Pharmacophore Fingerprinting. 2. Application to Primary Library Design

Malcolm J. McGregor and Steven M. Muskal*

Affymax Research Institute, 3410 Central Expressway, Santa Clara, California 95051

Received June 22, 1999

A methodology for pharmacophore fingerprinting (PharmPrint), previously described in the context of QSAR, has been used to address the issues involved in primary library design. A subset of the MDDR (MDDR9104) has been used to define a reference set of bioactive molecules. A statistic has been devised to measure the discriminating power of molecular descriptors using the target class assignments for this set, for which the PharmPrint fingerprint outperformed other descriptors. A principal components analysis (PCA) of the fingerprints for the MDDR9104 produces a low dimensional representation within which molecular properties and other libraries can be visualized and explored. PCA calculations on subsets of classes show that this space is robust to the addition of new classes, suggesting that pharmacophoric space is finite and rapidly converging. We demonstrate the application of the PharmPrint methodology to the analysis and design of virtual combinatorial libraries using common scaffolds and building blocks.

INTRODUCTION

Recent advances in combinatorial chemistry and high throughput screening have generated interest in analyzing calculated properties of large collections of compounds.[1−7] In the field of drug design, two broad applications can be identified: (i) design of targeted (or focused) libraries, where the main activity is prediction of binding to a particular protein target (enzyme or receptor), and (ii) design of exploratory primary libraries, to be screened across a number of targets that may be structurally unrelated. In addition there may be an intermediate case, where compounds are to be screened against a family of structurally related targets.

Targeted library design is essentially an extension of the areas of computational chemistry and molecular modeling which utilize quantitative structure−activity relationships (QSAR) for scaffold design and building block selection. This involves calculating molecular descriptors, using them in a model to predict biological activity, and selecting building blocks to maximize a library's performance against the target of interest. We have previously reported on the development of the pharmacophore fingerprinting methodology named PharmPrint.[8] It has been applied to some examples of activity prediction for a single target, generating QSARs for estrogen receptor ligands. The results were shown to be superior to previous methods applied to the same data. However, the versatile and information-rich nature of this descriptor means that it is also useful in addressing the issues of primary design, to which we now turn.

The goal with primary library design is to generate active compounds for one or more targets when there is little or nothing known about the protein structures or their ligands. In addition there may be the desire to optimize early in the design process other drug properties not related to binding such as absorption, distribution, metabolism, excretion (ADME), and toxicity.

A starting point for this kind of analysis is the calculation of descriptors to characterize molecular structures. Many kinds of descriptors have been used.[9−13] They can be broadly classified according to how they treat molecular structures. Many descriptors can be calculated from a molecule connection table that specifies the atom types and their connectivity (1D/2D). Examples are molecular weight, calculated logP (clogP), and descriptors that contain information about chemical functionality (e.g., H-bonding groups). A widely accepted rule of 5 has been established using such properties to define the requirements for molecules to be successful as drugs.[14] Calculation of 3D properties involves generating an energetically reasonable 3D structure. In addition, some methods incorporate a treatment of multiple conformations. Sometimes descriptors are chosen based on features observed to be important in ligand binding, or descriptors are used which have been shown to correlate with desirable properties. Other times many descriptors are calculated and statistical methods are used to establish a minimal set that is important.

The pharmacophore concept is widely used in the field of computer aided drug design.[15−17] It is based on the kinds of interactions observed to be important in ligand-protein interactions: hydrogen bonding, charge, and hydrophobic interactions. A pharmacophore is a set of functional group types in a particular spatial arrangement; traditionally the goal has been to find one, or a small number, that represent the interactions made in common by a set of small molecule ligands with a protein receptor. Pharmacophore fingerprinting extends this concept by constructing a basis set of pharmacophores by enumerating a set of pharmacophoric types with a set of distance ranges, and determining which ones are present in a molecule.[8,18−24] The PharmPrint methodology has been developed at Affymax to fingerprint large libraries of compounds. The PharmPrint binary bitstring is a compact but information-rich descriptor, containing information about 3D molecular structure and multiple conformations.

There are several desirable properties that a calculated molecular descriptor should have. Ideally a descriptor should allow for a quantitative measure of molecular similarity. A calculated molecular descriptor has better utility if it cor-

* To whom correspondence should be addressed. E-mail: smuskal@molseek.com.

10.1021/ci990313h CCC: $19.00
Published on Web 12/08/1999

relates with an experimentally measurable property. Clearly a calculated logP should correlate as closely as possible with the measured value. In drug design an important property is binding to a protein target. Although this can be calculated explicitly (e.g., with docking calculations, if the structure of the target is available) usually it is done as a function of more easily calculated properties which are regarded as independent variables. It would be expected that descriptors that contain conformational information would be more predictive of biological activity than ones that do not, and that 3D descriptors would be better than 1D/2D descriptors. However this has been difficult to demonstrate, and sometimes 2D descriptors actually outperform 3D ones.[10,11]

It is assumed that the goals of primary library design will be achieved by a collection of compounds that have a property distribution which is close to that of all compounds demonstrating some level of biological activity. We have used the PharmPrints of compounds from the MDDR to define the properties of bioactive molecules, and we have used these results in the design of combinatorial libraries that have optimum property distributions. Thus we make a conceptual distinction between chemical space and bioactive space, and between maximizing molecular diversity and optimum coverage of bioactive space.

