ELSEVIER

Novel algorithms for the optimization of molecular diversity of combinatorial libraries

Marvin Waldman, Hong Li,[1] and Moises Hassan

Molecular Simulations Inc., San Diego, California, USA

Various approaches to measuring and optimizing molecular diversity of combinatorial libraries are presented. The need for different diversity metrics for libraries consisting of discrete molecules ("cherry picking") vs libraries formed from combinatorial R-group enumeration (array-based selection) is discussed. Ideal requirements for diversity metrics applied to array-based selection are proposed, focusing, in particular, on the concept of incremental diversity, *i.e., the change in diversity as redundant or nonredundant molecules are added to a compound collection or combinatorial library. Several distance and cell-based diversity functions are presented and analyzed in terms of their ability to satisfy these requirements. These diversity functions are applied to designing diverse libraries for two test cases, and the performance of the diversity functions is assessed. Issues associated with redundant molecules in the virtual library are discussed and analyzed using one of the test examples. The results are compared to reagent-based diversity optimizations, and it is shown that a product-based diversity protocol can result in significant improvements over a reagent-based scheme based on the diversity obtained for the resulting libraries.* © *2000 by Elsevier Science Inc.*

Keywords: molecular diversity, combinatorial library algorithm

INTRODUCTION

The advent of combinatorial chemistry and high-throughput screening (HTS) technologies are revolutionizing the process of drug discovery in the pharmaceutical industry.[1,2] Whereas medicinal chemists previously sought to address the question of what is the next compound to synthesize for biological testing, today's chemists are faced with the broader question of what is the next library to make. In the early stages of a drug discovery project where the emphasis is on lead generation (rather than lead optimization), this question is often addressed by attempting to optimize the molecular diversity of the initial libraries produced.[3] This follows from the implicit assumption that maximizing the diversity of a library enhances the probability of finding active compounds of differing structural types in HTS experiments. These assumptions follow from applying principles of experimental design to the library design problem.[4,5] In the analysis presented herein, it is assumed that this is the primary goal of library design from molecular diversity optimization, namely, to maximize the likelihood of finding as many chemically distinct active compounds as possible in an HTS experiment. This does not speak to additional issues such as ensuring that the actives found are suitable leads from a medicinal chemistry standpoint or attempting to optimize other aspects of the problem such as minimizing the cost of producing the library or the number of reagents or compounds used. Although these types of issues are increasingly being recognized as an important aspect of the library design problem, they are outside the scope of the present article and are instead dealt with in an accompanying article in this issue by Brown et al.[6] Even with these additional factors to consider, it is generally acknowledged that optimizing diversity remains an important criterion for a well-designed library and, as such, is the primary focus of this work.

To date, there have been a number of approaches to address the problem of diversity optimization. Generally, they focus on the choice of molecular descriptors used and the way one attempts to cover the descriptor space. The choices of descriptors have ranged from 2D fingerprints, 3D pharmacophores, topological and graph theory indices, BCUT values, and various molecular properties (log P, volume, dipole moment, etc.).[3,7–9] Reducing the dimensionality and correlations among the large number of descriptors considered through the use of principal component analysis also has been investigated.[10] Assessing how well the descriptor space is covered usually is

Color Plates for this article are on pages 533–536.

Corresponding author: Marvin Waldman, Molecular Simulations Inc., 9685 Scranton Road, San Diego, CA 92121, USA. Tel.: 858-458-9990; fax: 858-458-0136. *E-mail address*: marvin@msi.com (M. Waldman)

[1]Present address: ChemInnovation Software, Inc., 8190-E Mira Mesa Boulevard, #108, San Diego, CA, 92126, USA.

Journal of Molecular Graphics and Modelling 18, 412–426, 2000

655 Avenue of the Americas, New York, NY 10010

1093-3263/00/$–see front matter
PII S1093-3263(00)00071-1

performed by clustering techniques, partitioning of space into cells, or various dissimilarity-based metrics.[7,8,10-14]

Most of the methods proposed and analyzed to date have focused on the problem of selecting discrete molecules from a possible combinatorial library reagent list or from a compound collection or database. More recently, attention has been given to the problem of attempting to design combinatorial libraries by optimizing the diversity of the products (product-based diversity) under the constraint that the optimal library satisfy the condition that it results from the full combinatorial enumeration of a set of selected reagents.[15] To solve this problem, it generally has been previously assumed that optimizing product-based diversity could be reasonably approximated by optimizing the diversity of the reagent R-groups. Gillett and coworkers[15] refer to this assumption as the "diversity hypothesis." If true, the diversity of each substituent R-group could be optimized by considering only the diversity of the possible reagents at each substitution point in a library scaffold without examining the diversity of the resulting products and without the need to enumerate a full virtual library and evaluate its descriptors. However, this assumption recently has been called into question.[15,16]

In this article, we conduct a further examination of the problem of optimizing product-based diversity subject to a library design that satisfies the combinatorial constraint (array-based design). In addressing this problem, we focus primarily on the issue of constructing a diversity function that properly analyzes how well a set of molecules covers a diversity descriptor space. We propose a set of requirements that should be satisfied by an "ideal" diversity function based on considerations of how diversity changes as redundant or nonredundant molecules are added to a system. We then discuss the behavior of some existing diversity functions with regard to these requirements and several new functions that are proposed herein. We then apply several of these diversity functions to two test systems where we can compare the results obtained to assess how well different approaches work and test the "diversity hypothesis" (optimizing diversity in reagent space only) on these systems. We show how the choice of a library scaffold can strongly influence the diversity of the resulting library as well as the performance of the "diversity hypothesis."

THEORY

In order to devise a procedure to optimize the diversity of a combinatorial library in product (rather than reagent) space, we need a protocol to assess the diversity of any possible library subset from the full virtual library being considered. We will assume that the choice of descriptors to use has already been decided, and we focus here on the problem of deciding how well the descriptor space is covered by a particular subset library. We refer to any protocol that quantitatively assesses the coverage of descriptor space as a *diversity function.* To guide us in constructing a diversity function, we first propose a set of theoretical requirements that a "perfect" diversity function should satisfy.

As already indicated, we suggest that optimizing diversity is equivalent to optimizing coverage of our descriptor space. As such, we equate diversity with coverage (or sampling), which suggests a set of "rules" that should be satisfied based on considerations of coverage. The rules mainly arise from considerations based on adding redundant or nonredundant molecules into our descriptor space. The proposed requirements for a diversity function are as follows:

1. Adding redundant molecules to a system does not change its diversity.
2. Adding nonredundant molecules to a system always increases its diversity.
3. Space-filling behavior of diversity space should be preferred.
4. Perfect (i.e., infinite) filling of a finite descriptor space should result in a finite value for the diversity function.
5. If the dissimilarity or distance of one molecule to all others is increased, the diversity of the system should increase. However, as this distance increases to infinity, the diversity should asymptotically approach a constant value.

We now proceed to discuss and analyze each of these requirements in turn.

Requirement 1 follows from the simple consideration that adding redundant molecules to a system does not increase our coverage of the descriptor space. By redundant, we refer to any molecule that has the same values for its descriptors as a molecule already in the system. In this context, redundant molecules do not necessarily have to be identical to each other but rather merely have the same values for their descriptors. For example, suppose we wished to measure the diversity of molecules in the context of a two-dimensional descriptor space consisting of logP and dipole moment. If a molecule in the system had values of (3.0, 2.0 Debye) for logP and dipole moment, respectively, then another molecule with the same values of logP and dipole moment would be considered redundant to it, whether or not the second molecule was in fact identical to the first. Thus, redundancy is determined in the context of the descriptor space. Now, while the coverage of descriptor space is not increased by the presence of redundant molecules, it also is not decreased. We are sampling exactly the

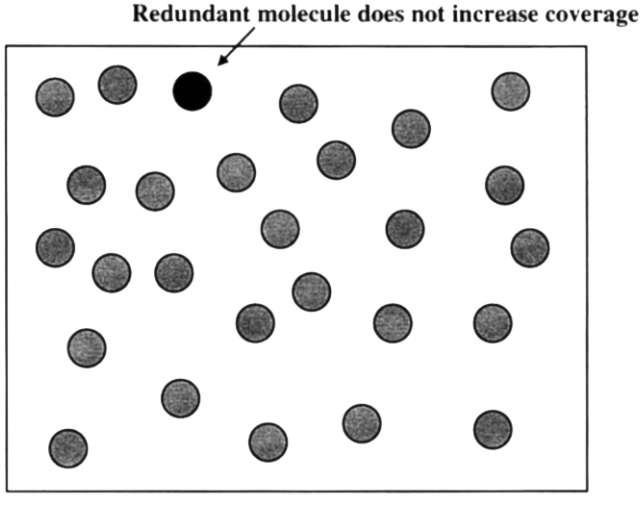


Figure 1. Illustration of diversity coverage with a redundant molecule in a two-dimensional descriptor space. The molecules (points) in light gray are nonredundant. The region in black is occupied by two molecules in this space, but does not result in any increase in diversity (coverage) relative to a single molecule occupying this region as shown by the circles (sampling regions) surrounding the points.

same regions of descriptor space with or without the redundant molecule present. This concept is illustrated in Figure 1, where an example of molecules spanning a two-dimensional descriptor space is shown. The solid black circle signifies the location of a pair of redundant molecules. The presence of redundant molecules does not increase the coverage of the space, because the same region would be covered by just a single nonredundant molecule occupying that location. As such, the diversity (i.e., coverage or sampling) of the system is unchanged. Note that we are not referring to the diversity per molecule. Rather, we are considering the diversity of the system as a whole, and this is not changed by the addition or removal of redundant molecules from the system. The diversity of the system also does not reflect the cost of making the library. Although the diversity is unchanged as redundant molecules are added, the cost of producing or purchasing such a library would be expected to increase. Criteria for combining cost with diversity in the context of library design is the subject of a companion article.[6] Nevertheless, although no explicit penalty is imposed on the diversity score for redundant compounds, library subsets with redundant compounds will tend to have lower diversity scores relative to subsets without redundancy when requirement 1 is satisfied by the diversity function. For example, an N compound library with a pair of redundant molecules will yield a score for the corresponding N−1 nonredundant compound library. Thus, library subsets optimized for diversity will tend to avoid the presence of redundant compounds when requirement 1 is satisfied, as will be seen in examples studied in the Applications and Discussion section.

