Assessing the Scaffold Diversity of Screening Libraries

Mireille Krier,[†,‡,§] Guillaume Bret,[†,‡] and Didier Rognan*[,†,‡]

CNRS UMR7175-LC1, Institut Gilbert Laustriat, 74 route du Rhin, F-67401 Illkirch Cédex, France, and
Idéalp'Pharma, Bâtiment CEI, 66 Bd Niels Bohr, F-69603 Villeurbanne Cédex, France

Medicinal chemists have traditionally realized assessments of chemical diversity and subsequent compound acquisition, although a recent study suggests that experts are usually inconsistent in reviewing large data sets. To analyze the scaffold diversity of commercially available screening collections, we have developed a general workflow aimed at (1) identifying druglike compounds, (2) clustering them by maximum common substructures (scaffolds), (3) measuring the scaffold diversity encoded by each screening collection independently of its size, and finally (4) merging all common substructures in a nonredundant scaffold library that can easily be browsed by structural and topological queries. Starting from 2.4 million compounds out of 12 commercial sources, four categories of libraries could be identified: large- and medium-sized combinatorial libraries (low scaffold diversity), diverse libraries (medium diversity, medium size), and highly diverse libraries (high diversity, low size). The chemical space covered by the scaffold library can be searched to prioritize scaffold-focused libraries.

INTRODUCTION

With the advent of initiatives such as the CNRS "National Chemical Library"[1] or the NIH Molecular Libraries Initiative,[2] public research rejoined the pharmaceutical industry in its effort to organize and to curate small molecular-weight molecules for the purpose of drug discovery and target deorphanization. The drastic and steady increase of commercially available compounds beginning in the early 1990s[3] provided chemical information scientists the opportunity to enhance the diversity of their proprietary compound collections. The main challenge that remained over the years, which is regularly revisited,[4] is the way of measuring the diversity of a compound library. Very informative testimonials about this key aim were shared with the community elsewhere.[5] Nevertheless, molecular diversity is heavily dependent on the descriptors, metrics, and multivariate methods used to assess it. Most studies on commercially available compound libraries[6−8] have traditionally used physicochemical and topological descriptors, summed up into a score[9] or encoded into fingerprints[10] or hash codes,[11−13] to evaluate the uniqueness and diversity of such libraries. Although fingerprints can be quickly computed for large collections of compounds, it results in classifications of molecular libraries that are not very intuitive for medicinal chemists because a single class of compounds may contain quite different molecular scaffolds accessible by very different synthetic routes. Traditionally, medicinal chemists mining high-throughput screening (HTS) data have organized hits into homogeneous chemical series. Why not use the same partitioning method before the virtual or real screening process? Archiving compounds by scaffolds is much more natural but computationally more

demanding if calculated ad hoc. However, various definitions of a scaffold are possible. For a given compound (e.g., dopamine D3 receptor antagonist BP-890; Figure 1), a scaffold may be considered, among many possibilities, as a maximum common substructure (MCS),[14] the largest rigid fragment or rings,[15] molecular frameworks or Murcko's scaffolds[16] with or without descriptors (e.g., topological torsions),[17] and molecular fragments as generated by the RECAP method.[18] According to the chosen definition, the scaffold may be unique (superstructure) or multiple (two to three substructures for BP-890). Therefore, depending on the way a scaffold is regarded, very different substructures (eight in the present case) could be stored as representative of the cognate compound.

During a medicinal chemistry project, it is not uncommon that structural parts of the scaffold are redefined by either extension or reduction. If the limit of one extreme is reached by setting the full compound equal to the scaffold, how far can one reduce the compound structure to obtain a chemically meaningful scaffold? In the present study, we classified 17 commercially available screening collections according to graph-based maximum common substructures[19] and joined the resulting classification into a single library of nonredundant classes. A new metric (PC50C) is proposed to assess the diversity of a screening collection, by computing the percentage of classes accounting for 50% of the classified compounds. Since this metric is independent of the size of a library, it can be used to compare collections of different sizes.

METHODS

The overall workflow for reading, processing, and extracting molecular scaffolds out of commercial libraries is illustrated in Figure 2 and further detailed in the following paragraphs.

* To whom correspondence should be addressed. Phone: +33-3-90 24 42 35. Fax: +33-3-90 24 43 10. E-mail: didier.rognan@pharma.u-strasbg.fr.
† Institut Gilbert Laustriat.
‡ Idéalp'Pharma.
§ Current address: Merck KGaA, D-64293 Darmstadt, Germany.

