Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry

Miklos Feher* and Jonathan M. Schmidt

SignalGene Inc., 335 Laird Road, Unit 2, Guelph, Ontario, N1G 4P7, Canada

Received July 3, 2002

The differences between three different compound classes, natural products, molecules from combinatorial synthesis, and drug molecules, were investigated. The major structural differences between natural and combinatorial compounds originate mainly from properties introduced to make combinatorial synthesis more efficient. These include the number of chiral centers, the prevalence of aromatic rings, the introduction of complex ring systems, and the degree of the saturation of the molecule as well as the number and ratios of different heteroatoms. As drug molecules derive from both natural and synthetic sources, they cover a joint area in property space of natural and combinatorial compounds. A PCA-based scheme is presented that differentiates the three classes of compounds. It is suggested that by mimicking certain distribution properties of natural compounds, combinatorial products might be made that are substantially more diverse and have greater biological relevance.

INTRODUCTION

Over the past decade, combinatorial chemistry has become the major source of NCEs in drug discovery.[1] However, despite the increased speed of synthesis this changeover from traditional to combinatorial synthesis has not yielded any real increase in the number of lead optimization candidates or drugs.[2] Although the reasons underlying this apparent lack of productivity remain unclear, we propose that it may in part reflect significant deficiencies in the types of chemical structures generated using combinatorial approaches. Since Lipinski first published his "rule of 5",[3] describing statistically the optimal combination of properties of drug-like molecules, additional methods for identifying distinguishing features between drugs and other organic molecules have been sought. Approaches have included the characterization of molecular frameworks and substituents,[4,5] the statistical analysis of different drug databases,[6] the application of artificial neural networks to differentiate drugs from reagents,[7−9] and the introduction of a drug-like index to rank compounds.[10]

Traditionally, natural products have been a major source of new drugs, and many successful drugs were originally synthesized to mimic the action of molecules found in nature.[11] Natural compounds are highly diverse and often provide highly specific biological activities. This follows from the proposition that essentially all natural products have some receptor binding capacity.[12] Natural molecules, however, differ substantially from synthetic ones. These differences appear to be amplified when the products of combinatorial synthesis are considered. Although an important aim in combinatorial chemistry is to generate highly diverse libraries, the need for speed and automation introduces new structural idiosyncrasies into the method. An algorithm for the automatic classification of natural and synthetic compounds has been recently published using Shannon entropy analysis and binary QSAR[13] using different topological, distance, charge distribution, and other descriptors. Statistical comparisons between synthetic drugs and natural products have been performed for a limited set of properties.[14,15] The comparison with these works is difficult to make because of the different philosophy in defining the different compound classes. Henkel et al.[14] investigated the category 'synthetic compounds', but the selection criteria, diversity, and whether some of these were from combinatorial origin cannot be established, only that they are a 'representative pool of synthetic test compounds from Bayer AG'. Also, the natural compound selection in Lee and Schneider[15] did not distinguish natural products and natural product derivatives, molecules that contain both natural and synthetic elements. None of the previous works has investigated the differences between compounds from combinatorial synthesis and natural products or drugs. The object of this paper is to perform a systematic and simultaneous statistical comparison between the three classes of compounds, marketed drugs, combinatorial compounds, and natural products, based on a wide variety of simple properties. Although some of the obtained distinguishing features, such as the rigidity and the high number of chiral centers in natural products may appear somewhat obvious, this work is the first systematic attempt to compare these and other properties quantitatively among the three major sources of novel compounds in modern drug discovery. To evaluate the differences in the occupation of diversity space, we also developed a PCA model based on structural and physicochemical descriptors. It is hoped that the identification of the differences will help to find approaches to bring the combinatorial class closer to the other two. This could serve to address deficiencies in the diversity of combinatorial libraries and potentially increase the hit rate in such compound collections.

* Corresponding author phone: (519)823-9088; fax: (519)823-9401; e-mail: miklos.feher@signalgene.com.

10.1021/ci0200467 CCC: $25.00
Published on Web 12/11/2002

COMPUTATIONAL SECTION

Four different sets of molecules were assembled for this study. Drug molecules were taken from the Chapman and Hall Dictionary of Drugs,[16] the content of which essentially corresponds to The Merck Index from the same publisher.[17] The combinatorial database was assembled from the following publicly available collections: Maybridge HTS database,[18] the ChemBridge EXPRESS-Pick database,[19] the ComGenex collection,[20] the ChemDiv International Diversity Collection,[21] the ChemDiv CombiLab Probe Libraries,[21] and the SPECS screening compounds database.[22] Natural compounds were assembled from the following sources: the BioSPECS natural products database,[22] the ChemDiv natural products database,[21] and the Interbioscreen IBS2001N and HTS-NC databases,[23] with only those compounds kept that were marked as natural products or from natural sources. The source databases for the compound class 'natural products and their derivatives' (the derivatives are also often referred to as 'seminatural' compounds) were the same as those used for the natural compounds, except that all molecules in these collections were retained.

After assembling these four databases, the duplicates in each were removed. During this process, chirality was also taken into account and racemic molecules (or cases where the stereosymmetry was not indicated) were considered as one of the possible stereoisomers. The following numbers of unique compounds were present in each database: 10 968 drug molecules, 670 536 combinatorial compounds, 3287 natural products, and 27 338 molecules in the category of 'natural and seminatural' molecules. As some drug molecules originate from natural sources, there is some overlap between the drug and natural products databases. Overall, 635 molecules (5.8% of the drug database) appear in both the natural product and the drug collections, and a further 214 drug molecules (1.9% of the drug database) are listed among natural product derivatives.

The calculation of all molecular properties was conducted with the MOE modeling suite,[24] using dedicated programs written in SVL. The flexibility of molecules was expressed similarly to Oprea,[7] except that single bonds in any ring were ignored. The number of rotatable bonds was calculated from the full number of rotatable single bonds, with freely rotatable end groups (such as CH_3, NH_3, OH, etc.) excluded from the calculations. All references to the number of rings in this paper relate to the smallest set of smallest rings (SSSR). Ring systems are sets of rings with fused or spiro connections that are linked to other ring systems with only single and double bonds (not with ring bonds). The number of ring systems (NRS) was calculated using the following equation that we derived from the Euler formula

$$\mathrm{NRS} = (e_t - e_r) - (v_t - v_r) + 1$$

where e_t and v_t are the total number of edges (bonds) and vertices (atoms) in the molecule, and e_r and v_r are the number of edges and vertices that are in rings. The normalized number of ring systems was obtained from the number of ring systems through division by the number of rings. The degree of ring fusion is calculated by dividing the number of rings by the number of ring systems.[11] The chain length in this paper was defined as the longest heavy-atom chain in the molecule with none of the constituent atoms of the chain belonging to rings and was determined by SMARTS-type pattern matching.