Requirement 2 is essentially a corollary to requirement 1. Because adding nonredundant molecules covers a region of the space not already sampled, it follows that the diversity of the system as a whole should increase. The amount by which it should increase may be difficult to assess, but it would certainly appear reasonable that the diversity function increase whenever nonredundant molecules are added. This requirement is illustrated in Figure 2, wherein a nonredundant molecule is added to the library of Figure 1. It can be seen that the spatial coverage (diversity) of the system is increased by the presence of the new molecule. Because requirements 1 and 2 refer to the incremental change in diversity as molecules are added to the system, we have termed them *incremental diversity requirements.*

Requirement 3 seems reasonable, although a precise definition of the meaning of space-filling behavior is somewhat difficult to provide. In Figure 3, we offer an example that illustrates the intent. Two clusters of points (i.e., molecules) represented by the small circles are shown in a two-dimensional descriptor space such that the intracluster distances are much less than the intercluster distances. Thus, the points are tightly clustered, and the clusters are far apart from each other. Now, adding a new point approximately midway between the two clusters should increase the diversity of the system more than adding a point in the vicinity of either cluster. This is shown by the larger circle midway between the two clusters that does not overlap either of the clusters compared to the larger circle near the lower right cluster that overlaps with the circles of that cluster. The circle size represents a coverage region or sampling radius of the points (*vide infra*). However, because the coverage region of a point is not generally known, we minimize the overlap (in a probabilistic sense) and maximize the coverage (diversity) by adding the new point as far as possible from existing points. In other words, a diversity function should always prefer to fill in the largest holes or voids in diversity space. Note once again that this criterion is not necessarily aimed at finding compounds that are most suitable from a medicinal chemistry standpoint, but simply aims to ensure that all regions of descriptor space are adequately spanned such that if one or more active compounds are to be found somewhere within the diversity space, this strategy optimizes the *likelihood* of finding such a compound. An excellent discussion of the merits of sampling nonoverlapping regions of property space to satisfy space-filling behavior has been made by Patterson et al.[17]

Requirement 4 is somewhat more subtle, but it also derives from a consideration of diversity as coverage. Consider two libraries, A and B, whose diversity we are analyzing in a

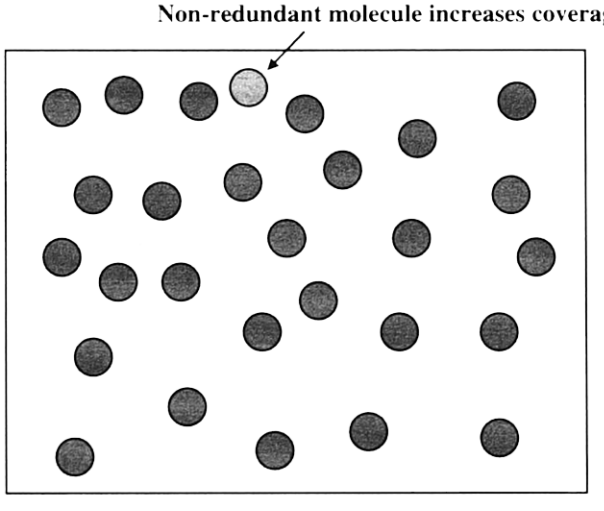


Figure 2. Illustration of how adding a non-redundant molecule to the library of Figure 1 increases the coverage or diversity of the system. The non-redundant molecule is indicated by the arrow.

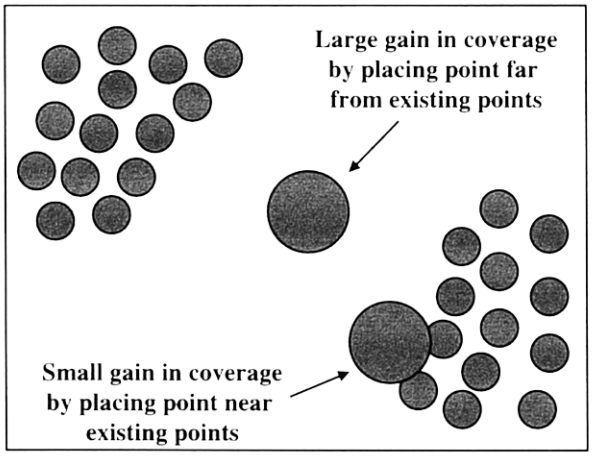


Figure 3. Illustration of space-filling behavior of diversity functions. Increase in diversity is greater by filling in holes far from existing points than by adding points near existing points.

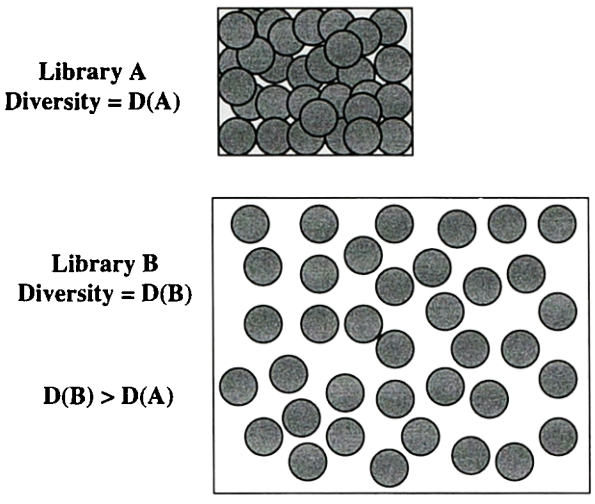


Figure 4. Illustration of finite coverage property of diversity. Filling in a small region with many points does not provide as much coverage as a well-distributed set of points occupying a larger region, no matter how many points occupy the smaller region. Thus, infinite filling of a finite region provides finite diversity or coverage.

two-dimensional descriptor space. The libraries are illustrated in Figure 4. All molecules of library A are restricted to a small area of the space corresponding to a square of side 1. Library B covers an area of the diversity space equal to a square of side 2. We will assume that library B covers the larger square with "reasonable" coverage (although this term is not yet well defined, it is discussed in more detail later). Assume that the diversity of library A is evaluated as D(A), and the diversity of library B is evaluated as D(B). Then, once library B covers its larger space "sufficiently" well, we should always find that D(B) > D(A), because A is confined to the smaller square. This places an upper bound on the diversity of library A. Thus, no matter how many molecules (even an infinite number) are present in library A, its diversity cannot grow infinitely, but should approach some finite value. Otherwise, it would always be possible to "flood" a poor coverage library (A) with many molecules and have it appear to be more diverse than a better coverage library (B) with fewer molecules. This requirement is important for cases in which we wish to compare the diversity of two systems with very different numbers of molecules. However, this kind of comparison generally is not needed when optimizing the diversity of a combinatorial library, because the possible library subsets being compared will have the same number of molecules. Consequently, this requirement is not necessary for diversity functions whose primary use is for performing combinatorial library optimization, which is the main focus of this study. We state the requirement here for completeness in considering the design of new diversity functions in the future.

In several of the previous examples, we have somewhat loosely made use of terms such as "reasonable" or "sufficient" coverage of space. We also have been careful to state that although library diversity should change under certain conditions (i.e., adding nonredundant molecules), we have not stated by how much it should change. This looseness in our definitions stems from the need to introduce an additional parameter into our library diversity function. This additional parameter is a "sampling radius" that provides some indication of how much of descriptor space is sampled by a particular molecule (this concept has been implied by the circles used to represent molecules in the figures). This concept of a sampling radius in descriptor space is intimately related to requirement 5. Recall that the ultimate goal of designing diverse libraries is to maximize the probability of finding one or more biologically active compounds. It is assumed that we have already chosen a set of descriptors that correlate with biological activity in the following sense. Similar molecules (based on the values of their descriptors) are assumed to be *probabilistically* more similar in their biological activities than would be observed by selecting molecules at random. More precisely, one should *tend* to observe an increase in the density of active compounds in the vicinity of an active compound in descriptor space compared to the density of active compounds in a randomly chosen region of descriptor space. Although examples in which small chemical changes can lead to large changes in biological activity are well known throughout medicinal chemistry, this desired behavior (similar molecules have similar activities) should occur on a statistical basis when sampled across a large number of molecules. This behavior often is referred to as the similarity property principle,[18] and various criteria for validating descriptors with regard to this behavior have been proposed by a number of authors.[17,19-21] Thus, molecules with similar diversity descriptors should *tend* to exhibit a correlation of their biological activities (i.e., molecules in the vicinity of an active compound will tend to be active with a probability greater than random), whereas molecules with very different values for their diversity descriptors have essentially uncorrelated biological activities (molecules far away from an active have a random probability of being active). The transition from correlated to uncorrelated biological activities occurs when molecules no longer sample a common region of descriptor space, i.e., they are outside each other's "sampling radius." A similar concept has been discussed by Matter,[19] who uses the term "similarity radius." Once this transition to uncorrelated behavior occurs, there is no further gain in diversity by making the molecules any more dissimilar, because once the molecules do not overlap in their coverage of descriptor space, there is no further gain in coverage by any additional increase in their separation from each other. These points are illustrated in Figure 5, where the gain in coverage (diversity) as two points having fixed sampling radii are separated from each other is shown. Once the circles no longer overlap, there is no further gain in coverage. However, it is actually more correct to consider the sampling radius parameter as having a "soft" rather than "hard" value because (ideally) the transition from correlated to uncorrelated biological activity behavior occurs gradually and continually (i.e., the density of actives in the vicinity of an active decays gradually back to random). Consequently, rather than state that the diversity remains constant once molecules are outside each other's sampling radius, we propose the requirement that the diversity should increase monotonically and asymptotically approach a constant as a molecule becomes very distant from all others in a diversity descriptor space. One desirable feature of this asymptotic behavior of diversity with distance is that it reduces the likelihood that a few

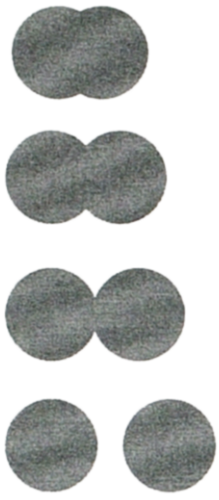


Figure 5. Illustration showing that diversity increases as points separate from one another, but once their sampling radii no longer overlap, there is no longer any increase in diversity (coverage) with increasing separation.

diversity space outliers in a library can dominate its diversity score.

