ELSEVIER

Combinatorial library design for diversity, cost efficiency, and drug-like character

Robert D. Brown, Moises Hassan, and Marvin Waldman

Molecular Simulations, Inc., San Diego, California, USA

Most computational techniques for the design of combinatorial libraries have concentrated solely on maximizing the diversity of the selected subset or its similarity to a known target. However, such libraries can produce high-throughput screening hits with properties that make them unsuitable to take forward into medicinal chemistry. This article describes software that allows the design of library subsets to simultaneously optimize a library's diversity or similarity to a target, properties (such as drug likeness) of the library members, properties (such as cost) of the reagents required to make them, and efficiency of synthesis in arrays or mixtures. Example are given showing that libraries can be designed to contain drug-like molecules with only a small trade-off in terms of the maximum possible diversity, and that the cost of the library, in terms of the reagents required to make it, can be contained. Other examples show that libraries can be designed to minimize the deconvolution problem or to maximize the number of molecules predicted to be active while also being designed for efficiency of synthesis. © *2000 by Elsevier Science Inc.*

Keywords: library design, diversity, similarity, drug-likeness, Lipinski's Rules

INTRODUCTION

Combinatorial chemistry is now an established part of the drug discovery process, and computational methods are playing an increasingly important role in several stages of this process. Computational library design typically starts with the specification of a *virtual combinatorial library* that contains all the molecules that might be made given the constraints of the chemistry being employed and the reagents that are commercially or synthetically accessible and thought to be compatible with that chemistry. Virtual libraries frequently are far too large to consider synthesizing in their entirety and often contain a high level of near redundancy based on their chemical properties. Thus, a computational *library subsetting* process is employed to analyze the virtual library and select a subset for synthesis and screening; this subset is termed hereafter the *design library.*

Early computational methods for library subsetting concentrated on a single aspect of design, namely, the diversity (or similarity) of the products to be made or even more simply of the reagents that would be used.[1-3] However, there has been a growing realization that this is a somewhat naïve view. Its application has led to libraries that, although often producing hits in high-throughput screens, have not identified leads amenable to follow-up lead optimization studies or have produced leads that subsequently suffered from problems of bioavailability. It now is clear that the library design process should incorporate other factors that contribute to producing a library whose hits can be optimized into good drug candidates. Both properties of the product molecules and properties of the reagents that will be used to make them should be considered in the design of an appropriate library to synthesize.

Library Molecule Properties

Design libraries intended for lead identification should seek to sample all the variation of the molecules in the virtual library. Concepts such as *diversity, coverage,* and *representativeness* are used to ensure a good sampling of the virtual library using the minimum number of molecules. All methods are dependent on the similar property principle, which states that structurally similar molecules tend to have similar activities.[4] It should be noted that this is a probabilistic argument with an emphasis on the *tend* in the previous sentence. One is always aware of individual situations in which very small changes in structure can lead to major changes (typically loss) of activity. It also is important to note that the correct selection of a chemical space into which the compounds are embedded is essential for this *neighborhood* behavior to hold.

Diversity- or coverage-based selection aims to ensure that all

Color Plates for this article are on page 537.

Corresponding author: Robert D. Brown, Molecular Simulations, Inc., 9685 Scranton Road, San Diego, CA 92121, USA. Tel.: 858-458-9990; fax: 858-458-3752. *E-mail address*: rbrown@msi.com (R.D. Brown).

Journal of Molecular Graphics and Modelling 18, 427–437, 2000

655 Avenue of the Americas, New York, NY 10010

1093-3263/00/$–see front matter
PII S1093-3263(00)00072-3

selected molecules are maximally different from each other so as to sample/cover as much chemical property space as possible, thereby increasing the likelihood of finding a set of quite different leads in a screening experiment. Representativeness aims to sample the virtual library in such a way that the distribution of compounds in chemical space is similar to that found in the virtual library, i.e., densely populated areas of space are sampled more than sparse areas of space.

Design libraries intended for lead optimization may seek to sample the local neighborhoods around each active structure of interest, using a similarity search. Alternatively, they may be designed to maximize the predicted activity based on the use of a model. Such models may be built from high-throughput screening data using pharmacophore modeling or classification techniques such as recursive partitioning or linear discriminant analysis. Alternatively, models for estimating the binding energy may be available based on knowledge of the protein active site and affinities of known compounds.

Whether libraries are designed for lead identification or optimization, it is important to consider the molecular properties of the molecules selected from the virtual library. Considerable emphasis has been placed lately on identifying the characteristics of compounds that make them "lead like" or "drug like" (i.e., good medicinal chemistry or development candidates). The intent is to preferentially select molecules for screening that, if found to be active and selected for further development, are not likely to exhibit problems in absorption, delivery, metabolism, excretion, or toxicology (ADME/Tox). In this way, the drug discovery and development cycle should be shortened and the failure rate of compounds should be reduced.

Other constraints on whole library properties may arise from the way in which the library will be deconvoluted or decoded to identify the active molecules following high-throughput screening. Again, this will be dependent on the methodology by which the library will be synthesized and screened and whether a tagging strategy is being used. As an example, Brown and Martin[5] describe a library decoding strategy for mixtures in which the molecular weight of the active molecule(s) is identified during the screening process. In order to identify the active molecules themselves, every molecule in the screening sample with molecular weight equal to that of any of the active samples must be resynthesized and rescreened individually. Constraining the design library to have only a few occurrences of molecules with any one molecular weight minimizes the effort necessary for the deconvolution process.

Reagent Properties

It is assumed in this discussion that prior to specification of the virtual library, reagent lists have been assembled to be compatible with the chemistry to be used. Furthermore, it also is assumed that the lists have been filtered to remove reagents with reactive or other undesirable functionalities. In this way, every member of the reagent list is assumed, within the constraint of the current knowledge of the chemistry, to be appropriate on a chemical compatibility basis and available for use in the design library. These assumptions notwithstanding, economic and supply considerations suggest that not all reagents may be equally desirable for use in a library.

Factors to consider in selecting reagents include the total cost of the reagents that will be used, a preference for reagents that are already available in in-house stockrooms or from preferred suppliers, and even a chemist's intuition about the desirability of using each reagent. Furthermore, it may be desirable to minimize the number of reagents used. (Note that this is not the same as using the minimum possible number of reagents to make a given number of products, which is a concept captured in the combinatorial constraint discussed later). Finally, it may be desirable to minimize the number of different suppliers that are used, both to ensure timely delivery and reduce the complexity of the ordering process.

The efficiency of synthesis of a library is dependent on the methodology (either automated or manual) used in its production. Often, a *combinatorial constraint* is applied to the library design. This constraint requires that every reagent used at each diversity position be used in all combinations with all reagents at every other diversity position and be used in the design of full arrays and mixtures. The combinatorial constraint ensures that the maximum number of products is obtained from the minimum number of reagents and that the library will be synthetically efficient on array synthesis automation. It does, however, place a restriction on the products that can be made, because individual products can no longer be "cherry picked" to be included in the library. Instead, the selection of a reagent implies that all products of that reagent with the selected reagents at all other diversity sites must be included in the design. For example, a combinatorial subsetting problem can be expressed as the selection of a $10 \times 8 \times 12$ library, with the implication that all 960 products will be made, whereas the equivalent cherry-picking problem would be expressed as the selection of a 960-member library.