The degree of unsaturation of a molecule was quantified using a new formalism.[25] It is functionally equivalent to the properties 'double bond equivalents' and 'index of hydrogen deficiency', except that it is independent of the graphical representation of the molecule (such as resonance structures). The degree of unsaturation (DU) was calculated using the following formula[25]

$$\mathrm{DU} = \mathrm{DB} + 2\mathrm{TB} + \mathrm{RING} + (1\text{-}\mathrm{DIS}) + 1/2\ \mathrm{ELE}$$

where DB is the number of double bonds, TB is the number of triple bonds, RING is the number of rings, DIS is the number of disjoint parts, and ELE is the number of excess localized electrons. The degree of unsaturation without aromatic bonds was calculated by removing from the above sum those DB increments that were introduced as a result of including aromatic bonds.

Hydrogen bond acceptors and donors were taken into account using two different approaches. The Lipinski-type donors (expressed as the sum of OH and NH groups) and Lipinski-type acceptors (expressed as the sum of N and O atoms) were calculated as defined by Lipinski et al.[4] In contrast, $n_{acc,solv}$ and $n_{don,solv}$ were evaluated as the number of acceptors and donors in a solvated environment, respectively, taking protonation and deprotonation of the molecule into account.[24]

Octanol−water partition coefficients were calculated using the Wildman and Crippen SlogP method.[26] Bond statistics were obtained by SMARTS-type pattern matching and counting within the database. Histograms for step variables were calculated by counting the number of occurrences within each bin. In case of continuous variables, bin boundary effects were accounted for by assuming a normal distribution over the boundary with a width corresponding to a quarter of the width of the bin. Values falling into these boundary regions were distributed between the two bins according to their normalized weight, determined from the distribution. To represent the property distributions numerically, the mean was calculated for different properties. Although variance or standard deviation would be useful in judging the dispersion of the property distributions, these would be of little value as most distributions in this work are very far from normal. In such cases, the median is more representative than the mean[27] and hence is also given.

The principal component analysis was performed using the MOE suite.[24] Some of the descriptors in Table 1 that were used to characterize the databases are highly correlated with each other. Although this typically does not present a problem for principal component analysis itself, it makes it more difficult to understand the significance of the variables. Thus the intercorrelation of all variables on the three data sets was calculated. Variables were then removed one by one from the set until none of the squared correlation coefficients between the variables was above 0.7. Although the order in which variables were removed and the final value for intercorrelation are somewhat arbitrary, this process was aimed primarily at removing variables that carry the same information (e.g. number of bonds vs ratio of the same bonds, number of atoms vs number of bonds, etc.). In this work, the selection of variables was not optimized in any way. The

Table 1. Mean/Median Values of Different Properties among Natural, Drug, and Combinatorial Compounds as Well as in the Category 'Natural Products and Derivatives'[a]

property	combinatorial ($n = 670536$)	drugs ($n = 10968$)	natural ($n = 3287$)	seminatural ($n = 27338$)
molecular weight	393/389	340/312	414/362	381/134
number of heavy atoms	27.2/27	23.5/22	29.6/26	27.3/26
number of chiral centers	0.4/0	2.3/1	6.2/4	2.2/1
number of rotatable bonds	6.4/6	5.6/5	4.4/3	5.3/4
chainlength	4.22/4	4.52/4	3.09/2	3.75/3
degree of unsaturation	12.0/12	8.0/8	8.8/8	10.8/11
degree of unsaturation, excl. aromatic bonds	5.4/5	4.7/4	6.3/6	6.2/6
number of rings	3.2/3	2.6/2	4.1/4	3.6/4
ratio of aromatic atoms to ring atoms	0.80/0.82	0.55/0.60	0.31/0.27	0.60/0.60
ratio of in-ring and out-of-ring heavy atoms	0.63/0.64	0.54/0.58	0.65/0.66	0.65/0.67
number of ring systems	2.6/3	1.7/2	1.7/1	2.0/2
normalized number of ring systems	0.85/1.0	0.75/0.83	0.47/0.4	0.59/0.57
ring fusion degree	1.27/1	1.67/1.2	2.83/2.5	2.04/1.75
number of carbon atoms	20.48/20	17.2/17	22.8/21	20.9/20
number of nitrogen atoms	**2.69/3**	**1.64/1**	**0.84/0**	**1.87/2**
number of oxygen atoms	**2.77/3**	**4.03/3**	**5.9/5**	**4.13/4**
number of sulfur atoms	**0.45/0**	**0.23/0**	**0.03/0**	**0.14/0**
number of halogen atoms	**0.80/0**	**0.34/0**	**0.02/0**	**0.24/0**
ratio of carbon atoms to all heavy atoms	0.75/0.75	0.72/0.74	0.78/0.79	0.77/0.77
ratio of nitrogen atoms to all heavy atoms	**0.10/0.10**	**0.08/0.07**	**0.03/0**	**0.07/0.06**
ratio of oxygen atoms to all heavy atoms	**0.10/0.10**	**0.16/0.15**	**0.19/0.18**	**0.15/0.14**
ratio of sulfur atoms to all heavy atoms	0.02/0	0.01/0	0.00/0	0.01/0
ratio of halogen atoms to all heavy atoms	0.03/0	0.02/0	0.00/0	0.01/0
number of Lipinski-type acceptors	5.5/5	5.7/5	6.8/5	6.0/5
number of Lipinski-type donors	1.0/1	1.9/1	2.6/2	1.4/1
$\mathbf{n}_{acc,solv}$	3.7/4	5.7/5	6.8/5	4.7/4
$\mathbf{n}_{don,solv}$	1.0/1	1.9/1	2.6/2	1.5/1
number of C−C bonds	18.5/18	15.6/15	22.5/21	19.4/19
number of C−N bonds	5.3/5	3.3/3	1.88/0	4.16/3
number of C−O bonds	3.1/3	4.89/3	8.2/6	5.65/5
number of C-halogen bonds	0.80/0	0.34/0	0.02/0	0.24/0
number of C−S bonds	0.75/0	0.35/0	0.05/0	0.23/0
ratio of C−C to all heavy atom bonds	0.62/0.63	0.61/0.62	0.69/0.69	0.64/0.66
ratio of C−N to all heavy atom bonds	0.18/0.17	0.15/0.13	0.07/0	0.15/0.12
ratio of C−O to all heavy atom bonds	0.11/0.10	0.18/0.17	0.24/0.23	0.19/0.18
ratio of C-halogen to all heavy atom bonds	**0.03/0**	**0.02/0**	**0.00/0**	**0.01/0**
ratio of C−S to all heavy atom bonds	**0.03/0**	**0.02/0**	**0.00/0**	**0.01/0**
SlogP	4.3/4.2	2.2/2.3	2.4/2.7	3.3/3.3
number of in-ring Lipinski acceptors	1.6/2	1.3/1	1.6/1	2.0/2
number of out-of-ring Lipinski acceptors	3.8/4	4.4/4	5.1/4	4.0/3
ratio of out-of-ring and in-ring Lipinski acceptors	2.3/2	3.4/4	3.2/4	2.0/1.5

[a] The properties that show important differences between these compound classes are boldfaced.

