Diversity Space and Its Application to Library Selection and Design

Sara H. Fitzgerald, Michal Sabat, and H. Mario Geysen*

Department of Chemistry, University of Virginia, Charlottesville, Virginia 22904

Received February 28, 2006

To promote more productive combinatorial endeavors, the Diversity Space methodology introduced here enables similarity comparisons at the *library* level. Particularly at an early screening stage, when little or no information is available regarding the pharmacophoric entities necessary for binding, it is more efficient to select or discard an entire ensemble of molecules at once, rather than focus on individual compounds. Also described are applications of the methodology to a form of scaffold hopping, herein categorized as "soft" scaffold hopping, and to a newly introduced approach called surrogate synthesis, both of which are furthered by library-level information that is absent in more traditional molecular similarity calculations.

INTRODUCTION

With the advent of combinatorial chemistry and automated synthesis technologies, it is feasible for a single research group to produce and screen hundreds of thousands of compounds in a single year. Despite the efficiency with which molecules can be synthesized, however, the pharmaceutical industry, as a whole, has neither the time nor the resources to search the universe of potential drug molecules exhaustively. It has been proposed that the maximum number of molecules in this universe may range anywhere from 10^{14} to 10^{30}, depending on the specific criteria employed.[1] Therefore, library design involves complex decision making, considering such factors as the choice of a scaffold or core structure, the availability of monomers, and the synthetic feasibility implied therein. The library's members must also provide an adequate level of molecular similarity or diversity, depending on the task at hand. Though chemists using combinatorial methods are becoming increasingly adept at contemplating these design issues, the challenge that has remained somewhat elusive is that of establishing comparisons between libraries in order to make an appropriate selection for synthesis. For this purpose, researchers require the answers to questions that do not always afford prompt, straightforward responses:

- How do two libraries compare in terms of the chemical space accessible to their monomer decorations?
- What factor should be used to rank a set of proposed libraries according to their overlap in chemical space?
- Can a library be chosen such that it is expected to mimic another library?
- Is a proposed library sufficiently different from those previously synthesized such that it will provide new information regarding binding?
- Given a molecular structure of interest, such as a successful drug molecule, which libraries should be synthesized and screened in order to maximize the possibility of obtaining a hit on the same target?

The methodology described here was designed to aid in answering these types of questions.

Various approaches have been established to quantitatively assess the degree of similarity between two molecules.[2−7] Though the overall goal of each of these methods is the same, they differ in the properties that are considered to have the strongest effect on similarity. Most commonly applied are the two-dimensional approaches, which focus solely on properties resulting from a molecule's connectivity table, with no emphasis on the three-dimensional molecular conformation. Generally, a binary bit string is used to compare two structures, where a "1" designates the presence of a particular descriptor of interest and a "0" indicates its absence. The full bit string, as determined by the entire set of chosen descriptors, makes up the "fingerprint" of the molecule. Descriptors can include atom or element types, structural fragments, or more abstract features such as hydrogen-bond donor/acceptor capability and hydrophobic/hydrophilic tendency. Because of its relative ease of use and quantitative outcome, the 2-D fingerprinting methodology has been implemented in a number of commercially available software packages, including UNITY,[8] MDL MACCS,[9] and MOE.[10] Nevertheless, the effectiveness of bit-string-based similarity methods has been called into question.[11] Because atomic identity and local structure are emphasized, global properties such as molecular size and shape are not accounted for adequately. In addition, for large molecules, saturation of the bit string can lead to an increased probability of obtaining a high similarity score. To overcome these issues, it seems important to turn to a more global approach. Some attempts have been made in this direction, including the methodologies of shape matching[12−14] and pharmacophore searching.[15−20] However, in both of these instances, the biologically active conformation is typically designated as the query structure. When this conformation or its relevant pharmacophoric entities are unknown, these similarity measures lose a great deal of their utility. Even if the lowest-energy conformation or a small sampling of conformations in the energy well surrounding the global minimum are employed instead, they are not likely to correlate with the bioactive conformation,[21,22] devaluing the biologically sig-

* Corresponding author phone: (434) 243-7741; fax: (434) 243-8923; e-mail: geysen@virginia.edu.

10.1021/ci060066z CCC: $33.50
Published on Web 07/08/2006

nificant conclusions that can be drawn and hindering the goal of promoting a more efficient drug discovery pursuit.

Perhaps the biggest drawback to currently available techniques is the necessity to draw comparisons on a *structure-by-structure* basis. In a well-designed library, the monomer selection itself is very diverse, so besides verifying this internal diversity, a similarity comparison of library members does not provide particularly helpful information. Rather than focusing attention at the molecular level, chemists designing libraries would be better aided by a methodology that allowed them to answer questions regarding the degree of similarity or diversity *between* libraries, not just within libraries. In this way, they could decrease the probability of synthesizing libraries that are sufficiently similar to those already explored and, in general, could ensure that the pursuits of combinatorial chemistry are providing access to as much chemical space as possible. A methodology intended to serve these objectives was previously proposed by Pickett et al.[23] However, a molecular-level comparison was retained, and library design decisions were based on the maximum dissimilarity of the final products. (This approach was also heavily dependent on the assignment of pharmacophoric features, which, as mentioned above, can be somewhat ambiguous in the absence of abundant and detailed biological data.)

Described here is a new similarity/diversity measure that addresses the lingering need for an adequate *library-level* approach. Rather than being limited to the information available in a molecule's connectivity table, the Diversity Space approach encompasses 3-D information specific to each of the molecule's accessible conformations. The bioactive conformation is not necessary as query input, and as implied, the methodology is designed to allow for the comparison of libraries rather than discrete molecules. Ideally, a library is designed such that elements of diversity are only introduced at positions predicted or known to be critical to the binding interaction with a protein target. Regardless of the identity of the diversity elements, or *decorations*, their spatial orientation along the contact surface with the protein is a key determinant in the success or failure of the binding interaction. This spatial display is imposed by the *scaffold*, or the common template on which the decorations are attached. To establish the degree of similarity between libraries, then, it is proposed that one can simply compare scaffolds, with respect to the similarity or difference in the display of decorations. (Note that the term "decorations" refers to the diversity elements that are attached to a scaffold. This should not be confused with the term "monomers", which is frequently used to identify the building blocks of a combinatorial library synthesis. The synthesis often dictates that some portion of each monomer is used to build the scaffold, while the unincorporated, diverse portion of each monomer becomes a decoration that is "displayed" by the resulting scaffold.)