As an aside, an alternative way of handling the combinatorial constraint is to independently run a diversity analysis on each reagent list and then combine the results. However, studies have suggested that this usually will not lead to as diverse a subset of products as a product-based analysis.[6,7]

In summary, we have shown the library subsetting problem to be a complex one in which a multitude of competing factors should be optimized simultaneously, namely,

1. Library product properties

- Overall diversity or similarity to a target
- Predicted activity
- Drug-like character
- Deconvolution/decoding strategy

2. Reagent properties

- Cost of reagents
- Availability of reagents
- Total number of reagents or suppliers
- Combinatorial constraint

An additional problem that must be faced in library subsetting is that the search spaces are extremely large, because the number of appropriate reagents that are available for many typical chemistries is high, and so the problem cannot seemingly be solved deterministically. For example, consider a library with three positions of diversity in which there are 100 reagents available for use at each position, giving a virtual library of one million members. Consider the problem of selecting a library of 1,000 molecules from this virtual library. If the library is not combinatorially constrained, then there are

$C_{1000}^{1000000} \approx 10^{3400}$ possible libraries. However if the combinatorial constraint applies and the subsetting problem is actually to select a $10 \times 10 \times 10$ library, then there are $C_{10}^{100} \times C_{10}^{100} \times C_{10}^{100} \cong 5 \times 10^{39}$ possible libraries. In its favor, this does mean that, in general, there are many possible design libraries of any particular size to choose from.

This article describes methodologies available in the Cerius[2] software from Molecular Simulations Inc.[8] for stochastically searching the space of possible design libraries to identify one or more that satisfy a complex set of requirements on both the reagents and products. Example library designs will be presented showing the application of various aspects of the possible set of constraints that can be employed. Specifically, examples are given for which a design library is selected to be

- both diverse and drug like
- diverse and optimized for deconvolution
- diverse, drug like, and optimized for the cost of the reagents required for its production
- drug like and optimized to contain the largest number of members predicted to be active using a virtual high-throughput screen.

METHODOLOGY

Evolutionary algorithms have been applied successfully to many problems in which there is both a vast search space and a number of factors that must be optimized simultaneously.[9,10] In this work, we made use of a Monte Carlo procedure similar to one we previously applied to the "cherry-picking" problem.[11] Two versions of the algorithm are implemented, one for simple product selection and the other that incorporates the additional problem of mapping between reagents and products in the case of combinatorially constrained selection.

Figure 1 shows a schematic view of the two versions of the algorithm. An initial selection of the required number of products (noncombinatorial) or reagents (combinatorial) is made at random. In the latter case, the corresponding products are identified. An objective function is evaluated for this design library. A product or reagent is selected at random and replaced with another. The objective function is re-evaluated, and the replacement is either accepted if it improves the current result or conditionally accepted according to a Monte Carlo Metropolis criterion (involving a user-definable "temperature" factor). Otherwise, the step is rejected. This process is repeated until convergence has been achieved (no improvement after a given number of cycles) or until a maximum number of cycles has been run. At this point, the resultant library products and, optionally, reagent lists are reported.

The procedure depends on the specification of an objective function to determine the suitability of the candidate library proposed at each iteration. The objective function consists of a weighted combination of terms, each of which accounts for one of the constraints on the products or reagents previously described in the introduction:

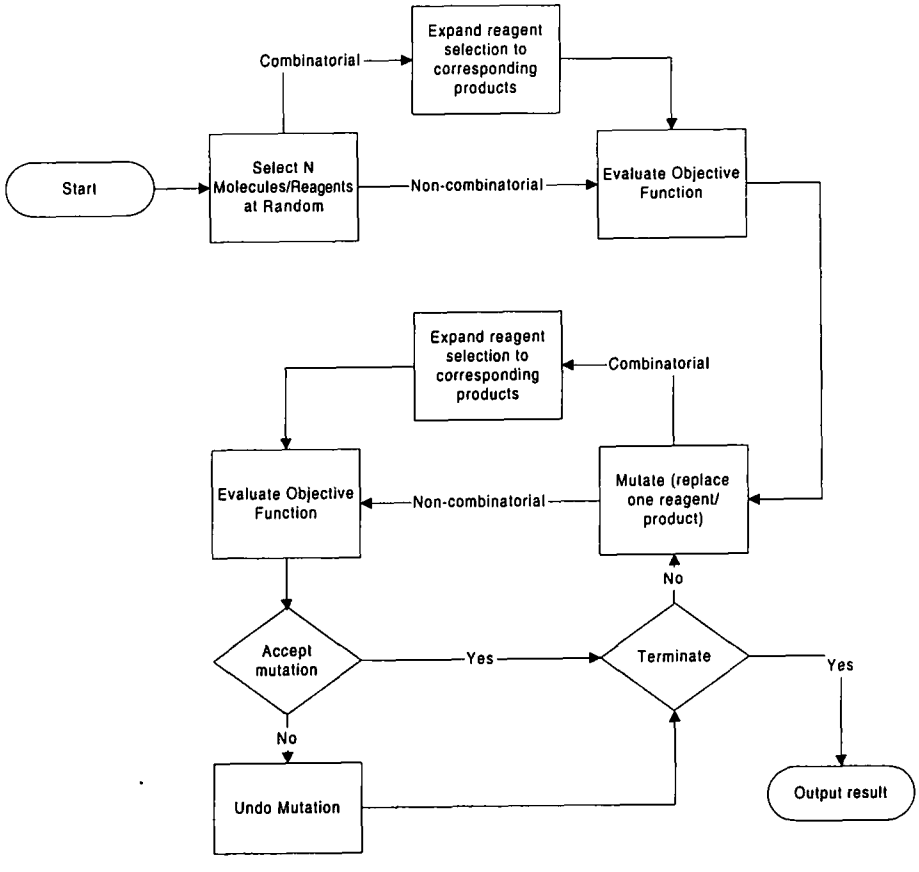

Figure 1. Program flow of the combinatorial and non-combinatorial Monte Carlo optimization procedures.

$$F = w_{ds}(DS) + \sum_{i=1}^{n} w_i \, ProductPenalty_i + \sum_{j=1}^{m} w_j \, ReagentPenalty_j$$

where DS = Diversity or Similarity score; w_{ds}, w_i, *and* w_j are weighting factors; and n and m are the number of product and reagent constraints respectively.

Each of these terms will be discussed. In order to be able to combine each term and assign reasonable weights for each of the factors, it is necessary to normalize each one. The normalization of each term also will be discussed.