Table 2. Normalized Loadings for the First Three Principal Components[a]

	PC1	PC2	PC3
number of chiral centers	−0.12	−0.04	0.01
number of rotatable bonds	−0.04	0.12	0.01
ratio of aromatic atoms to ring atoms	0.71	0.98	−0.59
ring fusion degree	−0.13	−0.46	0.27
$n_{acc,solv}$	−0.14	0.07	0.00
$n_{don,solv}$	−0.19	0.09	−0.03
number of C−N bonds	0.04	0.12	0.12
number of C−O bonds	−0.10	0.02	−0.03
number of C-halogen bonds	0.13	0.16	−0.52
number of C−S bonds	−0.09	0.24	0.62

[a] The principal component analysis was performed on a database containing drugs ($n = 10\,968$), natural products ($n = 3287$), and a random selection of combinatorial compounds ($n = 13\,506$). The first two and three principal components explain about 54% and 66% of the variance, respectively.

selected variables used in the PC analysis are displayed in Table 2. To reduce the number of compounds to be plotted, a random selection of 2% of the full combinatorial collection was used for the analysis (a total of 13 506 compounds). To ensure that the selection was representative, the means of all marked properties were calculated for this subset. All were found to be within 1% of the value for the full set.

RESULTS AND DISCUSSION

The distributions for selected properties among drugs, natural, and combinatorial compounds are shown in Figures 1−8. The average values of these properties (mean and median) for these databases as well as for the compound class 'natural products and derivatives' are displayed in Table 1.

Molecular Weight. The molecular weight distribution for drugs follows a Gaussian distribution (see Figure 1), and the distribution characteristics are similar to those in the CMC database[7] in accordance with previous observations.[7] In agreement with the results of Lee and Schneider,[15] the weight distribution for natural compounds peaks at a similar position as drugs and is skewed toward higher molecular weights. (Obviously natural product databases do not contain entire DNA and protein chains, otherwise the molecular weight distribution would be shifted toward much greater

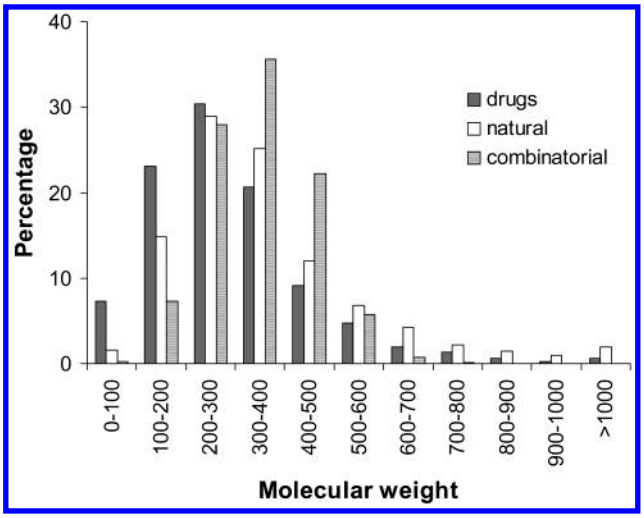

Figure 1. The molecular weight distribution among drug molecules, natural products, and compounds from combinatorial synthesis.

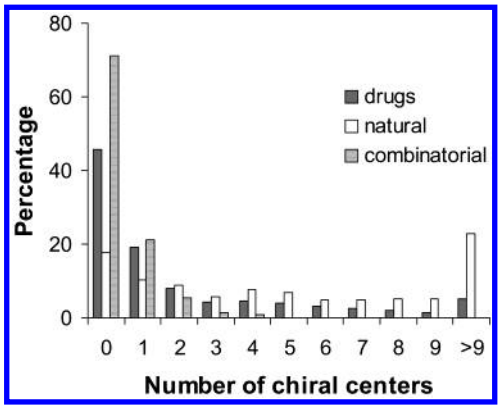

Figure 2. The distribution of the number of chiral centers among drug molecules, natural products, and compounds from combinatorial synthesis.

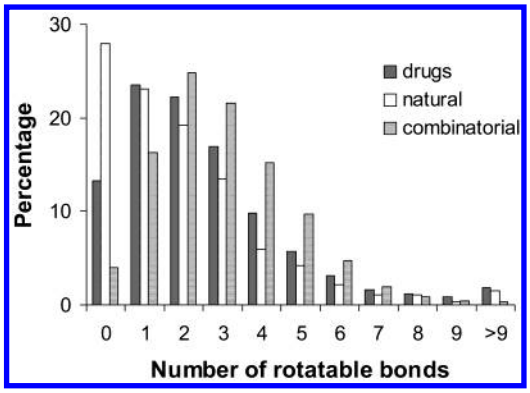

Figure 3. The number of freely rotatable bonds among drug molecules, natural products, and compounds from combinatorial synthesis.

weights.) The molecular weight distribution for combinatorial libraries peaks at a higher value and has a narrower distribution than that of drugs.

The calculated average values for molecular weight (see Table 1) reflect the observations above. Because of the wider distribution in natural compound libraries, the mean of molecular weight is highest for these compounds. However, based on the median, the combinatorial collection appears to contain heavier molecules. These observations are also reflected by a related property, the average number of heavy

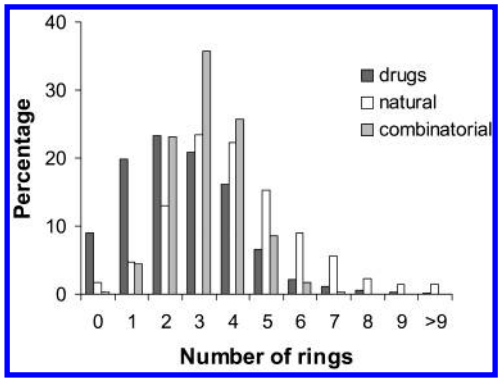

Figure 4. The distribution of the number of rings among drug molecules, natural products, and compounds from combinatorial synthesis.

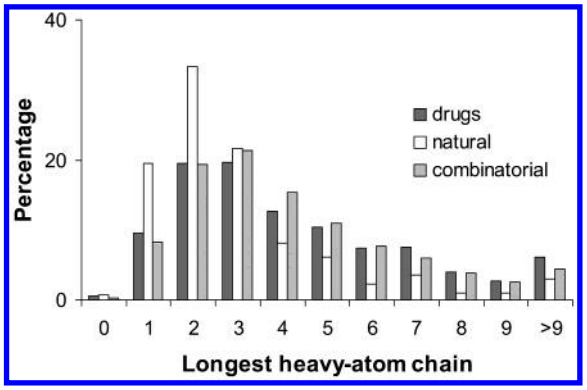

Figure 5. The distribution of the longest heavy-atom chain lengths among drug molecules, natural products, and compounds from combinatorial synthesis.

atoms (see Table 1). It must be noted that the averages and distribution curves for molecular weights within combinatorial data collections cover substantial differences between such databases. For example, in the ExpressPick database the mean and median of the molecular weight are 353 and 347, whereas the corresponding values in the ComGenex database are 472 and 479, indicating that this latter database is much more skewed toward high molecular weight compounds.