It is worthwhile to mention, at this point, that the Diversity Space methodology is most relevant when the scaffold serves as the aforementioned "template" and is not itself involved in the binding interaction with the protein target. Considering a small molecule to have a *back* and *front* is helpful, where the back is the noninteracting surface and the front is the interacting surface (Figure 1a,b). (Note that *front* and *back* here are not identified from the two-dimensional representation of a library but rather from a three-dimensional perspective, defined by the binding interaction.) Scaffolds that aim all their decorations in the same general direction play the role of a *backside* template, projecting only the elements of diversity to the front. Presumably, it is this class of scaffolds for which one can imagine trading a given scaffold for another without affecting the binding interaction, provided the diversity elements are displayed in a similar orientation. In this case, the protein effectively sees no change in the contact surface. (This, of course, does not imply that the measured biological activity will be equivalent, as this may be impacted either positively or negatively by a change in the solvation properties of the new scaffold.) On the other hand, planar scaffolds or those in which the decorations are not all projected in the same direction cannot be considered just backside templates. Now, the scaffold is much more likely to be a significant component of the contact surface rather than merely the handle from which the diversity elements protrude.

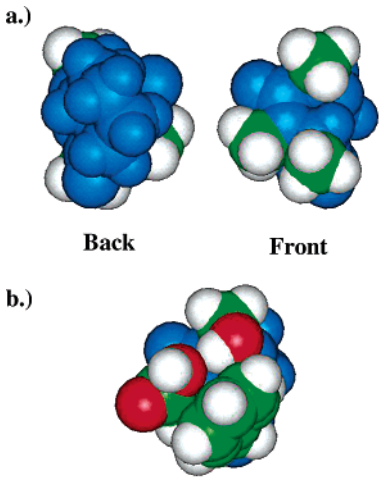

Figure 1. (a) Image showing the *back* and *front* of the 245-21 scaffold (structure shown in Figure 5a) as modeled for the Diversity Space analysis with methyl groups at each R position. (b) Image showing the front of the 245-21 scaffold with typical R-group functionality, here represented by the side chains of serine, aspartic acid, and phenylalanine. For this library, the scaffold (shown in blue) is clearly not a significant part of the front contact surface. To some extent, this is evident even in the methyl-substituted model shown in **a**, but it becomes even more obvious when typical monomer decorations are employed. Substituting an alternative scaffold, in this case, would not be expected to interrupt any binding interactions involving the diversity elements, provided they could be displayed in a similar orientation. (Note: The minimum energy conformation of 245-21 is shown here, but for this particular library, 30 out of 31 total conformers preserved an arrangement consistent with the notion of a scaffold that serves primarily as a backside template.)

Having defined the scaffolds to which our approach is most applicable, then, the tenets of the Diversity Space methodology can be delineated as follows:

(1) The structures investigated will be scaffolds with three points of diversity or decoration, all of which are expected to contribute to the structure−activity relationship (SAR). (If a three-point diverse library contains a site of decoration that does not contribute to the SAR, one could argue that it would be more accurately represented as a two-point diverse library, for which the current methodology is not relevant.)

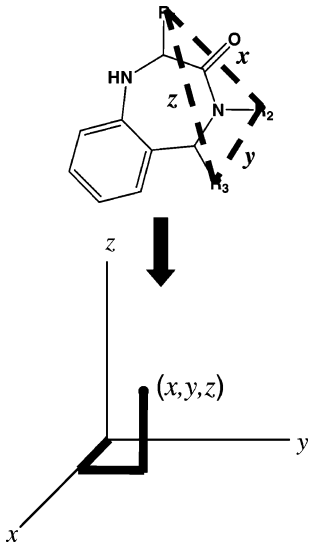

Figure 2. Diversity triangle for a selected library and its corresponding translation to a diversity space point. The R−R distances are assigned to the *x*, *y*, and *z* coordinates of diversity space starting at the 12:00 position and moving in a clockwise direction around the triangle.

(2) The spatial orientation of the decorations on a scaffold will be designated as the primary indicator of the similarity or diversity between libraries. As currently implemented, the methodology specifically focuses on the *distances* between the decorations, as the angular component of the spatial orientation is proposed to be secondary to the distance component. (An appropriate angular measure is being investigated, however, as this information may be particularly informative in cases where the distance analysis results in several viable alternative libraries.)

(3) The distances between a library's three diversity elements will be combined to produce a *diversity triangle*.

(4) The lengths of the sides of a library's diversity triangle will be translated into the *x*, *y*, and *z* coordinates designating a point in *diversity space* (Figure 2).

(5) Diversity space will be divided into boxes in order to quantify the coverage of the space.

(6) The degree of similarity or diversity between two libraries will be assessed on the basis of the number of diversity space boxes the two libraries have in common (given as a percent overlap).

The choice of three-point diverse libraries serves a purpose beyond simply the ease of translation to the three-dimensional diversity space environment. Two-point diverse libraries, though often utilized, do not offer the numerical advantage desired in synthesizing large combinatorial libraries. That is, much larger monomer sets are required for a two-point diverse scaffold to reach the same number of overall library members as a three- or four-point diverse scaffold. However, one is unlikely to ensure that all of the positions of diversity project toward the same contact surface if the number of decorations exceeds three. Therefore, if it is to be assumed that all of the positions of diversity on a scaffold contribute to the structure−activity relationship for that library, it is practical to consider a three-point diverse system as the optimal choice for library selection and design.

As will become obvious, the Diversity Space approach naturally lends itself to exercises that are central to the goals of combinatorial science, including the following two tasks:

(1) ***Scaffold hopping*** rests on the tenant that preserving a few key features of a bioactive molecule is enough for activity to be witnessed in other molecules that are structurally very different. It is a well-established approach, often employed to replace a component of a molecule that is undesirable (in terms of solubility, toxicity, binding affinity, etc.) or to avoid intellectual property (IP) infringements. To distinguish our approach to scaffold hopping from more traditional methodologies, we have chosen to identify it as "soft" scaffold hopping. This choice of terminology will be further explained below.

(2) ***Surrogate synthesis*** is an in-house term we have used to designate the act of identifying the optimal library to synthesize from a structurally related set of potential libraries. This "surrogate" library is chosen on the basis of the chemical space it shares in common with other members of the set as well as on its own chemical tractability, which includes such significant factors as monomer availability, synthetic fidelity, and average yield.

These two applications of the described Diversity Space methodology are explained in more detail in the Results and Discussion section below.