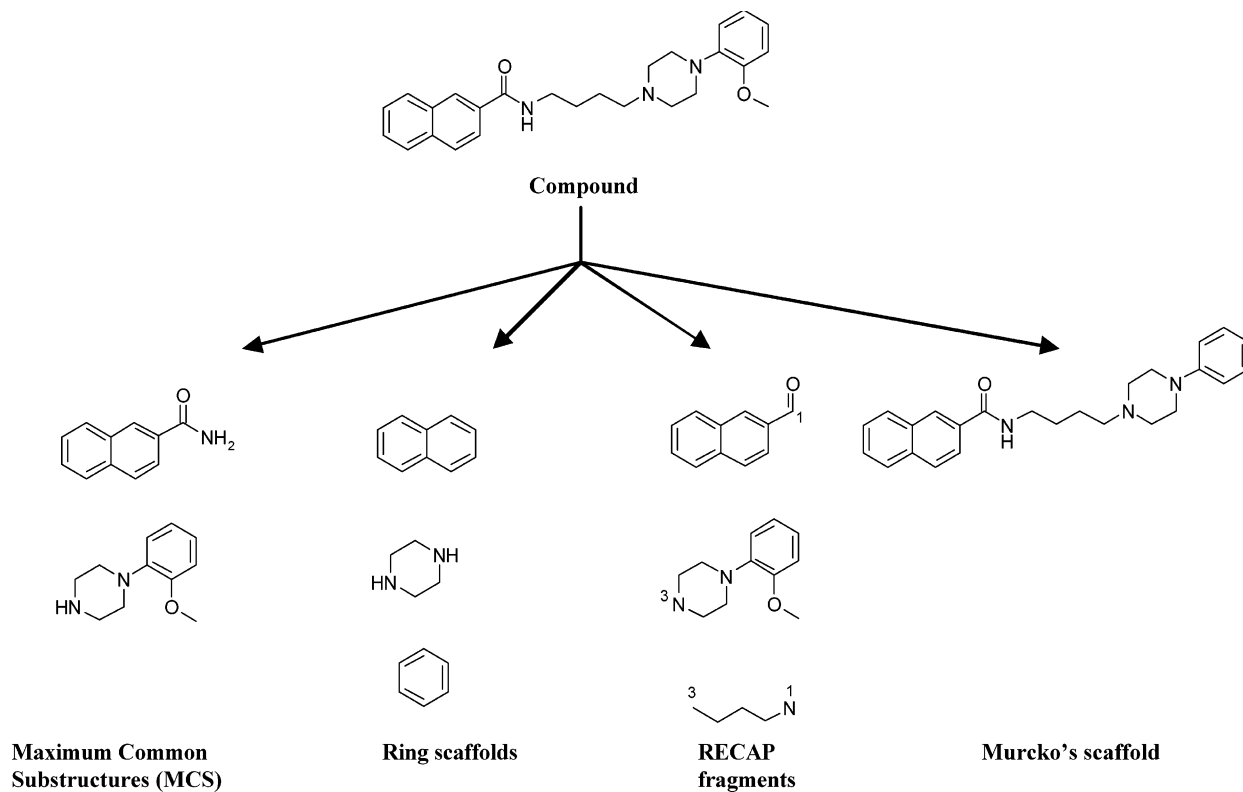SCAFFOLD DIVERSITY OF SCREENING LIBRARIES

J. Chem. Inf. Model., Vol. 46, No. 2, 2006 **513**

Figure 1. Possible representations of molecular scaffolds for the dopamine D3 antagonist BP-890.
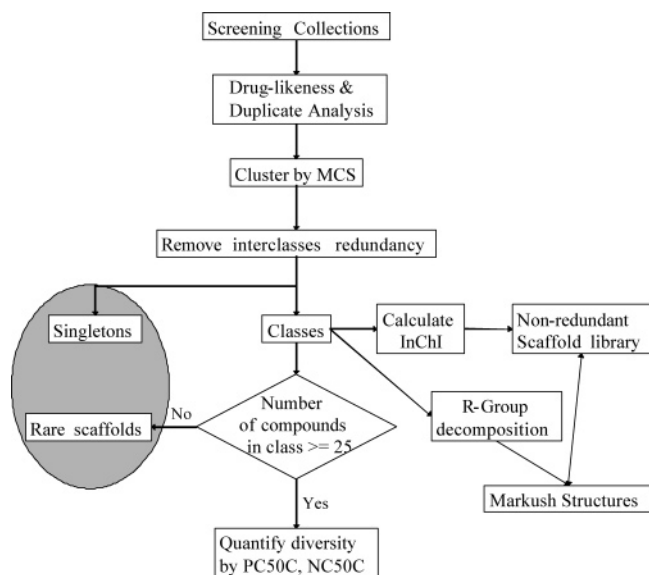

Figure 2. General workflow for processing screening libraries.

Database Processing. The screening collections used in this study date from the last quarter of 2003 except the MDDR, for which the first 2004 release has been used. A total of 17 libraries from 12 suppliers plus the MDDR describe the commercially available chemical space addressed by the present paper. It covers 2 410 857 compounds easily available as powders in vials. The collections need also to have a computer-readable counterpart, delivered as an SDFile on a CD-ROM or downloadable from the supplier's webpage.[9] The very first processing steps consisted of standardizing the structure and data headers of SD files using an in-house Perl script. Property- or functional group-based filtering rules[20] implemented in OpenEye's Filter[21] program were then used to select the most suitable com-

pounds for each library (see filtering rules in the Supporting Information). In this step, counterions were removed and the ionization state of each compound at physiological pH was assigned. For each collection, an additional step consisted of eliminating remaining redundant compounds, taking stereochemical information into account using the CLIFF program[22] (please note that CLIFF has recently been split into several separated routines).