METHODOLOGY

Fingerprint Generation. The methodology used was the same as that described previously.[8] Briefly, a basis set of 3-point pharmacophores was constructed by enumerating 7 pharmacophoric types and 6 distance ranges. The types are: H-bond acceptor (A) and donor (D), formal negative (N) and positive (P) charges, hydrophobic (H), aromatic (R), and a default type (X) for any atom that is not labeled with any of the first six types. The distance ranges are: 2.0−4.5, 4.5−7.0, 7.0−10.0, 10.0−14.0, 14.0−19.0, and 19.0−24.0 Å. The pharmacophores are filtered to eliminate duplicates related by symmetry, and ones that violate the triangle rule, resulting in a basis set of size 10 549. The PharmPrint program was developed in-house to rapidly fingerprint large numbers of compounds. It takes as input a single 3D molecular structure generated by the Corina program,[25,26] assigns the pharmacophoric types to atoms, rotates about bonds to generate multiple conformations, and builds the fingerprint by measuring distances between pharmacophoric groups. The output is a binary bitstring containing information about the pharmacophores presented by the molecule. The program can run in batch mode accepting MDL SD files as input and providing fingerprints for each structure as output.

Preparation of the MDDR9104 Subset. The MDDR (MDL Drug Data Report)[27] was used as a reference for bioactive molecules. A subset of the total 92 604 entries in version 98.1 was prepared as described previously.[8] Structures were extracted according to the assigned activity class, where the class indicates a single protein target (as opposed to a general therapeutic area). Molecules were further filtered based on molecular weight range (200−700) and atom type content (only C, N, O, H, S, P, F, Cl, Br, and I allowed), and a measure of 2D similarity (MDL 166 keys) was applied to eliminate close analogues. This procedure resulted in 9104 compounds in 152 classes. A compound may belong to more than one class; however only 1083 (11.9%) do so.

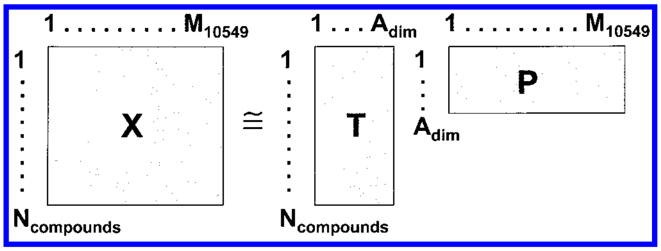

Figure 1. Reducing PharmPrint dimensionality: the principal component transformation illustrated in matrix form. **X** is the data matrix (fingerprints), **T** is the scores matrix (coordinates in low dimensional space), and **P** is loadings matrix, which defines the transformation between them.

Table 1. *t*′ Statistic Using Class Assignments in the MDDR9104 Set and Various Molecular Descriptors

mol. wt.:	*t*′ = 321.3
MDL 166 keys Tanimoto:	*t*′ = 301.8
PharmPrint Tanimoto:	*t*′ = 455.8

	MSI_{50}/PCA		PharmPrint/PCA	
dim.	*t*′	%var	*t*′	%var
1	330.1	63.5	306.0	22.9
1-2	344.5	72.8	403.2	30.2
1-3	359.7	79.1	445.1	35.4
1-4	351.1	84.8	455.2	39.2
1-5	372.1	88.9	442.1	42.6
1-6	365.9	92.0	434.9	45.2
1-7	369.9	94.0	434.6	47.0
1-8	371.7	95.8	440.3	48.6
1-9	374.0	96.8	440.9	49.9
1-10	374.9	97.6	441.9	51.0
1-11	374.9	98.1	442.7	52.0
1-12	375.7	98.5	446.3	53.0
1-13	375.3	98.9	447.2	53.8
1-14	374.8	99.2	446.8	54.5
1-15	374.7	99.4	447.9	55.2
1-16	374.6	99.5	448.4	55.8
1-17	374.6	99.6	448.7	56.4
1-18	374.6	99.7	447.8	56.9
1-19	374.6	99.7	448.1	57.5
1-20	374.7	99.8	447.3	57.9

Principal Components Analysis (PCA). PCA was performed on the MDDR9104 set to produce a low dimensional space suitable for visualization. The bits in the fingerprint were converted to the real numbers 0.0 and 1.0 for the calculation. The NIPALS algorithm was used, which calculates one component at a time.[28] The data were mean centered but not variance scaled. Figure 1 illustrates the process whereby the data matrix **X** (fingerprints) is broken down into the scores matrix **T** (new coordinates in reduced dimensional space) and loadings matrix **P**, which can be applied to any new fingerprints to transform them to the new space. The variance accounted for by the addition of each component is included in the rightmost column of Table 1.

Molecular Similarity. It is important that molecules which are judged to be similar according to a calculated property are similar in biological activity. The following method was established as a measure of the discriminating power of a molecular descriptor, using any data set that is classified into activity classes, such as the MDDR9104 set. (Previous analyses of this kind have usually used one target at a time.[13]) If all the $(n^2 - n)/2$ pairwise intermolecular comparisons are made, then these can be divided into two types: comparisons made within classes and those made between classes. (For compounds belonging to several classes, if a

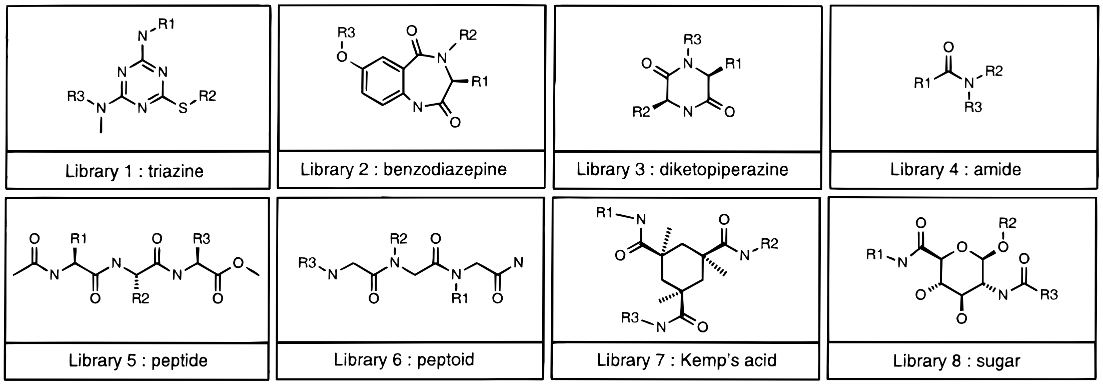

Figure 2. The eight combinatorial scaffolds analyzed in this study.