METHODS

All of the members of a library can be represented by a Markush structure, in which the scaffold is depicted with its assumed diversity at the points of decoration, labeled R_1, R_2, R_3, and so forth. As previously mentioned, the R−R distances for all three-point diverse libraries create a triangle, the size and orientation of which represent the spatial display of the library's decorations. To establish a reference for measuring these diversity triangles, the monomer decorations on each scaffold were modeled as methyl groups, with the carbon atom of the methyl substituent representing the point of attachment of the decoration to the scaffold. This strategy reduces each library to the core structure common to all members (Figure 3a−c). To account for the flexibility in each library, conformational searches were completed in Catalyst 4.10 (BEST algorithm; maximum of 255 conformers; energy range of 10 kcal/mol from the global minimum conformation).

In addition to taking into account the conformational flexibility of the scaffolds, their rotational capability was considered as well. It was assumed that any scaffold acting only as a backside template is free to rotate in the plane of the diversity triangle, provided there is no flip about this plane that orients the decorations making up the contact surface away from the protein target. With this rotational capability, the size of the scaffold's diversity triangle does not change, only its orientation with respect to the binding site on the protein. However, as alluded to in the caption for Figure 2, the orientation of the scaffold clearly defines the assignment of the R−R distances to diversity space coordinates. By rotation, any one of the decorations could be at the 12:00 position, thereby producing three nonredundant diversity space points for each scaffold (Figure 4). It is essential to account for this rotation in order to adequately distinguish between scaffolds that truly differ in the spatial display of the decorations and those that differ only by rotation and could therefore be expected to bind to the same protein target. When the nonredundant diversity

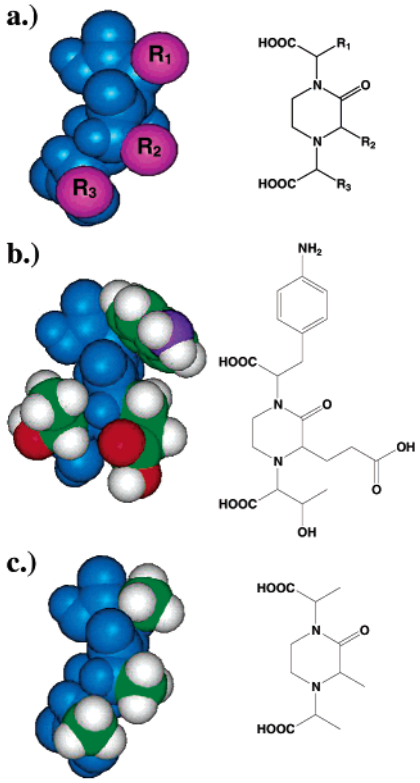

Figure 3. Space filling and line models for (a) the Markush representation of a library, where the scaffold is depicted in blue and the monomer decorations are depicted with the R_1-, R_2-, and R_3-labeled pink spheres; (b) the same scaffold substituted with three typical decorations, here a 4-aminotoluene substituent in the R_1 position, a glutamic acid side chain in the R_2 position, and a threonine side chain in the R_3 position; and (c) the core structure common to all members of the depicted library, regardless of the identity of the decorations (assuming R_1, R_2, and $R_3 \neq H$). Reducing the numerous members of each library to this core structure thereby enables the desired library-level similarity comparison.

space points are included, the latter group will produce an overlap in diversity space that implies their similar binding capacity, while the scaffolds that have a different spatial display altogether will not. Consider the hypothetical set of R−R distance values given in Table 1. Once rotated, library

Table 1. Hypothetical R−R Distance Values that Illustrate the Need to Account for Scaffold Rotation

	R_1-R_2	R_2-R_3	R_3-R_1
library A	2.5	4.8	3.2
library B	4.8	3.2	2.5
library C	5.6	2.9	3.7

B will produce a diversity triangle identical to that of library A and could therefore be expected to display similar binding properties. Library C, on the other hand, has a diversity triangle with completely different dimensions. Clearly, by considering all three nonredundant points for each of these libraries, the similarity of A and B will emerge as overlap in diversity space, while library C will appropriately retain its distinction.

Due to the accessible rotations, a library with 10 conformers, for instance, produces a three-dimensional diversity space plot of 30 points. As expected, then, the diversity space plot for an entire set of libraries quickly becomes too dense to be useful for library comparisons. However, if this space is

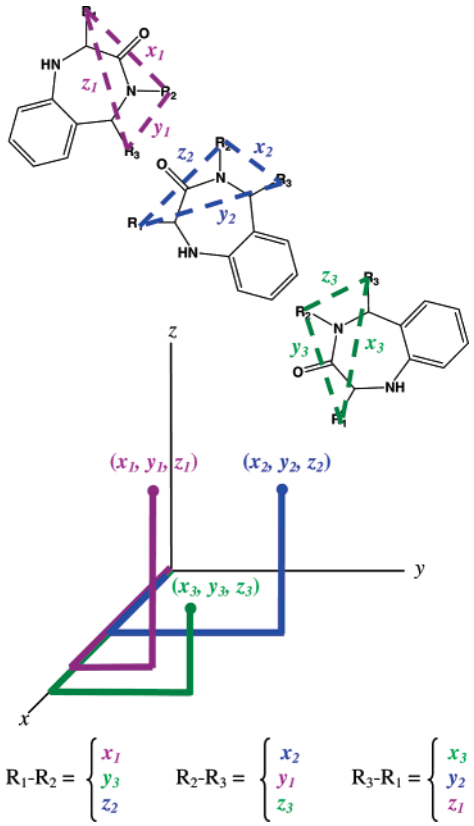

Figure 4. Image showing that the rotational capability of a small molecule scaffold enables any one of the monomer decorations to be at the 12:00 position. The chosen scaffold is rotated such that R_1 is at the 12:00 position for the orientation corresponding to the purple diversity triangle, R_2 is at the 12:00 position for the orientation corresponding to the blue diversity triangle, and R_3 is at the 12:00 position for the orientation corresponding to the green diversity triangle. The clockwise assignment of the R−R distances to diversity space coordinates thereby produces three nonredundant points in diversity space.

considered to be divided into boxes, comparing the coverage of diversity space becomes a much more meaningful exercise. And, because the points in diversity space correspond to the diversity triangle and therefore to the spatial display of the decorations, libraries that fill many of the same boxes can be considered to produce contact surfaces that are highly conserved. Conversely, libraries that fill none of the same boxes can be considered to access separate regions of chemical space. (For the research described here, the boxes were chosen to have a dimension of 1 Å, but this value can vary depending upon the level of sensitivity desired in the resulting overlap information.)