Diversity and Similarity

Assessing diversity and similarity is a multistep process in which

1. A series of descriptors are computed for each compound under consideration.
2. The compounds are projected into a common chemical space of either the raw descriptors or a (usually smaller) set of descriptors derived from the originals using techniques such as principal components analysis or factor analysis.
3. A subset of compounds is selected based on their proximity within the chemical space. This may be based on either a distance computation or the division of space into cells and the comparison of cell occupancies of compounds.

An excellent review of all of these methods is given in a special issue of *Perspectives in Drug Discovery and Design*,[12] and it is not the intention of this article to discuss this aspect of the methodology in detail. Although we used a specific chemical space and diversity selection method in this work, many other methods would be equally applicable.

The chemical space used in the studies herein is the default descriptor set in the Cerius^2•Diversity[13] module. The descriptors used are chosen to be reasonably fast to calculate while still tending to group biologically similar molecules together.[14] The descriptors are a set of 50 physicochemical properties including molecular weight, AlogP,[15] number of hydrogen bond donors and acceptors, number of rotatable bonds, surface area and volume, Balaban,[16] PHI,[17] Kappa,[17] CHI,[17] Weiner,[18] and Zagreb[19] topological indices. A principal components analysis is applied to reduce the dimensionality to a level such that 90% of the variance of the original descriptors is explained by the most significant principal components.

Two classes of diversity metrics have been widely used. The first are based on measuring the distance between pairs of compounds and then maximizing the distances between the selected set for diversity or minimizing the distances for similarity. The second are cell-based metrics that divide each dimension in the chemical space into bins, thereby forming hyperboxes in the space and then sampling compounds from the boxes. For diversity, compounds are selected to sample different boxes; for similarity, compounds are drawn from the box(es) closest to the target. In either case, metrics are available that vary in the way in which they aim to achieve diversity, coverage, and representativeness.

The simplest sampling scheme for a cell-based selection is *cell-based fraction.* The space is first binned in such a way that there are sufficient occupied cells (in the virtual library) so that an ideal selection would allow each selected compound to occupy a different cell in the space (in the absence of the combinatorial constraint). A selection then is made to maximize the number of cells sampled by the sublibrary. Formally,

$$Diversity = \frac{N_{filled\_cells}}{N_{selected}}$$

The diversity score is normalized so a diversity score of 1.0 would be achieved if a perfect selection were made. A selection made by this method is diverse but is not necessarily representative of the distribution of all the compounds in the virtual library in chemical space. *Cell-based density* considers this original distribution and attempts to draw a sample that maintains the same relative distribution by drawing more compounds from highly occupied cells and fewer from sparsely occupied cells. Formally,

$$Diversity = \frac{\sum_i N_i \ln(N_i/M_i)}{N_{sel} \ln(N_{sel}/N_{mol})}$$

where N_i is the number of compounds in cell i for the candidate sublibrary, M_i is the number of compounds in cell i in the virtual library, and N_{mol} is the total number of molecules in the virtual library. Once again, the score is normalized such that a perfectly representative set would score 1.0, occurring when each N_i is proportional to M_i with a constant proportionality factor. In this work, cell-based fraction was used for the selections; however, cell-based density or alternative distance-based metrics also would be applicable.

Product-Based Properties

There are two methods implemented in the Cerius^2•LibProfile software[20] that allow for specification of desirable or undesirable features in the selected molecules, thereby allowing the selection to be directed toward the former. Restraints may either be applied as property ranges in which penalties are assigned to any selected molecule whose properties lie outside the desired range or libraries may be designed to mimic one or more prespecified distributions of various properties.

In the case of selecting drug-like molecules, the best known set of formal rules are those described by Lipinski et al.,[21] who propose that drug molecules typically observe these rules:

1. Molecular weight should not exceed 500
2. LogP should not exceed 5
3. The number of hydrogen bond acceptors should not exceed 10
4. The number of hydrogen bond donors should not exceed 5.

Teague et. al[22] note that these rules were derived from drugs and that the properties required of library compounds intended to provide leads suitable for further optimization may be rather different. Their analysis suggests that libraries with molecules having a molecular weight range 100–350 and logP range (as measure by clogP) of 1–3 will provide better leads than the drug-like molecules found by Lipinski's rules. They suggest that libraries designed to meet these criteria should provide micromilar affinity hits in high-throughput screens, and that these hits should allow for the discovery and exploitation of additional interactions at the lead optimization phase.

To incorporate Lipinski's or Teague's rules and other range-

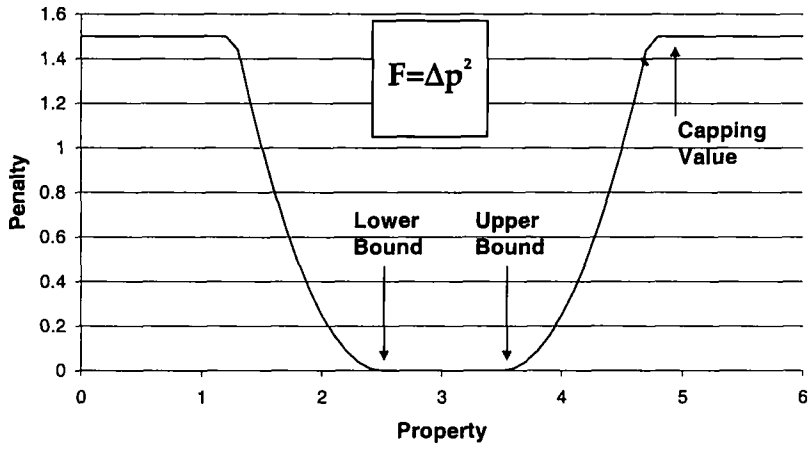

Figure 2. Penalty function for a penalty range.

based rules into a library design, product-based penalties are imposed by specifying an allowed range on any calculated or measured properties of the molecules. A penalty function (shown graphically in Figure 2) then is applied for each property in term. No penalty is incurred by any molecule that falls within the range, and the penalty increases with the square of the difference between the value and the nearest bound, up to a maximum at a user-specified value. Formally,

$$\Delta p = 0 \quad for \quad l < x < u$$

$$\Delta p = Min[Cap,(l - x)^2] \quad for \quad x < l$$

$$\Delta p = Min[Cap,(x - u)^2] \quad for \quad u < x$$

where x is the calcualated or measured value of the property, l and u are the lower and upper bounds, and *Cap* is the maximum penalty. When a number of range-based penalties are imposed, a relative weight may be set on each one, allowing user control over their relative importance.

It should be noted that this function provides a "soft" limit on the application of the rules. What this means in practical terms is that molecules that have somewhat unfavorable characteristics in some properties but favorable characteristics in others can still be selected. Consider the case of a diversity selection constrained by Lipinski's rules. A molecule with molecular weight 525 that is at a considerable distance from any other molecules in the diversity space (or, equivalently, occupies an otherwise unoccupied cell) may still be desirable to include in the selected set and by making a significant contribution to the diversity score may be selected despite its penalty score. Such an inclusion would not be possible with a hard cut-off, i.e., an elimination of all molecules above 500. Another molecule with molecular weight 525 possessing a high degree of similarity to another with molecular weight 325 is less desirable because it can be represented by its neighbor, thus avoiding the violation of the Lipinski rule for molecular weight while not materially affecting the diversity of the set.