ANALYSIS OF EXISTING DIVERSITY FUNCTIONS

In this section, we analyze several currently used diversity functions with regard to their ability to satisfy the requirements discussed earlier. The two main approaches taken to develop diversity functions can be divided into two general categories: distance (or dissimilarity)-based methods and cell (or partition)-based methods. Clustering techniques potentially can be viewed as a third approach, but from the point of view of this analysis, they exhibit essentially the same properties as cell-based methods. In both cases, each molecule is assigned to a partition (either cell or cluster), and the goal is to design a library with an optimal coverage of the partitions.

Distance-based methods have been found to be very useful when applied to the problem of selecting discrete molecules (as opposed to selecting libraries in the context of a combinatorial constraint) from compound collections or in designing parallel libraries. In particular, we found that the MaxMin approach described by Hassan et al.[10] is very effective in selecting sets of molecules that provide optimal coverage of a descriptor space. The MaxMin method can be described as follows. Given a set of descriptors associated with each molecule, calculate the distance between each pair of molecules in descriptor space as $D_{ij}^2 = \Sigma(x_{ik} - x_{jk})^2$, where D_{ij} is the distance between molecule i and molecule j, x_{ik} is the value of the k'th descriptor for molecule i, and the summation runs over all descriptors k. Now, for a given subset of molecules, find the minimum value of D_{ij} for all i,j pairs. Finally, attempt to find the subset (usually through a stochastic optimization procedure) that maximizes the (minimum) value of D_{ij}.

Although this method works well in selecting highly diverse molecules, it is not suitable for a constrained library optimization problem. In particular, it does not satisfy the incremental diversity requirements. For example, if we add a redundant molecule to the system, the MaxMin diversity value goes to zero (rather than staying constant). If we add a nonredundant molecule to the system, the MaxMin value either stays constant (the minimum distance between all possible pairs may be unchanged) or decreases (the new molecule may have a smaller distance to an existing molecule than the previous minimum value), rather than increasing as required. These deficiencies cause MaxMin to be unsuitable for dealing with combinatorial library optimization problems. The MaxMin function finds the closest pair of points in the system and assigns a diversity value based on the distance between the closest pair. This is fine when we are free to take one of the molecules involved in the closest pair and replace it with another, more diverse molecule. However, in combinatorial optimization, one does not replace a single molecule at a time, but rather attempts to find a subset of R-group choices that optimize the diversity of the enumerated sublibrary. In this case, one needs to compare the diversity of different combinatorial subsets. For example, it may be that one subset has a good overall coverage of descriptor space, but has a single pair of molecules quite close to one another that causes the MaxMin function to score the diversity as low. A second subset may have a poorer overall coverage of the space, but does not have any pair of molecules very close to one another. MaxMin would score the second subset as more diverse than the first, and this is clearly an undesirable result. It stems from the failure of MaxMin to satisfy (even approximately) the incremental diversity requirements.

Another distance-based method that has been applied to combinatorial library optimization is the sum of pairwise dissimilarities over all molecules proposed by Turner et al.[22] and applied by Gillett et al.[23] to the library design problem. The pairwise dissimilarity is evaluated as $1 - \cos(i,j)$ where $\cos(i,j)$ is the cosine similarity between molecules i and j using a 2D fingerprint key calculated using the Daylight toolkit. The main advantage of this function is that it can be evaluated in order N (number of molecules) time, which makes it attractive for dealing with large libraries. Unfortunately, this function also fails to satisfy the incremental diversity requirements as has been previously noted.[24] In this case, adding a redundant molecule will cause this diversity function to increase, because the redundant molecule will have nonzero dissimilarities to other molecules in the system (assuming that not all molecules of the system are redundant). Although adding nonredundant molecules also will cause this function to increase, cases can arise where adding redundant molecules causes it to increase more than nonredundant molecules. If these cases arise in practice, the resulting libraries may show a tendency to contain redundant or nearly redundant (i.e., highly similar) molecules. The consequences of such behavior will be seen in one of the examples studied in the Applications and Discussion section.

In clustering or cell-based methods, a library subset is selected, usually with the goal of maximizing the number of occupied cells or clusters. We will refer to this diversity metric as either the cell counts or cell fraction (because it is equivalent to maximizing the fraction of occupied cells) diversity function. This diversity function, which simply counts the number (or fraction) of occupied cells or clusters, tends to reasonably obey requirements 1–5. Because redundant molecules will occupy the same cell or cluster as an existing molecule in the system, the cell count remains unchanged when redundant molecules are added, thereby satisfying requirement 1. A nonredundant molecule may belong to a previously unoccupied

cell or cluster, in which case the cell count will increase. If it belongs to an already occupied cell, then the cell count is unchanged. Thus, we see that requirement 2 is partially satisfied (diversity increases sometimes but not always). At least, the diversity does not decrease when nonredundant molecules are added. Requirement 3, preferring space-filling behavior, is partly adhered to in the sense that the function increases whenever an unoccupied cell becomes occupied. Perfect filling of a finite space leaves the diversity function finite, assuming that the number of cells partitioning the finite space is finite, thereby satisfying requirement 4. As two molecules become very far apart, they will occupy separate cells or clusters, but once they do so, the diversity function will not increase further, partly satisfying (because the approach is not asymptotic) requirement 5. Furthermore, the cell size (actually the edge size) effectively serves as the sampling radius parameter. Thus, most of the requirements tend to be satisfied or partly satisfied. The drawback to this approach is that it is limited by the level of resolution provided by the cell divisions. Thus, a set of moderately similar molecules all falling within the same cell or cluster will appear to be less diverse than a set of molecules that are very similar but happen to lie across one or more cell boundaries thereby spanning several cells. Clustering methods also can experience this problem, because similar molecules sometimes can be assigned to different clusters. There is no guarantee that molecules within a cluster are all more similar to each other than to any molecules outside the cluster (see, for example, Figure 4 of Patterson et al.[17] and the accompanying discussion). Thus, sets containing very similar molecules can happen to occupy more cells or clusters than a set containing only moderately similar compounds that all happen to lie within the same cell or cluster. This drawback is related to the failure to satisfy requirement 2. In these cases, we see examples where adding nonredundant molecules to an already occupied cell or cluster can result in incorrect assessment of the diversity of the system.

Another diversity function that can be used with cells or clustering methods is the χ^2 statistic, employed in C2.LibSelect.[25] The definition of the diversity function in this case is:

$$D_{\chi^2} = -\Sigma(N_i - N_{sel}/N_{cells})^2 \qquad (1)$$

where N_i is the number of molecules in cell i, N_{sel} is the total number of molecules in the sublibrary, and N_{cells} is the total number of cell partitions. The sum runs over all cells. This function attempts to produce a uniform distribution of cell occupancies, rather than simply maximize the number of occupied cells. The minus sign in front of the sum is used so that more uniform distributions are scored as being more diverse. For example, suppose diversity space was divided into two cells (or clusters). Suppose there were two libraries each containing 10 molecules. Let the cell occupancies for the first library be (5,5). Let the occupancies for the second library be (7,3). If we simply count cells, both libraries are scored as equally diverse because there are two occupied cells in each case. However, the χ^2 diversity function considers the (5,5) library to be more diverse, which is a desirable result. Nevertheless, the χ^2 function does not satisfy incremental diversity (requirements 1 and 2), and this may result in some problems in its behavior.

Our investigations of the χ^2 function (see following) have led us to conclude that it may be too strongly biased toward producing uniform distributions while not placing enough emphasis on the goal of simply maximizing cell occupancies. For example, consider a three-cell system containing six molecules. The ideal distribution is (2,2,2). The χ^2 function ranks the two distributions (3,3,0) and (4,1,1) as equivalent, whereas cell counting would prefer (4,1,1) because all three cells are occupied. Another function that tends to show behavior intermediate between the χ^2 function and the cell occupancy count is one that measures the entropy (often termed information content in the computer science and information theory literature) of the system[26,27]:

$$D_{entropy} = -\Sigma(N_i/N_{sel})\ln(N_i/N_{sel}). \qquad (2)$$

This also favors more uniform distributions, but it would rank the (4,1,1) distribution ahead of (3,3,0) in the example mentioned earlier. This function also does not strictly obey requirements 1 and 2. However, it can be shown that requirement 2 is partially satisfied in that adding a molecule to an empty cell will increase the cell entropy diversity value. This behavior is not generally satisfied by the χ^2 function, suggesting an additional theoretical reason for preferring the cell entropy function. Performance of the various cell-based functions will be examined in the Applications and Discussion section.

