ELSEVIER

Oriented Substituent Pharmacophore PRopErtY Space (OSPPREYS): A substituent-based calculation that describes combinatorial library products better than the corresponding product-based calculation

Eric J. Martin and Thomas J. Hoeffel

Chiron Corporation, Emeryville, California, USA

Initial combinatorial library designs were based on 2D substituent properties. Subsequently, two important extensions were introduced to improve the approach: use of pharmacophores to introduce 3D information, and performing calculations on the enumerated library products rather than just on the substituents. Unfortunately, practical compromises due to the large number of possible products, the large number of conformations per product, and the explicit dependence on the scaffold limit the application of these extensions in five important ways: (1) to small virtual libraries, (2) to only 3- or 4-point pharmacophores, (3) to inadequate conformational sampling, (4) to simplistic diversity measures, and (5) to requiring a complete new calculation for every new library. The 3D oriented substituent pharmacophores have been developed to overcome these limitations. These add two additional points and corresponding distances to each substituent pharmacophore. This adds little additional computation beyond a normal 3D pharmacophore calculation on the substituents, but recaptures most of the orienting information lost in breaking up the enumerated products into fragments. Two main approximations are still implicitly required: the combinatorial conformer assumption and the template alignment assumption. In turn, however, they are designed to account not just for the 3- and 4-point pharmacophores, but for pharmacophores with up to 9 points in enumerated products with three sites of diversity. Perhaps more importantly, pharmacophore calculations are shown to be very sensitive to conformational sampling. The small number of substituents, plus the small number of rotatable bonds per substituent, permits very thorough conformational sampling. For a rigid scaffold with three diversity sites of 1,000 candidate substituents each, the number of molecules to analyze is reduced by a factor of 10^6, and the number of conformations per molecule is reduced by another 10^4. In addition, the modest number of pairwise substituent similarities permits the creation of a Euclidean property space by MDS. This allows for sophisticated experimental design methods that require coordinates, rather than just the counting of the number of set bits in a library union fingerprint. Finally, oriented substituent calculations are scaffold independent and transferable. They can be stored in a database and need not be repeated for every new library. Thus, there are some approximations in the correspondence between oriented substituent pharmacophore similarities and enumerated product pharmacophore similarities. However, these errors are minor compared to the five advantages that the new method enables: large virtual library sizes, thorough conformational sampling, accounting for 1- to 9-point pharmacophores, creation of a Euclidean property space, and a reusable database of precomputed substituent values. © 2000 by Elsevier Science Inc.

Keywords: combinatorial library design, 3D pharmacophore, 3D similarity, oriented substituent, descriptor, molecular diversity, molecular similarity, combinatorial chemistry, multidimensional scaling

Corresponding author: Eric J. Martin, Chiron Corporation, LSC 4.250, 4560 Horton Street, Emeryville, CA 94608, USA. Tel.: 510-923-3306; fax: 510-923-3360. *E-mail address*: martine@chiron.com (E.J. Martin).

Journal of Molecular Graphics and Modelling 18, 383–403, 2000
© 2000 by Elsevier Science Inc.
655 Avenue of the Americas, New York, NY 10010

1093-3263/00/$–see front matter
PII S1093-3263(00)00064-4

BACKGROUND

The 2D Substituent-Based Library Design

Initial combinatorial library designs were based on computing a "property space" using 2D descriptors computed for the library's substituents.[1,2] A point in this property space, where proximity between the points reflected similarity between the substituents, represented each substituent. Designing diverse or focused libraries was thus mapped to the geometry problem of selecting sets of points that either efficiently covered the entire space, or that were focused in a small region of the space, or some combination of the two.

The quality of the library design is therefore a direct consequence of the quality of the property space, i.e., of the accuracy in which the relations between the points reflect biologically relevant relations between the library products. Two important trends have since been pursued to improve this correspondence: using 3D instead of 2D descriptors, and performing the calculations on the enumerated library products rather than on the substituents. The former assumes that ligand-protein interaction is a three-dimensional problem. Just as 3D QSAR and 3D database searching often have been effective improvements over 2D QSAR and 2D database similarity searching, so adding 3D similarity to combinatorial library design should work better than using only 2D similarity. The enumerated product-based (EPB) design approach argues that although substituent-based (SB) analysis takes advantage of the inherent structural similarities between all of the members of a combinatorial library, it implicitly assumes that diverse substituents generate diverse products. A design based on substituent properties does not explicitly account for interactions between the fragments in the assembled molecules, so assembling diverse substituents might not result in diverse products. Both of these extensions are computationally expensive, turning a relatively modest computational task into a much more formidable one.

The 3D N-Point Pharmacophore Descriptors

Numerous 3D descriptors have been offered for library design, including CoMFA topomeric fields, Compass surfaces, Hook-Space indices, and 3-point pharmacophore (3PP) or 4-point pharmacophore (4PP) fingerprints.[3–11] Each 3D method has its proponents, but the 3PP and 4PP fingerprint methods are supported by many past successes in 3D database searching. They are computationally tractable, are available in several commercial computational chemistry packages, and have been widely adopted by drug discovery groups. The molecule's functional groups are abstracted into a set of "features," such as H-bond donors, charges, and aromatic ring centers. The distances between features typically are binned to give a fixed number of possible pharmacophores, each of which is assigned to a bit in a bit-string or "fingerprint." For flexible molecules, the "molecule fingerprint" is the union across all conformations. A "library fingerprint" typically is taken as the union across a set of molecules.

Most pharmacophore-based diversity calculations have used 3PPs. However, although they are based on 3D structures, 3PPs themselves are only 2D objects. Furthermore, because a union typically is taken across all conformations, the ensemble of 3PPs does not contain the information necessary to infer the geometry of 4PPs or higher from sets of overlapping 3PPs. This is because the potentially overlapping 3PPs in the fingerprint may have come from different conformations and thus, would never be simultaneously present in any single conformation. Therefore, 4PPs have been introduced recently for diversity analysis.[12]

Even 4PPs however, are a great simplification given the number of interactions typically found in a ligand-protein interface. Kahn[13] found that although 4PPs give some enrichment over 3PPs in distinguishing actives from inactives in 3D database searching, 5PPs and 6PPs give successively higher enrichments. Some database search hypotheses have used 7, 8, and even 9PPs. This suggests that PPs are a good practical choice for 3D diversity analysis, but that higher PPs might be an improvement over mere 3 or 4PPs.

Union Fingerprints

Another consideration in the trade-off between fast and convenient 2D SB diversity analysis and the 3D or EPB improvements is the method of computing a property space and calculating the diversity for a subset of structures, i.e., given a set of structures and their corresponding PP fingerprints, how does one find a diverse subset and compute its diversity score? The most rigorous property space and diversity calculations have been amenable to the simple 2D SB descriptors, whereas less rigorous methods have been necessitated by the added complexity of 3D descriptors and EPB designs.

For a set of 2D or 3D fingerprints, where each feature within each structure is represented by a bit in a binary string, the simplest measure of diversity for a collection of molecules is simply to count the total number of features displayed by the entire library. The number of set bits in the library fingerprint, which is the union of all the individual bit-strings, gives total the number of features presented by the ensemble of molecules. This diversity measure was first employed using Daylight 2D substructure fingerprints, to compare the number of molecular substructures presented by libraries of biopolymers, peptoids, and small molecule collections.[2,14] Davies and Briant[7] later employed it with 3PP fingerprints to compare small molecule screening collections and to find small subsets of compounds that presented most of the possible pharmacophores. For 3 or 4PP fingerprints, the library union fingerprint includes all of the pharmacophores for all the conformations for all of the molecules in the library. This simple diversity measure is very quick to calculate for a large number of compounds. It requires minimal storage space even for large libraries, because the union can be accumulated as the fingerprints are computed. Thus, the fingerprints for each individual molecule need never be stored. Furthermore, a somewhat diverse subset can be found rapidly by selecting only those new molecules that set a large number of new bits in the accumulating fingerprint, until all or most of the bits are set. This subset may be far from optimal, because this is a one-pass order-dependent algorithm, but it avoids having to test the similarity of every pair and can even determine that most possible bits have already been set and quit before computing the 3PPs for all of the compounds. Due to these simplifications, it has become the standard for 3 and 4PP fingerprint diversity analysis of large libraries.

Property Space from Similarities

Although counting the set bits in a union fingerprint is computationally convenient, it is a single number taken over all conformations of all molecules in the accumulating library. It does not measure or optimize similarities or distances between the individual library members, or include any consideration of a property space and the geometry of how it is covered.

More sophisticated "distance-based" diversity measures do take into account the similarities between the individual compounds. Clustering methods simplify the problem by grouping proximal members that can then be treated as a single super member.[10] Algorithmic "spread designs" try to maximize the distances between the selected molecules to eliminate redundancy. "Coverage designs" try to minimize the distances between selected subset members and remaining unselected candidates to give good representation of the full candidate set.[15] These are difficult optimization problems. The sophistication of diverse subset selection can vary from very fast but crude one-pass Greedy algorithms to very thorough global simulated annealing optimizations.[4,16] Although these methods do take into account similarity between pairs of members, which can be calculated directly from the molecule 3PP fingerprints, they do not require the explicit calculation of a property space, i.e., they do not require Cartesian coordinates for the individual members.

The most sophisticated diversity calculations do require the calculation of an explicit property space with actual coordinates for each member. Analyzing the design in an explicit property space gives the dimensionality of the problem, how well all the dimensions of the space are covered, whether the selected points lie near the center or the edges of the space, and whether the selected points are well balanced or orthogonal. Both grid-based and covariance-based designs require an explicit property space. Covariance-based designs, like D-optimal design, are ideal for SB selection where the number of molecules is not much larger than the number of property space dimensions.[17] Grid-based designs are ideal for EPB selection where the number of molecules is much larger than the property space dimension. Unfortunately, comparing fingerprints only gives pairwise distances, not coordinates. Obtaining a Euclidean property space from pharmacophore fingerprints, in order to use these sophisticated sampling methods, requires some form of multidimensional scaling (MDS). This adds an additional compute intensive step.

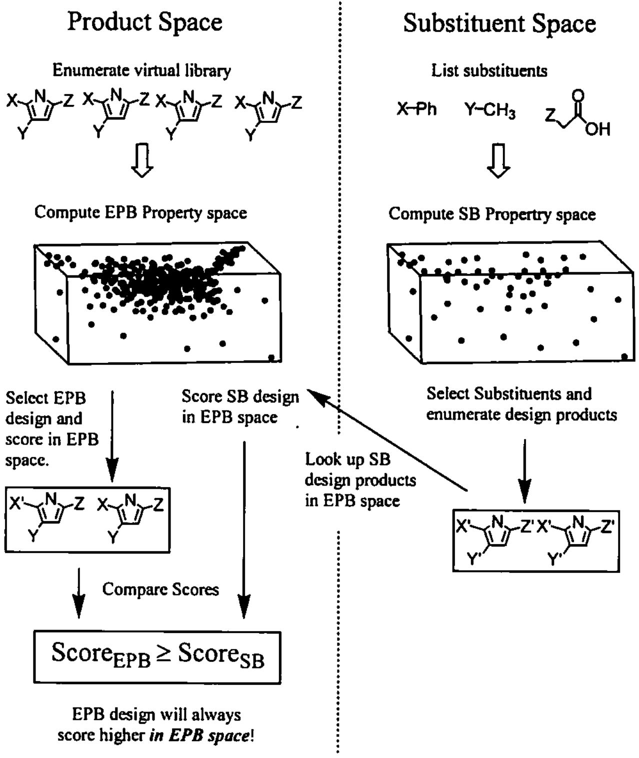

Product Space
Enumerate virtual library
X
N
Z
Y
Compute EPB Property space
Select EPB design and score in EPB space.
X'
Compare Scores
Score SB design in EPB space
$\text{Score}_{\text{EPB}} \geq \text{Score}_{\text{SB}}$
EPB design will always score higher *in EPB space*!
Substituent Space
List substituents
X–Ph
Y–CH$_3$
OH
Compute SB Propertry space
Select Substituents and enumerate design products
Look up SB design products in EPB space
Z'
Y'

Figure 1. Typical strategy for comparing substituent-based and enumerated product-based library design. First, the entire virtual library is enumerated, the most diverse combinatorial subset of products is found, and its score is recorded. Second, a property space is computed for all candidate substituents, a diverse subset of substituents is selected, and the products of just the selected subset of substituents are enumerated. Finally, the enumerated substituent-based products are scored in the product-based property space. The result has to be that the substituent-based design will score worse than the product-based design in the product-based space. This proves whether the designs are different, but not which design is better.