For a library subset, the total penalty is calculated by summing the contribution from each molecule. This total then is normalized such that the total penalty will be one if each molecule violates each penalty range by one standard deviation. Formally,

$$P = \frac{1}{N_{prop}} \frac{1}{N_{mol}} \sum_{i=1}^{N_{prop}} \frac{w_i}{\sigma_i^2} \sum_{j=1}^{N_{mol}} \Delta p_{ij}^2$$

where N_{prop} is the number of properties being restrained, N_{mol} is the number of molecules being selected, w_i is the weight assigned for property i, σ_i is by default the standard deviation for property i evaluated over the full virtual library, and Δp_{ij} is the penalty (as defined earlier) for molecule j and property i. Although σ can, of course, be folded into the weight w_i, its use in the formula is intended to allow the weights to be determined more easily and transferred between different virtual libraries. The capping value is also set in terms of number of standard deviations from the upper and lower bounds. The software also allows the user to specify a value for σ that may be based, for example, on another distribution (e.g., an in-house compound collection or chemical database such as the World Drug Index [WDI]).

Distribution-based rules can be inferred from an analysis (similar to Lipinski's[21]) of a collection of molecules known to have desirable property distributions. These might be known drugs in the appropriate class, a list of which could be extracted from drug databases such as WDI, or a set of known actives from screening. This type of analysis was previously proposed by Gillett et al.[23] In the Cerius[2] implementation, a set of properties is first identified as important, and these are computed for the set of known drugs. A frequency histogram is obtained by binning each property across this set of molecules. These histograms, one for each property, provide the rules that the design library should follow.

To score a subset of the virtual library during the optimization process, each histogram from the subset is compared to the equivalent histogram from the drug database. An upper and lower bound is assigned for each bin in terms of its relative frequency (percentage of compounds occupying the bin). A penalty is calculated based on the difference between either the upper and lower bound and the relative frequency of compounds for that bin in the library subset. The amount of the penalty increases with the square of this difference. User-controlled weights can be set on the relative importance of each bin in each histogram. The penalty function is shown graphically in Figure 3.

Again, the penalty score must be normalized so that it can be appropriately weighted against all other factors in the optimi-

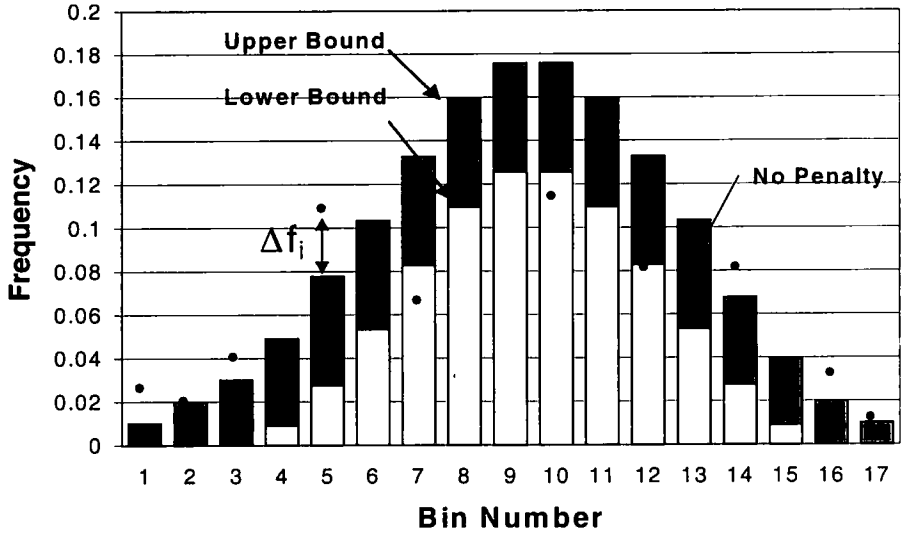

Figure 3. Penalty function for penalty profile histograms. Solid bars represent the required profile. Points represents values for the subset to be scored.

zation. In this case, we have chosen the score to be approximately 1 if each histogram bin of each profile violates the bin range by 10%. Formally,

$$P = \frac{1}{N_{prop}} \sum_{i=1}^{N_{prop}} \frac{1}{N_{bins}(i)} \sum_{j=1}^{N_{bins}(i)} w_{ij} \left(\frac{\Delta f_{ij}}{0.1} \right)^2$$

Although this type of analysis typically is applied to calculated physical properties, it can be applied equally well to any computed or experimentally measured property of each member of the library. For example, a series of virtual high-throughput screening models, derived from techniques such as recursive partitioning or hypothesis models, could be used to predict activities for each molecule in a virtual library. A library then could be designed to be simultaneously focused toward molecules predicted to be active, but at the same time still satisfy the combinatorial constraint and drug-like or cost requirements.

Reagent-Based Properties

At the reagent level, information about the source and cost of reagents is required to allow selections to be biased toward more desirable reagents. Specifically, the unit cost for each reagent from its preferred supplier (including in-house if applicable) should be specified along with an optional, user-defined penalty that can be used to encode criteria such as ease of synthesis, toxicity, or chemist preference. In addition, a relative penalty can be assigned to each supplier, allowing the user to indicate preferences in the choice of supplier. Such preferences might be used to encompass reliability, speed of delivery, or geographical location, for example.

In evaluating a subset library during the optimization, it first is necessary to compute the total quantity of each reagent that will be required to make the library, based on the amount of each product required. Various terms then can be scored, namely,

- Total monetary cost of the reagents required
- Total number of reagents used
- Total of the user defined reagent penalties
- Total number of suppliers used
- Total of the user defined supplier penalties.

Any combination of these can form the basis for a total penalty based on the reagents used in that subset, and relative weights can be assigned to each term.

Again, a normalization factor is applied to the total score to allow it to be assigned a relative weighted against the product properties:

$$P_{Reagents} = 1/N_{Pen} \, (w_1 P_{cost} + w_2 P_{number\_reagents} + w_3 P_{number\_suppliers} + w_4 P_{reagent\_penalties} + w_5 P_{supplier\_penalties})$$

where w_1 to w_5 are weight factors, and N_{Pen} is the number of penalty terms used.

EXAMPLES

A number of examples are given to illustrate the following design scenarios:

- A combinatorially constrained library designed for diversity and drug likeness defined by (a) Lipinski's rules and (b) the distribution of some physical properties in a set of known drugs
- A combinatorially constrained library designed for diversity and ease of deconvolution by affinity selection mass spectroscopy
- A combinatorially constrained library designed for diversity, drug likeness, and cost of reagents
- A combinatorially constrained library designed to maximize the number of molecules predicted to be active and minimize the cost of reagents.