NEW DISTANCE-BASED DIVERSITY FUNCTION

Given the problems with the existing distance-based diversity functions discussed earlier as well as the limitations of the cell- or cluster-based methods, we sought to construct a distance-based function that satisfies or nearly satisfies most of the proposed diversity requirements and which will prove suitable for combinatorial diversity optimization. In attempting to construct a new function, we found that a diversity function that exactly satisfies all requirements 1–5 can be constructed for a one-dimensional system (i.e., where there is only a single descriptor used to characterize each molecule). In this case, each molecule can be represented as a point on a one-dimensional line where the position of the point corresponds to the value of its descriptor. A diversity function for this system satisfying requirements 1–5 is illustrated in Figure 6. Above each point (molecule) on the line is drawn a normalized Gaussian curve. The Gaussians of all the points are allowed to

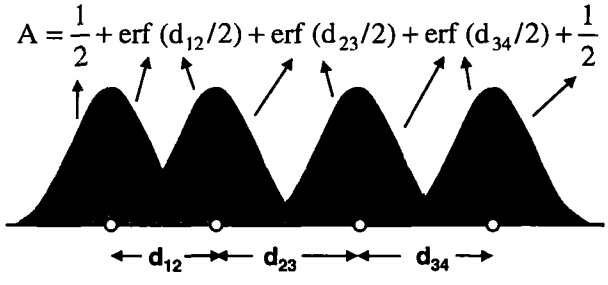


Figure 6. Representation of diversity function for a one-dimensional system based on the envelope of overlapping Gaussian functions.

overlap, and the area under the *envelope* of the Gaussians is taken as the diversity of the system. In this case, the width of the Gaussians (taken to be equal for all the points) corresponds to the sampling radius parameter previously discussed. We can calculate the area under the curve exactly by partitioning the system into a set of Gaussian segments. This is shown by the vertical lines going through each point and through each midpoint of neighboring pairs. The areas under the far right and far left Gaussians are each equal to 1/2, as indicated on the figure. The area under each of the remaining Gaussian segments is equal to 1/2 the error function, erf(),[28] evaluated over the length of the segment. As shown in Figure 6, these segments occur in pairs of equal area, so the net result is a sum of error functions plus one.

The area under the Gaussian envelope satisfies all the diversity requirements 1–5. Adding redundant points makes no change to the area under the Gaussian *envelope*, because two coincident Gaussian will overlap into the envelope of a single Gaussian. Conversely, adding nonredundant points always increases the area of the Gaussian envelope. The area under the envelope is maximally increased by adding points where the largest gaps between points exist, satisfying requirement 3. The envelope of an infinite number of Gaussians over a finite line segment is a rectangle over the line segment plus two half Gaussians extending at the right and left of the segment. The area under this region is equal to one (for the two Gaussian halves at the far left and far right) plus the height of the rectangle times the length of the filled line segment. This value remains finite, thereby satisfying requirement 4. Finally, as a point is moved to either the extreme left or right of the line, the area monotonically increases but asymptotically levels off, as the Gaussian of this point no longer overlaps with any of the other points. As already mentioned, the width of the Gaussians acts as a sampling radius. Thus, this representation of diversity satisfies all the stated requirements.

We can represent the area in the general case of an arbitrary number of points on a one-dimensional line with the following equation:

$$A = 1 + \Sigma \, \mathrm{erf}(\alpha d_{i,i+1}/2). \qquad (3)$$

Here, the sum is over pairs of adjacent points, (i,i+1). This function obeys requirements 1–5. For example, in the case of adding a redundant point, the distance d involving the redundant point pair is zero, and erf(0) = 0, so there is no change in diversity. This is the mathematical equivalent of the statement that the envelope of two Gaussians centered at the same point is a single Gaussian. We also have introduced explicitly the width (or radius) parameter of the Gaussian into the argument of the error function via the parameter α. The use of the sampling radius parameter allows us to tune the behavior of the diversity function. Small values of the α parameter correspond to very wide Gaussians and lead to a near-linear dependence of diversity on distance. Large values of α correspond to very narrow Gaussians and lead to a dependence of diversity on distance that rises rapidly and then quickly levels off to an asymptotic value. These properties are illustrated in Figures 7 and 8.

Thus, we see that Equation 3 provides a mathematical representation of diversity that perfectly satisfies all of the recommended requirements 1–5. However, it is only valid for a one-dimensional system. The problem arises as to how to

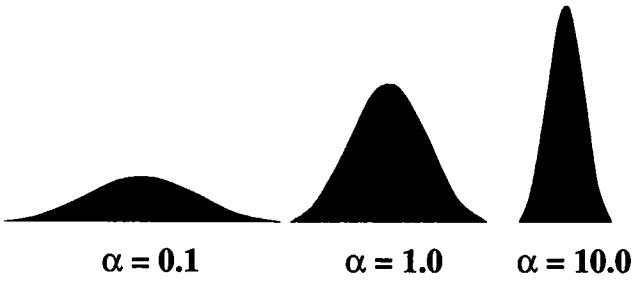


Figure 7. Effect of the α parameter on the shape of the Gaussian function. Small values of α lead to a broad Gaussian having a large sampling radius, whereas large values of α lead to narrow Gaussians with a small sampling radius.

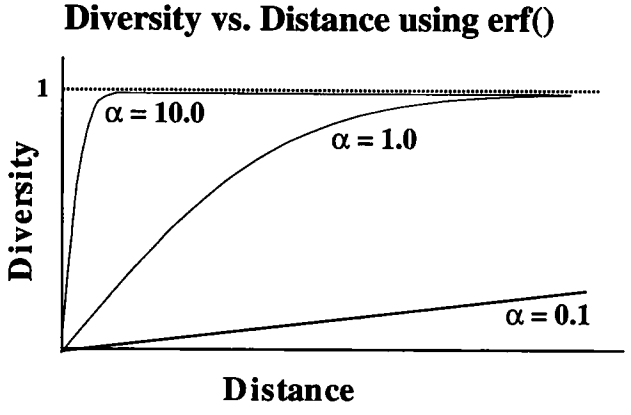


Figure 8. Representation of diversity as a function of distance using the error function, erf(), with different values of the α parameter. Small values of α exhibit a nearly linear dependence of diversity on distance, whereas large values produce a very steep initial rise with distance followed by a rapid leveling off.

generalize the result to a multidimensional system (i.e., representing molecules in diversity space using multiple properties or descriptors), which is the situation encountered in practical applications. In principle, this could be done by attempting to calculate the hyperdimensional volume of the envelope of multidimensional Gaussians centered at each of the molecules in the space. This volume can be represented as a multidimensional integral, and it does, in fact, satisfy all five diversity function requirements. However, unlike the one-dimensional case, this multidimensional Gaussian envelope volume does not have an analytical representation. One possibility would be to evaluate the integral by approximate numerical quadrature, but such an approach may lead to problems in high-dimensional spaces where numerical quadrature is difficult to perform accurately. Instead, we have sought a different approach that results in a closed form (but approximate) expression for the diversity of a multidimensional system. This alternate approach stems from the idea of treating the multidimensional case as a pseudo one-dimensional system and then to make use of the one-dimensional diversity result of Equation 3. One possible way to do this would be to connect all the points via a path and treat the line segments of the path as

the quantities to sum over in Equation 3. Because we wish redundant points to make no contribution to the diversity, the path should directly connect points that are very close (or redundant) to each other. This implies that we should seek the shortest path connecting all the points (usually referred to as the Traveling Salesperson Problem)[29] and use the segments of this path in the diversity function of Equation 3. However, problems arise with this approach because computation of the shortest path through the points is a difficult problem known to be NP-complete.[29] However, an alternative treatment turns out to be available that still allows use of the one-dimensional formula, Equation 3, while being far more computationally tractable. Instead of requiring that the points be connected via a *path* (which implies a set of edges with no branching), we can relax this restriction and consider a set of connections in which the edges are allowed to branch. A set of edges that connect a set of points and are allowed to branch (and do not form any cycles) is known as a spanning tree.[29] Instead of trying to find the shortest path through the points, we seek instead to connect the points with a minimum spanning tree, which is the spanning tree that has the smallest value for the sum of its edge lengths.[29] Thus, the minimum spanning tree is the analogue of the shortest path through the points, but with the additional flexibility of allowing for branching. An excellent discussion of minimum spanning trees was provided recently by Mount et al.[24] An illustration of the minimum spanning tree for a two-dimensional set of points is shown in Figure 9.

Using the minimum spanning tree, we can formulate a diversity function for the multidimensional case as follows. First, connect the molecule points with a minimum spanning tree. Assign a value for the sampling radius parameter α (a protocol for this will be discussed in the Methodology section). For each edge of the minimum spanning tree, calculate its error function, erf(αd), where α is the sampling radius and d the edge length, and sum over all the edges. The diversity function can be represented by the following equation:

$$D = \Sigma \, \mathrm{erf}(\alpha d_i). \qquad (4)$$

Here, the sum is over all the edges of the minimum spanning tree, and d_i is the length of each edge. Here, we have absorbed the factor of 2 from Equation 3 into the radius parameter α, and we have omitted the additive constant, 1, in front of the sum, because it does not affect the relative diversity between systems. A similar function for diversity using the minimum spanning tree was proposed previously by Mount and coworkers.[24] The main difference between their function and ours is the use of the error function in Equation 4 (which stems from considerations of overlapping Gaussians in one dimension). Their function involves summing up the edge lengths of the spanning tree. As such, it relates diversity linearly with distance and only satisfies requirement 1.