Compound Classification. One of the major challenges was to obtain an organization of the screening collections into chemically meaningful classes. ClassPharmer Suite's proprietary clustering methodology[19] was adopted. To make a clear distinction with the common association of clustering with fingerprints in chemoinformatics, the grammatical root "class" (classes/classification) is preferred over "cluster" (clusters/clustering). But, in strictly algorithmic terms, the method used herein happens to be a clustering algorithm and not a classification where one starts from predefined scaffolds.[23]

Two parameters mainly influence the outcome of the classification: the homogeneity and the redundancy level. Homogeneity is related to the size (heavy atom count) of the scaffold divided by the size (heavy atom count) of the largest compound in the class. Redundancy describes to what extent a compound is allowed to appear in multiple classes. Hence, the classes are represented by a scaffold assimilated to the MCS. The underlying algorithm is covered by trade secret but can, however, be approximately described as follows: given a data set of N compounds, (i) find topologically aware (approximated) MCSs for all pairs, triplets, quadruplets, ..., and $N - 1$ groups of compounds passing the user-defined homogeneity level; (ii) select the smallest number of MCSs that fulfills the user-defined redundancy level while giving the minimal number of singletons, and if

the option is selected, (iii) generate subclasses with larger (exact) MCSs where subsets of a class with higher homogeneity can be found. The implementation of the algorithm is preceded by a normalization process of the input structures. For the present classification, the homogeneity and redundancy were set to medium and low, respectively. Exact ring closure and exact atom match parameters were chosen to define classes. No subclasses were computed. Hereafter, the term scaffold will, thus, be restricted to Bioreason's MCS.

Scaffold Distribution. The nonhierarchical disjunctive algorithm which was used allows that a compound belongs to more than one class. To compare the scaffold distribution of different libraries, the interclass redundancy of a compound was removed using a Python script based on the OpenEye OEChem1.3 library.[24] This task was achieved by computing the central scaffold score (CSS) of each compound/class pair and assigning the compound to the class presenting the lowest CSS, calculated by the following equation:

$$CSS = \frac{MW_{compound} - MW_{scaffold}}{N_R}$$

where $MW_{compound}$ is the molecular weight of a compound, $MW_{scaffold}$ the molecular weight of the scaffold, and N_R the number of substitution points (R groups).

For every screening collection, the classes were ordered by decreasing size and two metrics (NC50C and PC50C) computed. NC50C describes the number of nonredundant classes covering 50% of classified compounds. PC50C features the percentage of classes covering 50% of classified compounds. Two classifications were analyzed. In the first one, all classes of at least two unique compounds were investigated. In the second one, a threshold of 25 was assigned to the minimal size of a class (number of unique compounds). Classes populated by less than 25 unique compounds will be referred to as "rare scaffolds" (Figure 2).

R-Group Decomposition. The topology around the scaffold and generation of the corresponding SMILES strings[25] were obtained by R-group decomposition (see general procedure in Figure 3). A compound subset and its corresponding ClassPharmer scaffold were the input for finding the minimum common supergraph[26] under the form of a Markush structure. For each compound/scaffold pair, the substitution points were determined and a scaffold with R groups was generated. Among this R-group-labeled scaffold, isomorphs were eliminated and the pairwise substructure relationship was checked. This reduced considerably the number of structures to compare. The remaining N Markush structures were then used to find the minimum common super-Markush structure. The process is similar to the one described by Brown et al., which identifies the hyperstructure,[27] and is outlined as follows:

```
SuperStructure := MarkushStructure(1)
FOR n :=2 to N DO
BEGIN
        COMPARE(SuperStructure, MarkushStructure(n))
        UPDATE_CT()
END
```

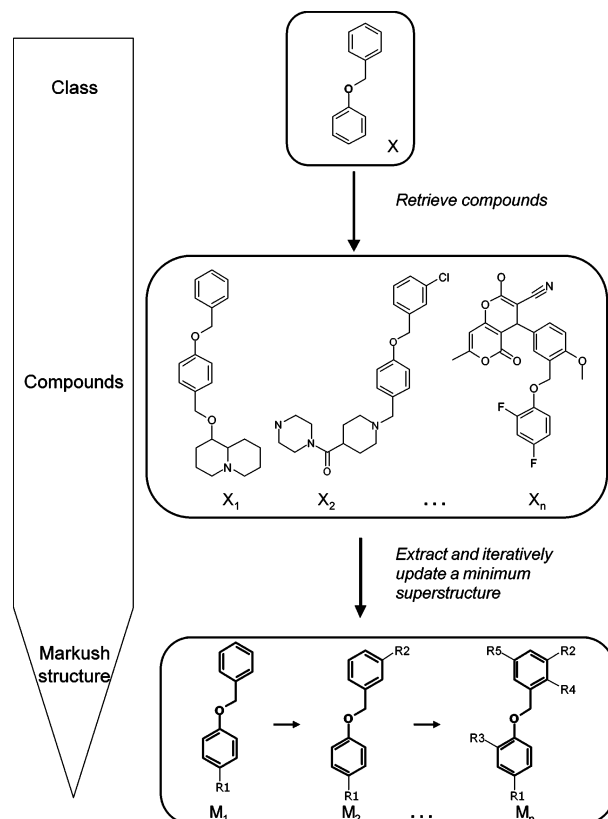As for the removal of interclass redundancy, the R-group

Figure 3. R-group decomposition procedure from Classpharmer classifications. For each class X, all compounds (X_1, X_2 to X_n) are extracted and a Markush structure is extracted from the first compound (M_1) and then iteratively refined to a minimum superstructure (M_2) by adding the next molecule until the last compound in the class has been processed to give the final superstructure (M_n).

decomposition was implemented in Python on the basis of OpenEye OEChem.[24]

Setting Up a Scaffold Library. All classes (excluding the singletons) were assembled from the generated classifications to form a scaffold library. Computing InChI[28] representations ("Mobile H Perception" option ON) for all scaffolds gave the possibility to identify tautomeric forms and group them together. All structural data were deposited in a relational database (MySQL 4.1; for database structure, see the Supporting Information, Scheme A). Each scaffold was annotated with molecular properties (AlogP, polar surface area, hydrogen bond donor and acceptor count, rotatable bonds, and ring systems count) and the Markush structure SMILES. The main scaffold structure table can be browsed and queried by similarity, substructure, or superstructure using JChemBase[29] and is freely accessible at http://bioinfo-pharma.u-strasbg.fr/scaffolds.