pair share at least one class, they are regarded as being in the same class.) Compounds in the same class are assumed on average to be more similar in biological activity than ones in different classes (although this is not strictly implied in the data, i.e., it is not the case that every compound has been tested against every target). This produces two distributions of molecular similarities. The difference in the means of the distributions can be expressed in units of standard error by the formula

$$t' = (\bar{X}_1 - \bar{X}_2)/\sqrt{(s_1{}^2/n_1 + s_2{}^2/n_2)}$$

where for samples 1 and 2, $\bar{X}$ is the mean, s^2 is the variance, and n is the sample size. For small samples this follows the Student t distribution; for large samples it goes toward a normal distribution. This statistic is sometimes used as a test of significance for the difference between two distributions. With the results presented here the statistic is always highly significant, so the absolute value of the statistic is presented; the larger the absolute value the better. It can be calculated for any data set that is assigned to classes, and for any measure of similarity.

MDDR9104 Coverage of Bioactive Space. This set was designed to be as representative as possible of bioactive molecules in general, given current data. A test was devised to investigate whether the space produced by the PCA calculation on the MDDR9104 set is a true universal space, or if it is highly dependent on the content of the database at any point in time. Thus PCA calculations were done on randomly selected subsets of the 152 classes. Growing subsets of compounds which belong to 19, 38, 57, 76, 95, 114, and 133 classes were created, where the larger sets are supersets of the smaller sets. This simulates the situation where over time new targets are discovered, whose active compounds are added to the MDDR database. The PCA transformation is defined by the loadings matrix **P** (Figure 1). A comparison of the **P** matrix was made for each subset with the preceding smaller subset and reported as a root-mean-square value (referred to as ΔP), for the first four PCA dimensions. For example, a PCA was performed on the compound set from 19 randomly selected classes. Another 19 were added and the PCA calculation was repeated. The ΔP(19,38) value was calculated between them. Another 19 classes were added and the ΔP(38,57) calculated, and the process repeated until the full set with 152 classes was reached. The whole process was repeated 20 times with different random number seeds. A low ΔP value as classes are added, especially in the later stages of the calculation, would indicate that adding further classes in the future will not substantially change the nature of this space.

Scaffolds and Building Blocks for Combinatorial Library Analysis. Eight scaffolds were used, illustrated in Figure 2. They have been chosen to be a diverse set of commonly used scaffolds. They have all been reported in the literature and/or used in our laboratories. Each one has 3 positions of diversity, and were enumerated with the same set of 20 surrogate building blocks in all positions, to give libraries of 8000 molecules each. This simplifies the comparison between them. The building blocks are based on the side chains of the 20 coded amino acids (the exception was proline, for which we substituted cyclopentyl glycine). In reality the building blocks would be chosen for each scaffold based on synthetic feasibility and availability, and would be of different chemical classes, e.g., amines and aldehydes. However we would expect that most of these building blocks would be available with the same or equivalent functionality. For example, if amine building blocks are required, then phenylalanine would be represented by the available reagent benzylamine. The amino acid side chains are chemically diverse, biologically relevant, and easy to report using the one letter code.

Building Block Selection. A methodology was implemented to select subsets of building blocks to optimize a function. The selection is done for each position in each scaffold; i.e., a subset out of a total of 480 (20 building blocks in 3 positions for 8 scaffolds). Initially 50% (240) of the building blocks were randomly selected. A combinatorial constraint was implemented such that all selected building blocks were enumerated for each scaffold, giving a subset of approximately 8000 selected molecules out of the fully enumerated 64 000.

The algorithm proceeds as follows. Starting from a random selection of building blocks the function is calculated on the enumerated products. Then a randomly selected building block from the included set is excluded, and a randomly selected building block from the excluded set is included, and the function is reevaluated. A Metropolis (probability) function is used to decide if the step is accepted or rejected, and the method proceeds iteratively until no further improvement is possible.

Two functions were explored. The first function was an overlap between the compound subset and the MDDR9104 compounds in components 1−3 of the MDDR9104/PCA space, referred to as the overlap function. Maximizing this