Chiral Centers. The percentage distribution for the number of chiral centers is shown in Figure 2. There is a marked difference in the distribution among the three classes of compounds. Although about 45% of drug molecules have no chiral centers and 19% have one, the distribution drops only slowly for higher numbers. This sharply contrasts with combinatorial collections: the difficulty of chiral separation in combinatorial synthesis heavily favors molecules with no chiral centers (over 71%) and the distribution falls off rapidly. Natural products, on the other hand, behave remarkably differently. Biological processes, in which stereospecific reagents and catalysts are most common, frequently generate active molecules with high numbers of chiral centers. In many cases, the presence of such chiral centers contributes to the selectivity of these molecules for their predominantly stereospecific binding sites. The median for the number of chiral centers among natural compounds is 4, in contrast to 1 in drugs and 0 in combinatorial compounds. The averages also reflect this enormous difference between the three classes of molecules: the mean number of chiral centers is 6.2 in natural compounds, 2.3 in drugs, and a mere 0.4 in combinatorial molecules. A similar ratio between natural,

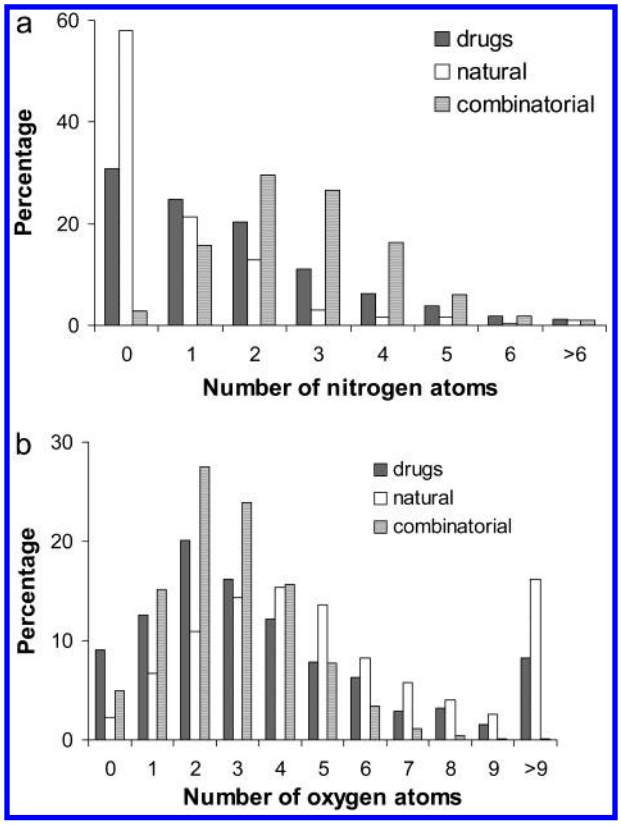

Figure 6. The distribution of the number of a) nitrogens and b) oxygens among drug molecules, natural products, and compounds from combinatorial synthesis.

drug, and synthetic compounds was found by Henkel et al.,[14] although the actual means were different (3.2, 1.2, and 0.1, respectively), reflecting the different origin and compound selection criteria in that work. The avoidance of chiral centers in synthetic compounds is the greatest single difference between the three data collections and has a profound impact on a number of other molecular properties, such as the prevalence of aromatic rings, as described later.

Rotatable Bonds, Unsaturations, Rings and Chains. The distribution of the number of rotatable bonds is shown in Figure 3, with the average numbers given in Table 1. Natural compounds display a wide range in flexibility, with a steadily decreasing distribution from the peak at zero rotatable bonds. In contrast, such rigid structures are present in fewer than 4% of the combinatorial compounds. The average numbers also reflect this difference: molecules from combinatorial synthesis have on average two more rotatable bonds (mean) than natural compounds. As usual, drugs deriving from both natural and synthetic sources overlap both distributions.

Generally, if a flexible and a rigid ligand can form the same pattern of hydrogen bonds and hydrophobic interactions with the protein, the rigid ligand will exhibit much stronger binding due to lower entropic losses.[28] Hence the flexibility of the molecule is an important factor in determining ligand binding. The presence of a large proportion of rigid natural compounds suggests that at least some of these compounds may exploit the thermodynamic advantages conferred by rigidity to achieve superior binding properties. In contrast, the process of combinatorial synthesis generally introduces new rotatable bonds joining the basic building blocks. The combinatorial synthesis of highly constrained, fused ring systems is not generally feasible, and hence molecules

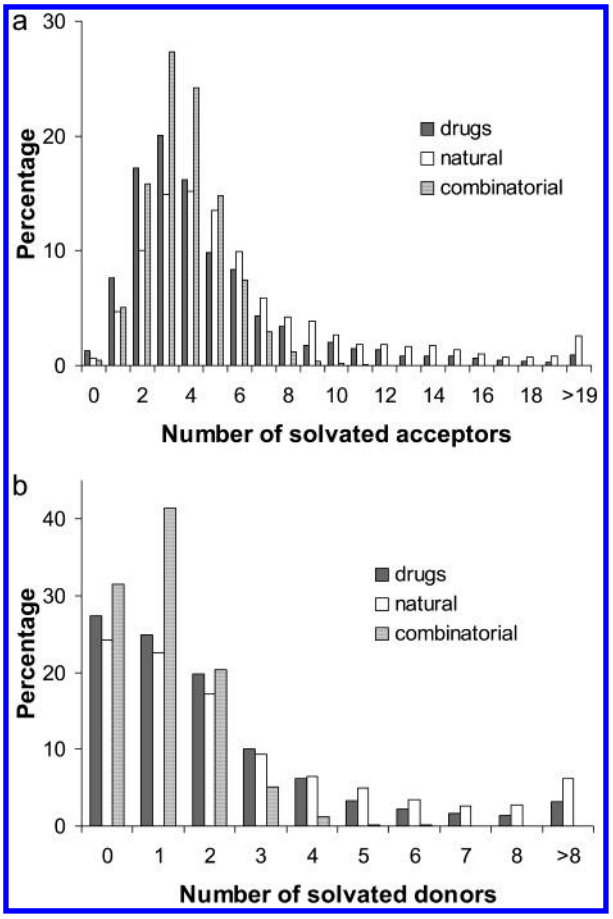

Figure 7. Distribution of the number of solvated hydrogen-bonding functionalities among drug molecules, natural products, and compounds from combinatorial synthesis a) hydrogen-bond acceptors and b) hydrogen-bond donors.

