405
2	36	40	0.409
4	72	10	0.423
4	36	20	0.431
4	18	40	0.430
8	72	5	0.421
8	36	10	0.433
8	18	20	0.433
8	9	40	0.430

[a] Investigating the effect on diversity of changing the configuration of the thiazoline-2-imine library with a library size fixed to 2880 compounds. Pool 1 corresponds to isocyanates, Pool 2 corresponds to amines, and Pool 3 corresponds to haloketones.

be chosen according to how easy they make the automated chemistry, without affecting the overall diversity of the final library. Performing this type of experiment enables the chemist to make an informed decision on which configuration to choose, and allows a balance of the ease of synthesis against the loss of library diversity.

CONCLUSIONS

The program SELECT has been described for the design of combinatorial libraries that have user-specified structural and physical properties. SELECT is based on a genetic algorithm with a multicomponent fitness function. Typical properties that can be included in the fitness function are diversity, which can be measured using Daylight fingerprints and D_{SUM} or D_{NN}, and physical properties that are thought to influence the ability of a compound to act as a drug, for example, the number of rotatable bonds. SELECT can also be used to suggest libraries, that while being internally diverse, are also structurally distinct from some existing collection of compounds.

SELECT is current being used to design combinatorial libraries at GlaxoWellcome.[16] It has been used to determine the optimum configuration of a three-component library consisting of approximately 12 000 compounds selected from a 100 × 40 × 200 virtual library. Subsequently, libraries have been designed that are both optimally diverse and that have physical properties that are similar to compounds in the WDI. The libraries are currently being synthesised.

ACKNOWLEDGMENT

We thank GlaxoWellcome Research and Development for funding and Daylight Chemical Information Systems Inc. for software support. The Krebs Institute for Biomolecular Research is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES AND NOTES

(1) Dunbar, J. Jr.; Cluster-Based Selection. *Perspect. Drug Discov. Design* **1997**, *7/8*, 51−63.
(2) Lajiness, M. S. Dissimilarity-Based Compound Selection *Perspect. Drug Discov. Design* **1997**, *7/8*, 65−84.
(3) Mason, J. S.; Pickett, S. D. Partition-Based Selection *Perspect. Drug Discov. Design* **1997**, *7/8*, 85−114.
(4) *Combinatorial Chemistry*; Czarnik, A. W.; DeWitt, S. H., Eds. American Chemical Society: Washington, 1997.
(5) Willett, P. Computational Tools for the Analysis of Molecular Diversity *Perspect. Drug Discov. Design* **1997**, *7/8*, 1−11.
(6) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity; Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.
(7) 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.
(8) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification Of Biological Activity Profiles Using Substructural Analysis And Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.
(9) Goldberg, D. E. *Genetic Algorithms in Search, Optimisation, and Machine Learning*; Addison-Wesley: Reading, MA, 1989.
(10) Clark, D. E.; Westhead, D. R. Evolutionary Algorithms in Computer-Aided Molecular Design. *J. Comput.-Aid. Mol. Design* **1996**, *10*, 337−358.
(11) Sheridan, R. P.; Kearsley, S. K. Using a Genetic Algorithm to Suggest Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 310−320.
(12) Weber, L.; Wallbaum, S.; Broger, C.; Gubernator, K. Optimization of the Biological Activity of Combinatorial Libraries by a Genetic Algorithm. *Angew. Chem. Int. Ed., Engl.* **1995**, *34*, 2280−2282.
(13) Singh, J.; Ator, M. A.; Taeger, E. P.; Allen, M. P.; Whipple, D. A.; Soloweji, J. E.; Chowdhary, S.; Treasurywala, A. M. Application of Genetic Algorithms to Combinatorial Synthesis - A Computational Approach to Lead Identification and Lead Optimization. *J. Am. Chem. Soc.* **1996**, *118*, 1669−1676.
(14) Brown, R. D.; Martin, Y. C. Designing Combinatorial Library Mixtures Using a Genetic Algorithm. *J. Med. Chem.* **1997**, *40*, 2304−2313.
(15) Good, A. C.; Lewis, R. A. New Methodology for Profiling Combinatorial Libraries and Screening Sets: Cleaning Up the Design Process with HARPick. *J. Med. Chem.* **1997**, *40*, 3926−3936.
(16) Green, D. SPICE Up Your Life: Computational Aids for Smarter Library Synthesis. Presented at *Computational Approaches to the Design and Analysis of Combinatorial Libraries*, Molecular Graphics and Modelling Society and the Chemical Structure Association; Sheffield, UK, 1998.
(17) *Daylight Theory Manual*; Daylight Chemical Information Systems: Mission Viejo, CA.
(18) Beasley, D.; Bull, D. R.; Martin, R. R. A Sequential Niche Technique for Multimodal Function Optimization. *Evol. Comput.* **1993**, *1*, 101−125.
(19) Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm for Selecting Sets of Dissimilar Molecules from Large Chemical Databases. *Quant. Struct.-Act. Relat.* **1995**, *14*, 501−506.
(20) Matter, H.; Lassen, T. Compound Libraries for Lead Discovery. *Chimica Oggi-Chemistry Today* **1996**, *14*, 9−15.
(21) 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.
(22) Pickett, S. D.; Luttmann, C.; Guerin, V.; Laoui, A.; James, E. DIVSEL and COMPLIB - Strategies for the Design and Comparison of Combinatorial Libraries Using Pharmacophoric Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 144−150.
(23) Watson, S. Solution Phase Synthesis of Libraries Based on Thiazole Templates. 3rd Annual Random and Rational Conference; Geneva, 1996; Strategic Research Institute: New York.
(24) The SPRESI database is produced by the All-Union Institute of Scientific and Technical Information of the Academy of Science of the USSR (VINITI) in Moscow, and the Central Information Processing for Chemistry (ZIC) in Berlin. It consists of data extracted from approximately 1000 journals, and also patents, books and other sources from 1975−1990. It is distributed by Daylight Chemical Information Systems, Inc., Mission Viejo, CA.
(25) The World Drug Index (WDI) is maintained by Derwent Publications Ltd., London.
(26) Andrews, P. R.; Craik, D. J.; Martin, J. L. Functional Group Contributions to Drug-Receptor Interactions. *J. Med. Chem.* **1984**, *27*, 1648−1657.
(27) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of Algorithms for Dissimilarity-Based Compound Selection. *J. Mol. Graphics. Model* **1997**, *15*, 372−385.

CI980332B