EPB Diversity Analysis

As mentioned earlier, several groups have recommended EPB rather than SB similarity analysis. This is an expensive proposal for three reasons: the number of products in an enumerated library can be vastly greater than the number of substituents; the number of conformations in each product can be vastly greater than those in the substituents; and, even if the same substituents are reused, the entire calculation must be repeated for each new scaffold.

Some groups have tried to quantify the improvement gained by the extra effort of performing the design on the enumerated products rather than on the substituents.[18-22] Figure 1 illustrates the typical approach. First, one computes a SB property space and finds the most diverse possible substituent subset by some measure of diversity in that space. Next, one computes the "same" property space for the enumerated products and finds the most diverse possible subset by the "same" measure of diversity. One additional complication is that the EPB design should be constrained to correspond to a combinatorial set, i.e., one that is made from all possible combinations of the substituents of which it is comprised. Finally, one enumerates products of the SB design and scores the SB library in the EPB space. The EPB design inevitably scores higher, because it was the optimal solution in the EPB space, and the difference is reported as the error due to SB selection. Some groups have found a significant improvement in going to EPB design, some have found no improvement, and some have found that it depends on the details of the property space.

One thing is always certain about the result of this approach. EPB selection always scores higher than SB selection. This foregone conclusion is guaranteed because EPB selection is, by definition, the highest scoring design in the EPB property

Table 1A. Default Daylight Tanimoto distance matrix for the tripeptoids and tetrapeptoids made from two building blocks: glycine, and the N-substituted analog of asparagine

NasnNasnNasnNasn	0																							
GlyNasnNasnNasn	0	0																						
NasnGlyNasnNasn	0	0	0																					
GlyGlyNasnNasn	0	0	0	0																				
NasnNasnGlyNasn	0	0	0	0	0																			
GlyNasnGlyNasn	0	0	0	0	0	0																		
NasnGlyGlyNasn	0	0	0	0	0	0	0																	
GlyGlyGlyNasn	0	0	0	0	0	0	0	0																
NasnNasnNasnGly	0	0	0	0	0	0	0	0	0															
GlyNasnNasnGly	0	0	0	0	0	0	0	0	0	0														
NasnGlyNasnGly	0	0	0	0	0	0	0	0	0	0	0													
GlyGlyNasnGly	0	0	0	0	0	0	0	0	0	0	0	0												
NasnNasnGlyGly	0	0	0	0	0	0	0	0	0	0	0	0	0											
GlyNasnGlyGly	0	0	0	0	0	0	0	0	0	0	0	0	0	0										
NasnGlyGlyGly	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0									
GlyGlyGlyGly	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.02	0								
NasnNasnNasn	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0.04	0							
GlyNasnNasn	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0.04	0	0						
NasnGlyNasn	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0.04	0	0	0					
GlyGlyNash	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0.04	0	0	0	0				
NasnNasnGly	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0.04	0	0	0	0	0			
GlyNasnGly	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0.04	0	0	0	0	0	0		
NasnGlyGly	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0	
GlyGlyGly	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.02	0	0.04	0.04	0.04	0.04	0.04	0.04	0.02	0

Table 1B. Substituent-based Euclidian distance matrix for the tripeptoids and tetrapeptoids made from two building blocks: glycine, and the N-substituted analog of asparagine

NasnNasnNasnNasn	0																							
GlyNasnNasnNasn	0.50	0																						
NasnGlyNasnNasn	0.50	0.71	0																					
GlyGlyNasnNasn	0.71	0.50	0.50	0																				
NasnNasnGlyNasn	0.50	0.71	0.71	0.87	0																			
GlyNasnGlyNasn	0.71	0.50	0.87	0.71	0.50	0																		
NasnGlyGlyNasn	0.71	0.87	0.50	0.71	0.50	0.71	0																	
GlyGlyGlyNasn	0.87	0.71	0.71	0.50	0.71	0.50	0.50	0																
NasnNasnNasnGly	0.50	0.71	0.71	0.87	0.71	0.87	0.87	1.00	0															
GlyNasnNasnGly	0.71	0.50	0.87	0.71	0.87	0.71	1.00	0.87	0.50	0														
NasnGyNasnGly	0.71	0.87	0.50	0.71	0.87	1.00	0.71	0.87	0.50	0.71	0													
GlyGlyNasnGly	0.87	0.71	0.71	0.50	1.00	0.87	0.87	0.71	0.71	0.50	0.50	0												
NasnNasnGlyGly	0.71	0.87	0.87	1.00	0.50	0.71	0.71	0.87	0.50	0.71	0.71	0.87	0											
GlyNasnGlyGly	0.87	0.71	1.00	0.87	0.71	0.50	0.87	0.71	0.71	0.50	0.87	0.71	0.50	0										
NasnGlyGlyGly	0.87	1.00	0.71	0.87	0.71	0.87	0.50	0.71	0.71	0.87	0.50	0.71	0.50	0.71	0									
GlyGlyGlyGly	1.00	0.87	0.87	0.71	0.87	0.71	0.71	0.50	0.87	0.71	0.71	0.50	0.71	0.50	0.50	0								
NasnNasnNasn																	0							
GlyNasnNasn																	0.58	0						
NasnGlyNasn																	0.58	0.82	0					
GlyGlyNasn																	0.82	0.58	0.58	0				
NasnNasnGly																	0.58	0.82	0.82	1.00	0			
GlyNasnGly																	0.82	0.58	1.00	0.82	0.58	0		
NasnGlyGly																	0.82	1.00	0.58	0.82	0.58	0.82	0	
GlyGlyGly																	1.00	0.82	0.82	0.58	0.82	0.58	0.58	0

space. Thus, any other selection method must produce a lower score.

INTRODUCTION

A Dubious Assumption

Comparing SB and EPB designs by scoring the SB design in the EPB space does indeed indicate whether SB and EPB selection is the same or different when evaluated in the EPB property space. However, the conclusion that is generally drawn is "This is the amount by which EPB selection is superior to SB selection." That conclusion requires an additional assumption, which seems so obvious that it has not been challenged. It assumes that the EPB calculation gives a better description of the library *products* than the corresponding SB calculation. That is, it implicitly assumes that applying the same property algorithm to the substituents and to the enumerated products produces two spaces that have the same "meaning," or at least that the EPB space has a more realistic correspondence to the biologically relevant features of the products. (i.e., not just the selection of molecules, but the very SB space computed on the right side of Figure 1 is inferior to the EPB space on the left side.) The SB selection is assumed to be a poor imitation of the EPB selection: an EPB selection "wannabe." This article challenges that seemingly obvious assumption.

Contrary evidence is that SB QSAR correlations for homologous series often are better than the corresponding whole-molecule–based correlations. Patterson et al.[23] described this phenomenon when they found that 2D fingerprint similarity for substituents correlated with biological activity consistently better than the corresponding whole-molecule descriptors across 20 QSAR data sets from the literature in near neighbor analysis. They pointed out that fragments in the substituents that also are present in the scaffold are always set, so they lose their ability to distinguish activity in the variable positions. The same is true for 3D pharmacophore fingerprints.

Table 1A presents a dramatic example of how this problem can plague EPB library design. It shows the default EPB Daylight Tanimoto distance matrix for the 16 tetrapeptoids, and eight tripeptoids made from glycine and the N-substituted analog of asparagine.[24] All pairs are computed to be nearly identical, with all 20 branched compounds computed to be exactly identical. The problem, of course, is that all of these compounds are made from different numbers and arrangements of glycine fragments. Because the default Daylight fingerprints enumerate paths only up to seven bonds long, and because they only record presence and absence of fragments, it is completely insensitive to the numbers and arrangements of the substituents.

Virtually any SB analysis would treat each position as a binary variable. The tetramer products are described by the concatenation of the four substituent spaces and thus would be described by a four-dimensional binary property space. The tripeptoids would be described by a three-dimensional binary space. Table 1B shows the Euclidean distance matrices based on these SB product descriptions. Experimental design based on these distance matrices typically would produce two minimal factorial designs: one for the tetramers and one for the trimers. This sensible design could pick all points that by the EPB space in Table 1A are identical. Clearly, the SB description of these peptoids is a better description of similarity in the products than is the corresponding product-based description. This is a clear (although admittedly contrived) violation of the usual "obvious" assumption about product-based calculations.

Although this is an extreme example, the point is not mute. Due to similarities among readily available reagents, combinatorial library products typically have many of the same fragments in several parts of the same molecule. Taking advantage of the common scaffold to break the molecule into substituents adds information. In the case of binary fingerprints, it adds information about where the fragments occur and about whether the same fragments occur simultaneously in different parts of the molecule. That is, it resolves the information spatially. Because 3D pharmacophore fingerprints are unions over all conformations, breaking out the scaffold and substituents not only adds spatial resolution, but also adds information about sets of distal 3PPs that can occur simultaneously in a single conformation rather than possibly having come from two different incompatible conformations, i.e., diversity libraries based on substituent 3PPs incorporate information on up to three features for each substituent, so libraries with three substituent positions describe each product molecule by a *triplet* of 3PPs. This is information about nine features that can be simultaneously present in a single conformation. It is somewhat akin to a 9PP analysis, although it is clearly missing information that orients the three 3PPs relative to each other. Nevertheless, based on Kahn's work mentioned earlier indicating the value of six or higher PP searches, the SB 3PP property space that includes some information about 9PPs might well be more relevant than the EPB 3PP analysis. Given this, one would be equally justified in rescoring the EPB design in the SB space and reporting that as the error in the EPB design!

Clearly, it begs the question of which design is better, to judge either method by testing it in the other's property space. The only fair way to compare the EPB and SB designs would be to come up with a third diversity measure which is more complete than either the SB or EPB analyses, and test both methods against that "gold standard." *E.g.*, since the 3PP SB design aspires to approximate a "poor man's" 9PP EPB design, one ideally might compute the best 3PP SB design and the best 3PP EPB design, then score them both in a 9PP EPB design space to determine which is better. Whichever design scored better in the 9PP EPB space would be the more diverse library.

Ideal 3D Pharmacophore Diversity Analysis

Based on the earlier analysis, the ideal 3D pharmacophore-based diversity analysis would compute all pharmacophores from 1 to 9 points for every conformation of each enumerated product of a virtual library from perhaps a thousand substituents for each of perhaps three positions of substitution. Conformation union fingerprints would be computed and stored from all conformations for each of the 10^9 individual product molecules. A complete $10^9 \times 10^9$ distance matrix would be calculated, and MDS would generate a Euclidean property space. Grid-based selection, constrained to produce only fully combinatorial subsets, would find the library that covers the entire space as completely, orthogonally, and nonredundantly as possible.

Unfortunately, the current state-of-the-art EPB pharmacophore diversity analysis computes a combinatorial subset that maximizes the number of bits in a library union fingerprint of all 1- to 4-point pharmacophores (1-4PPs) for libraries with three positions with only tens to hundreds of candidate substituents.

Table 2 shows that the difficulty in advancing from the current state of the art to the ideal is that every part of this problem scales badly. The simplest problem would be 1PP analysis of substituents, or an EPB library with just a single site of diversity, and a single rotatable bond per molecule. Assuming five feature types and a three-fold rotational barrier, there are only five pharmacophore types to identify, three conformations to compute, the fingerprint is only five bits long, there are only 1,000 candidate molecules, and there are no distances to measure. This is a simple calculation. However, there are 15 2PP types, 35 3PP types, etc., up to 715 9PP types. The number of distances to determine also grows with the number of features, so the fingerprint size grows dramatically. The problem also grows exponentially with the number of rotatable bonds per molecule and with the number of diversity sites on the scaffold. Furthermore, the number of distances in the distance matrix scales with the square of the number of candidate product molecules, and computing a property space by MDS scales roughly with the number of distances. The last column of Table 3 shows that it would require millions of years to compute the fingerprints for this hypothetical 1-9PP EPB library, and millions more to compute a Euclidean property space from the fingerprints. For that matter, Table 3 shows that it takes nearly as long for EPB 3PPs.