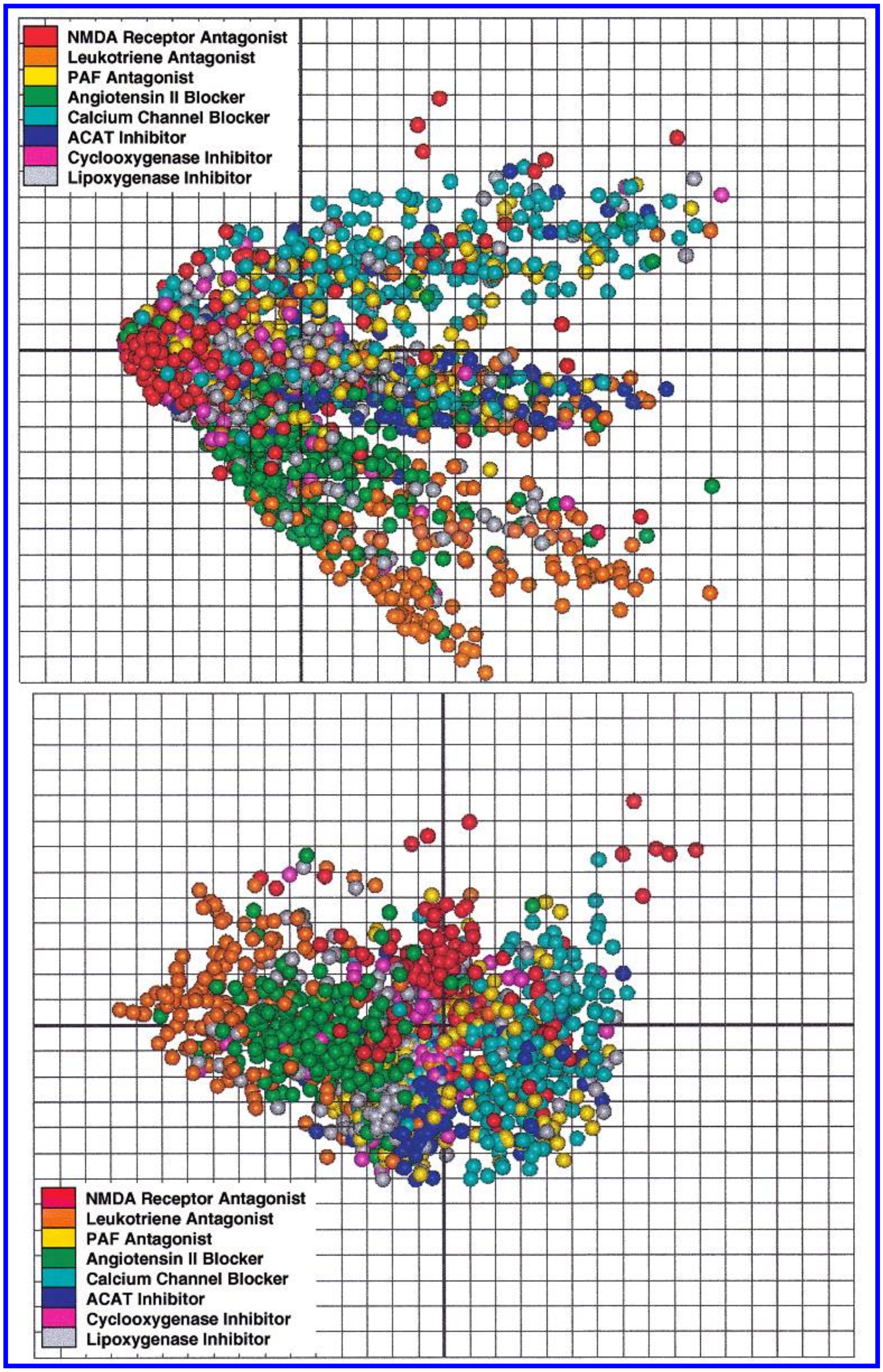

Figure 3. The eight largest target classes in the MDDR9104 set, color coded, shown using principal components 1−2 (a, top) and 2−3 (b, bottom).

function optimizes the distribution of the enumerated compounds to most closely resemble that of the MDDR9104. The coordinate space resulting from the PCA calculation on the MDDR9104 set was divided into cubic cells of size 2.0 units in 3 dimensions. Counts of the number of points with coordinates in each cell were made and scaled according to library size. Then a measure of the overlap of the distributions was made as follows:

$$\text{overlap} = \sum_i \{n1_{\mathrm{i}} + n2_{\mathrm{i}} - \mathrm{abs}(n1_i - n2_i)\}/(\mathrm{N1} + \mathrm{N2}) \times 100.0$$

where N1 = total number in set 1, N2 = total number in set 2, $n1_i$ = number from set 1 in cell i, and $n2_i$ = number from set 2 in cell i.

The second function explored was the maxmin function which sums, for each molecule, the distance to its nearest