RESULTS AND DISCUSSION

ClassPharmer MCS and Classification. To illustrate the MCS and classification concepts in ClassPharmer, a very simple data set of 20 known dopamine D$_2$ antagonists (Figure 4) randomly chosen from the Hert data set[30] was taken as a reference. When medium homogeneity and redundancy settings were used, seven classes and four singletons were generated. Interestingly, computed MCSs feature either classical MCS (e.g., class 1), ring scaffolds (e.g., class 2),
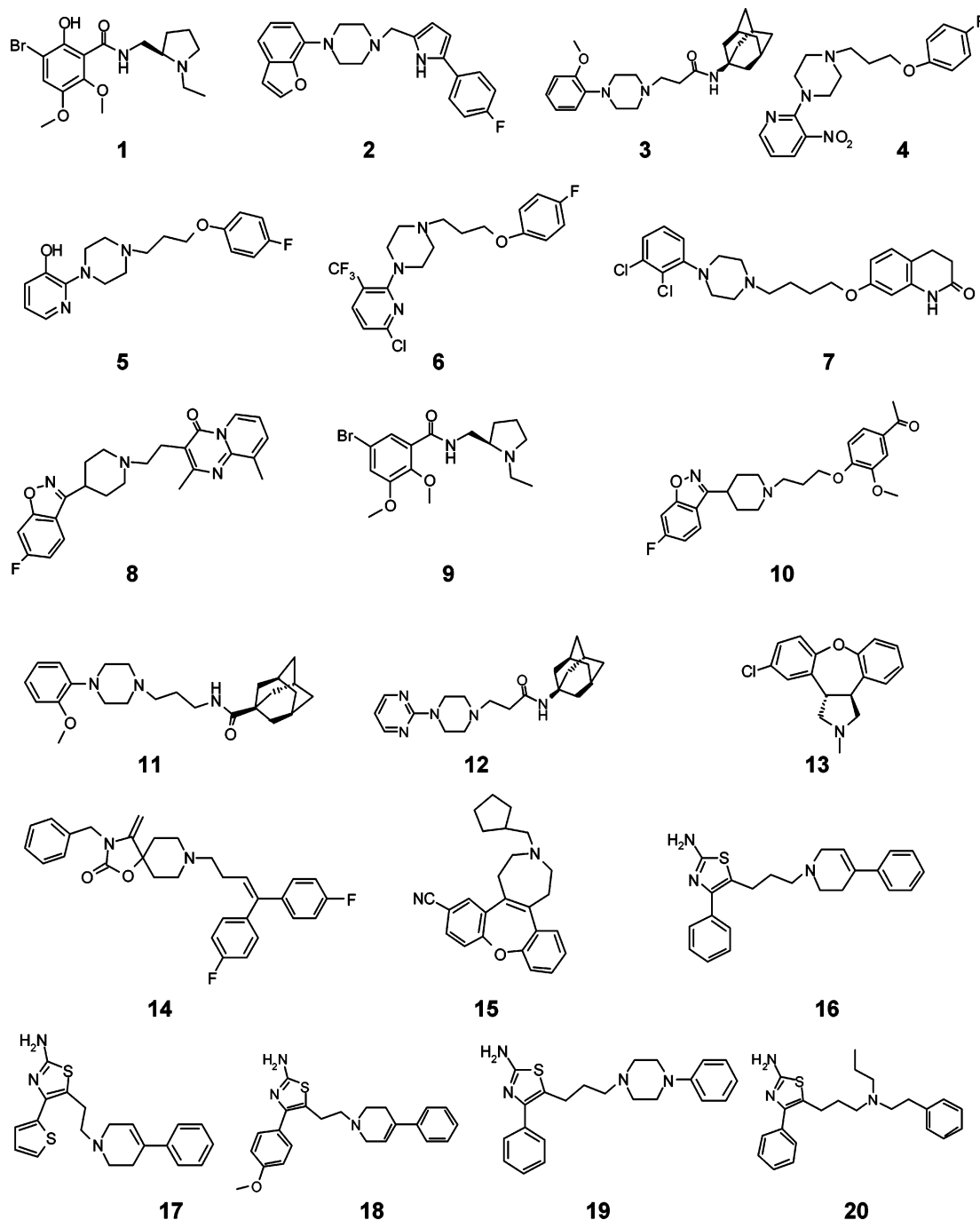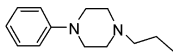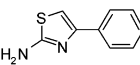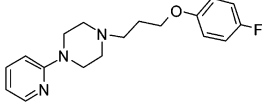
SCAFFOLD DIVERSITY OF SCREENING LIBRARIES
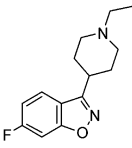
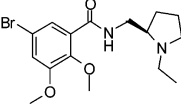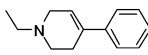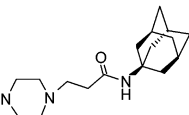*J. Chem. Inf. Model., Vol. 46, No. 2, 2006* **515**

Figure 4. Structure of 20 true dopamine D-2 antagonists randomly extracted from the Hert data set.[30]

or even Murcko's scaffolds (e.g., classes 3 and 5; Table 1). ClassPharmer MCSs span at least half of the largest compound's size (heavy atom count) in the class (Table 1) according to the user-defined homogeneity level. Therefore, generating larger or smaller MCSs is easily customizable by adjusting the homogeneity setting. Here, singletons are compounds which either bear a MCS not present in any other compound (e.g., compounds 13 and 15; Figure 4) or present a known MCS but fail to pass the homogeneity criterion (e.g., high-molecular-weight compound, no case in Table 1) or for which a common MCS could be found (e.g., phenyl-piperazine) but with a different chemical environment (the benzofuranyl-piperazine 2 is not classified with the phenyl-piperazines 3, 7, 11, and 19 in class 1; Figure 4 and Table 1). ClassPharmer MCSs are, therefore, more complex than

simple maximum common substructures. Not only the MCS but its chemical environment should be conserved to group compounds in the same class. Fuzzier definitions are of course possible (e.g., disabling exact atom matches and ring closures), but resulting classifications would be more difficult to interpret and, therefore, have not been investigated in the present paper.