Oriented Substituent Pharmacophores

Given the impracticalities of achieving the ideal EPB 9PP similarity and MDS analysis, what might best be done to simplify the problem? Combinatorial libraries are nothing like random collections of molecules; they have a highly constrained intrinsic structure. Rather than ignoring that structure and treating the enumerated products as an arbitrary compound collection, it can be used to advantage. Recall from before that for a library with three positions of substitution, maximizing SB 3PP diversity has some kinship to maximizing EPB 9PP diversity. The obvious shortcoming is that two compounds whose substituents contain the same 3PPs typically would present them in different orientations, and, thus, the corresponding 9PPs could be completely different. Within a combinatorial library, if the SB fingerprints not only accounted for the 3PPs, but also for their orientations relative to the scaffold and therefore to each other, then two compounds whose substituents contained the same *oriented* substituent (OS) 3PPs would present the same 9PPs as well! This would reduce the very difficult problem of comparing many-featured pharmacophores on the huge number of very flexible enumerated products to the fairly simple problem of comparing *oriented* 3PPs on a small number of relatively rigid substituents. The first simple attempt to introduce some orienting information was to define a special site-of-attachment feature for the substituent atom nearest the template and compute the resulting 3PPs of each.[25] This anchored one point of a triangle, which contained only two "real" pharmacophore features, because one point was "wasted" on the site of attachment. It fixed the distance from the ends of the pharmacophore pairs to the scaffold, but the triangles were still free to pivot in all three dimensions around that anchor point without the method distinguishing between them. At best, they were only crudely oriented 2PPs.

Table 2. Every part of the pharmacophore problem scales badly: going from few features per pharmacophore to many, few rotatable bonds per molecule to many, and few sites of substitution per scaffold to many

N	Pharms[a]	FP bits[b]	Confs[c]	Prods[d]
1	5	5	3	1000
2	15	150	9	1E+06
3	35	35000	27	1E+09
4	70	7.0E+07	81	1E+12
5	126	1.3E+11	243	
6	210	2.1E+14	729	
7	330	3.3E+17	2187	
8	495	5.0E+20	6561	
9	715	7.2E+23	19683	

[a] Number of pharacophore types for five feature types and N features/pharmacophore.
[b] Number of bits in the fingerprint assuming 10 bins/distance.
[c] Number of conformations for N (threefold) rotatable bonds.
[d] Number of products for N diversity sites.

The work that follows describes an improved approach to defining and computing OS 1-3PPs and a corresponding Euclidean property space. The full process is illustrated in Figure 2. The template attachment point, "Tm," is added to each substituent smiles. Each substituent then is searched for all conformations. Each conformation is searched for all 1, 2, and 3PPs. An additional orienting point, "Pt," is placed 5 Å from Tm along the bond attaching the substituent to the scaffold. Additional distances are measured from Tm and Pt to each vertex of each of the 1-3PPs. This orients the pharmacophores to within 1 degree of freedom, rotation around Tm-Pt vector. The union fingerprint is taken for all pharmacophores in all conformations of each substituent. Computing the pairwise Tanimoto coefficients produces a distance matrix. MDS converts the distance matrix into a property space where proximity between points reflects 3D pharmacophore similarity between the corresponding substituents. Efficiently sampling this property space corresponds to choosing a pharmacophorically diverse library design. The article will argue that combinatorial libraries designed from sampling this SB property space yield a very reasonable experimental design for efficiently sampling the 1-9PPs of the virtual library of enumerated products.

METHODS

MOE

As mentioned earlier, PP analysis is already a well-established methodology. The goal is to augment ordinary 1, 2, and 3PPs to orient them relative to the scaffold, creating OS 1-3PPs. To minimize the programming burden, we started with MOE, a commercial molecular modeling package that already included a pharmacophore analysis module.[26] The MOE software is distributed with source code in the SVL language, so it is completely customizable. SVL is very compact and well suited to rapid prototyping. Our new module is called "Oriented Substituent Pharmacophore PRopErtY Space" (OSPPREYS).

Table 3. Breakout of estimated time to fingerprint and compute a Euclidean property space, using oriented substituent 3-point pharmacophores, enumerated product based 3-point pharmacophores, and enumerated product based 9-point pharmacophores, for a library with three diversity sites and 1,000 candidate substituents per site

	OS 1-3PP		EPB 3PP		EPB 9PP	
	N	Time units	N	Time units	N	Time units
Fingerprinting						
Conformations[a]	120	0.7 s/cnf	450	2 s/cnf	450	2 s/cnf
Fingerprint[b]	4E+09	0.1 s/cnf	4E+04	0.01 s/cnf	7E+23	3 s/cnf
Molecules[c]	14	11.2 s/mol	186000	4.3271 d/mol	186000	10.764 d/mol
Libraries[d]	3000	**9.333 hr/lib**	1E+09	**1E+07 yr/lib**	1E+09	**3E+07 yr/lib**
Property space						
Similarities[e]	4E+09	0.003 s/pair	4E+04	0.001 s/pair	7E+23	0.01 s/pair
Dist. matrix[f]	2E+06	1.25 h/lib	5E+17	2E+07 yr/lib	5E+17	2E+08 yr/lib
Property space[g]	2E+06	1.5 h/lib	5E+17	6E+07 yr/lib	5E+17	6E+07 yr/lib
Total		**12.08 hr/lib**		**8E+07 yr/lib**		**2E+08 yr/lib**

N is the number of objects in that step that accounts for why the times are different.

[a] Seconds to compute each conformation in Rubicon with MMFF minimization. N is the average MW. (Distance geometry scales with the square of the number of atoms.) Without minimization in MOE, the time decreases to 0.1 s/cnf for OS 1-3PP, reducing the time to fingerprint the library to 2.3 hr. Using Omega without minimization drops the time to 1.1 hr, roughly the time for fingerprinting alone.

[b] Seconds to compute the fingerprint each conformation. N is the numer of fingerprint bits.

[c] Time to repeat for N conformations. N is the mean number of conformations per molecule (or substituent), assuming four rotomers per bond.

[d] Time to repeat for N molecules, where N is number of molecules (or substituents) in the virtual library.

[e] Seconds to compute similarity between a pair of molecules. N is the number of virtual "bits."

[f] Time to compute distance matrix for all pairs of molecules. N is number of distances.

[g] Time to compute a property space from the distance matrix by MDS. N is number of distances.

[h] All times are on an SGI Origin 200 using a single R12000 CPU. The OS 3PP problem is three separate sets of 1,000 substituents each, so with three CPUs and 3 sets of licenses, the whole OS1-3PP computation takes about 4 hr.

The MOE pharmacophore module has some limitations for our purposes. Most obviously, it does not compute OS 1-3PPs. It does include the ability to define pharmacophore feature types, to search a single conformation of a molecule for features, and to create fingerprints for 2 and 3PPs. MOE also includes several algorithms for conformational searching, but it does not intrinsically include routines to take the union fingerprints for a flexible molecule across conformations. Rather than storing the actual fingerprint bit-string, MOE stores indices to the set bits of a virtual bit-string, with each set bit encoded as a single integer. This is a big advantage, because creating a property space requires that individual fingerprints be saved for each substituent, each of which has only a few set bits, rather than accumulating a single library union fingerprint which would contain a large fraction of set bits. MOE can compute a Tanimoto similarity matrix and can cluster for diversity analysis, but it will not export the distance matrix and does not perform MDS or algorithmic experimental designs. In addition, due to the six extra distances (see later), the number of OS 3PPs is too large to encode in a single integer as MOE usually does. We store these as an ordered triplet of integers. The Tanimoto similarity routines therefore needed to be modified to work with these triplets.

OS 1-3PP Fingerprints

Of course, the most important customization was to add the orientation information to each pharmacophore to make OS 1-3PP fingerprints. Makespace, an external Tcl/Daylight toolkit program for property space calculation,[25] supplies an input file where each substituent structure is a SMILES in which a Tm atom (for "template") has been added to mark where the first atom of the library scaffold attaches to the substituent. The user supplies OSPPREYS with a SMARTS query to identify the first substituent atom bound to Tm, called the "bearing" atom. Usually this is just [*][Tm], unless the substituent contains two atoms attached to Tm, as in a spyrofusion. OSPPREYS then places a dummy Pt (for "point") atom 5 Å along a vector from Tm through the bearing atom. Tm and Pt are reference points used to orient the substituent.

Figure 3 illustrates how the OS 1-3PPs are generated for an example substituent with three pharmacophoric features (2 H-bond acceptors and an aromatic ring center). Each of the three OS 1PPs consists of a feature type and two distances: from the feature to Tm and from the feature to Pt. Each of the three OS 2PPs consists of two feature types and five distances: the distance between the feature pair and the distances from each end to Tm and to Pt. The single OS 3PP consists of three feature types and nine distances: the three edges of the pharmacophore triangle, the three distances from each corner to Tm, and the three distances from each corner to Pt. Thus, the full OS 1-3PP for a single conformation of a substituent with three features sets seven virtual "bits" in the virtual OS 1-3PP fingerprint. They are encoded as three single-integer indices in an OS 1PP list for the three feature bits, three single-integer indices in the OS 2PP list for the three feature-pair bits, and one ordered triplet in the OS 3PP list to index the one OS 3PP bit. For each conformation, seven more virtual bits are computed and appended to the three growing lists (1, 2, and 3PP lists). The three lists then are sorted and duplicates removed, to yield the final OS 1-3PPs.

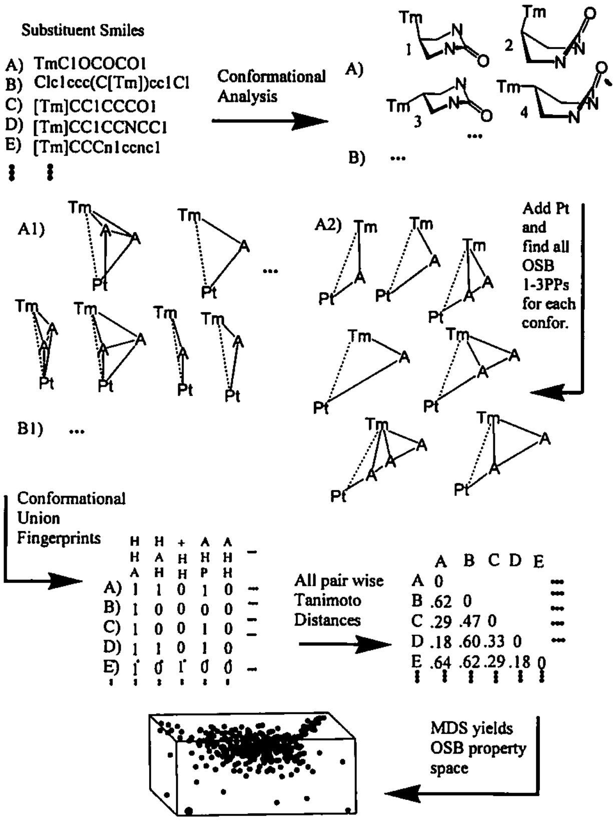


Figure 2. Schematic overview of the entire OSB 1-3PP property space calculation. Substituents are represented by smiles, where the template has been replaced by "Tm." Conformational analysis produces on average 14 conformations per molecule. Orienting point "Pt" is added, and OSB pharmacophores are identified in each conformation. The union fingerprint across all pharmacophores of all conformations produces a single bit-string "fingerprint" for each substituent. Computing all pairwise Tanimoto coefficients yields a distance matrix. MDS converts the distance matrix into a property space where proximity between points reflects 3D pharmacophore similarity between the corresponding substituents. Efficiently sampling this property space yields a pharmacophorically diverse library design.

A minor detail is that only 3N-6 distances are needed to define the OSPP geometries, where N is the number of pharmacophore features + 2 (for Tm and Pt). The Tm-Pt distance is already fixed at 5 Å, so for the respective OS 1PPs and OS 2PPs, an additional two or five distances are required, which are the numbers computed. For the OS 3PP this leaves eight distances, but nine are computed, so one is redundant. The shortest distance to Pt is not used.

Pharmacophore Parameters

MOE uses modifiable SMARTS rules to define the feature types and ionizable groups. Each substituent is typed in two passes. The first pass ionizes any acids or bases. The second pass assigns the feature types. We used six feature types: Hbd, Hba, +, −, aromatic center, and hydrophobic center. Distances were divided into 10 bins. This may seem like coarse sampling, but each substituent is small, and the OS 3PPs contain eight distances to match up.