Example 1: Constraining a Library by "Lipinski-Like" Rules

A virtual library of 34,596 dipeptides was prepared from the combination of 186 amino acids from the Available Chemicals Directory[24] with themselves. From this, the software was used to select a subset of 400 compounds that was both

Table 1. Diversity and penalty scores for a subset of the dipeptide library designed for diversity and/or drug likeness according to Lipinski's rules

Optimization	Diversity score	Penalty score
Diversity only	0.65	0.30[a]
Diversity-penalty (×1)	0.57	0.058
Diversity-penalty (×10)	0.46	0.023
Penalty only	0.35	0

[a]Calculated with weight = 1.

diverse and drug like to be made in a 20 × 20 array. The diversity was determined by first calculating the Cerius[2] combinatorial chemistry default set of 50 descriptors described previously. A principal component analysis showed that five components were needed to explain 92% of the variance of the original descriptors. These principal components were scaled to zero mean and unit variance to reduce further the influence of highly correlated descriptors.[25] The cell-based fraction metric was used as the selection method. The Monte Carlo optimization was set to run for 1,000,000 steps with at least 100,000 idle steps (i.e., no improvement in the optimal result found) and a temperature factor of 100K. Drug likeness was judged in a manner similar to that described by Lipinski, namely, molecular weight should not exceed 500, logP (computed using the AlogP method[15]) should not exceed 5, the number of hydrogen bond donors should not exceed 5, and the number of hydrogen bond acceptors should not exceed 10. In our protocol, we did not couple these rules. To precisely duplicate Lipinski's method, one could derive a column in the Cerius[2] Study Table (spreadsheet) that summed the total number of rules violated by each compound and then select for compounds that violated no more than one rule.

Subsets were selected for diversity only, for diversity weighted equally against the drug-likeness score, for diversity weighted as 10 times less important than drug likeness, and for drug likeness only. Table 1 shows the diversity and penalty scores for both libraries.

Examination of the diversity scores in Table 1 shows a range of 0.65 for the subset designed only for diversity to 0.35 for the subset designed only to minimize the penalties without regard for diversity. Even in the case in which the set is optimized for diversity, 35% of theoretically possible diversity is lost, i.e., not all the possible cells of the diversity space are occupied by the subset. This is due the imposition of the combinatorial constraint, which prevents the algorithm from simply picking one compound per cell. Without this constraint (and without the penalty constraints), a diversity score of 1.0 is achieved.

The imposition of drug-like restraints on the products results in the loss of only an additional 8% of diversity (measured as fraction of occupied cells). Simultaneously, the property profile with respect to drug likeness has been greatly improved, the penalty score lowering from 0.54 to 0.06. Increasing the weight of the penalty 10 times has resulted in a further loss of diversity of 9% but an almost perfect drug-like score. Color Plate 1 shows the three subsets plotted in the first three principal components of the five-dimensional principal components analysis space. Ex-

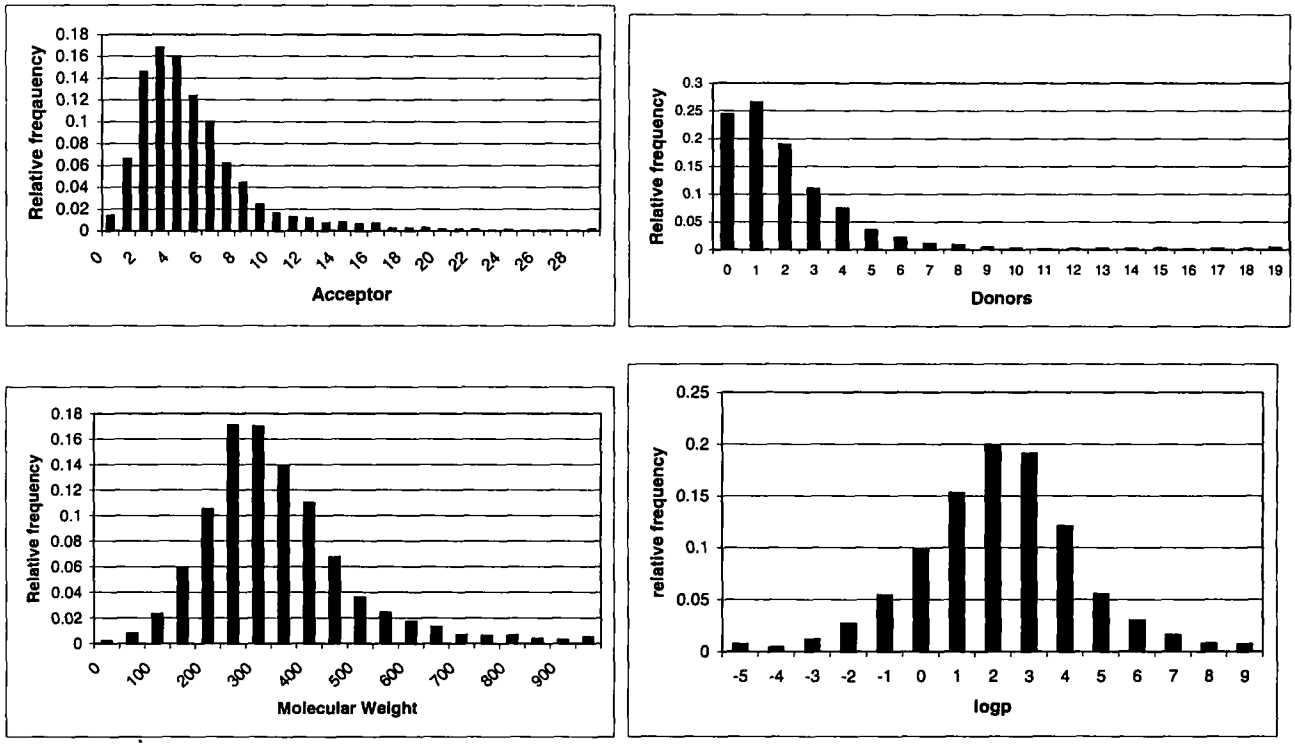

Figure 4. Frequency distributions of number of donors, number of acceptors, molecular weight, and logP derived from the WDI subset.

Table 2. Diversity and penalty scores for a subset of the dipeptide library designed for diversity and/or drug likeness defined by property profiles of known drugs

Optimization	Diversity score	Penalty score
Diversity only	0.65	0.97
Diversity-penalty	0.61	0.75
Penalty only	0.41	0.67

amination of the penalty only based selection (yellow) shows that only a portion of the entire property space is druglike, according to the simple rules in use. It also shows that the library selected in this region is highly redundant, having many close near neighbors for most of the subset. A diversity only subset (red) does not have this redundancy but extensively samples areas outside the drug space. The subsets chosen to satisfy both criteria are concentrated toward the drug-like area (particularly the set weighted toward this restraint) but do not exhibit the near redundancy problems.