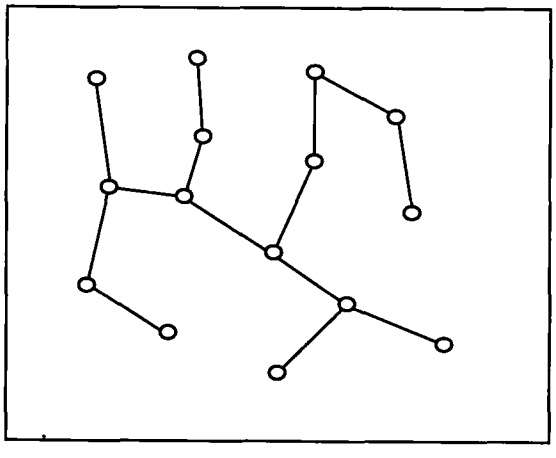


Figure 9. Illustration of a minimum spanning tree for a two-dimensional system.

This diversity function of Equation 4 also satisfies requirement 1, because adding redundant points adds edges of zero length to the minimum spanning tree, which consequently adds zero to the diversity value [because erf(0) = 0]. This function tends to satisfy requirement 2, but it is possible to devise cases where adding points can lower the score. However, requirement 2 is strictly satisfied for the following cases. Adding a nonredundant point in the vicinity of an existing point will always increase the score. Adding a nonredundant point along one of the edges of the spanning tree (or in the vicinity of such an edge) also will always increase the score. The violation of requirement 2 occurs when the distances between points are very small (or α is very small), so that the error function reduces to a linear function of distance, and we return to the diversity function of Mount et al.[24] In this case, one can find examples where adding a point into the center of several points can cause the diversity value to decrease. Fortunately, examples of this unusual behavior tend not to arise in practice, as will be seen in the Applications and Discussion section. Requirement 3 also is partly satisfied in essentially the same sense as requirement 2. It can be shown that requirement 4 is not generally satisfied in the multidimensional case, but this is not important for the applications presented here, because the combinatorial library optimizations all involve subset libraries of a fixed size. Finally, requirement 5 is fully satisfied, because erf(x) approaches 1 as x approaches ∞ (Figure 8). In the next section, we describe how we apply this function (as well as the cell-based methods) in practice, and then proceed to several applications using real combinatorial libraries.

PROTOCOLS

Diversity Optimization

To apply these various diversity functions to the design of combinatorial libraries, a protocol is needed for optimizing the diversity function to find an optimally diverse combinatorial library subset. To this end, we have implemented a procedure similar to the Monte Carlo optimization strategy previously described, which was applied to the problem of selecting a subset of molecules with the similar goal of finding optimally diverse subsets (without the constraint applied here of choosing a combinatorial subset).[10] In that work, a random subset was initially chosen, a single molecule was randomly replaced with one of the unselected molecules, and the diversity of the system was re-evaluated. If the diversity increased, the substitution was accepted. If it decreased, the substitution was accepted with a Metropolis criterion[30] controlled by a user selectable "temperature."

In this study, we implemented a similar Monte Carlo strategy with the aim of optimizing the choice of R-groups so as to perform true R-group array-based optimization of the diversity

of the resultant product library subset. Our goal is to choose an optimally diverse subset corresponding to a combinatorial library with a prescribed number of n_1 reagent choices for the R1 substituent, n_2 for R2, and so on. We begin the optimization by randomly selecting n_1 R1 reagents, n_2 R2 reagents, and so on from a larger virtual pool. The diversity of this subset is evaluated based on the resulting enumerated subset library. We then take a Monte Carlo step, but this time instead of randomly "mutating" one molecule, we "mutate" one R-group by randomly replacing one of the currently selected R-group choices in our subset with an unselected R-group reagent. We then evaluate the diversity of the corresponding enumerated library subset using one of the diversity functions and accept or reject the trial R-group mutation with a Monte Carlo Metropolis criterion as before. We iterate this process for many Monte Carlo steps until convergence (i.e., no further improvement in the value of the diversity function) has been achieved. This procedure has been described previously by Jamois et al.[16] A similar procedure using a genetic algorithm has also been previously described.[23]

Descriptors

The next protocol we discuss is the choice of descriptors. Here, we have followed mainly from our previous work and others in which a variety of 2D descriptors: topological, information content, AlogP, molecular weight, etc., were used in combination with principal component analysis.[3,10] The default 2D descriptors available in the C2.Diversity[31] product were used for the first test system. For the second test system examined (*vide infra*), we also used the electrotopological E-state descriptors developed by Kier, Hall, and colleague.[32-34] In each case, principal component analysis was applied (to the full virtual library) such that the number of components extracted corresponds to at least 90% of the variance of the data.[35]

Cell Divisions and Sampling Radius

We also need protocols for choosing how to partition each property for the cell-based methods and for the choice of the sampling radius parameter α for the minimum spanning tree method. Ideally, one would like to use information on how the descriptors correlate with biological activity data to guide these choices, as discussed earlier. However, this information usually is unavailable for lead generation libraries, and a detailed protocol employing activity data is outside the scope of this work in which our main goal is to examine the performance of the diversity functions as a means for sampling/covering a descriptor space under a combinatorial constraint.

After some investigation, we settled on the following default choice for the number of cells. The value is chosen such that the number of occupied cells in the full virtual library is equal to (or just slightly greater than) the number of molecules we are selecting for our subset library. This choice means that if a combinatorial constraint was not imposed on the solution, each selected molecule could occupy a different cell. We find in practice that the combinatorial constraint rarely permits this ideal solution, and thus the fraction of occupied cells in the sublibrary (relative to the ideal value of each molecule occupying a different cell) serves as a reasonable measure of how much diversity is lost by the imposition of the combinatorial constraint. To partition the descriptor space, each descriptor is uniformly partitioned (ranging from its minimum to maximum values) such that the cell edge length for each descriptor is as uniform as possible while requiring that the resulting number of occupied cells (in the virtual library) is equal to or just greater than the target value (number of molecules to select). This partitioning is achieved by an iterative algorithm that increases the number of bins by one for the descriptor having the currently largest cell edge until the resulting binning scheme results in the number of occupied cells in the virtual library being equal to or just exceeding the number of molecules to select for the sublibrary.

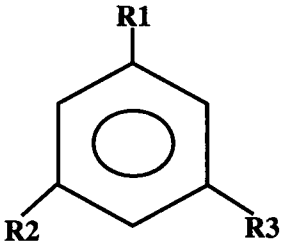
In choosing the α parameter for the minimum spanning tree method, we used a similar type of reasoning. Ideally, we would like each selected molecule in our subset to occupy a different region of descriptor space. In the case of the cell methods, this (ideally) corresponds to a different cell for each molecule. In the case of the spanning tree method, we need to choose a value of α that produces a reasonable result for the diversity calculated by Equation 4. Because the distances in our descriptor spaces are not bounded, a fixed choice for α may behave differently for different libraries and/or descriptor sets. We deal with this issue by "renormalizing" our descriptor space as follows. We determine the "hypervolume" of the descriptor space from the minimum and maximum values of each descriptor (for a given virtual library). A single scale factor is applied to all the descriptors such that the hypervolume of the space is set to the total number of molecules to be selected. In an ideal case, each molecule would now occupy its own region of descriptor space with a hypervolume of one and be separated from its nearest neighbors with a distance of about one. With this choice for rescaling the descriptors, we then simply set α to one. This protocol is equivalent to leaving the descriptors unscaled and setting α as:

$$\alpha = \left[N_{\mathrm{sel}} \Big/ \prod_{i=1}^{n} R_i \right]^{1/n} \tag{5}$$

where R_i is the range for each of the n descriptors. For some of the test cases that follow, we examine the influence of the number of cells or value of α on the results to help insure that these default protocols are reasonable.

APPLICATIONS AND DISCUSSION

Two test example libraries were chosen to study the performance of the various diversity functions discussed earlier. The first library was selected to represent an extreme example with regard to the presence of redundant molecules in a library. This library is shown in Figure 10. It is a 1,3,5-trisubstituted phenyl ring with the substituents being identical at each of the three positions. The full virtual library consists of a 15 × 15 × 15 array with substituents listed at the bottom of the figure. This library contains 3,375 molecules, but only 680 nonredundant molecules. A set of 44 1D and 2D descriptors were calculated for this library using the Cerius^2 Analog Builder, QSAR+, and Descriptor+ software modules.[36-38] The descriptors comprise molecular weight, molar refractivity, AlogP, and a default set of topological and information content indices. They have been discussed previously in the context of individual molecule-based diversity selection.[10] Principal component analysis was applied to these 44 descriptors for the 3,375 molecules, and the first five principal components were extracted to represent the



R1,R2,R3 = (acetyl,amino,bromo,chloro,fluoro, iodo,methyl,ethyl,methoxy,carboxy, hydroxy,cyano,nitro,benzyl,t-butyl)

Figure 10. Representation of symmetrically trisubstituted phenyl library used to test various diversity functions. The 5 × 5 × 5 subsets from the 15 × 15 × 15 virtual library were selected according to the various diversity protocols described in the text.

diversity space of the system. These five components contained 90% of the variance present in the original descriptors.

For this library, various experiments were conducted using different diversity functions and protocols with the goal of selecting an optimally diverse 5 × 5 × 5 combinatorial subset. To help provide motivation for why we resorted to some of the more elaborate approaches presented here, such as the spanning tree method, we first present the results of several approaches to this problem that proved unsuccessful in yielding a diverse nonredundant subset. The first approach attempted was to select a subset based on just the diversity of the R-groups rather than the product molecules. Of course, this library constitutes a system where an R-group–based selection cannot yield a reasonable result due to the equivalence of the R-groups caused by the topological symmetry present in the scaffold. In this case, the five most diverse R-groups were chosen by selecting five molecules out of the 15 molecules corresponding to the choice of R1 = any, R2 = H, and R3 = H. The five most diverse molecules from these 15 were chosen using the five principal component descriptors and the MaxMin method. The five diverse R-groups found by this procedure were amino, benzyl, iodo, t-butyl, and ethyl. Because symmetry dictates that the same procedure applied to the R2 and R3 positions would yield the same result, the resulting library using independent selection of each R-group position then yields R1 = R2 = R3. A plot of this library is shown in Color Plate 1. The three-dimensional space shown here corresponds to the first three principal components with the full library shown in white spheres and the selected subset in larger red spheres. The number of nonredundant molecules in this library is 35 (forced by the nature of the selection procedure) out of a possible maximum of 125. Thus, this library clearly is highly redundant. In addition, the plot shows certain regions of the space (notably the upper left region) are not well sampled. Clearly, the protocol of independent selection of R-groups is inadequate for this type of library.