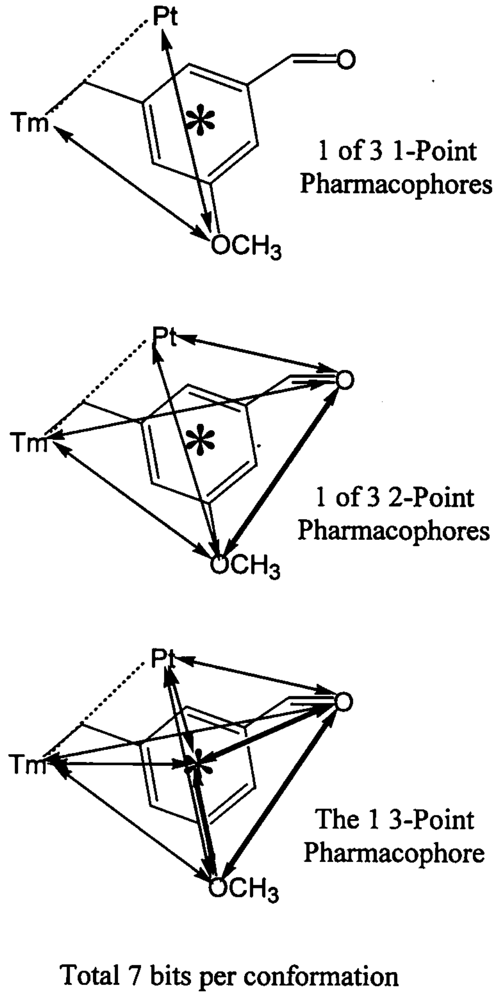

Pt
Tm
O
OCH3
1 of 3 1-Point Pharmacophores
1 of 3 2-Point Pharmacophores
The 1 3-Point Pharmacophore
Total 7 bits per conformation

Figure 3. Construction of oriented substituent 1, 2, and 3-point pharmacophores. Tm is the template atom to which the substituent is attached. Pt is placed 5 Å from Tm along the vector from Tm to the first substituent atom. Distances from each feature to Tm and Pt are added to the usual pharmacophore definitions. As shown, a substituent with three features typically would contribute a total of seven oriented pharmacophore "bits" to the fingerprint for each conformation.

Example Data Set

Four hundred ninety-seven amine substituents were chosen as an example candidate set for this study. The compounds were chosen to have molecular weight <175, ClogP <1.75, and most have two or fewer rotatable bonds, with an average of 1.2. These are strict filters, but are not unreasonable criteria for a combinatorial library with three diversity positions intended to produce orally available drugs. The frequency distributions of several relevant properties are given in Table 4. Generating 4^R conformations per substituent, where R is the number of rotat-

Table 4. Distributions of several properties for the 497 substituent test set

Values	−3	−2	−1	0	1	2	3	4	5	6	Avg/mol
Rot. bonds[a]	0	0	0	110	206	153	18	10	0	0	1.2
Features[b]	0	0	0	0	19	141	102	147	83	5	3.3
MW*30[c]	0	0	0	0	2	24	79	221	171	0	4.1
CLOGP[d]	21	65	83	162	166	0	0	0	0	0	−0.2

[a] Number of substituents with given number of rotatable bonds.

[b] Number of substituents with given number of pharmacophoric features.

[c] Distribution of molecular weights in increments of 30 Daltons.

[d] Distribution of calculated log P.

able bonds, the average number of conformations is 14.3. This is more than $4^{1.2}$ because the number of conformations scales exponentially with the number of rotatable bonds.

Conformational Search

Three conformational sampling methods of varying computational expense were studied: systematic search in MOE, random coordinates distance geometry in Rubicon,[27] and state-of-the-art rule-based search in Omega.[28,29] Minimizations also were performed in MOE using both the AMBER[30] and MMFF[31] force fields.

Similarity

The Tanimoto coefficient is a good definition of similarity between the pharmacophore fingerprint bit-strings because it ignores unset bits, and it is normalized to require more common bits as the number of pharmacophores increases. However, because distances are binned and bits are discrete, a small movement of a feature can change the Tanimoto similarity between two pharmacophores from identical to completely unrelated. Brown and Martin[32] addressed this problem by setting two bits whenever a distance was near a bin boundary. We chose to smooth the sharp edges by creating for each substituent, in addition to the ordinary OS 1-3PP fingerprints (FP), a "near neighbor" OS 1-3PP fingerprint (NNFP). The NNFP perturbs each OS 1-3PP pharmacophore by changing each of the distances, one at a time, by ±1 bin. Because each OS 1PP has two distances, the OS 1PP contribution to the NNFP sets about four times as many bits as the ordinary OS 1PP FP. Similarly, the OS2PP and OS 3PP contributions to the NNFP sets about 10 and 16 times as many bits as the ordinary OS2PP and OS 3PP, respectively. The overall similarity is now defined as:

$$\text{Similarity} = 0.65 \cdot \text{T(fp)} + 0.35 \cdot \text{T(nnfp)}$$

where T(fp) is the Tanimoto similarity between the ordinary FPs for two substituents and T(nnfp) is the Tanimoto similarity between the NNFPs. Similarity thus varies from 1 for identical FPs to 0 if there are no common bits either between the FPs or even the NNFPs. This "smears out" the FP bits, allowing some "partial overlap" to contribute to the similarity from "adjacent" pharmacophores.

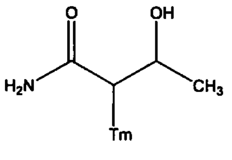
RESULTS

Computation Speed and Size

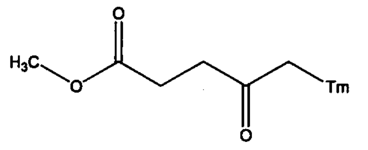
Table 3 breaks down the approximate CPU times, for a library with three sites of 1,000 candidate substituents each, required to compute each part of several PP property spaces: the sets of conformations, the pharmacophore fingerprints, the distance matrices, and the Euclidean property spaces. All of the steps in generating the fingerprints scale linearly with the number of conformations. Computing the distance matrix scales with the square of the number of substituents, and MDS scales slightly worse. The left section of the table is based on actual timings for OS1-3PP fingerprint similarities. These times assume that, on average, 14 conformations per substituent (as in our example data set) are generated in Rubicon, and minimized in MOE with the MMFF force field, the recommended protocol (see later). Without minimization, the time decreases to 0.1 s/cnf, lessening the time to fingerprint the library to 2.3 hours. Using Omega without minimization, by far the fastest method, drops the time to fingerprint the library 1.1 hours, roughly the time for fingerprinting alone. Systematic conformational search is unacceptably slow. All times are on an SGI Origin 200 using a single R12000 CPU. Because this OS 1-3PP problem is three separate problems of 1,000 substituents each, with three CPUs and three licenses for MOE and Rubicon, the whole computation takes 4 hours. The middle and right sections of the Table 3 are based on estimated times to compute EPB 1-3PP or EPB 1-9PP fingerprint similarities for the same library. They assume that the scaffold is completely rigid, but that a rotatable bond attaches each substituent. Given the distribution of substituents, therefore, the average product has 3 + 3*1.2 = 6.6 rotatable bonds. The average number of conformations in the enumerated library products is 186,000. This is much greater than $4^{6.6}$, because the number of conformations scales exponentially with the number of rotatable bonds, and some products have as many as 15 rotatable bonds. Notice that the EPB 3PP calculation is not estimated to be much faster than the EPB 9PP. This is because generating the conformations with minimization is the bottleneck when fingerprinting becomes faster. Using Omega to generate the conformations drops the time for EPB 3PPs by a factor of 50 to only 200,000 years.

Conformational Search

Two conformational tests were performed. The first test used a variety of search protocols on just two test substituents, structures I and II. Structure I has two rotatable bonds and four features: two H-bond donors and two H-bond acceptors. (The hydroxyl group is both a donor and acceptor.) Structure II has four rotatable bonds, and three features: all H-bond acceptors. (The methoxy is assumed to eclipse the carbonyl and does not move any pharmacophore features anyway.) Systematic search was performed at 30° and 15° increments on structure **I**, and 30° increments for structure **II**. For the stochastic distance geometry method, various numbers of conformers were generated to study the relationship between pharmacophore coverage and the number of rotatable bonds. Omega parameters were set at RMS cut-off of 0.1 Å and energy window of 100 kcal, values to encourage many conformations. The OS 1-3PP fingerprints were computed in each case, and similarities were determined to find which methods gave comparable results. Table 5 gives the number of fingerprint bits for each kind of pharmacophore, for 24 conformational search protocols, for structure I. Table 6 gives the Tanimoto distance matrix comparing all pairs of OS 1-3PP fingerprints from the 24 conformational search protocols. Tables 7 and 8 give the same corresponding information for structure II.



I



II

In the second test, OS 1-3PP Tanimoto distance matrices were generated for the test set of 497 substituents using seven different conformational sampling protocols: Omega without minimization, Rubicon without minimization generating 3^R, 4^R, or 5^R conformations per molecule, and the same Rubicon conformations minimized with the MMFF force field. Table 9A compares the seven conformational sampling protocols by taking the standard deviations of the differences between each pair of distance matrices. To appreciate the scale, comparing the 5^R minimized distance matrix to a matrix of ones gave a value of 0.136, so 0.07 is a big error. Table 9B–F gives the same information for subsets of the 497 substituents with 0 to

Table 5. Number of fingerprint bits set by several conformational search methods, with and without minimization for substituent I

	Confs[a]	1PP FP[b]	2PP FP	3PP FP	1PP NNFP[c]	2PP NNFP	3PP NNFP
3^2 dis. geom.[d]	9	4	7	5	12	50	48
4^2 dis. geom.	16	4	10	8	12	67	84
3^3 dis. geom.	27	4	8	9	12	57	84
4^3 dis. geom.	64	4	10	9	12	67	84
5^3 dis. geom.	125	4	10	9	12	67	84
30° sys. srch.[e]	37	4	11	17	12	70	168
15° sys. srch.	158	4	12	17	12	75	168
Rule-based[f]	12	2	3	4	8	24	36
3^2 dis. geom./AMBER[g]	9	4	7	7	12	50	84
4^2 dis. geom./AMBER	16	4	10	10	12	65	120
3^3 dis. geom./AMBER	27	4	10	10	12	65	120
4^3 dis. geom./AMBER	64	4	10	10	12	65	120
5^3 dis. geom./AMBER	125	4	10	10	12	65	120
30° sys. srch./AMBER	37	4	11	10	12	70	120
15° sys. srch./AMBER	**158**	**4**	**11**	**11**	**12**	**70**	**120**
Rule-based/ABER	12	4	10	7	12	66	72
3^2 dis. geom./MMFF[h]	9	4	7	7	12	52	84
4^2 dis. geom./MMFF	16	4	10	11	12	66	132
3^3 dis. geom./MMFF	27	4	10	12	12	66	144
4^3 dis. geom./MMFF	64	4	10	13	12	66	144
5^3 dis. geom./MMFF	125	4	10	12	12	66	144
30° sys. srch./MMFF	37	4	10	11	12	65	120
15° sys. srch./MMFF	**158**	**4**	**10**	**11**	**12**	**65**	**120**
Rule-based/MMFF	12	4	9	5	12	61	60

[a] Number of conformations generated.
[b] Number of 1-point pharmacophore primary fingerprint bits.
[c] Number of 1-point pharmacophore near neighbor fingerprint bits.
[d] Distance geometry stochastic search in Rubicon.
[e] Systematic search in MOE.
[f] Rule-based search in Omega.
[g] Distance geometry stochastic search in Rubicon followed by AMBER minimization in MOE.
[h] Distance geometry stochastic search in Rubicon followed by MMFF minimization in MOE.