Example 2: Restraining a Library Selection by the Property Profiles of Known Drugs

The experiments were repeated for the case where drug likeness is not predefined by rules, but is defined by profiling a set of known drugs and the design library is optimized to have similar property profiles. Drug likeness criteria for the selection were defined from an analysis of the WDI database[26] as follows. Of the 54,944 unique molecules in the March 1998 version of the database, 7,572 molecules were selected that had a United States Adopted Name (USAN) field, and 6,489 were selected that had an International Non-proprietary name (INN) field. The union of these two groups gave 8,504 molecules. A series of descriptors were calculated for this set, including structural descriptors such as molecular weight, logP (using the AlogP method), number of donors, and number of acceptors, and a frequency distribution histogram was produced for each. The distributions are shown in Figure 4. Table 2 shows the diversity and penalty scores for the subsets selected. Color Plate 2 shows the diversity only and penalty only subsets. Again, this shows that only a portion of the space is occupied by known drugs, and that there needs to be a balance between a

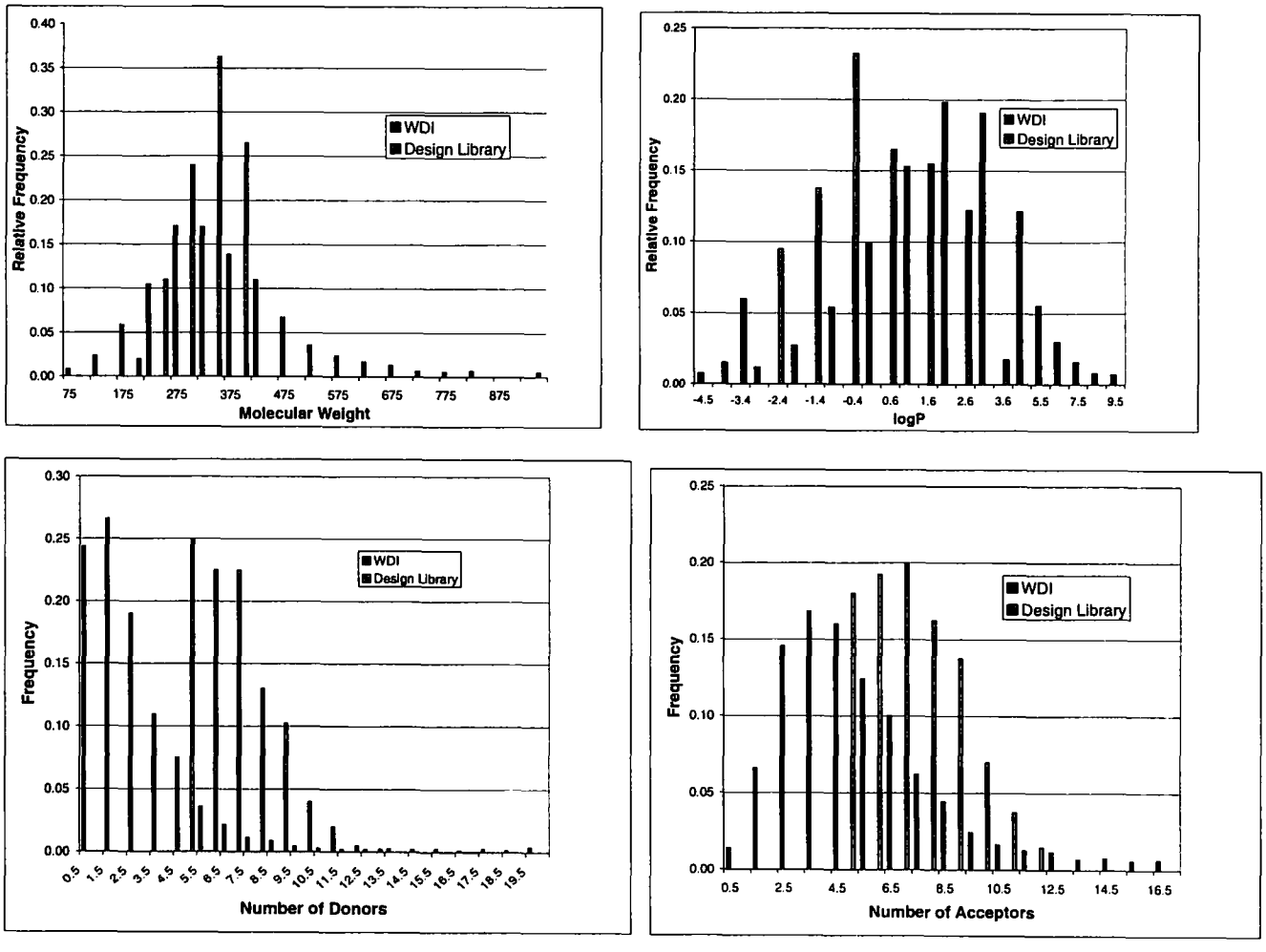

Figure 5. Relative frequency distributions of the design library compared to the target frequency distributions from WDI.

Table 3. Design of a subset of the dipeptide library optimized for diversity and efficiency of deconvolution

Optimization	Diversity score	Penalty score	Molecules over cut-off	Max redundancy
Diversity only	0.65	54.54	86	14
Diversity-Penalty	0.61	0.009	4	5
Penalty only	0.46	0	0	4