As mentioned above, the degree of similarity or diversity between two libraries can be assessed by comparing the coverage of the libraries in terms of their filling of the boxes of diversity space. This comparison leads to two separate, but equally important, forms of overlap. For the comparison of libraries A and B, the first overlap equation takes the following form:

$$\% \text{ overlap} = 100[AB/(A + B - AB)] \qquad (1)$$

where AB is the number of boxes libraries A and B have in common, A is the number of boxes filled by library A, and B is the number of boxes filled by library B. (As one will notice, this equation is of the same form as the Tanimoto

coefficient,[10] which is used to calculate bit string similarity in many of the two-dimensional molecular similarity measures discussed above. For the Diversity Space methodology, however, the "bit string" is now a function of the diversity triangles, or the chemical space accessed by the conformations of a given library, rather than a function of abstract descriptors or molecular fragments implied by the molecule's connectivity table. Therefore, it is anticipated that the data resulting from eq 1 will not suffer from the aforementioned pitfalls of the more traditional measures.) Because the value of eq 1 is the same with respect to both A and B, the matrix produced from the comparison of a set of libraries is symmetric in this case. The % overlap values from this first equation can be used to scaffold hop between different libraries, a high value indicating similar coverage of diversity space and therefore an increased likelihood of success in scaffold hopping.

The previous % overlap value, although useful for scaffold hopping, does not reflect the total number of boxes filled by each library. For example, if A fills 50 boxes, B fills 10, and they have 5 in common, a 9% overlap is obtained for both according to the above equation. It is obvious that scaffold hopping between A and B would not be predicted to be successful. But, regardless of this fact, it is also evident that the overlap is much less with respect to A than with respect to B. A second approach is therefore required to consider the overlap with respect to each individual library:

$$\%\ \text{overlap}_A = 100(AB/A) \tag{2}$$

$$\%\ \text{overlap}_B = 100(AB/B) \tag{3}$$

where % overlap$_A$ represents specifically the overlap with respect to library A and % overlap$_B$ represents specifically the overlap with respect to library B. Because the values of eqs 2 and 3 are now different, the matrix produced from the comparison of a set of libraries is asymmetric in this case. As will be demonstrated below, the matrix produced from these new overlap values can be used to further the exercise of surrogate library selection.

RESULTS AND DISCUSSION

As suggested thus far, the Diversity Space methodology is intended to enable a comparison between libraries, where the predominant factor contributing to overlap or similarity is the diversity triangle, represented by the distances between the three decorations on each library. A comparison such as this should facilitate library selection in a way that few current methodologies, with their molecular-level comparisons, are able. To establish the value of the approach, then, conformational searches have been run for five structurally related sets of in-house libraries: the 2,4,5-substituted benzodiazepine (BZD) rings, the 3,4,5-substituted BZD rings, the 1,3,5-substituted BZD rings, the diamide rings, and the tri-amino acid ring. This totals 50 libraries (Figure 5a−e). In cases where the scaffold contains an sp^3 bridgehead carbon, its hydrogen was retained in the axial position, as this arrangement was considered to be significantly more stable. However, all other stereocenters were considered variable, giving a total of 180 different scaffolds for the 50 libraries shown in Figure 5. The diversity space coverage and corresponding overlap matrices have been produced for all of these libraries (Figures 6−8).

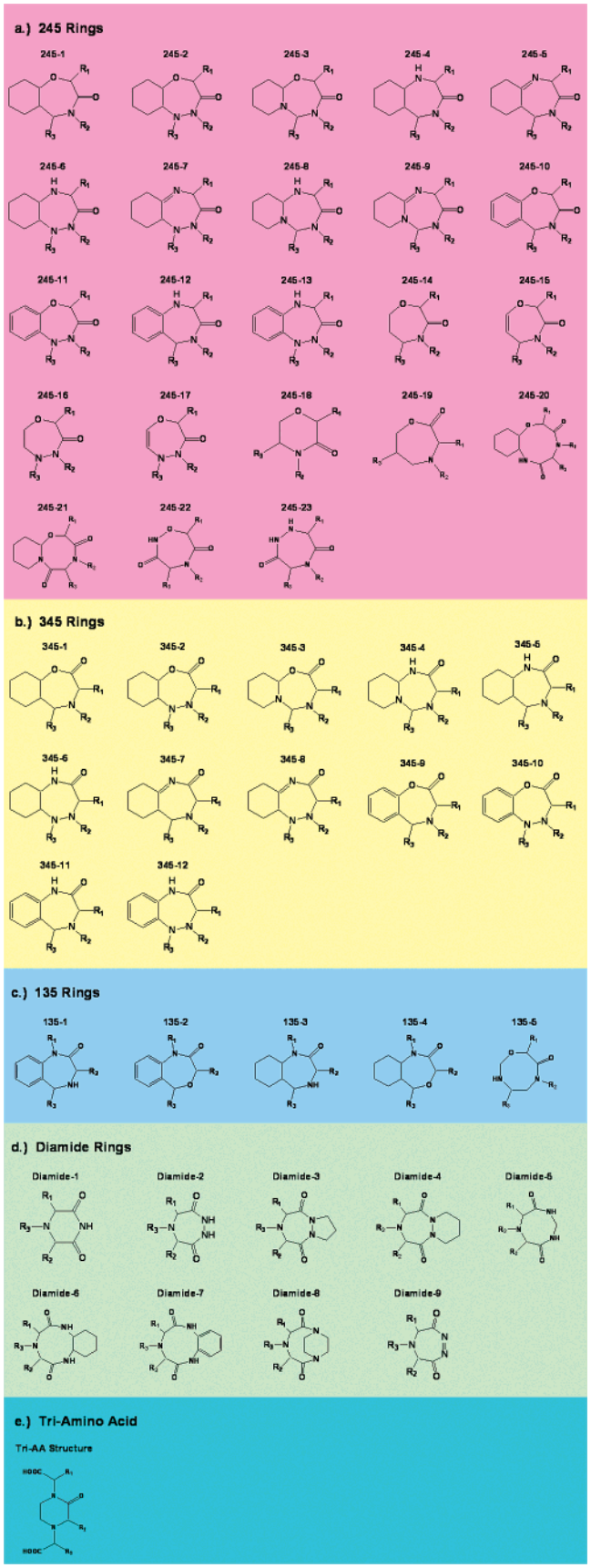

Figure 5. Structures for the 50 in-house libraries investigated using the Diversity Space methodology.