A second parameter that influences the classification is the level of accepted redundancy (how many times a compound may appear in multiple classes). Using a medium redundancy, several compounds are found in at least two classes (Table 1). For the general purpose of HTS data analysis, this is not really a problem and many medicinal chemists would indeed reproduce the herein-reported classification. However, to compare the scaffold diversity of

Table 1. Example of ClassPharmer Classification on a Small Data Set (20 Dopamine D_2 Antagonists)

Class	Compounds[a]	MCS	Homogeneity[b]
1	**3**, 7, 11, **19**		0.52
2	**16**, **18**, **19**, 20		0.51
3	4, 5, 6		0.89
4	8, 10		0.58
5	1, 9		0.98
6	**16**, 17, **18**		0.53
7	**3**, 12		0.75
Singletons	2, 13, 14, 15		

[a] Compounds found in multiple classes are indicated by bold numbers. [b] The homogeneity of a class is the size (heavy atom count) of the MCS divided by the size (heavy atom count) of the largest compound in that class.

heterogeneous screening collections with our new metrics (NC50C and PC50C) and facilitate the subsequent analysis, we preferred to simply remove redundancy by adjusting the redundancy parameter to low (the most recent release of ClassPharmer even allows to set this parameter to none) and postprocess the obtained classification as reported above. We do not claim that this kind of classification is the best one, but it allows a robust and chemically intuitive grouping of most drug-like compounds that we have seen up to now.

Processing the Libraries. In a first step, 17 commercially available screening collections were processed to retain unique druglike molecules. In addition, a prototypical collection of druglike compounds (MDDR) was taken as reference to delimit true druglike chemistry space. In agreement with previous reports,[6,20] the percentage of druglike molecules in these collections varies from ca. 30% (ChemStar) to 60% (Asinex Platinum) (see Table 2). No relationships could be established between size and druglikeness of the libraries. It should be noted that a set of very strict rules (see the Supporting Information, Chart A) especially regarding molecular weight ($250 < MW < 500$) and Lipinski's rule of five violations (none) was used herein. Considering the MDDR as a reference druglike data set, we can thus consider most of the screening collections investigated here to be druglike, reflecting the effort of vendors to produce higher quality collections.[3] Internal duplicates (compounds present several times within the same collection) ranged from none (ASIp) to 320 compounds (TRI). An exceptionally high number (146 425) was found for CBG but could be explained by the previous merge of two screening collections (EXPRESS-Pick and Hit2Lead) into a single data set. Retrospectively, only two compounds would have been duplicated in CBG Express-Pick. For the MDDR, there were still 2294 duplicates left, most of them arising from different counterions.

An exclusivity analysis of all screening collections shows that only five of them (ASIp, CNR, MAY, NET, and TRI) could be described as original as they contain more than 85% druglike compounds not present elsewhere (Table 2). Significant pairwise overlap exists between several libraries (e.g., ASIg, CBG, IBSs, CDIc, and VITs; see Tables A and B in the Supporting Information). However, having several commercial sources for a compound may be an advantage since it still guarantees a purchase even if the corresponding molecule is no longer available from a particular supplier.

What Is the Scaffold Diversity of Commercial Libraries? A first scaffold classification (classification 1, Table 3)

Table 2. Library Processing and Classification

supplier	collection	code	size	filtered[a]	% druglike	unique[b]	exclusive[c]	classified[d]
Asinex	Gold	ASIg	201 304	86 185	42.8	86 153	17 322	85 516
	Platinum	ASIp	120 563	71 255	59.1	71 255	69 716	70 978
ChemBridge	EXPRESS-Pick + H2LS	CBG	709 975	327 716	49.5	181 291	72 484	161 827
Chemical Diversity	CombiLab	CDIc	230 529	104 606	47.8	104 604	62 361	104 520
	International Diversity	CDIi	133 085	39 859	45.9	39 831	13 571	39 401
CNRS	National Chemical Library	CNR	12 670	4978	39.3	4806	4571	4770
ChemStar		CST	73 552	21 899	29.8	21 852	4857	21 758
InterBioScreen	Natural	IBSn	30 749	14 196	46.2	13 936	890	13 882
	Synthetic	IBSs	287 945	112 882	43.2	112 695	61 944	111 562
Maybridge		MAY	59 204	20 754	35.1	20 726	17 793	20 680
Bionet		NET	38 416	14 031	36.5	14 029	13 276	13 992
Specs		SPE	172 970	65 563	37.9	65 539	20 499	65 319
Timtec	Natural	TIMn	4202	2083	49.6	1945	147	1941
	Synthetic	TIMs	95 469	33 669	35.3	33 560	7873	33 408
Tripos		TRI	84 604	46 866	55.4	46 546	44 969	46 543
Vitas-M	Stock	VITs	134 167	52 583	39.2	52 544	8796	52 204
	Tulip	VITt	21 453	7190	33.5	7182	3778	7164
MDDR	2004.1	MDDR	98 880	37 857	38.3	35 563	35 142	35 033