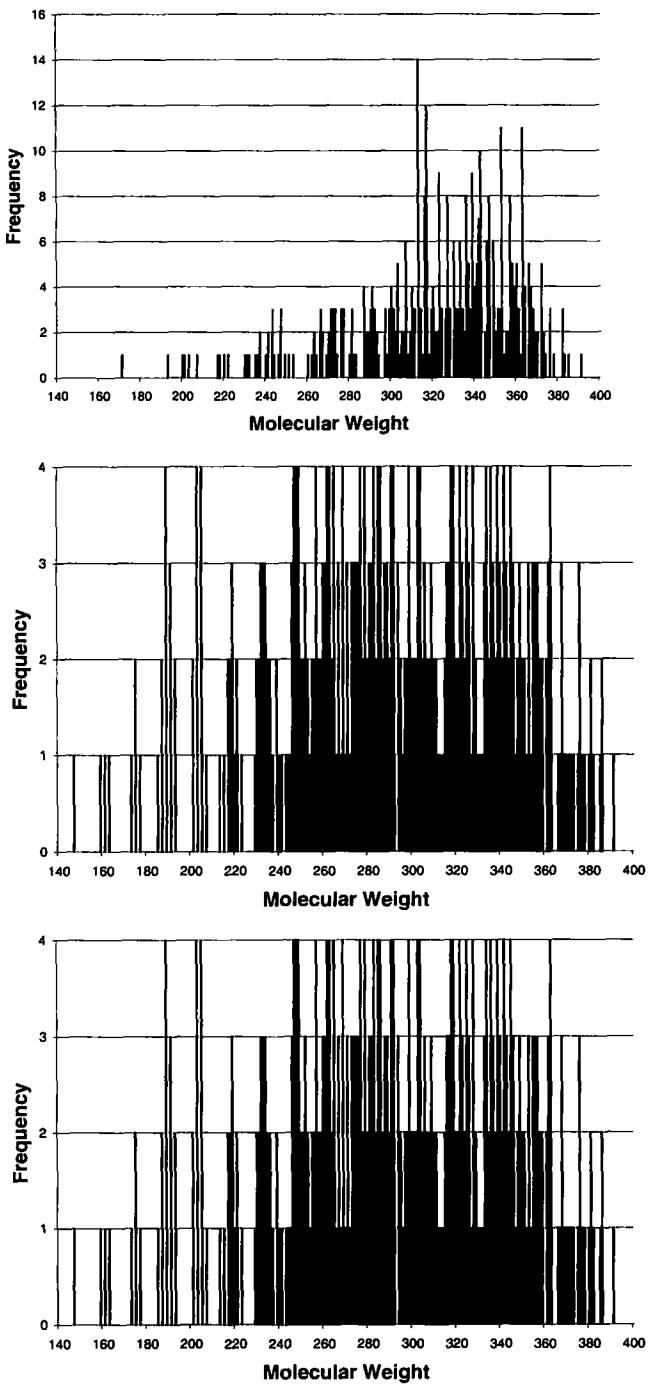

Figure 6. Molecular weight frequency distributions for design libraries optimized for (a) diversity only, (b) diversity-penalty, and (c) penalty only.

redundant sampling in this area alone and a diverse sample that is mainly outside this area. Table 2 shows that a design balancing both factors sacrifices only 4% of diversity over the most diverse set possible under the combinatorial constraint while dramatically improving the drug-like score. Figure 5 shows the relative frequency distributions of the design library and the target criteria.

Example 3: Design of a Library for Optimal Deconvolution Based on Affinity Selection Mass Spectroscopy

In this example, a subset of the dipeptide library is designed to be maximally diverse and to have an optimized profile for deconvolution using an affinity selection mass spectroscopy technique. As described in the Introduction, the requirement is to have only a few products sharing any one molecular weight. For this, a property profile histogram is prepared mimicking the required function. In the Abbott work,[5] an allowed frequency limit was defined for each experiment. In this case, a molecular weight frequency distribution is prepared with each bin one amu wide and with a relative frequency such that a frequency of more than the limit of four molecules in any one bin will incur a penalty. The nature of the penalty function is such that the penalty increases with the

Table 4. Property ranges for the molecules in the WDI subset.

Descriptor	Range in UGI library	90% cutoff in WDI_best
MW	399–669	550
Rotlbond	10–31	13
Hbond acceptor	2–13	9
Hbond donor	1–5	5
AlogP	−4.2–8.2	5
MolRef	102–191	120
Balaban Jx	1.3–6.6	2.8
PHI	7.3–23.2	8
Kappa-1A	22.1–40.4	25
Kappa-2A	10.1–26.3	12
Kappa-3A	5.3–19.5	8
CHI-V-0	16.3–29.3	20
CHI-V-1	9.31–17.7	12
CHI-V-2	6.13–15.4	10
CHI-V-3P	3.88–11.1	8
CHI-V-3C	0.42–3.81	2.2
Wiener	1737–9940	4000
Zagreb	118–238	175

Table 5. Design of a subset of the Ugi library for diversity, drug likeness, and cost of reagents

Optimization	Diversity	Penalty	Cost ($/mmol)	Cost ratio
Only diversity	0.597	1.248	94749	80
Diversity-penalty	0.438	0.062	5630	5
Only penalty	0.234	0.02	1184	1

square of the difference between the frequency and the limit for each bin.

The specification for the design was to select a 20 × 20 library with a maximum frequency of occurrence of any one molecular weight bin of four. The penalty was weighted at 1,000 times the diversity. Table 3 shows the diversity and penalty scores. Also shown are the total number of molecules, out of the 400, exceeding the molecular weight frequency limit and the number of molecules in the most frequently occurring molecular weight bin. The former gives an indication of the likelihood that a deconvolution problem will arise and the latter the worst case number of molecules that would have to be resynthesized. Figure 6 shows the molecular weight frequency distributions for each selection. For the library designed for diversity only, there are 86 molecules that are violating one of the molecular weight bin frequency bounds. In the worst case, 14 molecules would have to be resynthesized to identify the active molecule, which is 10 more than the acceptable limit. With a diversity loss of only 4%, only four molecules violate the bounds, and the worst case would be the resynthesis of five molecules, which is only one more than the prescribed limit.

Example 4: Diversity Analysis of an Ugi Library for Cost Effectiveness and Drug Likeness

A virtual library was prepared based on an Ugi reaction[27] using 10 acids, 10 aldehydes, 10 amines (one of which was bifunctional), and 10 isonitriles, giving a virtual library of 11,000 possible products. The design library was specified to be 4 × 4 × 4 × 4 (i.e., a combinatorially constrained selection) that would give either 256 products or 320 if the bifunctional amine was among the selected set. Each reagent was extracted from the Available Chemicals Directory, with a preferred supplier and unit cost.

The Cerius[2] default descriptor set described earlier was calculated, and a principal component analysis required five principal components to explain 93.8% of the total variance. Using the cell-based fraction measure (with mean/variance normalized principal components), the space was divided into a total of 576 cells, such that members of the virtual library occupied at least 256 cells. The Monte Carlo optimization was set to run for 1,000,000 steps at 100K, terminating after 100,000 idle steps.

In this case, drug-likeness criteria were defined by an analysis of a set of the WDI database[26] to derive a more extensive set of rules than those of Lipinki. This is designed to demonstrate how structural considerations as well as bulk molecular properties might be used to constrain a library to be like a set of desirable molecules and to show that any user-defined set of rules may be included in the design. The analysis proceeded as follows. The same 8,504 molecules were selected as previously described. A series of descriptors were calculated for this set, including structural descriptors (molecular weight, rotatable bonds, number of hydrogen bond donors, number of hydrogen bond acceptors, AlogP, and molar refractivity), and topological indices (Kappa indices [3], Phi index, Chi connectivity indices [5], Balaban Jx, Wiener, and Zagreb). An upper bound constraint for each property was set at the value exceeded by 10% of the molecules in the data set. These ranges were used as penalty ranges for the design of the Ugi library subsets. Table 4 shows the range of these properties in the Ugi virtual library and the bound established by the procedure described.

The design aimed to select a combinatorially constrained 4 × 4 × 4 × 4 library to be diverse (measured by the cell fraction), drug like (measured by having molecular properties within the 90% upper bound found in the WDI), and cost effective (established by the total cost of all the reagents required to make the library). In addition to a run to establish the best design library to satisfy these criteria, other runs were done using either diversity alone or the total penalty alone as objective functions to establish the optimal values for each criterion.