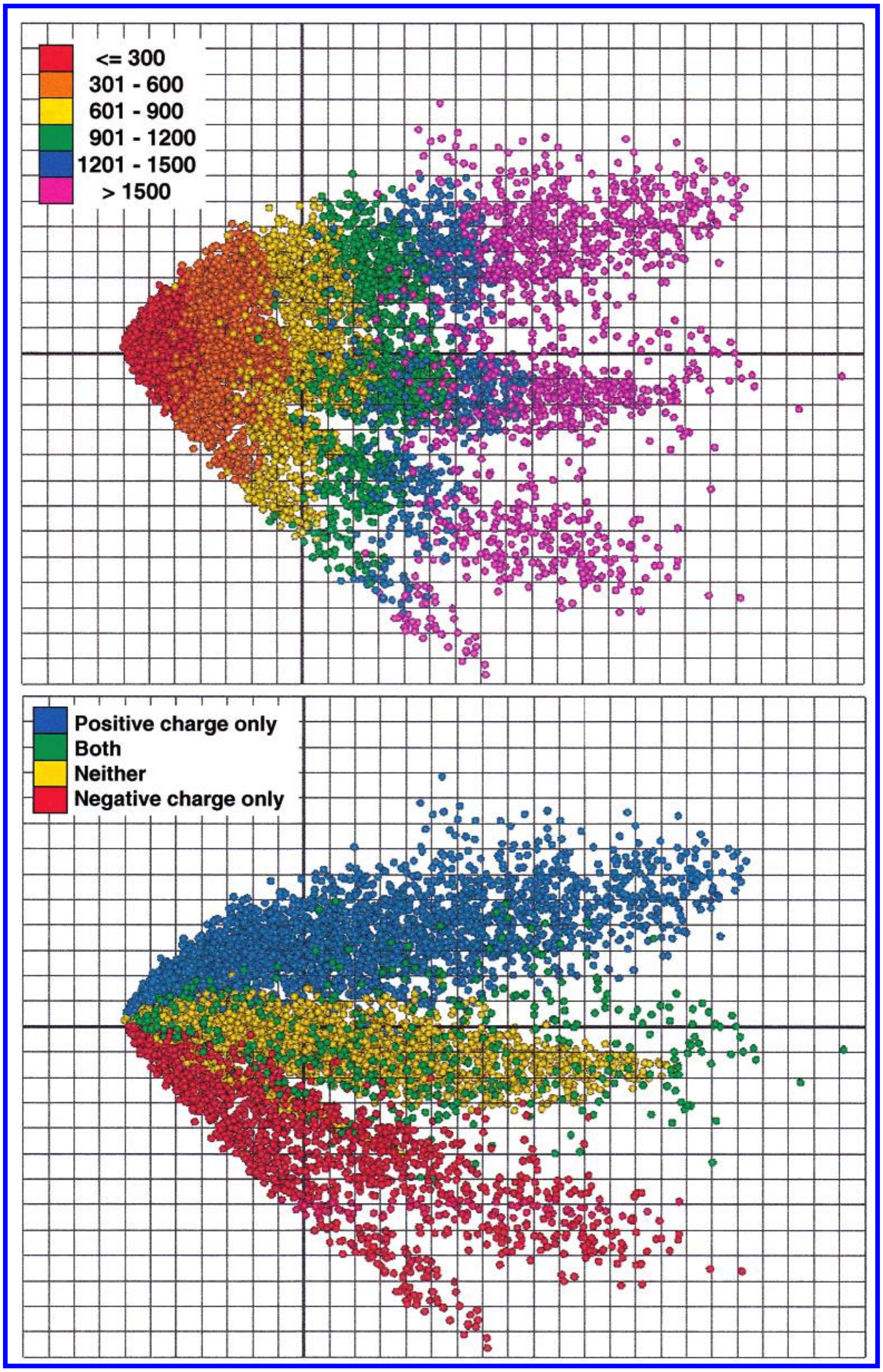

Figure 4. Principal components 1 and 2 of the MDDR9104 set, color coded according to (a, top) number of pharmacophores in the molecule, and (b, bottom) content of formal charges.

neighbor. When maximized, this produces a set which spreads points as far apart as possible in the accessible space.

RESULTS AND DISCUSSION

Molecular Similarity. The first property explored was the ability of the PharmPrint fingerprint to act as a measure of molecular similarity, as judged by the class assignments in the MDDR9104 set. The *t′* statistic (as defined in the Methodology Section) for the MDDR9104 set is presented in Table 1 for different molecular descriptors. The MDL 166 sskeys are used as an example of a 2D fingerprint, the Tanimoto coefficient[29] being used to compare them. The statistic is also calculated using the difference in molecular weight as a measure of molecular similarity. Molecular weight was regarded as a 1D descriptor that is generally descriptive of molecules, but we did not expect it to be highly predictive. However, molecular weight, with $t' = 321.3$, actually outperforms the MDL keys at 301.8. Both of these are outperformed by the PharmPrint/Tanimoto result at 455.8. Results are also presented for a PCA analysis of two descriptor sets: MSI_{50} and PharmPrint. The MSI_{50} are 50 default descriptors in the software package Cerius2 from MSI.[30] These descriptors vary in type; some are calculated on a single 3D structure, but none are calculated using multiple conformations. The set is typical of that used in many QSAR applications. The measure of similarity is the Euclidean distance calculated in up to 20 dimensions. The MSI_{50} result is a maximum of 375.7 for dimensions 1−12, though it is 372.1 for 1−5. The PharmPrint result reaches a maximum of 455.2 using the first four principal components, and with more components the value declines. Thus these data give the expected, but often difficult to demonstrate, result that 3D conformationally flexible descriptors outperform 3D one-conformer descriptors, which in turn outperform 2D descriptors. It also shows that the PharmPrint/PCA result is comparable to the PharmPrint/Tanimoto result. This implies that molecules can be meaningfully considered in a low dimensional space derived from fingerprints, which simplifies certain calculations and aids in visualization on paper, in 2D, and on the computer graphics in 3D.

PCA. Figures 3 and 4 show the results of a PCA calculation of the MDDR9104 set. The plots are the coordinates in the T matrix of Figure 1, and each compound appears as a single point. In the first two components the distribution is wedge shaped, and some regions appear to be more densely populated than others. The eight largest activity classes in the MDDR9104 set are shown color coded in Figure 3(a) for components 1 and 2, and in Figure 3(b) for components 2 and 3. This gives a qualitative and visual representation of the separation of activity classes as calculated by the t' statistic above.

A question that immediately arises is how do the individual pharmacophores contribute to each principal component. Figure 4(a) shows the complete MDDR9104 set (components 1 and 2) color coded according to the number of bits set in the fingerprint (i.e., number of pharmacophores hit by the molecule). A high count indicates large, flexible, highly functionalized molecules. This shows a strong separation in the first principal component with the bit counts increasing from left to right. Figure 4(b) shows the same plot color coded according to the presence of formal charges in the structures. This shows a separation in the second principal component, where compounds with only positive charges tend to have positive coordinates (top of plot); ones with only negative charges tend to have negative coordinates (bottom of plot); compounds which have both or neither are between these two (cluster toward the origin). When components 3 and 4 are viewed on a 3D computer graphics screen and colored appropriately, trends can also be seen in counts of H-bond, aromatic, and hydrophobic groups (data not shown), though these are not so clear-cut as bit count and charge.

The results of the ΔP calculation (as described in the Methodology Section) are shown in Figure 5. The value is an RMS of the first four principal components. There is a pronounced downward trend that approaches the baseline when the later sets of classes are added. This indicates that we may have reached the point where adding further classes in the future will not significantly change the nature of this space. This may be because the general features of the protein binding sites are well sampled by this set of 152 classes and 9104 ligands, at least within the PharmPrint description. (It is possible that with a more detailed description of molecules, e.g., 4-point pharmacophores, more sampling would be needed.)

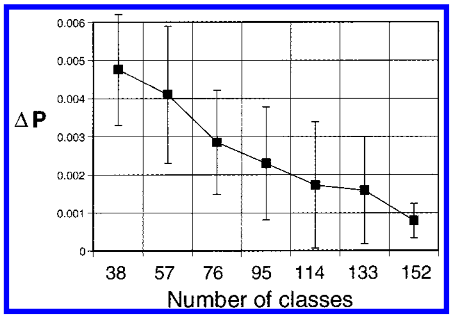

Figure 5. Results of the ΔP calculation (see text).