The next experiment corresponds to an attempt to avoid the redundancy problem by a modification of the procedure just described. After choosing the substituents for the R1 group, the substituents for R2 were next chosen by the same protocol after first removing the choices for R1 from the list available for R2. Finally, the substituents for R3 were chosen by the five remaining substituents not used for R1 or R2. This leads to the library shown in Color Plate 2. We term this protocol a nonoverlapping MaxMin selection (i.e., the choices for R1, R2, and R3 are completely orthogonal). It is guaranteed to yield a perfectly nonredundant library of 125 molecules as shown. However, it leaves the right half of the diversity space unsampled, and we shall see that it is possible to improve considerably on this result.

The third experiment arose from a choice of a diversity function that ultimately was found to be inadequate. We mention it here, because it bears some similarity to the Willett metric[22,23] described earlier. In this case, the diversity function is calculated by the sum of pairwise overlaps of Gaussian functions centered at each of the molecule points in diversity space (the choice of Gaussian width α is as described earlier). The overlap equals one for redundant molecules and continuously goes to zero as the molecules become well separated from each other. It is similar to the cosine similarity distance used by Willett, although it does not have the convenient property of being reducible to a sum of order N. The library diversity was optimized by minimizing the pairwise sum of overlaps using a Monte Carlo R-group mutation procedure as described earlier. The optimal result found is shown in Color Plate 3. Although visual inspection shows this library seems to do a reasonable job of sampling the space of the system, it is a highly redundant library containing only 49 out of a possible 125 nonredundant molecules. We will see that it is possible to do considerably better. The highly redundant library results from the failure of the diversity function to even approximately satisfy the incremental diversity requirement. With this function, highly redundant libraries can score better than nonredundant libraries because the addition of redundant molecules to the system often can increase the diversity calculated with this function. A redundant molecule will have many more weak overlaps with the distant, nonredundant molecules of the library than the strong overlap(s) it has with its redundant neighbor(s), making the system appear more diverse. This results in a bias toward libraries that can be highly redundant and cover the diversity space in the form of "clumps." This problem is likely to arise for all diversity functions that attempt to calculate the diversity as a pairwise sum of distances, overlaps, or similarities and reflects a general flaw with this approach. Similar criticisms of this type of metric have been offered by others.[24,39]

Before proceeding to the results with better behaved diversity functions, we note one other aspect of the diversity space representation of this system that can be seen in Color Plates 1–3. The first principal component (shown as the x-axis or horizontal direction of the space) has a considerably larger extent than the next two principal components, as seen by the extended rectangular nature of the space. (PC1 also is larger in extent than PC4 and PC5, which are not shown). This extension along the first principal component direction arises because many of the original descriptors (especially the topological and information content indices) are highly correlated with molecular weight. This correlation causes the first PC itself to be highly correlated with molecular weight and to be considerably larger in extent than the other PCs. Another manifestation of this behavior is that the four points along the lower right side of the space in Color Plate 3 correspond to choices of either benzyl or t-butyl for the R1, R2, R3 substituents. These sub-

stituents have the largest molecular weight in this library, and the resulting product molecules all lie along the larger values of the first principal component. A consequence of this aspect of the PCs is that diversity functions that measure distance in PC space will tend to overemphasize the role of molecular weight in contributing to molecular similarity or diversity. A similar behavior occurs with the cell-based metrics, because we partition the space to produce approximately uniform cells along each PC. Thus, molecules that are only slightly different in molecular weight will appear to be somewhat diverse due to the emphasis of distance (or cell partitioning) along the direction of the first principal component. To avoid this undesirable emphasis on molecular weight, we implemented an additional protocol whereby the principal components are rescaled so that each principal component has a unit variance for the virtual library. We term this choice of descriptors "normalized principal components." Most of the remaining results are obtained using these normalized descriptors. Diversity spaces shown using the normalized PCs tend to be less elongated along the direction of the first PC and present a more cubic shape when visualized using the first three principal components. Repeating the protocols used to generate Color Plates 1–3 with normalized PCs do not change the general conclusions and are not shown here. The three protocols used to produce the diversity spaces of Color Plates 1–3 still produce either highly redundant libraries or ones that do not sample the space well.

The next experiment is shown in Color Plate 4. It is the library optimized using the minimum spanning tree function described earlier. In this case, normalized principal components were used to represent the diversity space, accounting for the more cubic shape of the space shown. We see that the spanning tree function has managed to find a library that covers the space reasonably well (as can be seen from the figure) while producing a nearly nonredundant library (124 out of a possible 125 nonredundant molecules). It does this by giving up on including the all t-butyl and all benzyl combinations, both of which lead to many redundant molecules present in the system. However, by including benzyl in the R1 and R2 positions and t-butyl in the R2 and R3 positions, it does contain the R1 = R2 = benzyl, R3 = t-butyl, and R1 = benzyl, R2 = R3 = t-butyl choices, which still provide some coverage of the high molecular weight region of the space.

The next library is shown in Color Plate 5. It corresponds to an optimization performed with the cell counts diversity function. The system was divided into 1,600 cells (based on the default protocol described earlier: 1,600 cells yield 133 occupied cells in the full virtual library, which just exceeds the target value of 125 occupied cells). The resulting library covers most of the space reasonably well but misses the high molecular weight compounds involving all t-butyl/benzyl substituents altogether (neither benzyl nor t-butyl is present in the R1 substituent). The library contains only 118 nonredundant molecules compared to the 124 from the spanning tree result. By most criteria, it is not as diverse as the spanning tree library, but by the simple measure of cell occupancy, it does occupy more cells than any other $5 \times 5 \times 5$ library we found for this system. The library occupies 82 out of a theoretically possible 125 cells in the space. The spanning tree library occupies 76 cells in the space. We believe the spanning tree library has better coverage of the space based on its higher number of nonredundant molecules, visual inspection of the diversity space figures, and its incorporation of some of the all benzyl/t-butyl product molecules. This result tends to confirm that simple counting of cells, although providing a reasonable measure of diversity, may not necessarily be the best or only way to measure coverage.

In Color Plate 6, we see the result obtained using the cell-based entropy function. Once again, we obtain 118 nonredundant molecules in the library, but this library also includes two of the all benzyl/t-butyl combinations we found in the spanning tree library. This library occupies 80 cells of the diversity space. Overall, it appears to be a slight improvement from the cell-counts–based library, helping to validate the cell entropy function as a reasonable diversity metric. The library obtained with the cell chi-squared function only yielded 111 nonredundant molecules and contained one of the all t-butyl/benzyl product molecules. We have not shown this library with a figure, because it does not provide any significant improvement or insight relative to the libraries already shown. Due to the slight increase in library redundancy, this library appears slightly inferior in coverage/diversity to the ones in Color Plates 4–6.

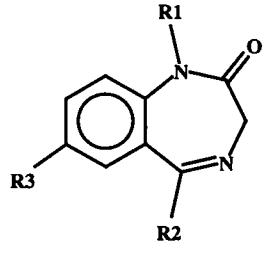
At this point, the question may be asked as to why the cell methods should ever be used, given that the spanning tree approach seems to yield better libraries in terms of their diversity or coverage as well as its ability to generate libraries with fewer redundant molecules. The main reason for using cell methods has to do with the simple practical matter of computational speed. The spanning tree computations require determination of the nearest neighbor for each member of the library subset, and although this can be made efficient through the use of k-d trees,[39,40] the spanning tree computation still is considerably slower than calculating any of the cell metrics. In our hands, we find the spanning tree optimizations for this system run from one to two orders of magnitude slower than the cell optimizations. Typically, the cell optimizations run from seconds to minutes (on an SGI R4K Indigo), whereas the spanning tree optimizations typically run in hours (several tens of thousands of Monte Carlo iterations are typically run to achieve convergence). In addition, the times increase as the problem size gets larger, with the ratio of the times for the spanning tree compared to the cell methods increasing with larger library subsets. This is because the spanning tree algorithm is of order N*log(N), where N is the size of the library subset while the cell methods typically increase as order N. Thus, the spanning tree method becomes impractical for dealing with very large libraries, and one of the goals of the studies conducted here is to see how close other approaches, such as the various cell methods, can come to the quality of the results obtained with the spanning tree.

In the next set of experiments, we seek to design a library that is completely nonredundant (125 nonredundant molecules) while still providing good coverage of the diversity space. The first of these libraries is shown in Color Plate 7. This library was designed by manually replacing the choice of iodo in the R1 position of the spanning tree library (Color Plate 14) with chloro. This results in a library that has no redundant molecules. In addition, it exhibits good diversity space coverage and continues to include two of the all benzyl/t-butyl product molecules. It has a spanning tree function value of 47.06 compared to the value for the optimized spanning tree library (Color Plate 4) of 49.28. It occupies 71 cells of the diversity space compared to 76 for the spanning tree library and 82 for the cell-counts–optimized library. Nevertheless, if a com-

pletely nonredundant library is a required criterion, this library constitutes a reasonable design satisfying this criterion and illustrates how the user can bring additional biases to bear and make adjustments to the answers provided by the computation.