[a] Using Filter 1.0.[21] [b] Using Cliff[22] with options "-unique 1 -usestereo 1". [c] Compounds not found elsewhere by comparison of SciTegic canonical SMILES generated by PipelinePilot 4.5.[48] [d] After normalization step in ClassPharmer.[19]

SCAFFOLD DIVERSITY OF SCREENING LIBRARIES

J. Chem. Inf. Model., Vol. 46, No. 2, 2006 **517**

Table 3. Classification Results
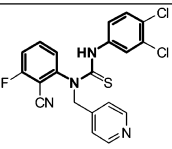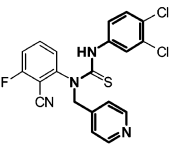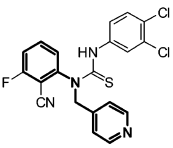
code	classification 1[a]					classification 2[b]		
	# classes	# singl[c]	% red[d]	NC50C[e]	PC50C[f]	# classes	NC50C	PC50C
ASIg	3491	5476	7.25	52	1.49	400	27	6.75
ASIp	1968	2907	9.27	27	1.37	252	19	7.54
CBG	3199	5269	15.79	45	1.41	709	32	4.51
CDIc	3430	5171	6.29	86	2.51	528	57	10.80
CDIi	2306	3447	7.99	62	2.69	219	27	12.33
CNR	391	662	2.74	26	6.65	33	7	21.21
CST	1011	1719	9.19	25	2.47	123	13	10.57
IBSn	757	1188	2.02	20	2.64	75	8	10.67
IBSs	3490	5370	5.25	68	1.95	492	48	9.76
MAY	1544	2501	12.59	84	5.44	151	30	19.87
NET	941	1230	5.72	58	6.16	107	21	19.63
SPE	3261	4971	8.11	59	1.81	313	27	8.63
TIMn	162	316	1.29	12	7.41	14	5	35.71
TIMs	1956	3445	7.23	67	3.43	207	28	13.53
TRI	1341	2041	11.55	33	2.46	282	22	7.80
VITs	2153	3134	8.85	35	1.63	237	20	8.44
VITt	402	513	6.59	16	3.98	48	9	18.75
MDDR	3058	4620	8.51	177	5.79	203	35	17.24

[a] Class defined as containing at least two unique compounds. [b] Class defined as containing at least 25 unique compounds. [c] Number of singletons. [d] Percentage of interclass redundancy (percentage of compounds present in multiple classes). [e] Number of classes accounting for 50% of classified compounds. [f] Percentage of classes accounting for 50% of classified compounds.

has been realized on the global set of 846 408 druglike molecules passing the ClassPharmer normalization step. A second one (classification 2) is a subset of the first one since it accounts for classes populated by at least 25 unique molecules. The second classification was undertaken to depict the optimization potential of each class. Hence, a class described by a low number of compounds might be of lower interest for a medicinal chemist because of a possible lack of synthetic tractability or insufficient statistics if the library has to be assayed experimentally. It should be noticed that there is still a lack of consensus on the minimal number of compounds that should be stored to accurately describe a class/cluster. McFayden et al.[13] proposed a minimal value of five compounds for postprocessing raw HTS data, whereas Nilakantan and Nunn suggested that many more compounds (100) should be selected for enumerating-scaffold focused libraries.[31]

Using our classification method, there are generally 10−30 times fewer classes (scaffolds) than molecules (Tables 2 and 3). Classification 1 afforded a total of 34 961 classes and 53 980 singletons. Interestingly, the number of singletons always exceeds that of classes for all libraries. Considering the homogeneity of the input libraries which was set to medium prior to the classification, most singletons do not describe unique scaffolds but rather compounds which failed to pass the homogeneity threshold level (i.e., the number of heavy atoms in the scaffold is too small in comparison to the overall size of the largest molecule in the class). Classification 2 (only classes populated by at least 25 compounds) led to a smaller set of 4390 classes. Since a single compound may be classified in different classes for a single library, there is a significant level of redundancy across the classes generated by ClassPharmer (about 10% on average, Table 3) which biases relationships between the number of classes and the number of compounds within a library. To get unbiased relationships and a clearer comparison of the scaffold diversity of input libraries, the redundancy was removed by a simple strategy aimed at selecting the most central scaffold for a compound appearing

Table 4. Example of Interclass Redundancy

Compound[a]	MW(scaffold)	N_R[b]	CSS[c]
	198	6	33
	218	4	55
	184	3	61

[a] The compound exemplified here (CD 05668, Maybridge) has a molecular weight of 413.32 and three proposed scaffolds highlighted in bold. [b] N_R: number of R groups. [c] CSS (Central scaffold score) = $(MW_{compound} - MW_{scaffold})/N_R$.

in multiple classes (Table 4). It is important to point out that class redundancy among different libraries has not been considered at this stage.