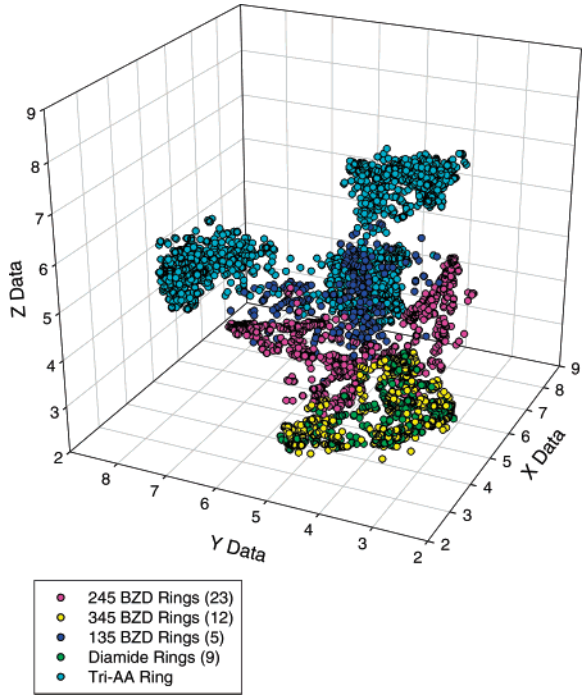

Figure 6. Diversity space points for all conformations of the BZD, diamide, and tri-amino acid rings shown in Figure 5. For each conformer, all three nonredundant points are plotted to account for the rotational capability of the scaffolds.

As evident in the matrices of Figures 7 and 8, the results for the stereoisomers of each library were collapsed in all cases except for the tri-amino acid library. For instance, the diversity space points produced for the conformers of scaffolds 245-1RR, 245-1RS, 245-1SR, and 245-1SS were combined into one file, and the diversity space coverage and overlap of library 245-1 were calculated from this combined file. But, to assess how different stereoisomers would actually compare, the diversity space coverage and overlap of the eight tri-amino acid scaffolds were calculated on their own. According to their overlap in diversity space, the isomerism was shown to contribute to a moderate level of distinction between the tri-amino acid structures, with an average overlap percentage of 61.2% in the symmetric case and 77.5% in the asymmetric case. It is not believed, however, that this distinction is significant enough to warrant the advanced data management and graphical organization that would be necessary to extract useful information from the diversity space analysis for a large number of libraries with multiple stereoisomers. Initially, then, the diversity space points for the stereoisomers can be combined into one file, with the understanding that a consideration of the optimal stereochemistry for any particular library can be re-examined at a later time.

The matrix cells have been color-coded so that any trends in overlap percentages can be easily spotted. The colors are as follows: purple $>$ 0%, blue $\geq$ 20%, green $\geq$ 40%, yellow $\geq$ 60%, orange $\geq$ 80%, and red $=$ 100%. As indicated by the black boxes in Figures 7 and 8, an analysis of the areas of high overlap on either matrix reveals that there are four distinct regions of chemical space being accessed by this collection, with the diamide rings falling in the same chemical space as the 3,4,5-substituted BZD rings. The two potential applications of this data are discussed in more detail below.

"Soft" Scaffold Hopping. To recognize the utility of the Diversity Space approach to scaffold hopping, particularly for early library screening rather than lead optimization, it is important to consider other techniques that are available at this time. Current methods of scaffold hopping employ many of the similarity search techniques previously discussed, including shape-, pharmacophore- and fingerprint-based strategies.[24] Gillet et al.[25] have also recently described the use of reduced graphs for scaffold hopping. In this approach, structural features pertinent to binding retain a topological relationship to one another but are encoded in 2-D rather than 3-D descriptors, making it unnecessary to consider conformational flexibility. In addition, fragment replacement methods such as those implemented in the CAVEAT program[26] are sometimes utilized, with the requirement that the bioactive or lowest-energy conformer be designated as the query structure. In general, however, it is believed that available scaffold hopping methods depend too heavily on the similarity of the scaffolds themselves. Despite the underlying assumption that only a few key features need to be conserved, most 2-D and shape matching approaches are dramatically affected by the structural fragments making up the scaffold. Pharmacophore searching and fragment replacement strategies are less biased toward the structural similarity of the scaffold, but these methods have drawbacks of their own regarding the aforementioned query requirements. In the Diversity Space methodology described here, both of these issues are avoided. First, the scaffold structure is only considered as the template for the display of the monomers and not as a measure of similarity on its own. Second, the analysis is undertaken for *all* conformers with a calculated energy that falls within a certain range, chosen here as 10 kcal/mol, of a library's global minimum structure. Thus, there is no need to know the bioactive conformation or otherwise depend on the quality of a given query structure.

As highlighted in the Introduction, the most significant advantage of our methodology is that it allows for a library-level assessment. Because of this, however, we consider it necessary to categorize our approach to scaffold hopping as "*soft*". Unlike the fragment replacement or reduced graph approaches, Diversity Space does not encompass the types of molecular comparisons frequently used in lead optimization. Rather, it allows for scaffold hopping either at an early screening stage, when little is known about important pharmacophoric features, or at a later stage, when there is a compelling reason to synthesize an alternative library that avoids absorption, distribution, metabolism, excretion, and toxicity (ADMET) or even IP issues. The lack of molecular-level information may mean that alternative scaffolds identified in Diversity Space are not as absolute as those identified by other approaches, but the library context and the diversity included therein should be more than enough to compensate for the "soft" nature of our methodology. Once this "soft" scaffold hopping approach has yielded adequate preliminary binding information and an appropriate SAR has been developed for the "new" scaffold, further optimization of the lead compound with a more absolute molecular-level scaffold hopping approach may be desired.