Table 2. Overlap of Fully Enumerated Libraries with Each Other and with the MDDR9104 Set

	MDDR	Lib1	Lib2	Lib3	Lib4	Lib5	Lib6	Lib7	Lib8
MDDR	100	30	22	29	31	7	8	7	8
Lib1		100	39	44	34	9	12	10	14
Lib2			100	32	18	18	18	22	23
Lib3				100	54	5	15	9	11
Lib4					100	2	6	4	5
Lib5						100	14	37	52
Lib6							100	13	19
Lib7								100	40
Lib8									100

Analysis of Combinatorial Libraries. Table 2 shows the overlap of the fully enumerated libraries with one another and with the MDDR9104 in PCA space (dimensions 1−3). Overlap with the MDDR9104 can be interpreted as a measure of the biological activity potential of the library. It can be seen that there is considerable variation, with the first four scaffolds overlapping in the region of 20−30%, whereas the last four have values less than 10. This would be an initial indication that these low scoring scaffolds are not good candidates for primary libraries, but better used in more specialized applications. The overlap between libraries can be interpreted as a measure of similarity. Once again there is a fair variation, and examination of these values can be made with reference to the scaffolds in Figure 2.

With the building block selection simulation, 10 independent runs were performed with different random number seeds for the two scoring functions, and the results are presented in Table 3 as mean and standard deviation for the 10 values. For optimization of the overlap function with MDDDR9104, the initial (random) overlap was 29.7(2.0)% and the optimized value was 52.6(0.3)%. As a point of reference, if the MDDR9104 set is split into two equal halves, the overlap between them is 68.1%, so it is difficult to approach 100%. Table 3 gives some general statistics for the initial and final combinatorial sets, and for the MDDR9104, including descriptors that were not part of the optimization calculation (molecular weight, clogP[31]). In addition two other reference sets, derived from MDL databases,[27] are included for comparison: (i) CMC (filters: mol.wt. 150 to 750, atom type filter as for MDDR, salts removed), (ii) ACD (filters: mol.wt. 1 to 1000, salts

Table 3. Statistics for Compound Sets[a]

	libraries[b]					
		final subset		databases		
	initial subset	overlap	maxmin	MDDR9104	CMC	ACD
overlap	29.7 (2.0)	52.6 (0.3)	26.4 (0.7)	100.0	57.9	48.0
compounds	7990 (286)	7992 (285)	7974 (287)	9104	6647	213968
MW	363 (85)	350 (87)	388 (74)	388 (104)	342 (111)	252 (122)
clogP	−0.22 (2.27)	1.80 (1.80)	0.11 (2.45)	3.7 (2.3)	2.6 (2.7)	2.4 (2.8)
atoms	25.4 (6.3)	24.5 (6.5)	27.3 (5.59)	27.4 (7.4)	23.7 (7.7)	20.4 (9.1)
bits	899 (622)	806 (633)	1137 (654)	790 (670)	529 (551)	317 (492)
rotbonds	9.43 (4.03)	7.83 (3.88)	9.79 (4.01)	6.74 (4.58)	5.43 (4.19)	4.76 (4.90)
X	13.82 (3.50)	13.71 (3.69)	15.09 (3.31)	13.68 (4.88)	11.88 (5.45)	9.33 (5.41)
A	4.31 (2.18)	3.58 (1.97)	4.38 (2.22)	3.49 (2.08)	3.44 (2.45)	2.97 (2.41)
D	3.69 (1.79)	2.77 (1.47)	3.67 (1.72)	1.57 (1.25)	1.66 (1.57)	1.01 (1.36)
H	3.83 (3.16)	4.65 (3.10)	4.16 (3.11)	8.80 (5.22)	6.96 (5.10)	7.13 (6.04)
N	0.30 (0.52)	0.28 (0.50)	0.41 (0.59)	0.24 (0.55)	0.23 (0.61)	0.17 (0.51)
P	0.58 (0.70)	0.37 (0.55)	0.70 (0.72)	0.42 (0.58)	0.52 (0.67)	0.13 (0.41)
R	0.70 (0.76)	0.97 (0.81)	0.98 (0.81)	1.76 (0.95)	1.24 (0.93)	1.32 (1.11)

[a] Mean and standard deviation for overlap function with MDDR9104 (see text), number of compounds, molecular weight, clogP, number of heavy atoms, number of bits (pharmacophores) in the fingerprint, number of rotatable bonds, and the number of atoms per molecule assigned to the pharmacophore types. [b] Results calculated for 10 simulations.

removed). The initial library subsets have several values close to that of the reference sets; the greatest discrepancies are an overabundance of H-bond donors, a relative lack of hydrophobic and aromatic groups, and clogP values that are somewhat low. In general the results of the overlap optimization bring the statistics closer to the MDDR9104 set than the optimization of the maxmin function, including descriptors that were not explicitly part of the simulation (e.g., clogP).

Table 4 shows the counts of the frequency of occurrence of the scaffolds and building blocks in the optimized libraries. The relatively small standard deviations indicate that the results are reproducible. With the libraries that have been optimized for overlap with the MDDR9104, the first four scaffolds have a much greater frequency than the last four, in agreement with the overlap of the completely enumerated libraries. The building block frequencies show a pronounced preference for hydrophobic and aromatic side chains, and a trend against charged and polar side chains. With the libraries optimized for the maxmin function, the scaffold and building block frequencies follow some of the same trends, but tend to favor the larger molecules in preference to the smaller ones.