Table 6. Distance matrix for several conformational search methods, with and without minimization, for substituent I

3^2 dis. geom.	0																							
4^2 dis. geom.	.29	0																						
3^3 dis. geom.	.25	.11	0																					
4^3 dis. geom.	.31	.03	.08	0																				
5^3 dis. geom.	.31	.03.	.08	0	0																			
30° sys. srch.	.6	.51	.54	.49	.49	0																		
15° sys. srch.	.61	.52	.55	.5	.5	.03	0																	
Rule-based	.82	.83	.83	.84	.84	.88	.88	0																
3^2 dis. geom./AMBER	.33	.42	.39	.43	.43	.43	.45	.84	0															
4^2 dis. geom./AMBER	.49	.44	.51	.45	.45	.24	.26	.87	.25	0														
3^3 dis. geom./AMBER	.49	.44	.51	.45	.45	.24	.26	.87	.25	0	0													
4^3 dis. geom./AMBER	.49	.44	.51	.45	.45	.24	.26	.87	.25	0	0	0												
5^3 dis. geom./AMBER	.49	.44	.51	.45	.45	.24	.26	.87	.25	0	0	0	0											
30° sys. srch./AMBER	.5	.46	.53	.47	.47	.21	.23	.88	.28	.03	.03	.03	.03	0										
15° sys. srch./AMBER	**.51**	**.47**	**.51**	**.45**	**.45**	**.19**	**.21**	**.88**	**.3**	**.06**	**.06**	**.06**	**.06**	**.03**	**0**									
Rule-based/AMBER	.6	.54	.58	.52	.52	.36	.38	.89	.36	.27	.27	.27	.27	.24	.22	0								
3^2 dis. geom./MMFF	.46	.53	.56	.54	.54	.43	.44	.88	.45	.39	.39	.39	.39	.35	.37	.46	0							
4^2 dis. geom./MMFF	.51	.46	.53	.48	.48	.2	.22	.88	.36	.19	.19	.19	.19	.16	.18	.37	.36	0						
3^3 dis. geom./MMFF	.53	.48	.55	.49	.49	.16	.18	.88	.39	.22	.22	.22	.22	.19	.21	.4	.32	.04	0					
4^3 dis. geom./MMFF	.54	.49	.56	.5	.5	.14	.16	.88	.4	.24	.24	.24	.24	.21	.23	.41	.33	.07	.02	0				
5^3 dis. geom./MMFF	.53	.48	.55	.49	.49	.16	.18	.88	.39	.22	.22	.22	.22	.19	.21	.4	.32	.04	0	.02	0			
30° sys. srch./MMFF	.5	.45	.52	.47	.47	.22	.24	.88	.27	.03	.03	.03	.03	.06	.08	.29	.4	.21	.24	.22	.24	0		
15° sys. srch./MMFF	**.5**	**.45**	**.52**	**.47**	**.47**	**.22**	**.24**	**.88**	**.27**	**.03**	**.03**	**.03**	**.03**	**.06**	**.08**	**.29**	**.4**	**.21**	**.24**	**.22**	**.24**	**0**	**0**	
Rule-based/MMFF	.62	.56	.63	.57	.57	.45	.46	.88	.37	.28	.28	.28	.28	.3	.32	.13	.53	.43	.46	.47	.46	.3.	.3	0

15° systematic search is assumed to be most reliable.

4 rotatable bonds. The sizes of these subsets are available from row 1 of Table 4.

MDS

The distance matrices from Table 9 were converted to a property space using nonlinear MDS. In all cases, eight dimensions were required to reproduce the distance matrix with a relative standard deviation of 10%, except with Omega, which required only seven. For comparison, the distance matrix for the same compounds for Daylight 2D fingerprint similarity required six dimensions, and 2D "atom-layer" similarities required only five dimensions.[2]

Three tests were performed to ascertain whether the information in the 3D pharmacophore fingerprints was already contained in the Daylight Tanimoto fingerprints or the atom layer tables. The correlation matrix showed that the first 1-3 PP dimension correlated with the first Daylight fingerprint dimension with $r^2 = 0.77$ and with the first atom layer dimension with $r^2 = 0.74$. The second 1-3 PP dimension correlated with the third Daylight fingerprint dimension with $r^2 = 0.28$ and with the fifth atom layer dimension with $r^2 = 0.23$. All other pairwise correlations were negligible. Besides those same correlated terms, all other variance inflation factors were <2, showing no multicollinearities with the Daylight and atom layer spaces either. Finally, when all three spaces were combined, the largest principal component explained only 15% of the variance, and 17 of 19 principal components were required to reproduce 99% of the variance. Thus, the 1-3 PP analysis adds new information that was not captured in the 2D analysis.

DISCUSSION

As suggested in the introduction, the proper standard to compare OS 1-3 PP diversity and EPB 1-3PP diversity would be to compare each to EPB 1-9 PP diversity, which should be a better diversity measure than either. Unfortunately, the latter is currently impossible and is unlikely to become possible in the near future. Even the EPB 4PP diversity is only possible for modest-sized libraries, unless serious compromises are made in the conformational sampling. Yet, one can still make some theoretical comparisons between the feasible methods and the hypothetical standard.

OS 1-3PPs *vs* EPB 1-9PPs

Figure 4 shows how the extra distances to Tm and Pt orient the pharmacophores, relative to the template, and thus to the other substituents in an assembled product molecule, but only to within rotation about the Tm-Pt vectors. The OS 1-3PP approach thus requires an implicit "combinatorial conformer assumption," i.e., that the same template and substituent conformations are available irrespective of the particular combination of substituents, and that the same rotomers around the Tm-Pt bond vector are available for each pharmacophore. The pharmacophores are not overlapping, so each conformation of each

Table 7. Number of fingerprint bits set by several conformational search methods, with and without minimization, for substituent II

	Confs	1PP FP	2PP FP	3PP FP	1PP NNFP	2PP NNFP	3PP NNFP
3° dis. geom.	81	3	14	11	8	75	66
4^4 dis. geom.	256	3	15	13	8	78	78
5^4 dis. geom.	625	3	16	15	8	81	90
6^4 dis. geom.	1296	3	16	13	8	83	78
Rule-based	100	3	11	18	8	62	107
30° sys. srch.	5822	3	18	38	8	90	228
3^4 dis. geom./AMBER	81	3	10	9	8	55	54
4^4 dis. geom./AMBER	256	3	10	9	8	55	54
5^4 dis. geom./AMBER	625	3	11	9	8	60	54
6^4 dis. geom./AMBER	1296	3	11	9	8	60	54
Rule-based/AMBER	100	3	9	6	8	51	36
30° sys. srch./AMBER	**5822**	**3**	**15**	**23**	**8**	**77**	**138**
3^4 dis. geom./MMFF	81	3	11	5	8	59	30
4^4 dis. geom./MMFF	256	3	11	5	8	59	30
5^4 dis. geom/MMFF	625	3	11	5	8	59	30
6^4 dis. geom./MMFF	1296	3	11	5	8	59	30
Rule-based/MMFF	100	3	9	5	8	51	30
30° sys. srch./MMFF	**5822**	**3**	**15**	**23**	**8**	**77**	**138**

Table 8. Distance matrix for several conformational search methods, with and without minimization, for substituent II

3^4 dis. geom.	0																	
4^4 dis. geom.	.16	0																
5^4 dis. geom.	.23	.14	0															
6^4 dis. geom.	.12	.2	.17	0														
Rule-based	.57	.53	.5	.54	0													
30° sys. srch.	.53	.48	.43	.47	.51	0												
3^4 dis. geom./AMBER	.48	.44	.49	.5	.5	.63	0											
4^4 dis. geom./AMBER	.48	.44	.49	.5	.5	.63	0	0										
5^4 dis. geom./AMBER	.45	.41	.46	.47	.51	.62	.04	.04	0									
6^4 dis. geom./AMBER	.45	.41	.46	.47	.51	.62	.04	.04	0	0								
Rule-based/AMBER	.55	.56	.56	.57	.57	.7	.18	.18	.22	.22	0							
30° sys. srch./AMBER	**.4**	**.41**	**.34**	**.38**	**.41**	**.34**	**.47**	**.47**	**.44**	**.44**	**.57**	**0**						
3^4 dis. geom./MMFF	.53	.53	.49	.55	.59	.69	.49	.49	.51	.51	.47	.58	0					
4^4 dis. geom./MMFF	.53	.53	.49	.55	.59	.69	.49	.49	.51	.51	.47	.58	.0	.0				
5^4 dis. geom./MMFF	.53	.53	.49	.55	.59	.69	.49	.49	.51	.51	.47	.58	0	0	0			
6^4 dis. geom./MMFF	.53	.53	.49	.55	.59	.69	.49	.49	.51	.51	.47	.58	0	0	0	0		
Rule-based/MMFF	.59	.59	.55	.6	.64	.72	.45	.45	.47	.47	.42	.63	.1	.1	.1	.1	0	
30° sys. srch./MMFF	**.47**	**.41**	**.38**	**.45**	**.45**	**.34**	**.51**	**.51**	**.51**	**.51**	**.57**	**.26**	**.55**	**.55**	**.55**	**.55**	**.59**	**0**

30° systematic search is assumed to be most reliable.

substituent can be simultaneously presented with each conformation of each other substituent. Thus, each of the ensemble of available 1-9PPs of each product can be represented either by a single EBP 1-9PP fingerprint, or (approximately) by a set of three OS 1-3PP fingerprints. (e.g., assume the template has two conformations, and each of the three substituent sites has a three-fold rotational barrier around the Tm-Pt bond.) For each triplet of 3PPs presented by any set of three substituents, 2*3*3*3 = 54 corresponding 9PPs will be presented by 54 corresponding combinatorial conformations of the associated product. The hypothetical set of all 1-9PPs that are implied in each product of a combinatorial library from all conformational combinations of all of the OS 1-3PPs will be called the "combinatorial 1-9PPs" (C1-9PPs). The differences between the set

Table 9. Comparison OS 1-3PP Tanimoto distance matrices for the test set of 497 substituents using seven different conformational sampling protocols: Omega without minimization, Rubicon without minimization generating 3^R, 4^R, or 5^R conformations per molecule, and the same sets of Rubicon conformations followed by minimization with the MMFF force field

	A			All				B			O rotatable bonds			
3^R Rubicon	0							0						
4^R Rubicon	.03	0						.01	0					
5^R Rubicon	.03	.02	0					.03	.02	0				
3^R Rubicon/MMFF	.08	.08	.08	0				.08	.08	.08	0			
4^R Rubicon/MMFF	.08	.07	.07	.03	0			.08	.08	.08	.02	0		
5^R Rubicon/MMFF	.08	.07	.07	.03	.02	0		.08	.08	.08	.03	.03	0	
Omega	.12	.12	.12	.11	.11	.11	0	.19	.19	.19	.21	.21	.21	0

	C			1 rotatable bond				D			2 rotatable bonds			
3^R Rubicon	0							0						
4^R Rubicon	.04	0						.04	0					
5^R Rubicon	.04	.02	0					.04	.02	0				
3^R Rubicon/MMFF	.09	.09	.09	0				.07	.07	.07	0			
4^R Rubicon/MMFF	.09	.08	.09	.04	0			.07	.07	.07	.04	0		
5^R Rubicon/MMFF	.09	.09	.08	.03	.02	0		.08	.07	.07	.04	.02	0	
Omega	.16	.16	.16	.17	.17	.17	0	.18	.18	.18	.18	.18	.18	0

	E			3 rotatable bonds				F			4 rotatable bonds			
3^R Rubicon	0							0						
4^R Rubicon	.01	0						.03	0					
5^R Rubicon	.01	.01	0					.04	.04	0				
3^R Rubicon/MMFF	.07	.07	.07	0				.09	.1	.09	0			
4^R Rubicon/MMFF	.07	.07	.07	.03	0			.09	.09	.09	.02	0		
5^R Rubicon/MMFF	.07	.07	.07	.03	.01	0		.09	.09	.09	.02	.02	0	
Omega	.29	.29	.29	.29	.29	.29	0	.39	.39	.38	.37	.37	.37	0

Table 9A compares the seven conformational sampling photocols by taking the standard deviations of the differences between each pair of distance matrices. Table 9B–F gives the same information for subsets of the 497 substituents with from 0 to 4 rotatable bonds, respectively. The maximum error, using a matrix of all ones, is 0.136.

of C1-9PPs and the true EPB 1-9PPs, which will be examined later, are errors in the method. Nevertheless, the claim of this article is that the C1-9PPs are still a better approximation of the true 1-9PPs than are the EPB 1-3PPs. Hence, similarities based on the OS 1-3PP calculation are better approximations of the library product similarities than are those from the corresponding EPB 1-3PP, even if the EPB 1-3PP calculation could somehow be performed with adequate conformation sampling. That is, the SB diversity calculation not only is faster, it also is better.

Errors in the combinatorial conformer assumption will lead to errors in the corresponding computed similarities, i.e., if two compounds, with a common scaffold and three sites of diversity, have the same set of OS 1-3PPs, but some of the pharmacophore rotomers around the Tm-Pt bond vector in one molecule are not available in the other, then the true EPB 1-9PP similarity will be lower than implied by the OS 1-3PPs. This is a failure of the combinatorial conformer assumption and is probably the biggest source of error in the method. It could be largely avoided, if so desired, by using template extended substituents (see later).