Again, potential libraries for "soft" scaffold hopping can be identified from eq 1, which corresponds to the symmetric matrix shown in Figure 7. For 10 of the 50 libraries analyzed

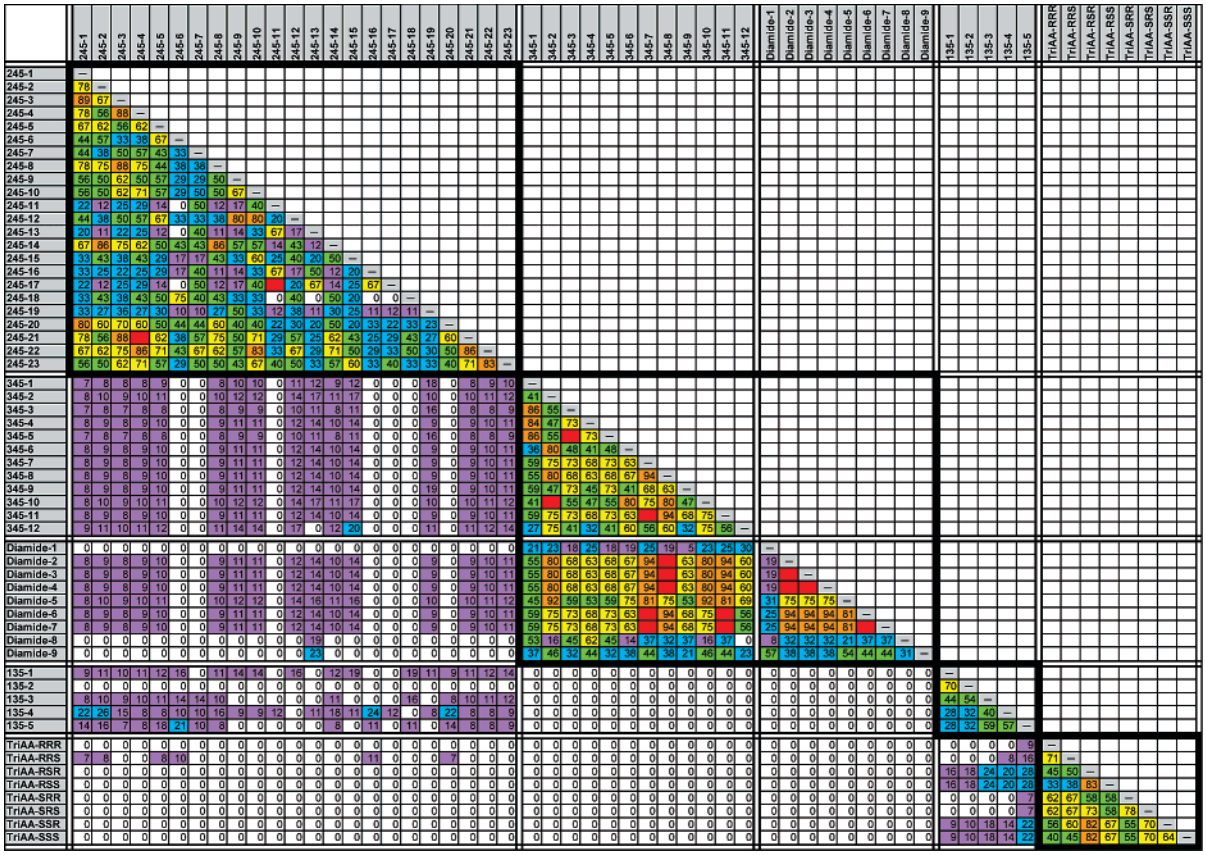

Figure 7. Symmetric matrix derived from the % overlap values of eq 1. The black boxes have been added to highlight the areas of high overlap.

to date, at least one other library can be selected that fills 100% of the same diversity space boxes, and almost half (24) of the libraries have at least one other library that fills ≥ 80% of the same diversity space boxes. Equally important is the identification of libraries with low overlap percentages that should be passed over when considering scaffold hopping alternatives. Even in a set of libraries with structurally similar scaffolds, such as the 2,4,5-substituted BZD rings, it is plain to see that one can make both good and bad choices for scaffold hopping (Figure 9). The Diversity Space methodology therefore appears to provide a novel approach by which to compare scaffolds and identify potential candidates for "soft" scaffold hopping.

Furthermore, because the Diversity Space approach is independent of the size of the libraries being compared, the limiting case can even be analyzed, that is, scaffold hopping from a *single* molecule of interest. If this molecule can be reduced to a structure representing a decorated scaffold, then scaffold hopping to other similar libraries becomes a feasible prospect. [This is, of course, with the understanding that the perceived scaffold is not expected to take part in the binding interaction with a protein target but is instead serving only as the *backside* template from which the interacting decorations protrude (Figure 1).] From a drug discovery perspective, this could be one of the most promising applications of the Diversity Space approach because it suggests a rational way of building upon current knowledge surrounding drug molecules, particularly those that have been successfully marketed, to develop novel, and hopefully more effective, treatments.

Surrogate Synthesis. As evident in the matrices of Figures 7 and 8, one can quickly determine if a particular region of chemical space is predominant among a set of libraries. If this is the case, resources should not be wasted in the study of multiple libraries that all provide access to the same region of diversity space, particularly not at an early screening stage in the drug discovery process. It becomes desirable, then, to determine which library from the set will most adequately provide information about the binding that would be accessible to all members of the set—in other words, the best "surrogate" can be chosen. If library A is a complete subset of library B (100% of the boxes filled by A are also filled by B), it is clear that there is no need to synthesize both. The decorations of library B will be able to access all of the spatial orientations accessible to library A, as well as possibly some additional orientations. Thus, for a set of libraries providing access to the same region of chemical space, the optimal choice of a surrogate library is the one for which the greatest number of other libraries are complete subsets of the surrogate or are at least highly contained in the surrogate. When such issues as monomer availability and synthetic feasibility for this surrogate library are considered, an alternative surrogate may need to be chosen if the first choice is chemically problematic. In this case, one can easily return to the asymmetric overlap matrix and consider the second best alternative.

To make a surrogate selection, it is important to know how the asymmetric overlap matrix should be read. As implied by eqs 2 and 3, the value given in each cell is the percent of diversity space coverage maintained if the library