In the next two experiments, we chose to drive the algorithm to produce nonredundant libraries through tuning of system parameters. The way to produce a bias toward nonredundant libraries using our diversity functions is to effectively reduce the sampling radius parameter so that any separation of the molecules causes them to appear diverse. In terms of cell-based selection methods, this is done by increasing the number of cell partitions. In Color Plate 8, we increased the cell partitioning to 97,200 and the resultant library was found to be completely nonredundant. It also contains one of the all benzyl/t-butyl product molecules in the high molecular weight region of diversity space. A more elegant way to influence the tendency toward nonredundant libraries is by adjusting the sampling radius parameter α of the spanning tree function. As discussed earlier, by increasing the value of this parameter, the Gaussians surrounding each point in diversity space are made more narrow, causing a rapid increase in diversity to occur as soon as the points separate from one another. In Color Plate 9, we show the result of a library obtained using an α value of 10 (compared to the default value of 1 used for the library of Color Plate 4). This library also contains no redundant molecules, occupies 69 cells (when the space is partitioned into 1,600 cells), has a spanning tree score of 47.59 (calculated using α = 1), and contains two of the all benzyl/t-butyl product molecules. It appears to be about as good as our manual choice constructed earlier. These examples show how the parameters of the algorithms can be tuned to influence the design of the library.

The second library was chosen to represent the case where no redundant molecules are present. Our goal in studying this second example was to ensure that the conclusions drawn from the first example did not result from the rather (intentional) pathological nature of the first example in terms of the high redundancy present in the system. As such, for the second example we chose a library with a benzodiazepine core together with a set of 20 R-group substituents chosen from the available default set in the Cerius[2] Analog Builder.[36] The library is illustrated in Figure 11. It has three R-group positions in the benzodiazepine core with (the same) 20 substituents



20 X 20 X 20 Library
8000 molecules

Select subsets from
5 X 5 X 5 through
10 X 10 X 10

Benzodiazepine Library (no redundant molecules)
R1,R2,R3 = (acetyl,amino,benzyl,bromo,carboxy,chloro, cyano,ethyl,fluoro,formyl,hydrogen,hydroxy, iodo,methoxy,methyl,phenoxy, phenyl,t-butyl,thiol,vinyl)

Figure 11. Benzodiazepine combinatorial library used for the second test example.

available at each of the three R-groups. As such, the full virtual library constitutes a 20 × 20 × 20 array totaling to 8,000 molecules. Library array subsetting experiments were conducted with subset sizes ranging from 5 × 5 × 5 to 10 × 10 × 10. In addition, three different descriptor sets were examined. The first set consisted of 43 1D and 2D descriptors described earlier for the 1,3,5-trisubstituted phenyl library (the Hosoya index was omitted from this descriptor set to save time in the descriptor calculations, as this index takes a very long time to calculate for larger molecules). These 43 descriptors were reduced to three descriptors using principal component analysis. The second set of descriptors consisted of the electrotopological E-state atom type indices. Of the 52 atom types proposed by Kier and Hall, the 35 atom types for the elements C,N,O,F,Cl,Br,I were used.[32–34] This yielded 35 descriptors for each molecule consisting of the summed E-state index value for each of these 35 atom types. For this set of E-state descriptors, 12 principal components were needed to capture 90% of the variance of the data set. Finally, a third set of descriptors was used in which the 43 1D, 2D topological descriptors were combined with the 35 electrotopological descriptors. This set then was reduced to 10 principal components. For each of these descriptor sets, we conducted subset selection experiments ranging from 5 × 5 × 5 arrays to 10 × 10 × 10 arrays using both normalized and unnormalized principal components and using each of the diversity metrics: spanning tree, cell counts, cell chi-squared, and cell entropy. This results in a total of 144 libraries that were obtained. Space does not allow for examining each one, so in the analysis that follows, we select a few of these libraries for detailed inspection and then present statistics analyzing some trends across the full set of libraries.

In Color Plate 10, we show the 5 × 5 × 5 library obtained using the spanning tree metric with normalized PCs from the topological descriptors. The red spheres show the selected molecules embedded in the space of the full virtual library. The most obvious omission from this library is the molecule at the top of the space, which corresponds to the choice of R1 = R2 = R3 = t-butyl. In this view, some of the selected molecules are obscured by other molecules of the virtual library. In Color Plate 11, we show just the selected molecules. It can be seen that the selected molecules do a reasonable job of covering the space, with the exception of the region near the top occupied by the single all t-butyl compound. In Color Plate 12, we show the equivalent selection using the cell counts function to optimize the diversity of the system. In this case, the all t-butyl molecule is selected (note that t-butyl is present in the selection set for each of the R-groups), so the question remains why this molecule was not chosen with the spanning tree method. It turns out that the cell counts selection gives a spanning tree score of 52.6, whereas the spanning tree score for the library optimized with the spanning tree function is 55.9, so the spanning tree optimized library is, in fact, more diverse based on the spanning tree function. Close inspection of the two libraries reveals that the spanning tree-based library actually has better coverage of the space in the lower regions of the diversity space, whereas the cell counts library has noticeable gaps in this region (note the layering or clustering effect present in the cell counts library in Color Plate 12 with the gaps between the clusters). This improved coverage of the lower part of the space turns out to be significant enough to offset the loss of diversity caused by the omission of the all t-butyl molecule for the spanning tree. Effects such as this are sensi-

tive to the choice of sampling radius, manifested as the α parameter for the spanning tree function and as the number of cells for the cell-based methods. Further investigations showed that inclusion of the all t-butyl molecule can be influenced by changing these parameters, just as we could influence the number of redundant molecules present for the symmetrical trisubstituted phenyl library discussed earlier.

In Color Plates 13–14, we show the corresponding libraries for the 10 × 10 × 10 array selection. Here, we see that both the spanning tree and cell counts libraries do an excellent choice of selecting 1,000 molecules out of 8,000 (subject to the combinatorial constraint) to cover the space of the system. Visual inspection of the plots shows that there appears to be more regions of unselected (gray) molecules visible in the cell counts plot (near the lower region of the space) than in the spanning tree plot. The effect is subtle, but it does suggest that the spanning tree is doing a slightly better job of covering the space.

In analyzing the cell-based metrics, we earlier commented that the cell chi-squared seems to be slightly inferior to both cell counts and cell entropy in terms of producing optimally diverse libraries. To quantify this observation, each of the benzodiazepine libraries obtained using a cell-based method was evaluated with regard to its spanning tree score normalized (i.e., divided) by the spanning tree score obtained for the corresponding library optimized using the spanning tree function itself (i.e., the optimal spanning tree score achievable for that library). Thus, if a cell method produced a library as diverse (assessed with the spanning tree function) as the spanning tree library, it would achieve a (normalized) score of 1.0. These normalized spanning tree scores were averaged across all the benzodiazepine libraries obtained for a given cell-based method. The results are as follows. The average score for the cell counts (or cell fraction) method was 0.926, or equivalently, on average it achieved 92.6% of the diversity of the spanning tree libraries. The cell entropy libraries achieved an average score of 0.920 and the cell chi-squared libraries achieved an average score of 0.910, showing very slightly improved results for the cell counts and cell entropy methods relative to cell chi-squared. However, the results are sufficiently close that it would be best to consider all the cell-based methods as yielding good results relative to the spanning tree libraries. Because the spanning tree libraries take significantly longer to calculate, these results tend to validate further the use of the cell methods for larger libraries where the spanning tree method becomes impractical due to its prohibitive cost in computer time.

One final set of experiments was performed to re-examine the conclusion from the first example that independent diversity selection at each R-group position (the "diversity hypothesis") is not a good protocol for selecting a diverse array-based library. Recall that this led to a highly redundant and low-diversity library for the trisubstituted phenyl library. For the benzodiazepine library, we performed R-group–based diversity selection using the MaxMin method in an analogous manner to the protocol described for the trisubstituted phenyl library (as illustrated in Color Plate 1). Each R-group was enumerated at a single position of the benzodiazepine core, and the MaxMin method was used to select either 5 or 10 substituents at that R-group position. The resulting R-group selections then were enumerated into combinatorial sublibraries for arrays of size 5 × 5 × 5 and 10 × 10 × 10. This protocol was performed using each of the three descriptor sets, topological, E-state, and topological plus E-state, with and without normalized principal components. The diversity for the resulting libraries was compared to the diversity for the corresponding library (same descriptors and same library size) obtained using product-based selection with either the spanning tree method or the cell counts method. In terms of either the spanning tree metric or cell counts, the diversity for the R-group–based selections ranged from 46%–85% of the optimal diversity obtained for the product-based libraries. Averaging over the libraries for each of the descriptor sets and averaging over the spanning tree and cell counts functions, the diversity for the 5 × 5 × 5 libraries averaged to 69% of the optimal product-based library diversity. The diversity for the 10 × 10 × 10 libraries averaged to 78% of the diversity for the corresponding product-based libraries. The increase with library size seems reasonable, because in the limit of selecting the full virtual library, 20 × 20 × 20, all methods would give the same result and achieve 100% of the optimal diversity. The poorest results occurred for the 5 × 5 × 5 libraries using either E-state or E-state plus topological descriptors with normalized PCs using cell counts as the diversity metric. In these cases, only 56% and 46% of the optimal cell counts diversity was achieved with the R-group–based selections. Thus, although the average diversity scores for the R-group–based selections ranged from ~70%–80% of optimal, the results could produce libraries as low as ~50% of the optimal diversity. These results tend to reconfirm that product-based diversity selection can yield significantly improved results compared to R-group–based selection, as has also been noted by others.[16,23]

SUMMARY AND CONCLUSION

In this article, differences in the problem of designing diverse libraries consisting of combinatorial arrays vs libraries consisting of discrete molecules ("cherry picked") were discussed. These differences arise because discrete molecule selection enables one to add or replace a single molecule at a time to the library subset, whereas array-based selection requires the selection of combinatorial subsets. Consequently, methods that have been found to work well for discrete molecule selection, such as the MaxMin algorithm,[10,13] encounter problems when applied to array-based selection. Thus, new methods are required for optimizing the diversity of such subsets. We proposed a set of five theoretical criteria for an ideal diversity function to be suitable for use in the problem of designing array-based libraries. We introduced the concept of *incremental diversity* and used it in formulating several of these criteria. The key points are that adding redundant molecules to a system should leave its diversity unchanged, adding nonredundant molecules should always increase the diversity, and space-filling behavior of diversity space should be preferred. Finally, we showed that the diversity function should incorporate a parameter that accounts for the sampling radius (coverage region) of a molecule in diversity space. Several literature diversity functions were assessed according to these criteria, and several new functions were proposed. An ideal diversity function satisfying these requirements was constructed for the simple problem of points in a one-dimensional descriptor space based on the envelope of overlapping Gaussian functions. In generalizing this approach to multidimensional systems, we employed minimum spanning trees to produce a pseudo one-dimensional representation of the multidimensional system. By

summing the error functions of the edge lengths of the spanning tree, we obtained a diversity function that approximately satisfies the proposed requirements. Three cell-based functions, cell counts, cell entropy, and cell chi-squared, were analyzed in the context of the five diversity requirements. The cell-based functions do not satisfy the criteria as well as the spanning tree metric, but they offer the benefit of significantly improved computational performance. The main deficiency with cell-based methods is that similar molecules separated by a cell boundary can yield a higher diversity score than molecules spaced further apart but lying within the same cell.