One method for identifying holes in such a property space was carried out as follows. For each cubic cell, a count was made of the number of MDDR compounds in cells that are devoid of library compounds. With the overlap-optimized subset the cell with the highest number had 44 such compounds, some of which are illustrated in Figure 6. It can be seen that these are generally neutral molecules with aromatic rings and H-bond acceptors but no H-bond donors. Visual inspection of the scaffolds shows that all except one (the amide scaffold #4) have at least one donor. Inspection of the building blocks shows that there are no neutral side chains that have acceptors but not donors. Therefore the origin of this lack of coverage can be easily appreciated, but was not self-evident at the start. New scaffolds and/or side chains can then be incorporated into the analysis to overcome this deficiency.

Table 4. Frequency of Occurrence of (i) Scaffolds and (ii) Building Blocks in the Library Subsets Optimized for the Overlap and the Maxmin Functions[a]

(i) scaffolds

	function	
scaffold	overlap	maxmin
1	1911 (157)	1455 (113)
2	1244 (139)	1694 (111)
3	1709 (217)	896 (168)
4	1444 (158)	463 (65)
5	463 (91)	1091 (114)
6	687 (75)	1389 (133)
7	219 (56)	302 (70)
8	313 (69)	684 (108)

(ii) building blocks

		function	
type	description	overlap	maxmin
D	charged	360 (129)	678 (101)
E	charged	258 (132)	662 (96)
H	charged	420 (92)	511 (130)
K	charged	124 (90)	539 (123)
R	charged	69 (53)	470 (135)
Q	polar	198 (123)	355 (125)
N	polar	191 (104)	188 (147)
C	polar	334 (89)	241 (103)
S	polar	149 (116)	144 (115)
T	polar	155 (119)	79 (100)
A	small neutral	514 (121)	247 (142)
G	small neutral	365 (140)	184 (90)
Y	aromatic polar	580 (150)	697 (64)
W	aromatic polar	486 (116)	756 (66)
F	aromatic hydrophobic	776 (70)	735 (88)
L	aliphatic hydrophobic	678 (101)	208 (123)
M	aliphatic hydrophobic	700 (100)	505 (158)
(P)	aliphatic hydrophobic	549 (129)	198 (119)
I	aliphatic hydrophobic	610 (109)	298 (164)
V	aliphatic hydrophobic	476 (121)	279 (13)

[a] Mean and standard deviation for 10 simulations.

These results validate the MDDR9104/PCA space as being useful for optimization of general properties of combinatorial libraries, and also for identifying deficiencies in them. Thus the 20 amino acid side chains, when fully enumerated, may not be an optimum choice for ligand design, as they produce a somewhat skewed distribution when compared to known bioactive compounds. We can think of two possible reasons

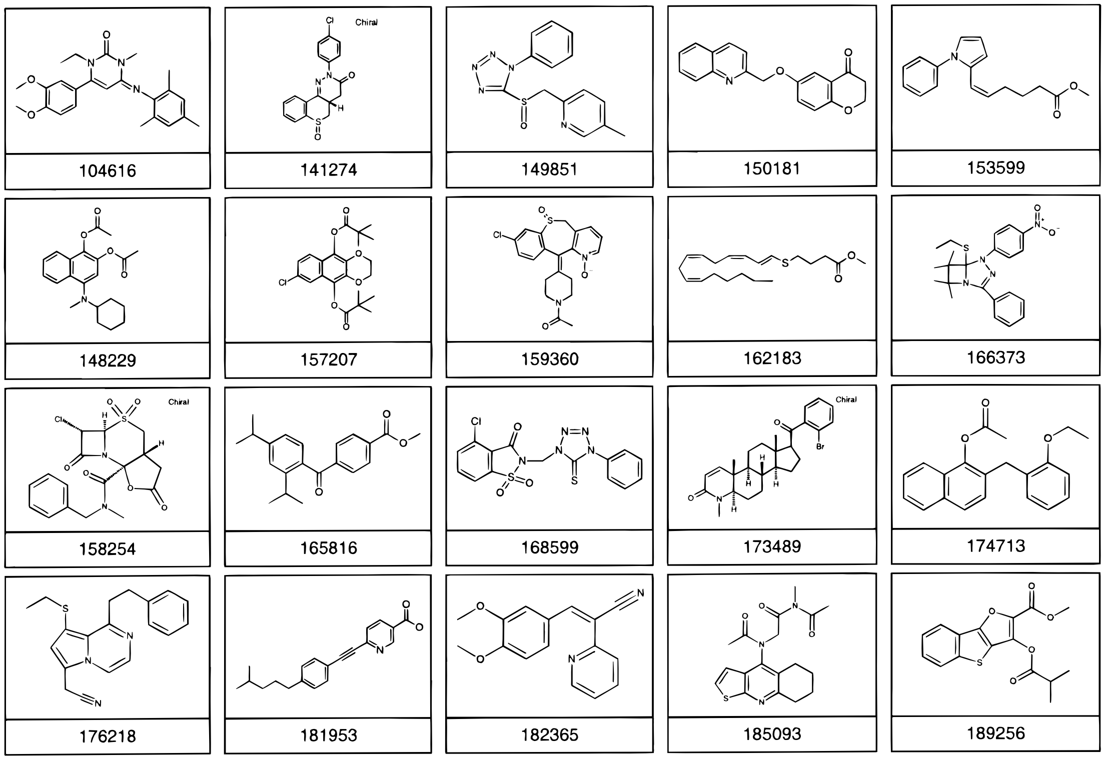

Figure 6. A sample of molecules from the MDDR9104 that occupy a region of PCA space not covered by the combinatorial libraries.

for this. First, protein binding sites have a tendency to be hydrophobic, with hydrophilic residues being reserved for the protein exterior. Second, ligands need to be complementary to the amino acids they interact with, not mimicking them, e.g., if proteins contain more H-bond donors, then ligands should contain more acceptors.