Two metrics (NC50C and PC50C) have been developed to measure and compare the scaffold diversity of screening collections. The first one (NC50C) is a simple measure of the number of classes accounting for 50% of the classified compounds for a particular collection. The NC50C descriptor has been derived from a first plot (Figure 5) describing the density (percentage of classified compounds) of each class which was then transformed into a cumulative plot (Figure 6) allowing interpolation of the number of classes required to describe 50% of classified compounds. The NC50C descriptor can be regarded as the absolute scaffold diversity of the collection. As expected, larger collections have higher

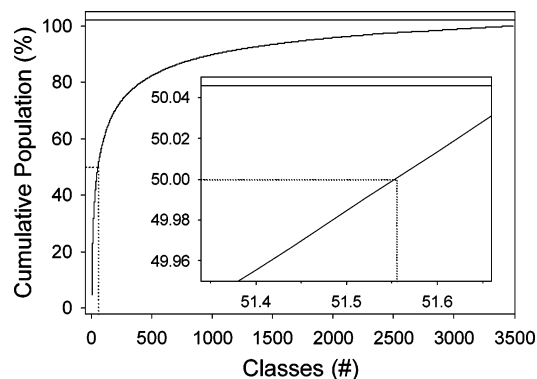**Table 5.** Classification of Collections According to Their Size and Relative Scaffold Diversity (PC50C)

category	libraries[a]	size[b]	PC50C[c]	PC50C_25[d]
large combinatorial libraries	CBG, CDIc, IBSs	>100 K	<3	<13
medium combinatorial libraries	ASIg, ASIp, SPE, TRI, VITs	50−100 K	<3	<13
diverse libraries	CDIi, CST, IBSn, TIMs	<50 K	<4	10−15
highly diverse libraries	CNR, MAY, MDDR, NET, TIMn, VITt	<50 K	>4	>15

[a] Libraries are indexed as shown in Table 1. [b] Number of drug-like and unique compounds passing the ClassPharmer normalization step. [c] PC50C value derived from all classes of a library. [d] PC50C value derived from classes populated by at least 25 representatives.
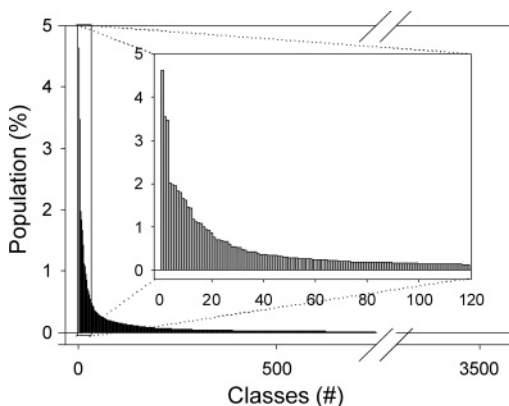
Figure 5. Density of ClassPharmer classes (ASIg collection) featuring the percentage of classified compounds for all classes. A zoom on the most populated classes is boxed within the graph.

Figure 6. Interpolating the NC50C value by plotting the number of classes versus the cumulative percentage of classified compounds (ASIg collection). A zoom around the NC50C value is boxed within the graph.

NC50C values (Figure 7A), except for four collections which either present a quite large panel of classes with respect to their size (MAY, NET, and especially MDDR) or a low number of classes (CBG). Discarding these four libraries, a significant correlation could be found between size (number of classified druglike compounds) and NC50C ($r = 0.75$, $n = 14$, and $p = 0.002$). Compared to the absolute scaffold diversity for classes containing at least 25 compounds (Figure 7B), all collections shift to lower NC50C values with the reference MDDR (Table 3, Figure 7B) performing the most notable shift to the left, thus joining commercially available collections. For classification 2, a significant correlation is also observed between size and NC50C for all but the CBG collection ($r = 0.70$, $n = 17$, and $p = 0.002$).

Since the first metric is dependent on the size of each collection, it cannot be used to compare the intrinsic scaffold diversity. We, therefore, computed a second descriptor (PC50C) estimating the percentage of classes accounting for 50% of the classified compounds (Table 3). It presents the advantage of being independent of the size of the library and, therefore, is more suitable for a comparative analysis (Figure 7C). Strikingly, plotting the PC50C versus the size of each collection allows segregation of the herein-analyzed 18 collections into four categories (Table 5). A first category (CBG, IBSs, and CDIc), in agreement with a previous report,[32] regroups large combinatorial libraries for which a very tiny percentage of the scaffolds (less than 3%) has been overrepresented. Corresponding scaffolds are usually very simple (e.g., N-benzylaniline; Table 6), account for over 10 000 unique compounds, and are available at a majority of suppliers. Promiscuous scaffolds are also found in the second category (e.g., N-phenylbenzenesulfonamide; Tables

Table 6. Example of Characteristic Scaffolds for the Four Categories of Screening Collections

Category	Scaffold	Identifier	Suppliers	Uniques compounds
Large combinatorial Libraries		SBI_4853	15	22 988
Medium combinatorial Libraries		SBI_2909	10	8 592
Diverse Libraries		SBI_2654	1	322
Highly Diverse Libraries		SBI_21089	1	106