Similarly, all of the EPB 1-9PPs shared by two enumerated product molecules in the same scaffold orientation will be reflected by corresponding OS 1-3PPs in the SB analysis. However, there also might be additional 1-9PP overlaps in other orientations where the scaffolds do not align, which will not be reflected in the OS 1-3PPs. For example, if two product molecules with three sites of diversity have no common OS 1-3PPs, then the corresponding two products will have no common 1-9PPs in the same scaffold orientation. However, there could coincidentally be other orientations of the products that would overlap some pharmacophores. Thus, the products from a diverse set of oriented substituents can be more similar than the SB analysis would imply, even if the combinatorial conformer assumption holds completely. These are failures of the template alignment assumption and cannot be repaired by using template extended substituents. Template aligned pharmacophore matches probably reflect biological similarity better than template aligned matches. Experience has shown that, in structure-based drug design, occasionally two compounds from the same congeneric series bind to the same target in different orientations. However, this failure only applies if the two compounds with the same template bind in different binding modes, but still use the same set of protein target

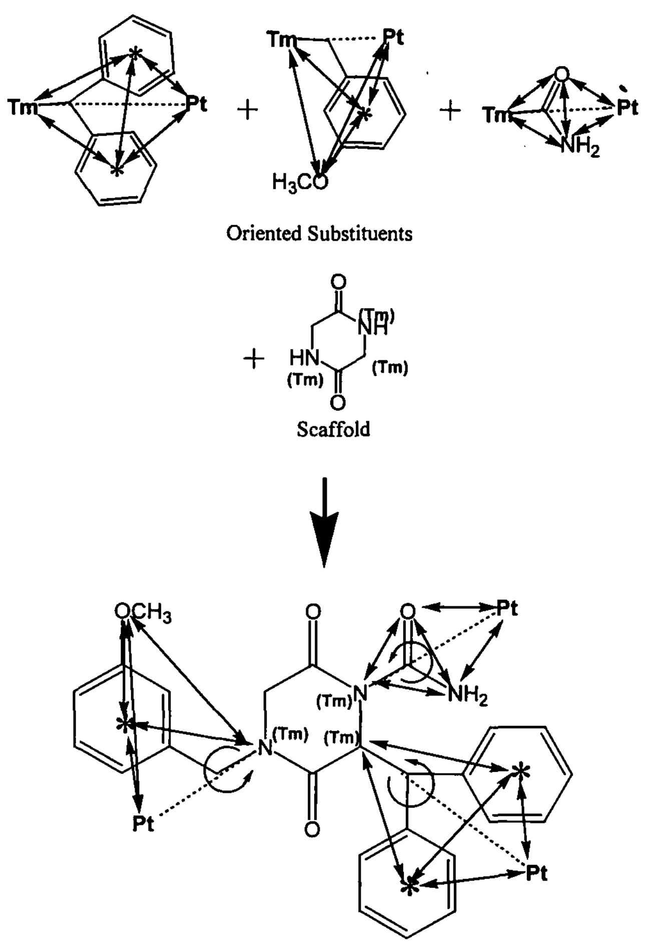


Figure 4. Assembling the substituents around the scaffold by overlapping the Tm atoms and Tm-Pt bond vector orients the substituents with respect to the scaffold and to each other, but only to within rotation around the Tm-Pt bond vector.

features, so that the EPB 1-9PP analysis would have detected the similarity. This is a less common occurrence.

One might try to argue that a similar analysis could be made from the sum of overlapping 3 or 4PPs in an EPB 1-4PP analysis. It is true that if a 9PP is present, then a corresponding ensemble of overlapping EPB 3 and 4PPs likewise must be present. However,in this case, the converse is not true. Recall that the EPB pharmacophore fingerprint is a union taken across all conformations. Unlike the nonoverlapping OS 1-3PPs, many of the overlapping 3 and 4PPs are likely to have come from different conformations of the product, such that no single conformation of the product presents them all simultaneously. Thus, two products that happened to have the exact same set of EPB 1-4 PP fingerprints would likely present very different sets of EPB 5-9PPs, depending on which subsets of bits came from each conformation.

Note that not all EPB 4-9PPs will be represented among the C4-9PPs implied by the oriented substituents, but only those that take at most three features from any one substituent, e.g., a 5PP that takes four points from one substituent, one from a second, and none from a third will not be represented. These "lop-sided" pharmacophores, which put most of the features in a single substituent, seem less interesting than those that cover the ligand with features. Still, this is a limitation of the method.

The combinatorial conformer assumption, that the same template conformations and Tm-Pt rotomers are available to all pharmacophores, obviously is best for a mainly rigid template that holds the substituents well spread out with a freely rotating attachment. For a very flexible, crowded template, the assumption will be worse and will introduce more error into the diversity analysis. Errors due to the combinatorial conformer assumption, then, equate to errors in conformational sampling of the products. Even so, this assumption probably is not bad compared to other errors introduced due to the practical difficulties of conformational sampling in general. In our opinion, the advantage of having vastly fewer conformations to explore, and therefore the luxury of careful conformational analysis, more than outweighs the problems due to missed similarities from matches that do not align the templates, or from failures in the combinatorial conformer assumption.

Experimental Design on the Products

Even granting the argument that similarities between products from EPB 1-9PPs are well reflected by similarities between the corresponding OS 1-3PPs, one still must ask the additional question of whether a "diverse" selection of substituents should lead to a "diverse" set of products. For purposes of screening, "diverse" means that the selected subset of products, assembled from the chosen substituents, forms a set that is not redundant, and that well covers all the dimensions of the hypothetical EPB 1-9PP property space. In other words, does a sound experimental design for the substituents generate a sound experimental design for the products? As shown in Figure 5, the presence of the common scaffold allows this product-based experimental design problem to be factored into within-substituent and between-substituent variability. Just making a full combinato-

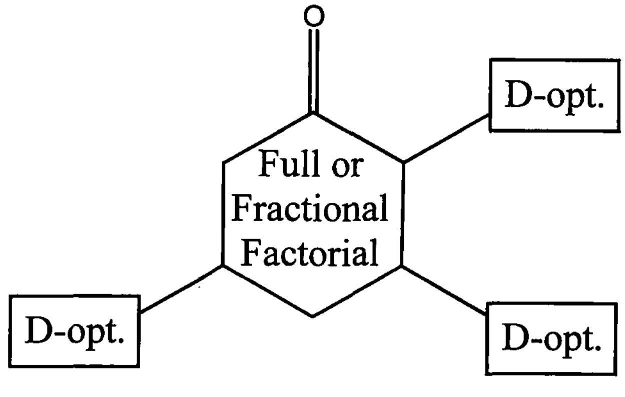


Figure 5. Making a full combinatorial library automatically samples between-substituent diversity by full-factorial design. The irregular distribution of within-substituent variability is best sampled by an algorithmic method such as D-optimal design. Thus, simply sampling the OSB 1-3PP property space by D-optimal design, along with synthesizing the full combinatorial library, produces a very reasonable experimental design to sample the 1-9PPs in the assembled products.

rial library automatically samples between-substituent diversity by full-factorial design, which is a sound classic design. The points in property space are irregularly distributed, so a classic design cannot be applied to the substituents. Therefore, within-substituent variability is best sampled by an algorithmic method. D-optimal design is a widely accepted algorithmic method that ensures that the points are well spread out *and* that they cover all the dimensions of property space. It is an excellent choice to sample within substituent variability where the number of substituents is not much larger than the dimensionality of the property space. Thus, simply sampling the OSB1-3PP property spaces by D-optimal design, along with synthesizing the full combinatorial library, produces a very reasonable experimental design for the 1-9PPs in the products.

If one wanted to make fewer compounds than a full combinatorial library (or to sample more substituents for the same number of synthesized products), one could still use an SB experimental design by using D-optimal design within the substituent sets and a fractional factorial design between the substituent positions. This approach not only produces a well-balanced experimental design, it also yields systematically arranged combinations of reagents that are easy to keep track of in either manual or automated parallel synthesis.

If one insists on making a noncombinatorial "cherry-picked" subset of individual products, still one can (and should) use the OSB 1-3PP calculation rather than the EPB 3PP calculation. An important premise of this article has been that the OSB 1-3PP library description characterizes each individual product. Thus, one can calculate pharmacophores for the substituents, but still perform selection on the products. With three diversity sites, each product molecule can be described either by a concatenation of three substituent fingerprints, or of three substituent property vectors, as was done for the peptoid example in Table 1B, e.g., for a library of three diversity sites with 1,000 candidate substituents each, one can perform the OSB 1-3PP calculation, exactly as in the left side of Table 3, which requires only 9 CPU hours. The OSB 1-3PP fingerprints are concatenated, for each of the 10^9 triplets of substituents, to get the "C1-9PP" fingerprints of the products (see earlier). This takes only an additional 5 hours (plus 4 more hours if one writes a file), rather than the 10^7 years required for full EPB 3PP calculation shown in the middle section of Table 3. As argued before, despite the enormous time savings, the C1-9PPs should still be a better approximation of the true EPB 1-9PPs than are the EPB 3PPs. Unions, intersections, and Tanimoto similarities can be computed from these C1-9PP fingerprints, just as with EPB fingerprints. Given these 10^9 fingerprints, one could apply simulated annealing to try to evolve the subset with the most set bits in the library union fingerprint (or greatest maximin distance, or other diversity criterion) from this enormous virtual library. (In practice, it is much easier to select triplets of OSB 1-3PPs and concatenate them "on the fly" to make new C1-9PPs as needed, rather than to prestore 10^9 values for lookup.) Performing the Monte Carlo perturbations by replacing random products evolves a noncombinatorial library. Performing the perturbations by choosing random substituents and replacing an entire "row" of products yields a combinatorial library. The only difficulty is that there are C(10^9, 10^4) possible "cherry-picked" libraries of 10,000 products, a daunting search space even with the simplest diversity measures and most powerful optimizers. It might be very hard to find a near optimal solution in a reasonable time.

A Euclidean product property space can be likewise obtained from combining the substituent spaces, which could be sampled by grid selection. MDS is performed on the substituent distance matrices to obtain the several substituent property spaces (an additional 3 hours for the example in Table 3). Concatenating the three substituent property data matrices creates the product property space, i.e., if substituents A, B, and C each have 3D substituent property vectors (A1, A2, A3), (B1, B2, B3), and (C1, C2, C3), respectively, then the corresponding product with those three substituents is described by the 9D property vector (A1,A2,A3,B1,B2,B3,C1,C2,C3). The Euclidean product property space, however, typically will be of many more than nine dimensions. The 497-substituent example from this article required eight dimensions to reproduce the distance matrix with a relative standard deviation of 10% (see earlier). A library that used this candidate set at each of three diversity sites would give a 24D property space. This space could not be sampled on a cubic lattice, because a 24D cube has 2^{24} = 16,000,000 corners alone. However, it might be feasible with more sophisticated lattice geometries.[33] The very effective 24D Leech lattice is almost 10^6 times more efficient than the cubic lattice, and a large fraction of cells will be empty.[34] Even so, the large numbers of coarse cells are discouraging. Of course, MDS can be forced to embed into fewer Euclidean dimensions, but only with a corresponding loss of fidelity.

Despite these difficulties of scale, the combination of SB pharmacophore calculations, coupled with product-based selection, is perhaps the most feasible way to perform pharmacophore diversity selection of noncombinatorial subsets for large virtual libraries. We still prefer, instead, to design full or fractional factorial libraries from diverse substituent sets.

Conformational Sampling

Conformational sampling is probably the biggest source of error in 3D pharmacophore analysis. The pharmacophore fingerprint should contain all pharmacophores presented by all accessible conformations of the molecule. If the conformational sampling misses accessible conformations or includes inaccessible ones, these will introduce corresponding errors in the pharmacophore similarities. First, there is the fundamental problem of recognizing which are the accessible conformations in complex biological environments when we usually must perform the calculations in vacuum, or at best in a very simplified polarizable continuum. This is one of the most important unmet needs of both experimental and computational physical organic chemistry. Throughout the discussion following, it is important to remember that all comparisons are limited by the inadequacies of even the best of the methods employed. It is important, however, not to confuse these fundamental difficulties with a license to do a "quick and dirty" conformational search, arguing speciously that if one cannot do the most desirable calculation, then all other calculations are somehow equally good. Whatever solutions are found to this complex problem, they undoubtedly will require even more computer power and more sophisticated algorithms, not less. Granting these fundamental limitations in current conformational searching, one must still work to ensure that one adequately samples the relevant conformations as thoroughly as possible within the limitations of present force fields.