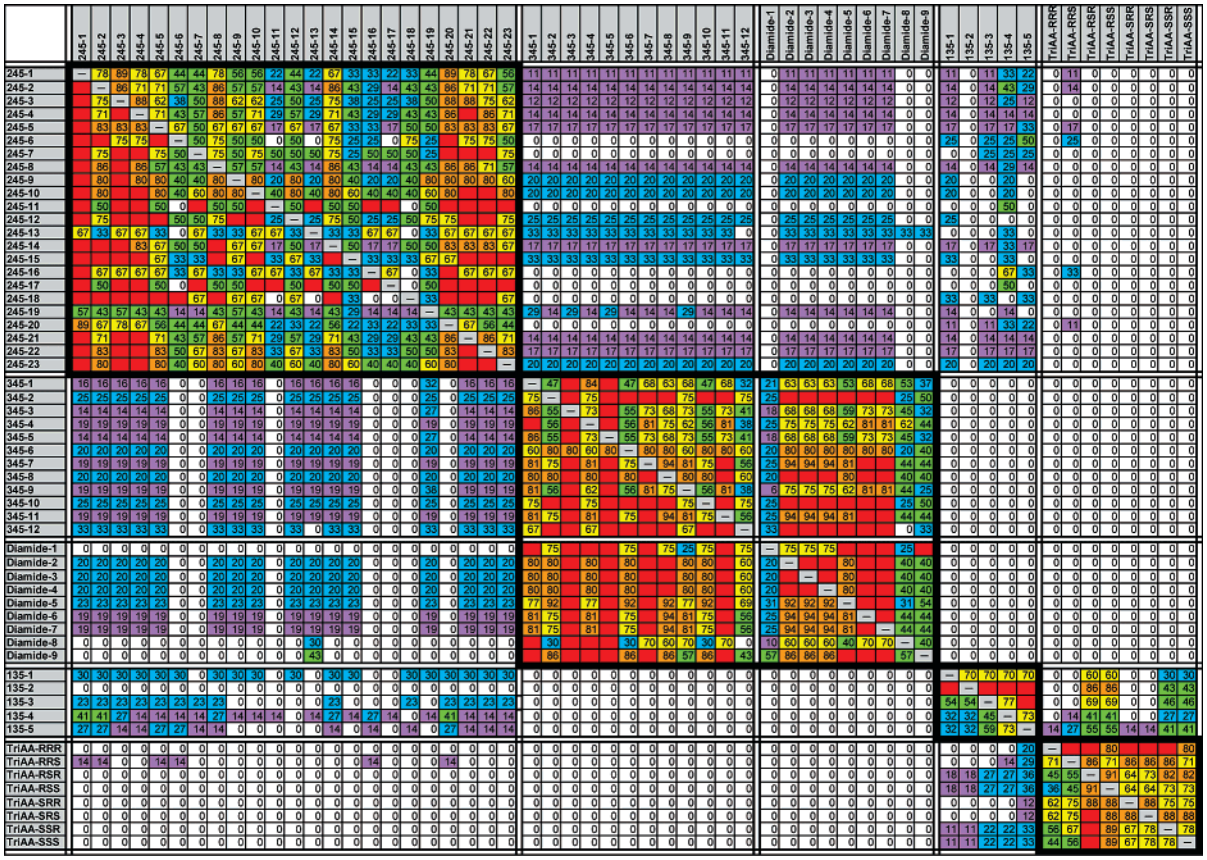

Figure 8. Asymmetric matrix derived from the % overlap values of eqs 2 and 3. Again, the areas of high overlap are readily apparent and have been highlighted with the surrounding black boxes.

listed in the row was replaced by the library listed in the column. Thus, the choice of a surrogate is made by comparing the % overlap values listed in each column. For the four regions of diversity space accessed by the 50 libraries studied to date, surrogate libraries can be easily selected. The column containing the greatest number of complete subsets (100% overlap) is represented by 245-1 for the first region, 345-3 or 345-5 for the second region, 135-5 for the third region, and TriAA-RSR for the fourth region. Keep in mind that the overlap of libraries in diversity space is linked back to the spatial display of the decorations on the scaffolds. To think, then, that one could access nearly all of the spatial orientations available to 50 libraries with the proper selection of just four libraries is an extremely promising finding, suggesting quite an efficient means by which to conserve synthetic resources.

As discussed in the Introduction, it is simply not feasible to synthesize all possible libraries of druglike molecules. Thus, the surrogate synthesis approach is intended to serve as a prioritization strategy, facilitating library selection and design by maximizing the likelihood of finding out about an entire region of chemical space with just one library synthesis. This enables subset libraries to be reserved and explored in a more informed manner in cases where the screening of the surrogate suggests promising bioactivity.

Before concluding, one important topic remains to be discussed. As illustrated in Figure 1, our approach is most applicable in cases where the scaffold serves as a backside template only, and not as part of the front contact surface. However, at this point, the notion of front and back is largely qualitative in nature. To quantify this concept, one needs a vector or angle for each diversity element that describes where it projects in relation to the remainder of the structure. In addition to reporting whether the decorations all project in the same direction, this angle may also provide additional criteria from which to establish a comparison between libraries, though, as mentioned, angular overlap is expected to be considerably inferior to the distance information already obtained. It was originally proposed that the vectors between the geometric center of each scaffold and its decorations would provide a good measure of whether all three decorations were pointing toward a common front surface. However, the results of this work revealed a bias toward large scaffolds. Because the geometric center is located farther from the points of decoration in the case of large scaffolds, the vectors obtained give a tighter projection than is warranted by the actual arrangement of the decorations (Figure 10). At this point, we are still investigating an angular measure that adequately defines the front/back concept in a nonbiased manner. Once this measure is established, it will be further analyzed for its potential utility as a second tier of information for making library comparisons.

CONCLUSION

As shown, the Diversity Space approach enables a more appropriate quantitative assessment of similarity/diversity at the library level, an assessment which makes possible two important applications: "soft" scaffold hopping and surrogate synthesis. It is predicted that scaffolds having a large percentage of overlap in diversity space could be substituted for one another in order to fine-tune the features of a given

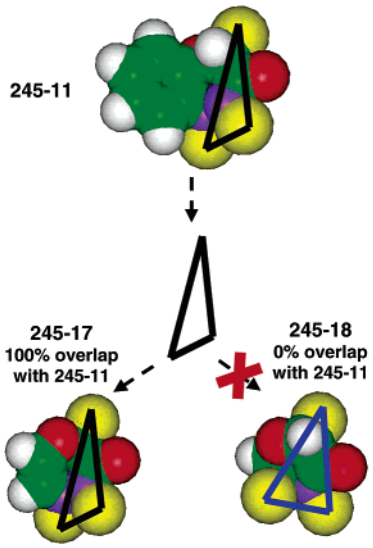

Figure 9. When "soft" scaffold hopping from library 245-11, 245-17 represents a candidate that would be expected to be successful (100% overlap in the symmetric matrix shown in Figure 7), while 245-18 represents a candidate that would be expected to perform poorly (0% overlap in Figure 7). This is supported by the observation that libraries 245-11 and -17 differ only in the backside portion of their scaffolds, whereas the structural change in library 245-18 affects the orientation of the diversity elements at the front as well. For each of these libraries, a single conformer is shown above. The methyl groups have been represented as yellow spheres, and the diversity triangles have been drawn in to emphasize the comparison. As expected, 245-11 and -17 share identical triangles, but 245-11 and -18 do not.