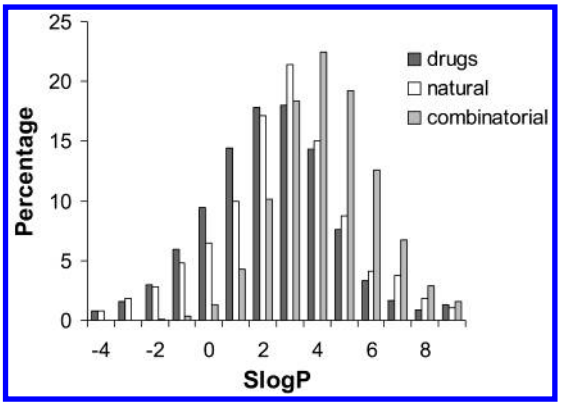

Figure 8. The distribution of calculated octanol−water partition coefficients among drug molecules, natural products, and compounds from combinatorial synthesis.

produced by synthetic methods are expected to have at least a few rotatable bonds.

Although the flexibility of molecules is highly dependent on the number of rotatable bonds, the distribution of these bonds within a molecule is also important. As an example, four rotatable bonds forming a chain is likely to lend more flexibility to a molecule than if these rotatable bonds are separately attached to a ring system. To characterize this property, the longest heavy-atom chain length was identified for each molecule. The values in Table 1 show that on average combinatorial compounds and drugs possess longer chains than natural products. The distribution of this property

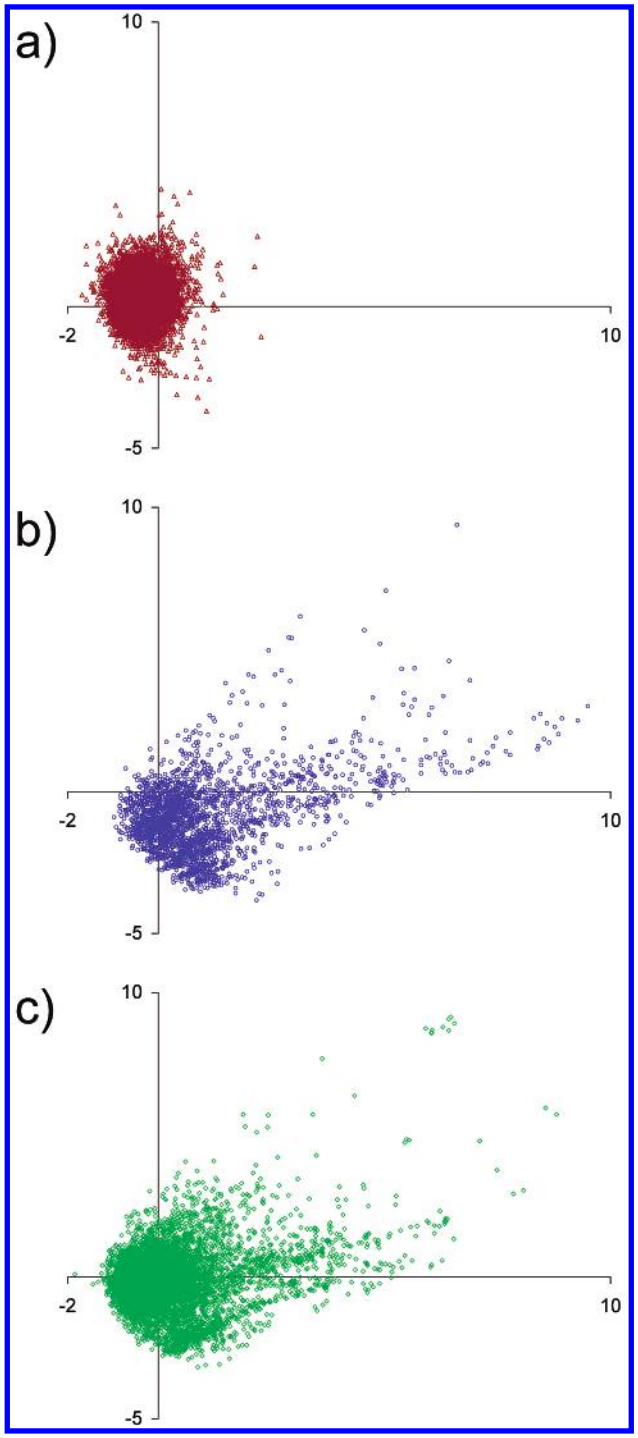

Figure 9. The plot of the first two principal components, obtained from a database containing a) a random selection of combinatorial compounds (n = 13 506), b) natural products (n = 3287), and c) drugs (n = 10 968). For clarity, the data points from the three databases are plotted separately but on the same axes. The properties used to derive the PC's and the obtained loadings are given in Table 2. The first two components explain about 54% of the variance. The figure shows that combinatorial compounds cover a well-defined area in the diversity space given by these principal components. In contrast, natural products and drugs cover almost all of this space as well as a much larger additional volume. Drugs and natural products have approximately the same coverage of this space.

in Figure 5 also demonstrates this. The distribution for natural compounds sharply peaks at 2-atom chains and quickly tails off, whereas the distributions for combinatorial products and drugs peak at 3-atom chains, run parallel to each other, and decrease more slowly. It should be noted that these distributions deviate substantially from a Gaussian distribution.

The level of unsaturation is an alternative method of characterizing the rigidity of the system. The degree of unsaturation, used in this study, quantifies the double bond equivalents independently of resonance-structure. As can be seen from Table 1, the mean of this quantity is 12 for the combinatorial database but only 8 for the natural compounds. Thus, at first sight, combinatorial compounds appear to be more unsaturated, which contrasts with the previous observations for flexibility. However, this difference is largely attributable to aromatic ring occurrence. The majority of rings in molecules produced by combinatorial synthesis are aromatic. Depending on ring size, aromaticity introduces 2−3 more unsaturations compared to the corresponding aliphatic rings. If the degree of unsaturation is calculated excluding aromatic bonds, the results obtained are quite different (see Table 1). Based on this measure, natural products are on average more unsaturated than combinatorial synthesis products. Interestingly, drugs do not appear to be intermediate between combinatorial and natural compounds in this respect, instead they are on average less saturated than either of the other classes.

The prevalence of rings is another measure for the rigidity of molecules. Figure 4 shows the distribution of the number of rings in the three databases. The most apparent difference between drugs and other molecules is the greater frequency of structures with a small number of rings. Although the distribution curves for natural compounds and combinatorial molecules peak at about the same value (n = 3), the means and the medians (see Table 1) indicate that natural products have on average one more ring. This difference results mainly from the extended tail of the distribution curve for natural compounds. In contrast, there is a sharp cutoff for molecules produced by combinatorial synthesis. As in the case of molecular weight distribution, there are marked differences among different combinatorial collections in the average number of rings. As an example, in the ExpressPick collection there are fewer rings (mean 2.93, median 3) than in the ChemDiv C-lab collection (mean 3.65, median 4).

The differences in the number of rings also relate to differences in the distribution of different ring systems. As previously mentioned, the design of combinatorial libraries generally avoids the introduction of chiral centers, to avoid expensive chiral separations. In consequence, aromatic rings tend to be favored over unsaturated ones in combinatorial syntheses with 80% of ring atoms being aromatic in these products. Such bias is clearly absent among natural compounds (only 31% of ring atoms are aromatic). The average for drugs is approximately halfway between the two collections (55%), again reflecting the mixed origins of these compounds from natural and synthetic sources.