Pharmacophore fingerprint analysis at least has the advantage that it is easy to define adequate sampling. Sampling in this context is complete when the generation of additional conformations ceases to generate additional pharmacophores, even if new geometries are still being produced. This is both because pharmacophores are discrete and because some atoms do not carry features. For example, a spinning phenyl substituent does not move its centroid. Adequate sampling, of course, will depend on the bin size, which determines the resolution of the fingerprint. Table 5 shows the number of fingerprint bits set for several different conformational sampling schemes on structure I, which has four features and two rotatable bonds. Differences due to the number of conformations within the same method are less than differences between methods. With Rubicon, $3^2 = 9$ conformations are not sufficient, but $4^2 = 16$ gave the same result as $5^3 = 125$, suggesting 4^R sampling is sufficient. However, 27 conformations missed some pharmacophores, indicating 4^R sampling may be optimistic. Without minimization, systematic search produced the most pharmacophores, and rule-based search produced the fewest. After minimization, discrepancies are less between the methods. The systematic search at 15° increments followed by minimization is assumed to be the most thorough and serves as our "standard." It yielded 38 pharmacophores in FP with both AMBER and MMFF, but AMBER minimization set 202 bits in NNFP whereas MMFF set 209. Rule-based search with minimization yielded substantially fewer bits, indicating that sampling is less complete. Minimization with either force field decreased the number of bits from systematic search, implying that it produced high-energy conformations, which collapse into a smaller number of common minima. Curiously, minimization increases the number of pharmacophores for both distance geometry and rule-based search. This might indicate multiple conformations that are similar enough to set a single bit relaxing into more distinct structures that set different bits.

If all conformational search protocols yielded the same pharmacophores, the pairwise Tanimoto distances between methods would all be 0. Table 5 shows that, without minimization, the fingerprints are similar within methods, regardless of the number of conformers, but are very different between methods. Again, 3^2 are too few conformations, but 4^2 and 5^3 gave virtually the same result. After minimization with AMBER, all of the Rubicon results with 4^2 or more conformations gave identical results and were virtually identical to the systematic search "standard," with rule-based search becoming more like the standard, but still missing a significant number of conformations. This indicates that Rubicon followed by minimization in AMBER is an excellent protocol for compounds where the force field applies. Minimization with MMFF brings all of the Rubicon results larger than 4^2 conformations into agreement, but with less agreement between Rubicon and systematic search. Systematic search minimized with MMFF gives virtually the same final fingerprints, as both systematic search and Rubicon search minimized with AMBER, indicating excellent agreement between the two force fields. However, the number of set bits is comparable for Rubicon minimized under either force field, indicating that Rubicon minimized with MMFF misses some conformations, but also finds some peculiar ones. The unfortunate overall conclusion is that minimization appears to be necessary to get results comparable to the standard. Given that, it is very welcome that in Rubicon, generating only 4^2 conformations is sufficient for two rotatable bonds.

Tables 7 and 8 are analogous to Tables 5 and 6, for the more flexible substituent II with four rotatable bonds but only three features. The "standard" in this case is systematic search at 30° increments. This was adequate for substituent I, and 15° increments generated over 90,000 conformations, which was an impractical number to minimize and fingerprint. Within each method, the results are similar to those for substituent I. The number of pharmacophores found by Rubicon without minimization varies over a small range with the number of conformations, but following minimization the numbers drop to a common set. This result is echoed in the similarity matrix. Systematic search without minimization generates a large number of pharmacophores, which is substantially reduced by minimization. For systematic search, both force fields gave the exact same number of pharmacophores, but this is purely coincidence, because the distance between them is 0.26, a nonnegligible discrepancy. Rule-based search produced 100 conformations, with a large number of pharmacophores before minimization, but the number after minimization was the least among all the methods. Unlike structure I, even after minimization, none of the faster methods found most of the pharmacophores produced by minimized systematic search. This appears to be mainly due to missing 3PPs. Furthermore, even systematic search produces different results under the two force fields for these rather ordinary structures. The unwelcome conclusion is that minimization is essential, but even with minimization, neither extensive stochastic search nor state-of-the-art rule-based search is sufficient to find all pharmacophores with the more flexible substituents. Substantial errors in the similarities between the substituents should be expected with any of these search methods. This problem would likely be compounded with enumerated products having many more rotatable bonds.

Table 3 shows that minimization is the computational bottleneck to fingerprinting, so it would be helpful to find an adequate method that does not require it. Table 6 showed that no method without minimization reproduced the fingerprints of the methods that used minimization for a single substituent. However, the important requirement for diversity design is to get the same Tanimoto distance matrix between molecules as the most thorough method. It is conceivable that systematic errors might cancel, and methods without minimization might produce the same similarities as the more rigorous methods, even though they produce different pharmacophore sets. Tanimoto distance matrices were calculated for the test set of 497 substituents, using seven different conformational search protocols. Systematic search was impractical for so many compounds, so the most thorough sampling was 5^R MMFF minimized Rubicon conformations per substituent. Table 9A shows once again that differences due to the amount of conformational sampling are less than the differences due to neglecting minimization. Note that the maximum expected error is only 0.136, so errors of 0.08 are substantial. Table 9B–F shows that the difference is fairly constant irrespective of the number of rotatable bonds, except for Omega, the fastest method, which showed an increasing discrepancy as the number of rotatable bonds increased. With minimization, there is a slight improvement from sampling more conformations, but 4^R are almost the same as 5^R.

Finally, MDS was used to extract property spaces from the seven distance matrices from Table 9A. Omega required seven dimensions; all other sampling protocols each required eight dimensions to reproduce all distances to within a relative standard deviation of 10%. The space from the most thorough method, 5^R MMFF minimized Rubicon conformations, was combined with each of the other spaces, and principal components (PC) analysis was performed, to see how many dimensions need be retained to cover 99% of the variance. Only 9 of 16 PCs, just one more than the original property spaces, were required for the combined space from either 3^R or 4^R MMFF minimized Rubicon conformations. This shows again that the minimized property spaces are very similar. In contrast, combining the space from 5^R MMFF minimized Rubicon conformations with the corresponding unminimized space required 13 of 16 PCs. The spaces from 5^R minimized Rubicon conformations and unminimized Omega required 13 of 15 PCs. This shows significant difference between the unminimized and minimized results. This result, together with Tables 5–9, indicate that, among the conformational search methods studied, for small substituents with few rotatable bonds, a stochastic distance geometry search sampling 4^R conformations, followed by minimization, is a sound protocol to produce reasonably accurate pharmacophore similarities. Even so, it is likely to introduce some errors with the more flexible substituents.

Thus, although the original motivation for OS 1-3PP analysis was to capture higher-order pharmacophores, our experience showing the sensitivity of pharmacophore analysis to conformational sampling reveals what is probably an even more important advantage. One can afford to do thorough, minimized, conformational sampling on a few thousand substituents, each of which has on average only 1.2 rotatable bonds and 14.3 conformers per substituent. One can afford only very limited conformational sampling on many millions or more enumerated products with an average 6.6 rotatable bonds that at 4^R sampling would produce on average 100,000 conformations per product. The luxury of performing high-quality conformational sampling is perhaps the most significant advantage of OS 1-3PP library design over EPB library design.

Importance of Quantitative Similarities

Some might question the need for careful quantitative diversity analysis, arguing that beyond a certain point compounds are so dissimilar that quantifying further distance might be meaningless. This remains an open question in the design of broad screening libraries that attempt to hit as many targets as possible. However, another equally important application of diversity analysis is to thoroughly sample a relatively small region of property space. For example, in a protein structure-based design, one might produce a short list of 200 related substituents, of which perhaps 20 are outstanding candidates, but the remaining 180 cannot be discounted. If one can only afford to synthesize 50 substituents, then it is best to make a design in which the 20 best candidates are included, and use property-biased diversity selection to chose 30 more that cover the property space of the remaining 180 as well as possible. One is now performing diversity analysis on a much more homogeneous set of compounds, where few of them are so different that the similarity measure would be considered meaningless. In this case, an accurate and complete diversity analysis tool is beneficial to increase confidence that all possibilities have been explored.

OS 1-3PP Similarity for 3D QSAR and Targeted "Lead Explosion"

Besides use as a dissimilarity measure for designing diverse broad screening libraries, OS 1-3PP substituent similarity should be an excellent tool for ligand-based combinatorial lead explosion and 3D QSAR. After hits are found in broad screening, OS 1-3PP similarity searching on each of the lead substituents will identify those remaining candidates that present many of the same 3D pharmacophores. One might select perhaps the 100 nearest neighbors to the lead in OS 1-3PP space and then perform D-optimal design to select a representative set of perhaps 20 of these "pharmacophore analogs." A second step of D-optimal design on the entire candidate set also might follow to select perhaps five more diverse substituents as a hedge to sample the remaining property space.

Combining with Other Kinds of Descriptors

It is unrealistic to expect that even EPB 1-9PPs carry all of the information required to determine similarity or difference in receptor recognition. One important advantage of using MDS to create a SB property space is that one can combine the OS 1-3PP property space with other kinds of properties like 3D shape descriptors, 2D Daylight fingerprint-space dimensions, or physical properties like log P, dipole moment, polar surface, or molar refractivity. These Euclidean substituent spaces can be combined and principal component analysis performed, to create a combined space for experimental design.

Similarly, diversity is not the only criterion for a good combinatorial library design. Given substituent points in a Euclidean property space, it is simple to add other constraints to the library design such as physical property distributions, synthetic difficulty, docking results, or cost.[35]

Future Directions: Template Extended Substituents and Elimination of Bins

If one were reluctant to make the combinatorial conformer assumption, one could determine a rule to orient the scaffold in space and then perform the conformational analysis on template extended substituents, with a "typical" rigid substituent in each of the other positions. This would eliminate the last unspecified degree of freedom, rotation around the Tm-Pt vector. In the current example, this would increase the average number of rotatable bonds from 1.2 to 4.2, increasing the average number of conformers from 14 to 900. However, it would not increase the number of substituents, so the calculation would remain tractable. It still would not account for possible idiosyncratic interactions between each pair of substituents, but it would account for idiosyncratic interactions between each substituent and a typical enumerated product environment, eliminating all but the most extreme exceptions to the combinatorial conformer assumption. Unfortunately, it also would make the substituent calculation nontransferable. The current method is independent of the scaffold, so a database of fingerprints can be precomputed and stored for a large number of candidate reagents. This database can be applied to any future library design problems without requiring additional calculations. In our current opinion, the additional gain in rigor has not justified the additional complications of defining a flexible template alignment rule, selecting the "typical" fixed substituents, increasing the number of required conformers due to additional rotatable bonds, and having to repeat the full calculation for each new scaffold. However, it would remove the most significant approximation in the method, at a modest additional computational cost, and therefore is worth consideration.

The use of binned distances and discrete pharmacophores was largely a matter of convenience to take fullest advantage of MOE's existing pharmacophore code. As mentioned earlier, this introduces an aliasing that is only partially removed by the use of near neighbor fingerprints. Alternatively, rather than binning the distances and assigning each possible OS 1-3PP to a discrete bit in a virtual bit-string, one could simply store each OS 1-3PP using the sorted feature types and corresponding distances. The conformational union pharmacophore "fingerprint" would now just be a list of all of the OS 1-3PPs presented by all the sampled (discrete) conformations for that substituent. Between two molecules, each pair of pharmacophores with the same feature types would be compared as the RMSD error of the 2, 5, or 8 distances (for OS 1PPs, OS2PPs, or OS 3PPs, respectively). Substituent dissimilarity would be taken as a normalized sum of the RMSDs for all comparable pairs of OS 1-3PPs. This more computationally involved continuous similarity function would increase the time to compute the distance matrix, which scales with the square of the number of substituents. In addition, one would lose the simplification that many pharmacophores are the same after binning. One also would lose the ability to combine fingerprints by logical operations, taking the union of fingerprints for active compounds and masking out bits for inactive ones. However, the more continuous similarities should yield a more accurate property space and might be expected to embed into fewer Cartesian dimensions by MDS.

These two modifications, template extended fingerprints and elimination of bins, could be combined. After conformational analysis on template extended substituents, align them all by the scaffold rule, which could just be least squares fitting to the atoms of a standard scaffold geometry. The pharmacophores now are stored using the sorted feature types and Cartesian coordinates for each feature (rather than interfeature distances). Intersubstituent dissimilarities between like pharmacophore types now would be determined directly as RMSD between corresponding features for each pharmacophore type. The double advantage of this modification would be both elimination of aliasing and amelioration of the combinatorial conformer assumption. The price again would be a more complicated calculation and the loss of scaffold independence.