We applied the new spanning tree diversity function and the cell methods to optimize the diversity of two test libraries. A Monte Carlo single R-group mutation algorithm was used to perform the optimization. The first test system examined was a symmetrical trisubstituted phenyl library with the same 15 R-group fragments at each of the three substituent positions. This virtual library thus contains 3,375 molecules (a $15 \times 15 \times 15$ array) for which $5 \times 5 \times 5$ library subsets were constructed. Only 680 nonredundant molecules were present in this library due to the high degree of symmetry in the scaffold, and thus it poses severe challenges to the problem of generating a diverse combinatorial subset. A set of 44 1D and 2D descriptors were calculated for the molecules of the library, and principal component analysis was used to project the result into a five-dimensional diversity space. Different approaches for selecting library subsets were analyzed. The minimum spanning tree method was found to perform well in producing a diverse library for this system while tending to avoid the selection of redundant molecules. In contrast, libraries designed using the "diversity hypothesis"[23] of independent R-group selection were shown to be significantly less diverse and produced a high proportion of redundant molecules in the library subset. The cell-based metrics, cell counts, cell chi-squared, and cell entropy, were found to yield nearly as diverse libraries as the spanning tree method. Although these libraries contained more redundant molecules than the spanning tree method, this approach offers the major benefit of significantly improved computational speed.

In a second test example, a benzodiazepine scaffold was used in generating a $20 \times 20 \times 20$ virtual library of 8,000 molecules. For this case, no redundant molecules were present in the virtual library. Three different descriptor sets consisting of 43 1D and 2D topological descriptors, 35 electrotopological E-state atom indices, and the union of the topological and E-state descriptors were used to define diversity spaces. Principal component analysis again was used to reduce the dimensionality of each of these descriptor sets. Library subsets ranging from $5 \times 5 \times 5$ to $10 \times 10 \times 10$ arrays were constructed using each of the descriptor sets with each of the diversity functions. Again, it was found that the spanning tree method worked well in selecting diverse subsets while the cell methods performed almost as well (achieving an average diversity of slightly more than 90% of the spanning tree result) while being significantly faster to calculate. In contrast, an approach using independent R-group selection ("diversity hypothesis") achieved only 46%–85% of the optimal diversity, with the higher percentages occurring for the larger size arrays (e.g., $10 \times 10 \times 10$).

In conclusion, the minimum spanning tree error function method was found to be suitable for selecting diverse combinatorial library subsets for both pathological (highly redundant) and more conventional (e.g., benzodiazepine) libraries. The cell methods work almost as well and are preferred for larger libraries where computational speed considerations make the spanning tree approach impractical. For smaller libraries (up to a few thousand compounds in the library subset), the spanning tree method may be employed in cases where one wishes to obtain optimal diversity. The proposed diversity criteria should serve as a guide in the construction of new diversity functions. As new approaches to diversity evaluation continue to develop, their quality now can be gauged by comparison with results obtained using the spanning tree method.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Scott Kahn, Rob Brown, and John Barnard for many illuminating discussions on molecular diversity.

REFERENCES

1 Gallop, M.A., Barrett, R.W., Dower, W.J., Fodor, S.P.A., and Gordon, E.M. Application of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries. *J. Med. Chem.* 1994, **37,** 1233–1251

2 Gordon, E.M., Barrett, R.W., Dower, W.J., Fodor, S.P.A., and Gallop, M.A. Application of combinatorial technologies to drug discovery. 2. Combinatorial organic-synthesis, library screening strategies, and future directions. *J. Med. Chem.* 1994, **37,** 1385–1401

3 Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K., and Moos, W.H. Measuring diversity: Experimental design of combinatorial libraries for drug Discovery. *J. Med. Chem.* 1995, **38,** 1431–1436

4 Kennard, R.W., and Stone, L.A. Computer aided design of experiments. *Technometrics* 1969, **11,** 137–148

5 Marengo, E., and Todeschini, R. A new algorithm for optimal, distance-based experimental design. *Chemometrics Intelligent Lab. Syst.* 1992, **16,** 37–44

6 Brown, R.D., Hassan, M., and Waldman, M. Combinatorial library design for diversity, cost efficiency, and drug-like character. *J. Mol. Graphics Modell.* 2000, **18,** 000–000

7 Brown, R.D. Descriptors for diversity analysis. *Perspect. Drug Discovery Design* 1997, **7/8,** 31–49

8 Brown, R.D., and Martin, Y.C. An evaluation of structural descriptors and clustering methods for use in diversity selection. *SAR QSAR Environ. Res.* 1998, **8,** 23–39

9 Pickett, S.D., Luttmann, C., Guerin, V., Laoui, A., and 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

10 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

11 Mason, J.S., and Pickett, S.D. Partition-based selection. *Perspect. Drug Discovery Design* 1997, **7/8,** 85–114

12 Pearlman, R.S. Novel software tools for addressing chemical diversity. *Network Science* 1996, **June,** http://www.awod.com/netsci/Issues/Jun96/feature91.html

13 Agrafiotis, D.K. Stochastic algorithms for maximizing

molecular diversity. *J. Chem. Inf. Comput. Sci.* 1997, **37,** 841–851
14 Schnur, D. Design and diversity analysis of large combinatorial libraries using cell-based methods. *J. Chem. Inf. Comput. Sci.* 1999, **39,** 36–45
15 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
16 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
17 Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D., and Weinberger, L.E. Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* 1996, **39,** 3049–3059
18 Maggiora, G.M., and Johnson, M.A., Eds. *Concepts and applications of molecular similarity.* John Wiley & Sons, New York, 1990
19 Matter, H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* 1997, **40,** 1219–1229
20 Pearlman, R.S., and Smith, K.M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* 1999, **39,** 28–35
21 Cheng, C., Maggiora, G., Lajiness, M., and Johnson, M. Four association coefficients for relating molecular similarity measures. *J. Chem. Inf. Comput. Sci.* 1996, **36,** 909–915
22 Turner, D.B., Tyrell, S.M., and Willett, P. Rapid quantification of molecular diversity for selective database acquisition. *J. Chem. Inf. Comput. Sci.* 1997, **37,** 18–22
23 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
24 Mount, J., Ruppert, J., Welch, W., and Jain, A. IcePick: A flexible surface-based system for molecular diversity. *J. Med. Chem.* 1999, **42,** 60–66
25 *C2.LibSelect*, version 3.8, 1998, Molecular Simulations Inc., San Diego, CA
26 Applebaum, D. *Probability and information: An integrated approach.* Cambridge University Press, Cambridge, UK, 1996
27 Roman, S. *Coding and information theory.* Springer-Verlag, New York, 1992
28 Abramowitz, M., and Stegun, I.A., Eds. *Handbook of mathematical functions.* Dover Publications, New York, 1972
29 Baase, S. *Computer algorithms: Introduction to design and analysis.* Addison-Wesley, Reading, MA, 1988
30 Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, M.N., and Teller, E. Equation of state calculation by fast computing machines. *J. Chem. Phys.* 1953, **21,** 1087–1192
31 *C2.Diversity*, version 3.5, 1997, Molecular Simulations Inc., San Diego, CA
32 Hall, L.H., and Kier, L.B. Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* 1995, **35,** 1039–1045
33 Hall, L.H., Kier, L.B., and Brown, B.B. Molecular similarity based on novel atom-type electrotopological state indices. *J. Chem. Inf. Comput. Sci.* 1995, **35,** 1074–1080
34 Kier, L.B., and Hall, L.H., *Molecular structure description: The electrotopological state.* Academic Press, San Diego, 1999
35 Everitt, B.S., and Dunn, G., *Applied multivariate data analysis.* Oxford University Press, New York, 1992
36 *C2.Analog Builder*, version 3.5, 1997, Molecular Simulations Inc., San Diego, CA
37 *C2.QSAR+*, version 3.5, 1997, Molecular Simulations Inc., San Diego, CA
38 *C2.Descriptor+*, version 3.5, 1997, Molecular Simulations Inc., San Diego, CA
39 Agrafiotis, D.K., and Lobanov, V.S. An efficient implementation of distance-based diversity measures based on k-d trees. *J. Chem. Inf. Comput. Sci.* 1999, **39,** 51–58
40 Bentley, J.L., and Friedman, J.H. Fast algorithms for constructing minimal spanning trees in coordinate spaces. *IEEE Trans. Comput.* 1978, **C-27,** 97–105