The importance of rings can also be described by the proportion of heavy atoms occurring in rings. Whereas over 65% of the heavy atoms are in rings in natural products, only 56% are in drugs. Interestingly, as it is relatively simple nowadays to connect available ring systems in an automated synthesis, the ratio of in-ring heavy atoms is also quite high in the combinatorial collections (65%). The major difference between natural and combinatorial products in this respect is in the topological distribution of rings in the molecule, which will be discussed in the next section.

Ring Topology. In general, natural products tend to be rigid and characterized by relatively few rotatable bonds, in large part due to the prevalence of ring systems. The automated method of producing combinatorial libraries introduces distinctive features in the topology of molecules. These arise in combinatorial syntheses because existing ring systems are connected by establishing single bonds between them. One method of quantifying the impact of this approach is by counting the number of ring systems, defined as a set of rings with fused or spiro connections. Such ring systems are counted as separate entities if they are linked to other ring systems with only single and double bonds, not with ring bonds. As shown in Table 1, natural compounds contain significantly fewer ring systems (mean: 1.7, median: 1) than combinatorial compounds (mean: 2.6, median: 3), despite the fact that natural compounds contain more rings. This reflects the occurrence of more fused, bridged, or spiro ring systems in natural products than among combinatorial compounds.

Normalizing the number of ring systems removes its dependence on the number of rings. Normalization is achieved by dividing the ring system count by the total ring number (in this case, molecules with no rings have to be removed from the set.). This normalized score ranges from 1 (for a molecule with unfused and nonspiro rings, connected by nonring bonds) to zero (which approximates the score of a multiring system with an infinite number of rings fused together). The average value for the combinatorial set is close to the first extreme (mean: 0.85, median 1), whereas the value for natural compounds is substantially closer to the other extreme (mean: 0.47, median: 0.4). The reciprocal of this normalized quantity is a measure of the degree of fusion[11] and describes the average number of fused rings (or those with spiro connection) in the molecule. This number is 2.8 (mean) among natural compounds and only about 1.3 (mean) among combinatorial ones. This greater rigidity in natural products contributes to their higher selectivity and binding free energy. It may also contribute in some cases to the induction of changes in receptor conformation, which may be important in receptor function, such as agonism-antagonism.[29] Obviously, mimicking this in synthetic compounds would introduce substantial synthetic costs, such as the need for multiple cyclization reactions and the formation of a number of carbon−carbon bonds.

Distribution of Different Atom Types. The average number of different atom types together with their ratio to the total number of heavy atoms is shown in Table 1, and their distribution is displayed in Figure 5. Combinatorial compounds, natural products, and drugs differ significantly with respect to elemental composition. On average, combinatorial products contain three times as many nitrogen atoms per molecule or per heavy atom as natural products. In contrast, natural compounds contain on average nearly twice as many oxygen atoms. This divergence probably reflects fundamental differences between combinatorial synthesis and biosynthesis. Many biosynthetic paths are based on reagents in which there are similar numbers of oxygen and nitrogen atoms, such as amino acids, or which contain more oxygen than nitrogen, for example in polyketide synthesis. (Reactions involving nucleotides would be one of the few counterexamples.) Ester and amide formation are key conjugation reactions in biosynthesis and again tend to equalize the net inclusion of oxygen relative to nitrogen.[30] Photosynthesis and the pathways leading to different carbohydrates are also responsible for the higher occurrence of oxygen in natural products.

Regardless of the synthetic basis of these differences in composition, they are likely to have important consequences for interactions with binding sites. This is particularly significant in the context of the evolution of receptor−ligand specificity. Specificity in binding interactions derives primarily from the correct matching of complementary polar and nonpolar surfaces in the bound complex. Conversely, a major requirement for high-affinity specific binding is the avoidance of mismatched pairings between polar and nonpolar regions or repulsive interactions between polar groups, for example hydrogen donors placed next to adjacent to each other. Hydroxyl groups play an especially important role in this regard, since they can act as both donors and acceptors.

Natural compounds and combinatorial synthesis products also differ greatly in the relative abundance of sulfur. Whereas sulfur is frequently introduced into combinatorial compounds in the form of sulfides, sulfonamides, and as heterocycles, it is rarely incorporated in natural products. This probably reflects the paucity of biosynthetic precursors such as amino and carboxylic acids containing sulfur. In light of the relatively limited polar character of sulfur in functional groups such as thiones and sulfides, it is unlikely that the occurrence of sulfur in combinatorial compounds compensates for the relative lack of other polar heteroatoms. It can be argued that nitrogen, sulfur, and halogens are often introduced in synthetic reactions for the sake of making them more amenable to combinatorial sythesis, for changing the pharmaceutical properties of compounds, or for generating series of molecules for SAR, and this partly explains the prevalence of these elements in synthetic compounds.

Acceptors and Donors. As discussed above, the major function of oxygen and nitrogen atoms in ligand−receptor binding is to provide specificity by participating in hydrogen bonding. A method often applied to quantify these effects is to count the number of hydrogen bond acceptors and donors, as described by Lipinski.[4] According to this definition, the number of acceptors is expressed as the sum of nitrogens and oxygens, whereas the number of donors is taken as the sum of NHs and OHs. These quantities have been shown to be critical in a drug development setting as they influence absorption and permeation.[4] The average numbers, given in Table 1, indicate that natural products have more acceptors and substantially more donors than combinatorial compounds, with the values for drugs being between these values.

Using Lipinski-type donors and acceptors is somewhat crude, as it does not properly describe the hydrogen-bonding properties of molecules in a solvated environment. As an example, using the Lipinski description, primary amines are simultaneous acceptors and donors. In reality, however, primary amines in aqueous solutions protonate and act solely as hydrogen bond donors. Considering protonation−deprotonation when calculating these numbers resolves this issue. The mean numbers of solvated acceptors ($n_{acc,solv}$) and donors ($n_{don,solv}$) are given in Table 1. On average, natural products contain 3 more acceptors and 1.5 more donors than combinatorial ones, with the corresponding values for drugs lying between the two.

The distribution of solvated hydrogen bond acceptors and donors is shown in Figure 7. For combinatorial compounds both distributions tail off rapidly, whereas there are numerous natural products and drug molecules with many acceptors and donors. The rapid tailing-off of the combinatorial curve may be partly the result of the extensive use of the Lipinski rules in compound selection. It is notable that the distribution curve for donors has its maximum at zero for natural compounds and drugs, in contrast to 2 for combinatorial products.