Limited to Within a Combinatorial Library

Of course, the biggest limitation is that this method can only be applied within a single combinatorial library. It cannot be applied to an arbitrary compound collection, or even between combinatorial libraries to maximize the diversity of a larger meta-library of individual combinatorial libraries. The method requires that there be a common scaffold by which it makes sense to assemble and orient the molecules. Just as it is easier to find a QSAR among a congeneric series of ligands than between structurally unrelated ligands for a given target, it is likewise easier to quantify similarity between members of a congeneric series than between molecules that are less closely related. That is, within-library diversity can be computed much more accurately and easily than between-library diversity. This argues that one should use a method specifically designed to work within a series whenever possible and resort to more general methods only when necessary. This argument would encourage one to design each library separately, ignoring any previously synthesized libraries.

However, if there was a lot of overlap between different combinatorial libraries in property space, perhaps one ought to compare the current library to all previously synthesized libraries, preferring members that are dissimilar to those in the existing archive. This would minimize the overall redundancy in a collection of libraries. Thus, an important question is whether within-library diversity is small or large compared to between-library diversity. This is a difficult question to answer, because it depends on the property space used in the analysis. Just as compounds that overlap in a 1D chromatograph are resolved by 2D or 3D chromatography, so the point clouds of libraries that overlap in a 3, 4, or 5D property spaces might resolve into discrete galaxies in more complete 20D or 30D property spaces. If there were large overlaps between the point clouds even in very complete and biologically relevant property spaces, then within-library diversity would be large with respect to between-library diversity. One then might be justified in using cruder, slower general diversity measures in order to perform cross-library design. If they spread out into discrete galaxies, then between-library diversity is much larger than within-library diversity, and one ought to use the faster, more rigorous methods that take advantage of the common scaffold. In the absence of such a rigorous general property space that can be used to test the overlap between libraries, one probably must choose a strategy based on intuition and experience about how activity clusters within and between compound series.

On the other hand, even if within-library diversity is large compared to between-library diversity, and libraries do overlap even in an ideal property space, one still might prefer to maximize the diversity of each library independently of the rest of the archive. Getting many initial hits from the same library is redundant in an initial screen. However, getting multiple hits, each based on a different scaffold, is a very valuable result. If one wants to maximize the chances of finding several different active series for each target, one would do well to take advantage of the ability to better compute diversity within a series and design each library independently using a scaffold-based approach.

Hence, if one either believes that between-library diversity is large, at least with well-chosen scaffolds, or if one wants to optimize an overall archive to discover several series per target, then library design is best factored into diversity within libraries and diversity between libraries. Within-library diversity can take advantage of the common scaffold to recast the problem as between-substituent diversity. Likewise, between-library diversity can take advantage of the common scaffolds to recast the problem as between-scaffold diversity selection. The Cliché program is a 3D computational tool for 3D scaffold diversity calculation.[36] Using a 3D SB analysis like OSPPREYS, coupled with a 3D scaffold diversity selection method like Cliché, gives an effective and practical way to optimize 3D diversity for an overall screening archive.

CONCLUSION

The 3D pharmacophore similarity has already demonstrated its biological utility in 3D database searching. It is, therefore, a known biologically relevant descriptor, attractive for combinatorial library design. Based on 3D searching experience, one would ideally like to find all pharmacophores up to 9 points, for all of the conformations of all of the enumerated products of a large virtual combinatorial library, create a Euclidean property space, and select from among them a diverse combinatorial subset for synthesis. This is an impossibly difficult computational task.

Taking advantage of the common scaffold to factor the problem into a substituent diversity analysis simplifies the problem by many orders of magnitude. For a rigid scaffold with three diversity sites of 1,000 candidate substituents each, the number of molecules to analyze is reduced by a factor of 10^6, and the number of conformations per molecules is reduced by 10^4, reducing the number of structures that must be analyzed by 10^{10}. Furthermore, the number of pairwise similarities is reduced by 10^8. Unfortunately, 9PPs cannot be simply broken into fragment 3PPs, because the information that orients them relative to the scaffold and to each other is lost. However, by making the combinatorial conformer assumption and adding two additional reference points, OS pharmacophores can be defined that preserve this information. In this way, a library design can be performed rapidly and conveniently, which is nearly equivalent to the intractable ideal of 9PP analysis on the enumerated products.

Of course, the correspondence between OS 1-3PP and EPB 1-9PPs is not perfect. There are implicit conformational sampling errors due to failures of the combinatorial conformer assumption, and there are implicit errors in the similarities due to neglecting pharmacophore matches in which the scaffolds do not align. On the other hand, phamacophore similarity is found to be very sensitive to conformational analysis. Tests indicate that some 4^R energy-minimized conformations are needed, where R is the number of rotatable bonds. The fact that there are at most a few thousand substituents to consider, each with only a few rotatable bonds, means that a very thorough conformational analysis can be performed on oriented substituents. Because of the exponentially greater number of enumerated products and the exponentially greater number of conformers per molecule due to the greater number of rotatable bonds in the products, it is impossible to perform a comparable conformational analysis on the enumerated products. Any product-based pharmacophore analysis must compromise greatly on conformational sampling. Given the sensitivity of pharmacophore similarity to conformational coverage, this error will be much greater than the errors in the OS 1-3PP approximations. This is probably an even more important reason to use OSPPREYS than capturing the additional pharmacophores with five to nine features.

Experimental design methods that use a Euclidean property space have many advantages over a simple library union fingerprint. A similarity matrix for 1,000 substituents can easily be converted to a Euclidean property space by MDS. This poorly scaling calculation is intractable on millions or more enumerated products. Finally, being independent of the scaffold, OS 1-3PP calculations are transferable. A precomputed database of substituent fingerprints can be applied to future library design problems without requiring additional calculations. For these reasons, OS pharmacophores offer great advantages over pharmacophore analysis of the enumerated products.

Someday, computers and algorithms may improve to the point that our hypothetical standard, an EPB 1-9 point property space analysis, may become feasible. At this point, should we begin to design combinatorial libraries in that way? Probably not. Our discussion suggests that oriented SB designs are reasonably equivalent to EPB designs. This should be true of other kinds of similarity besides just pharmacophore analysis. Pharmacophores are still just a very abstracted measure of molecular similarity. Additional computer power could be employed more profitably to compute more realistic similarity measures between oriented substituents, creating the corresponding property space, and sampling it by D-optimal experimental design.

ACKNOWLEDGMENTS

Paul Labute and Chris Williams at Chemical Computing Group contributed essential advice and valuable SVL utilities and routines throughout this entire project. Thanks to Matt Stahl for an evaluation copy of the very impressive and extremely fast Omega conformation generator.

REFERENCES

1 Simon, R.J., Martin, E.J., Miller, S.M., Zuckermann, R.N., Blaney, J.M., and Moos, W.H. Using peptoid libraries [oligo N-substituted glycines] for drug discovery. In: *Techniques in protein chemistry, Volume V*, Crabb, J.W., Ed., Acedemic Press, New York, 1994

2 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

3 Boyd, S.M., Beverley, M., Norskov, L., and Hubbard, R.E. Characterizing the geometric diversity of functional groups in chemical databases. *J-CAMD* 1995, **9**, 417–424

4 Mount, J., Ruppert, J., Welch, W., and Jain, A.N. IcePick: A flexible surface-based system for molecular diversity. *J. Med Chem.* 1999, **42**, 60–66

5 Jain, A.N., Dietterich, T.G., Lathrop, R.H., Chapman, D., Jr., R.E.C., Bauer, B.E., Webster, T.A., and Lozano-Perez, T. Compass: A shape-based machine learning tool for drug design. *J. Comput.-Aided Mol. Design* 1994, **8**, 635–652

6 Cramer, R.D., Clark, R.D., Patterson, D.E., and Ferguson, A.M. Bioisosterism as a molecular diversity descriptor: Steric fields of single "topomeric" Conformers. *J. Med. Chem.* 1996, **39**, 3060–3069

7 Davies, K., and Briant, C. Combinatorial chemistry library design Using pharmacophore diversity. *Network Science (http://www.awod.com/netsci/Issues/July95/feature6.html)* 1995, **1**

8 Good, A.C., and Lewis, R.A. New methodology for profiling combinatorial libraries and screening sets: Cleaning up the design process with HARPick. *J. Med. Chem.* 1997, **40**, 3926–3936

9 Ashton, M.J., Jaye, M.C., and Mason, J.S. New perspectives in lead generation, II: Evaluating molecular diversity. *Drug Discovery Today* 1996, **1**, 71–78

10 Brown, R.D., Bures, M.G., and Martin, Y.C. A comparison of some commercially available structural descriptors and clustering algorithms. In: *Proceedings of the First Electronic Computational Chemistry Conferenee-CDROM*, 1995, Landover, Maryland

11 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

12 Mason, J.S., Morize, I., Menard, P.R., Cheney, D.L., Hulme, C., and Labaudiniere, R.F. New 4-point pharmacophore method for molecular similarity and diversity applications: Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* 1999, **42**, 3251–3264

13 Kahn, S.D. Combinatorial libraries: Structure activity analysis. In: *The encyclopedia of computational chemistry,* PvR Schleyer, N.A., Clark, T., Gasteiger, J., Kollman, P.A., Schaefer, H.F. III, and Schreiner, P.R., Eds., John Wiley & Sons, Chichester, 1998

14 Spellmeyer, D.C., Blaney, J.M., and Martin, E. Computational approaches to chemical libraries. In: *Practical applications of computer-aided drug design,* Charifson, P.S., Ed., Marcel Dekker, New York, 1997, pp. 165–193

15 Higgs, R.E., Bemis, K.G., Watson, I.A., and J.H.W. Experimental designs for selecting molecules from large chemical databases. *J. Chem. Inf. Comput. Sci.* 1997, **37**, 861–870

16 Agrafiotis, D.K. Stochastic algorithms for maximizing molecular diversity. *J. Chem. Inf. Comput. Sci.* 1997, **37**, 841–851

17 Martin, E.J., and Wong, A. Sensitivity analysis and other improvements to tailored combinatorial library design. *J. Chem. Inf. Comput. Sci.* 2000, (in press)

18 Gillet, 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

19 Pearlman, R.S.S. Novel algorithms for the design of diverse and focussed combinatorial libraries. *Book of Abstracts, 217th ACS National Meeting*, 1999, COMP 197

20 Cramer, R.D., Poss, M.A., Hermsmeier, M., Caulfield, T., and Kowala, M. Recent applications of chemspace shape similarity searching. *Book of Abstracts, 217th ACS National Meeting*, 1999, COMP 182

21 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

22 Linusson, A., Gottfries, J., Lindgren, F., and Wold, S. Statistical molecular design of building blocks for combinatorial chemistry. *J. Med. Chem.*, 2000, (in press)

23 Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D., Weinberger, L.E. Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* 1996, **39**, 3049–3059

24 James, C.A., and Weininger, D. *DaylightTheory manual.* Daylight Chemical Information Systems, Inc., Irvine, CA, 1994

25 Martin, E.J., Critchlow, R.E., Spellmeyer, D.C., Rosenberg, S., Spear, K.L., and Blaney, J.M. Diverse approaches to combinatorial library design. In: *Pharmacochemistry library, Volume 29 (trends in drug research II),* Timmerman, H., Ed., Elsevier Publishers, Amsterdam, 1998

26 Inc., C.C.G. *MOE*, 1999.05, Montreal, 1999

27 Weininger, D. *Rubicon reference manual, v4.51.* Daylight Chemical Information Systems, Inc., Irvine, CA, 1997

28 Stahl, M. *Omega.* Santa Fe, NM, 1999

29 Hahn, M. Receptor surface models. 1. Definition and construction. *J. Med. Chem.* 1995, **38**, 2080–2090

30 Singh, U.C., Weiner, P.K., Caldwell, J., and Kollman, P.A. *AMBER 3.0A.* Available from Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143, 1989

31 Halgren, T.A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* 1996, **17**, 490–519

32 Brown, R.D., and Martin, Y.C. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* 1996, **36**, 572–584

33 Rush, J.A. Cell-based methods for sampling in high-dimensional spaces. In: *IMA Vol. Math. Its Appl. Volume 108 (Rational Drug Design),* 1999, pp. 73–79

34 Rush, J.A. Personal communication, 1996

35 Martin, E.J., and Critchlow, R.E. Beyond mere diversity: Tailoring combinatorial libraries for drug discovery. *J. Comb. Chem.* 1999, **1**, 32–45

36 Bartlett, P.A., and Lauri, G. The caveat vector approach for structure-based design and combinatorial chemistry. In: *1995 International Chemical Congress of Pacific Basin Societies*, 1995, Honolulu, Hawaii