Table 5 shows the results of these runs. Examination of the diversity scores in Table 5 shows a range of 0.6 for the subset designed only for diversity to 0.23 for the subset designed only to minimize the penalties without regard for diversity. Again, even in the case in which the set is optimized for diversity, 40% of possible diversity is lost due the imposition of the combinatorial constraint.

The imposition of additional restraints on the products and reagents results in the loss of an additional 16% of diversity. At the same time, the property profile with respect to drug likeness has improved greatly, the penalty score lowering from 1.2 to 0.06. Most remarkably, the total cost of the library has been reduced by a factor of over 15-fold and with the loss of an additional 20% of diversity can be reduced another 4-fold.

Example 5: Optimizing a Library for Predicted Activity and Cost

In this example, predictions of activity were made for each molecule in the Ugi virtual library. These predictions were based on a recursive partitioning model for plasmepsin activity that was de-

Table 6. Subset of the Ugi library designed to optimized number of predicted actives and/or cost of reagents

Optimization	Cost ($/mmol)	Cost ratio	No. of actives	No. of inactives
Cost only	1,184	1	8	312
Cost and actives	7,322	6	120	136
Active only	41,757	35	177	79

rived from high-throughput screening data from Pharmacopeia.[28] There were 1,060 predicted to be active and 9,940 predicted to be inactive. Again, a 4 × 4 × 4 × 4 library was designed to have the maximum number of members predicted to be active and/or minimize the cost of reagents.

Table 6 shows the results of these runs. Although there are sufficient predicted actives in the library to make a library containing only predicted actives, this will not be a synthetically efficient set to make. Inclusion of the combinatorial constraint allows a fully combinatorial 4 × 4 × 4 × 4 library to be made that still includes 177 molecules predicted active, which is a 69% hit rate compared to a 9% hit rate in the virtual library as a whole. However, the cost of this library is very high. Including the cost penalty can reduce the cost of the library approximately 6-fold while only reducing the number of actives to 120, thus still achieving a 47% hit rate. Simply selecting the least expensive library would give a hit rate of only 2%.

SUMMARY AND CONCLUSION

A good library design needs to account for a complex set of requirements of both the products that will be made and the reagents that will be used to prepare them. The examples in this article demonstrate a broadly applicable methodology that can consider simultaneously the diversity or similarity of the subset, the selection for drug likeness, predicted activity, ease of deconvolution, and/or the economics and convenience of obtaining the reagents that will be required to make it. With incorporation of these criteria, we believe that, unlike past approaches based solely on considerations of diversity or similarity, computer-aided library design becomes a tool with much more relevance for the chemists and biologists employing high-throughput techniques in the drug discovery process.

REFERENCES

1 Bures, M., and Martin, Y. Computational methods in molecular diversity and combinatorial chemistry. *Curr. Opin. Chem. Biol.* 1998, **2,** 376–380

2 Blaney, J., and Martin, E. Computational approaches for combinatorial library design and molecular diversity analysis. *Curr. Opin. Chem. Biol.* 1997, **1,** 54–59

3 Willett, P. Computational tools for the analysis of molecular diversity. *Perspect. Drug Discovery Design* 1997, **7/8,** 1–11

4 Johnson, M., and Maggiora, G., *Concepts and Applications of Molecular Diversity.* Wiley, New York, 1990

5 Brown, R.D., and Martin, Y.C. Designing combinatorial library mixtures using a genetic algorithm. *J. Med. Chem.* 1997, **40,** 2304–2313

6 Gillett, V.J., Willett, P., and Bradshaw, J. The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries. *J. Chem. Inf. Comput. Sci.* 1997, **37,** 731–740

7 Jamois, E.A., Hassan, M., and Waldman, M. Evaluation of reagent-based and product-based strategies in the design of combinatorial library subsets. *J. Chem. Inf. Comput. Sci.* 2000, **40,** 63–70

8 *Cerius*^{2}, version 4.0, 1999, Molecular Simulations Inc., San Diego, CA

9 Clark, D.E., and Westhead, D.R. A review of evolutionary algorithms in computer-aided molecular design. *J. Comput.-Aided Mol. Design* 1996, **10,** 337–358

10 Brown, R.D., and Clark, D.E. Genetic diversity: Applications of evolutionary algorithms to combinatorial library design. *Expert Opin. Ther. Patents* 1998, **8,** 1447–1460

11 Hassan, M., Bielawski, J.P., Hempel, J.C., and Waldman, M. Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Diversity* 1996, **2,** 64–74

12 Willet, P. *Perspectives in Drug Discovery and Design* 1997, **7/8**

13 *Cerius*^{2}*•Diversity*, version 4.0, 1999, Molecular Simulations Inc., San Diego, CA

14 Brown, R., Kahn, S., and Zhang, L. *The design of lead optimization combinatorial libraries.* In: *CHI Conference on Molecular Diversity*, 1998, San Diego, CA

15 Ghose, A., Viswanadhan, V.N., and Wendoloski, J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. *J. Phys. Chem.* 1998, **102,** 3762–3772

16 Balaban, A.T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* 1982, **89,** 399–404

17 Hall, L., and Kier, L., The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling In: *Reviews in computational chemistry, Volume 2,* Lipkowitz, K.B., and Boyd, D.B, Eds., VCH Publishers, New York, 1991, pp. 367–422

18 Muller, W., Szymanski, K., Knop, J.V., and Trinajstić, N. An algorithm for construction of the molecular distance matrix. *J. Comput. Chem.* 1987, **8,** 170–173

19 Bonchev, D. *Information theoretic indices for characterization of chemical structures.* Research Studies Press, Letchworth, England, 1983

20 *Cerius2•LibProfile*, version 4.5, 2000, Molecular Simulations Inc., San Diego, CA

21 Lipinkski, C.A., Lombardo, F., Dominy, B.W., and Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* 1997, **23,** 3–25

22 Teague, S., Davis, A., Leeson, P., and Oprea, T. The design of leadlike combinatorial libraries. *Angew. Chem. Intl. Ed.* 1999, **38,** 3743–3748

23 Gillett, V.J., Willett, P., and Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* 1998, **38,** 165–179

24 *Available Chemicals Directory,* version 1999, MDL Information Systems Inc., San Leandro, California

25 Waldman, M., Li, H., and Hassan, M. Novel algorithms for optimizating molecular diversity of combinatorial libraries. *J. Mol. Graphics Modell.* 2000, **18,**000–000

26 *World Drug Index,* version 1998, Derwent Information, London, UK

27 Ugi, I., Lohberger, S., and Karl, R., The Passerini and Ugi reactions. In: *Comprehensive organic synthesis: Selectivity for synthetic efficiency, Volume 2* Herausg, C.H.H., and Trost, B.M., Eds., Pergamon, Oxford, 1991, pp. 1083–1109

28 Brown, R. Virtual high-throughput screening of virtual libraries. In: *Cambridge Healthtech Institute's Third Annual Novel Bioactive Compounds, October 20–22, 1999,* Brussels, Belguim