An examination of the frequency of occurrence of nitrogen and oxygen in the context of acceptor−donor properties for combinatorial and natural compounds suggests that the "palette" of functional group types presented by these two groups differ substantially. Under physiological conditions of solvation and pH, nitrogens occur predominantly as hydrogen donors in the form of protonated amines. In contrast, oxygens occur predominantly as acceptors (e.g. carbonyl oxygen) or as simultaneous donor and acceptor groups (e.g. hydroxyls). On average, combinatorial compounds have fewer polar functional groups available for matching with the surfaces of binding site. In contrast, natural products have on average significantly more polar functional groups in total and, compared to combinatorial products, are substantially enriched with functional groups that can act as hydrogen bond acceptors. In this regard it is significant that drugs on average contain more oxygen and nitrogen atoms than combinatorial products and more closely resemble natural compounds in the relative abundance of solvated acceptors and donors. Since natural products and biological receptors have in part coevolved toward selective binding interactions, the availability of an appropriate mixture and supply of polar functional groups may be a key factor for achieving specific, high-affinity binding.

As described previously, combinatorial compounds contain on average fewer nitrogens and oxygens than natural products. However, there are also major differences in the distribution of these atoms inside and outside of rings. The sum of nitrogen and oxygen atoms is the number of Lipinski-type acceptors. As shown in Table 1, the number of nitrogens and oxygens in rings is similar in combinatorial and natural compounds. However, combinatorial compounds contain on average 1.3 more of these atoms outside rings. As the ratio of heavy atoms inside and outside of rings is similar in these two compound classes, this provides a clear demonstration that in combinatorial libraries these heteroatoms are more prevalent in rings than outside rings. This difference is probably related to the synthetic ease of building heteroatom-containing molecules. Heteroatoms are not usually built into rings during the combinatorial syntheses. Instead, heterocyclic rings are frequently used as building blocks, and carbon−heteroatom bonds are usually established during the condensation and conjugation reactions. In summary, further to the fact that natural products on average contain more acceptors and donors than combinatorial ones, these are predominantly concentrated in rigid structural elements in natural products.

Distribution of Different Bond Types. It has been suggested recently that "the examination of nature's favorite molecules reveals a striking preference for making carbon-heteroatom bonds over carbon−carbon bonds".[31] This is explained by the fact that "carbon dioxide is nature's starting material and most reactions are performed in water".[31] In view of this remark, it is interesting to consider the frequency of different bond types in the three groups of compounds. It must be emphasized, however, that the statistical analysis was carried out on a set of small molecule natural products, where the vast range of natural polypeptides and proteins are represented only by 130 entries, amino acids, and small peptides. Similarly, the rich selection of nucleic acids, carbohydrates, lipids, and other classes are only represented in the library by their elementary building blocks. Nonetheless, the results are interesting since they represent statistics on the natural products currently available as a library of potential test compounds for drug discovery.

The average number of different bond types and their ratio to the total number of heavy atom bonds is shown in Table 1. Both the number and the ratio of carbon−carbon bonds compared to other bonds are the highest among natural products, in contrast to the statement in the reference.[31] Even more striking, however, is how these carbon-heteroatom bonds are distributed among different heteroatoms. Whereas the ratio of C−N bonds appears to be surprisingly low among natural compounds, the ratio of C−O bonds among natural compounds is over twice that in the combinatorial collection. The same trend is shown by the total numbers: the average number of C−N bonds is just over a third, while the average number of C−O bonds is over 2.5 times greater among natural compounds than the combinatorial ones. The difference between the two collections is even more pronounced if the number of C-halogen or C−S bonds is compared (see Table 1): the ratio of average numbers between the combinatorial and natural databases is 17:1 and 33:1, respectively. These observations correlate with the ratios of different atom types (vide ante).

Lipophilicities. The distribution of calculated SlogP values among the three groups of compounds is shown in Figure 6. This distribution is comparable to the one obtained for experimental logP values from the CMC database.[7] As shown by this figure and the averages in Table 1 drugs are generally as lipophilic as natural products, whereas combinatorial compounds are decidedly more hydrophobic. This is directly related to the lower average number of nitrogens and oxygens present in combinatorial compounds. It is well-known that lipophilicity plays a decisive role in the ADME/T properties of molecules.[32] Hence the higher average hydrophobicity of combinatorial compounds may have profound effects on the success rate of converting initial hits from high-throughput screens to leads that can be further developed.

Principal Component Plot from Different Databases. The distribution curves of different properties displayed in Figures 1−8 and the mean and median values of many other properties in Table 1 amply demonstrate that natural products and combinatorial libraries have substantially different properties. To visualize how diverse these compound classes are based on these properties, principal components analysis (PCA) of these data sets was performed.

The plot of the first two principal components is shown in Figure 7. The first two components explain about 54% of the variance, the first three 66%. The normalized loadings, presented in Table 2, indicate that the first three principal components contain high loadings from the ratio of aromatic ring atoms to the total number of heavy atoms. In addition, the first PC has significant loadings from the number of

solvated hydrogen bond donors and acceptors, while the second and third PCs contain contributions from the ring fusion degree and the number of carbon−sulfur and carbon−halogen bonds. For reasons of clarity, the first two PCs are plotted on the same scale but separately for the three databases in Figure 7. The combinatorial set, plotted in Figure 7.a, occupies a much smaller part of this space than the other two databases. Most molecules fall into a well-defined region, little over half of which is also occupied by natural products and drugs. Much of the area covered by drugs and natural products contains no representative combinatorial compounds. This result is particularly interesting because it demonstrates that combinatorial compounds are substantially less diverse than either drugs or natural products. Furthermore, the diversity of combinatorial compounds is confined to an area where there appears to be little diversity shown by natural products. Interestingly, existing drugs seem to cover the joint volume in diversity space of combinatorial compounds and natural compounds. The current failure of combinatorial libraries to generate robust preclinical candidates at anticipated rates may reflect the restricted region of chemical space explored by products of combinatorial synthetic methods.

CONCLUSIONS

Natural products generally have high binding affinities for specific receptor systems and their biological action is often highly selective. It is common knowledge in medicinal chemistry that "the removal of chiral centers, introducing additional flexibility into the molecule and decreasing its size generally leads to a less specific and weaker activity".[33] In view of these facts, it is interesting to consider that the search for the replacement of natural compounds with synthetic ones is usually based on exactly these kinds of 'unfavorable' modifications. Given the stereospecificity of most biological targets, it is likely that many nonstereospecific synthetic analogues, created, for example by the introduction of aromatic rings, represent suboptimal compromises, especially in terms of selectivity. This situation appears to be amplified in the case of combinatorially synthesized compounds. In addition, the greater flexibility of combinatorial products is likely to have detrimental entropic consequences for the binding of these compounds, it may also adversely affect their ability to induce conformational changes in the receptor required for biological function. As well, the process of producing synthetic analogues radically alters the numbers and ratios of different atom types, such as nitrogens, oxygens, sulfurs, and halogens. These distributions in turn have a direct impact on the patterns of donors and acceptors available to complement receptor surface properties.