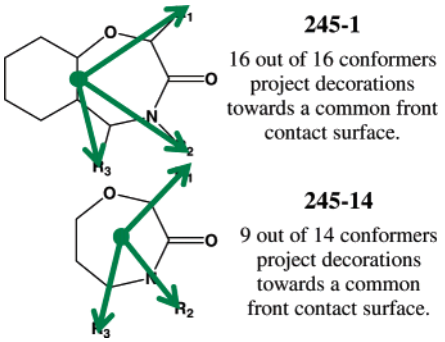

Figure 10. Example showing the bias toward large scaffolds that was discovered when the geometric center was used as the vector reference point. For this analysis, the geometric center of the structure was placed at the origin, and the center of the diversity triangle was placed on the $+z$ axis. In this way, any conformer with a decoration vector that pointed in the $-z$ direction was considered to be undesirable because all of its decorations did not point toward a common front contact surface. 245-1 and 245-14 are identical, with the exception of the additional ring. Thus, these two scaffolds should produce similar results with respect to the percent of total conformers that meet the above directional criteria. However, as shown, all of the 245-1 conformers project their decorations toward a common contact surface, while only ~64% of the 245-14 conformers do the same. The vectors have been drawn in to emphasize how the change in location of the geometric center affects the overall projection of the vectors enough to cause this discrepancy.

scaffold or to avoid IP issues surrounding a particular library of molecules. In addition, to avoid the waste of time and resources, sufficiently similar libraries should not be synthesized. In this case, it is believed that the results of a Diversity Space analysis can be used to identify an appropriate surrogate library. Certainly, an appropriate extension to the present work would focus on comparisons between libraries in the public domain, with respect to the similarity or difference in the display of their decorations and the resulting degree of overlap in diversity space. Because the synthetic approach has already been established for each of these libraries, the ability to define potential candidates for scaffold hopping and surrogate synthesis from this public domain set would be valuable indeed. As such, this research is currently underway, and the results will follow shortly.

As with any approach to library design, it is important to emphasize that we are not suggesting absolutes with the present methodology, only better probabilities. In other words, we do not trick ourselves into thinking the Diversity Space strategies for "soft" scaffold hopping or surrogate synthesis will work in every case. Future application of the methodology and subsequent experimentation will certainly highlight its advantages and limitations. However, if the Diversity Space approach can be used, in general, to guide the library design decisions of medicinal and combinatorial scientists and increase their chances of a more successful synthetic pursuit, then the overall goal has been met.

REFERENCES AND NOTES

(1) Geysen, H. M.; Schoenen, F.; Wagner, D.; Wagner, R. Combinatorial Compound Libraries For Drug Discovery: An Ongoing Challenge. *Nat. Rev. Drug Discovery* **2003**, *2*, 222−230.
(2) Downs, G. M.; Willett, P. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1996; Vol. 7, pp 1−66.
(3) Bures, M. G.; Martin, Y. C. Computational methods in molecular diversity and combinatorial chemistry. *Curr. Opin. Chem. Biol.* **1998**, *2*, 376−380.
(4) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(5) Mason, J. S.; Hermsmeier, M. A. Diversity assessment. *Curr. Opin. Chem. Biol.* **1999**, *3*, 342−349.
(6) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903−911.
(7) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338−345.
(8) *UNITY*; Tripos, Inc.: St. Louis, MO, 1995.
(9) *MACCS-II*; MDL Ltd.: San Leandro, CA, 1992.
(10) *MOE, Molecular Operating Environment*; Chemical Computing Group, Inc.: Montreal, Canada, 1997.
(11) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.
(12) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. New Method for Rapid Characterization of Molecular Shapes: Applications in Drug Design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79−85.
(13) Masek, B. B.; Merchant, A.; Matthew, J. B. Molecular Shape Comparison of Angiotensin II Receptor Antagonists. *J. Med. Chem.* **1993**, *36*, 1230−1238.
(14) Walker, P. D.; Maggiora, G. M.; Johnson, M. A.; Petke, J. D.; Mezey, P. G. Shape Group Analysis of Molecular Similarity: Shape Similarity of Six-Membered Aromatic Ring Systems. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 568−578.
(15) Van Drie, J. H. Strategies for the determination of pharmacophoric 3D database queries. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 39−52.
(16) Matter, H.; Pötter, T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.
(17) Kurogi, Y.; Güner, O. F. Pharmacophore Modeling and Three-dimensional Database Searching for Drug Design Using Catalyst. *Curr. Med. Chem.* **2001**, *8*, 1035−1055.
(18) Makara, G. Measuring Molecular Similarity and Diversity: Total Pharmacophore Diversity. *J. Med. Chem.* **2001**, *44*, 3563−3571.
(19) Abrahamian, E.; Fox, P. C.; Nærum, L.; Christensen, I. T.; Thøgersen, H.; Clark, R. D. Efficient Generation, Storage, and Manipulation of Fully Flexible Pharmacophore Multiplets and Their Use in 3-D Similarity Searching. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 458−468.

(20) Langer, T.; Krovat, E. M. Chemical feature-based pharmacophores and virtual library screening for discovery of new leads. *Curr. Opin. Drug Discovery Dev.* **2003**, *6*, 370−376.
(21) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. A. Conformational Changes of Small Molecules Binding to Proteins. *Bioorg. Med. Chem.* **1995**, *3*, 411−428.
(22) Perola, E.; Charifson, P. S. Conformational Analysis of Drug-Like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization upon Binding. *J. Med. Chem.* **2004**, *47*, 2499−2510.
(23) Pickett, S. D.; Luttman, C.; Guerin, V.; Laoui, A.; James, E. DIVSEL and COMPLIB − Strategies for the Design and Comparison of Combinatorial Libraries using Pharmacophoric Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 144−150.
(24) Böhm, H.-J.; Flohr, A.; Stahl, M. Scaffold hopping. *Drug Discovery Today: Technol.* **2004**, *1*, 217−224.
(25) Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold Hopping Using Clique Detection Applied to Reduced Graphs. *J. Chem. Inf. Model.* **2006**, *46*, 503−511.
(26) Lauri, G.; Bartlett, P. A. CAVEAT: A program to facilitate the design of organic molecules. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 51−66.

CI060066Z