CONCLUSIONS

We have explored the issues involved in primary library design using the newly developed PharmPrint methodology. First, we have shown that as a descriptor the PharmPrint fingerprint has superior ability to discriminate between compound classes as defined by binding to a protein target.

We have assumed that a desirable property of a primary library is that it should have property distributions close to that of known bioactive compounds. (It should be pointed out that reproducing general property distributions does not imply that the same "me-too" compounds are being generated.) The disadvantage of this approach is that if the reference set is not truly representative of all desirable compounds, then the types of compounds that are not already represented may be overlooked. While it may be argued that there are types of drug targets which are biologically attractive but underrepresented in the database, we have shown that the reduced dimensionality space from the PharmPrint/PCA calculation is relatively stable to the addition of new classes. However, we recognize that receptors which interact with protein ligands or with other receptors, which have so far defied traditional drug design efforts, may lie outside a property space defined by existing targets. Therefore the space defined by the MDDR/PharmPrints should be considered valid only for targets that can bind to small molecule ligands. The problem of mimicking macromolecular ligands may be best addressed in its own right as a structure-based design problem.

A goal generally considered desirable in primary library design is optimization of a measure of molecular diversity, here implemented by the maxmin function. Disadvantages of this approach are that it is prone to emphasize outliers, and it does not correct biases in property distributions arbitrarily imposed by initial selection of scaffolds or building blocks. In practice a prudent approach might be to optimize both types of functions at the same time.

These results also emphasize the need to consider libraries in a space that plots *molecules* as opposed to *pharmacophores*. A molecule is a collection of pharmacophores, usually several hundred of them in the PharmPrint fingerprint, and it is the particular combination of them (presence of some and absence of others) that defines the molecular properties. It is possible to have a library that contains all possible pharmacophores (100% coverage of "pharmacophore space") yet does not contain important classes of molecules (e.g., molecules that have an absence of a particular pharmacophoric type). Inspection of the pharmacophores not hit by the libraries or reference sets shows that they generally contain features uncommon or unlikely to be important in bioactive molecules, e.g., the largest distance ranges, or the two charge groups within the smallest distance bin (data not shown).

We are in the process of validating these results by the screening of large primary combinatorial libraries against many targets of different kinds (enzymes, receptors, so-called hard targets mentioned above). The mapping of pharmacophoric space will become increasingly important with the proliferation of new targets expected from the field of genomics, where initially there will be little structural or functional data available. Understanding and perhaps inferring regions of ligand pharmacophore space from a domain of biologically relevant targets will be an important step in the pursuit of bioactive molecules.

REFERENCES AND NOTES

(1) Agrafiotis, D. K.; Myslik, J. C.; Salemme, F. R. Advances in Diversity Profiling and Combinatorial Series Design. *Molecular Diversity* **1999**, *4*, 1−22.

(2) Eichler, U.; Ertl, P.; Gobbi, A.; Poppinger, D. Addressing the Problem of Molecular Diversity. *Drugs of the Future* **1999**, *24*, 177−190.

(3) Ghose, A. K.; Vellarkad, V. N.; Wendolowski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55−68.

(4) Martin, E. J.; Critchlow, R. E. Beyond Mere Diversity: Tailoring Combinatorial Libraries for Drug Discovery. *J. Comb. Chem.* **1999**, *1*, 32−45.

(5) Menard, P. R.; Mason, J. S.; Morize, I.; Bauerschmidt, S. Chemistry Space Metrics in Diversity Analysis, Library Design, and Compound Selection. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1204−1213.

(6) Lewis, R. A.; Mason, J. S.; McLay, I. M. Similarity Measures for Rational Set Selection and Analysis of Combinatorial Libraries: The Diverse Property-Derived (DPD) Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 599−614.

(7) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and Visualization of Molecular Diversity of Combinatorial Libraries. *Molecular Diversity* **1996**, *2*, 64−74.

(8) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569−574.

(9) Brown, R. D. Descriptors for Diversity Analysis. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 31−49.

(10) Brown, R. D.; 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.

(11) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand−Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1996**, *37*, 1−9.

(12) 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.

(13) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physicochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.

(14) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(15) Sprague, P. W. Automated Chemical Hypothesis Generation and Database Searching with Catalyst. *Persp. Drug Discovery Des.* **1995**, *3*, 1−20.

(16) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of Common Functional Configurations Amoung Molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 563−571.

(17) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.

(18) Murray, C. M.; Cato, S. J. Design of Libraries to Explore Receptor Sites. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 46−50.

(19) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; 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.

(20) Pickett, S. D.; Luttmann, C.; Guerin, V.; Laoui, A.; James, E. DIVSEL and COMPLIB−Strategies for the Design and Comparison of Combinatorial Libraries using Pharmacophoric Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 144−150.

(21) Mason, J. S.; Pickett, S. D. Partition-Based Selection. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 85−114.

(22) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−1223.

(23) Good, A. C.; Kuntz, I. D. Investigating the Extension of Pairwise Distance Pharmacophore Measures to Triplet-Based Descriptors. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 373.

(24) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. New Method for Rapid Characterization of Molecular Shapes: Applications in Drug Design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79−85.

(25) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D-Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547 (http://www2.ccc.uni-erlangen.de/software/corina/index.html).

(26) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.

(27) MDL Information Systems, Inc., 14600 Catalina St., San Leandro, CA 94577.

(28) Geladi, P.; Kowalski, B. R. Partial Least-Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.

(29) The Tanimoto coefficient is $N1\&2/(N1 + N2 - N1\&2)$, where $N1$ and $N2$ are number of bits set (=1) in bitstrings 1 and 2 respectively, and $N1\&2$ is the number of bits in the bitstring which results from the logical AND of the two bitstrings.

(30) Molecular Simulations Inc., 9685 Scranton Road, San Diego, CA 92121-3752.

(31) Daylight Chemical Information Systems, Inc., 27401 Los Altos, Suite 370, Mission Viejo, CA 92691.

CI990313H