The question arises how the results of this statistical analysis can be applied to improve the diversity of synthetic and especially combinatorial compounds. As has been indicated above, chirality is an essential issue. As chiral separation is rather expensive, the number of chiral centers could be somewhat increased by using readily available chiral reagents more often in combinatorial syntheses. Attempting to change the ratios of different heteroatoms is another promising way of increasing diversity in combinatorial products. This will also impact donor/acceptor properties as well as lipophilicities and hence raise the diversity of designs. Drug-like filters, such as the Lipinski rules, are very helpful in isolating likely problem molecules. However, overly strict adherence to it can have the adverse effect of restricting diversity in the important areas discussed above and hence also reducing similarity to natural products. Other methods that might help to bring some of the properties of combinatorial compounds closer to those of natural products, e.g. application of biocatalysis and biotransformations in the synthesis[34] or using natural product templates in library design,[35] have also been suggested. It is likely that by mimicking certain distribution properties of natural compounds, a substantially more diverse set of combinatorial products might be made that could also have greater biological relevance.

Analogously to different 'drug-like' filters, the concept of 'natural-product-like' filters can be introduced based on the statistical analysis of different properties described above. A large proportion of natural products is biologically active and has favorable ADME/T properties, despite the fact that they often do not satisfy 'drug-likeness' criteria. We can view the production of unique bioactive libraries by plants and organisms as analogous and complementary to the combinatorial generation of synthetic libraries, each approach providing access to very different types of lead compounds.[12] However, there is a significant difference between these processes. The generation of combinatorial libraries has been constrained primarily by the availability of reagents and suitable reactions. In contrast, the generation of natural product diversity has occurred not only within the constraints of available biosynthetic reactions and precursors but also in the context of biological utility. Most natural products have a function, and the synthetic routes generating these compounds have coevolved with the requirements of ligand functionality. This suggests that combinatorial synthesis must also evolve beyond synthetic feasibility to explicitly address the creation of compounds with proper biological function.

REFERENCES AND NOTES

(1) Szostak, J. W. Introduction: Combinatorial Chemistry. *Chem. Rev.* **1997**, *97*, 347−348.
(2) Leach, A. R.; Hann, M. M. The in silico world of virtual libraries. *Drug Discovery Today* **2000**, *5*, 326−336.
(3) 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 Del. Rev.* **1997**, *23*, 3−25.
(4) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. I. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2993.
(5) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095−5099.
(6) Oprea, T. I. Property Distribution of Drug-Related Chemical Databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.
(7) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn to Distinguish between Drug-like and Nondrug-like Molecules. *J. Med. Chem.* **1998**, *41*, 3314−3324.
(8) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(9) Frimurer, T. M.; Bywater, R.; Naerum, L.; Lauritsen, L. N.; Brunak, S. Improving the Odds in Discriminating Drug-like from Non Drug-like Compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1315−1324.
(10) Xu, J.; Stevenson, J.; Drug-like Index: a New Approach to Measure Drug-like Compounds and Their Diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177−1187.
(11) Kingston, D. G. I. Natural Products as Pharmaceuticals and Sources for Lead Structures. In *The Practice of Medicinal Chemistry*; Wermuth, C. G., Ed.; Academic Press: London, 1996.
(12) Verdine, G. L. The Combinatorial Chemistry of Nature. *Nature*, **1996**, *384*, 11−13.
(13) Stahura, F. L.; Godden, J. W.; Xue, L.; Bajorath, J. Distinguishing between Natural Products and Synthetic Molecules by Descriptor

Shannon Entropy Analysis and Binary QSAR Calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245−1252.
(14) Henkel, T.; Brunne, R. M.; Müller, H.; Reichel, F. Statistical Investigation of Structural Complementarity of Natural Products and Synthetic Compounds. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 643−647.
(15) Lee, M. L.; Schneider, G. Scaffold Architecture and Pharmacophoric Properties of Natural Products and Trade Drugs: Application in the Design of Natural Product-Based Combinatorial Libraries. *J. Comb. Chem.* **2001**, *3*, 284−289.
(16) Chapman and Hall Dictionary of Drugs for PC; Chemical Design Ltd., Chipping Norton: Oxfordshire, U.K., 1996.
(17) The Merck Index, Twelfth Edition on CD-ROM, Version 12:1; Chapman and Hall Electronic Publishing Division: London, 1996.
(18) Maybridge HTS database, June 2000, Maybridge Chemical Co. Ltd.: Tintagel, Cornwall, U.K.
(19) ChemBridge EXPRESS-Pick database, February 2001, ChemBridge Corporation: San Diego, CA.
(20) ComGenex stock databases, March 2001, ComGenex International: San Francisco, CA.
(21) ChemDiv Small Molecule Compound Collections, ChemDiv Inc.: San Diego, CA.
(22) SPECS and BioSPECS databases, April 2001, SPECS and BioSPECS B.V., Rijswijk: The Netherlands.
(23) Interbioscreen databases, February 2001, Interbioscreen Ltd.: Moscow, Russia.
(24) Molecular Operating Environment, Version 2000.02, Chemical Computing Group Inc.: Montreal, Quebec, Canada.
(25) Badertscher, M.; Bischofberger, K.; Munk, M. E.; Pretsch, E. A Novel Formalism to Characterize the Degree of Unsaturation of Organic Molecules. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 889−893.
(26) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(27) Sokal, R. F.; Rohlf, F. J. *Biometry*, 2nd ed.; W. H. Freeman: New York, 1981; pp 43−46.
(28) Klebe, G.; Böhm, H.-J. Energetic and Entropic Factors Determining Binding Affinity in Protein−Ligand Complexes. *J. Receptor Signal Transduction Res.* **1997**, *17*, 459−473.
(29) Triggle, D. J. The Transition from Agonist to Antagonist Activity: Symmetry and Other Considerations. In *The Practice of Medicinal Chemistry*; Wermuth, C. G., Ed.; Academic Press: London, pp 547−570.
(30) Mann, J. *Secondary Metabolism*, 2nd ed.; Clarendon Press: Oxford, 1987.
(31) Kolb, H. C.; Finn, M. G.; Sharpless, K. B. Click Chemistry: Chemical Function from a Few Good Reactions. *Angew. Chem. Int. Ed. Engl.* **2001**, *40*, 2004−2021.
(32) Tute, M. S. Lipophilicity: A History. In *Lipophilicity in Drug Action and Toxicology*; Pliska, V., Testa, B., van de Waterbeemd, H., Eds.; VCH: Weinheim, 1996; pp 7−26.
(33) Böhm, H.-J.; Klebe, G.; Kubinyi, H. *Wirkstoffdesign*; Spektrum Akademischer Verlag: Heidelberg, 1996; p 153.
(34) Zaks, A.; Dodds, D. R. Application of biocatalysis and biotransformations to the synthesis of pharmaceuticals. *Drug Design Discovery* **1997**, *2*, 513−531.
(35) Hall, D. G.; Manku, S.; Wang, F. Solution and Solid-Phase Strategies for Design, Synthesis and Screening of Libraries Based on Natural Product Templates: A Comprehensive Survey. *J. Comb. Chem.* **2001**, *3*, 125−150.

CI0200467