SCAFFOLD DIVERSITY OF SCREENING LIBRARIES

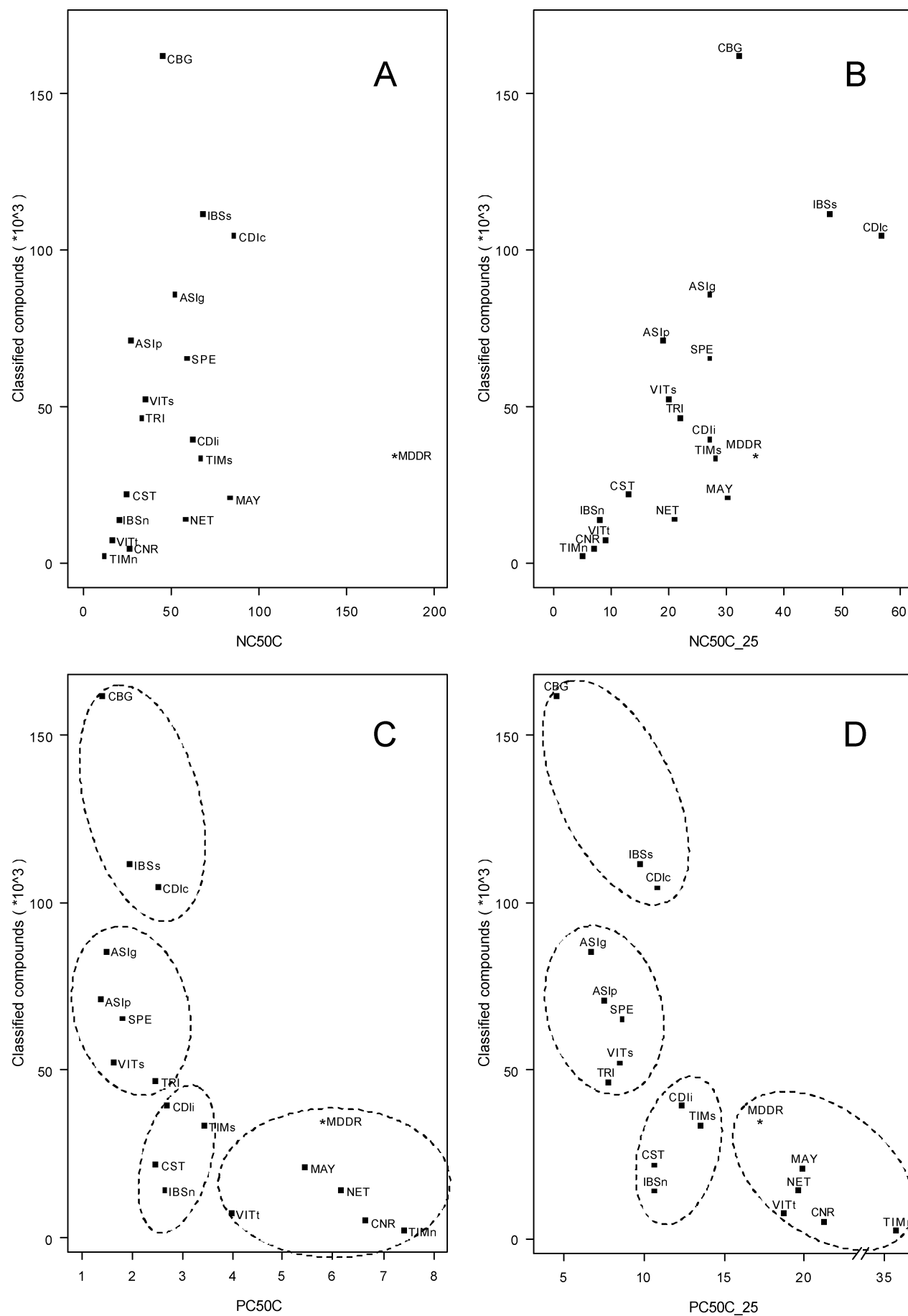J. Chem. Inf. Model., Vol. 46, No. 2, 2006 **519**

Figure 7. Scaffold diversity of screening collections.

Table 7. Distribution of Classes for the SBI Scaffold Library

library	classes[a]	exclusive classes[b]
ASIg	3485	1240 (36%)
ASIp	1964	1729 (88%)
CBG	3194	1213 (38%)
CDIc	3425	2325 (68%)
CDIi	2306	974 (42%)
CNR	391	299 (76%)
CST	1010	307 (30%)
IBSn	756	504 (67%)
IBSs	3484	1845 (57%)
MAY	1543	1052 (68%)
NET	941	722 (77%)
SPE	3260	1504 (46%)
TIMn	162	48 (30%)
TIMs	1954	700 (36%)
TRI	1338	1098 (82%)
VITs	2149	759 (35%)
VITt	402	264 (66%)
MDDR[c]	3057	2696 (88%)

[a] Nonredundant classes by comparison of INChI codes (Mobile H Perception option on). For duplicate classes, a single copy has been conserved corresponding to the first encountered library sorted by alphabetical order. [b] Classes not found elsewhere (by comparison of INChI codes). [c] MDDR-derived scaffolds are not included in the SBI scaffold library. Statistics are only given for comparison purposes.

5 and 6) of medium-sized combinatorial libraries (ASIg, ASIp, SPE, TRI, and VITs). In a third group are found libraries of smaller size (<50 000 druglike unique compounds; Table 5) with more original and less-populated scaffolds (CDIi, CST, IBSn, and TIMs). Last, a fourth category of highly diverse libraries (CNR, MAY, NET, TIMn, and VITt; Table 5) was identified nearby the reference MDDR data set (Figure 7C). The latter two categories of libraries are really diverse in terms of scaffold architecture and generally present a larger choice of proprietary low-populated scaffolds (see two prototypical examples in Table 6). These libraries are either collections of compounds from various origins (CNR and MDDR) or natural sources (TIMn and VITt) or have been synthesized by the supplier itself with the purpose of optimizing diversity versus size (NET and MAY). For example, the French National Chemical Library (CNR)[1] is a repository of compounds collected at 22 academic laboratories, each of them with a different medicinal chemistry history. Likewise, collections labeled

"natural products" (TIMn and VITt) are, in fact, synthetic compound libraries that are based on structures found in nature.[33] Acknowledging the high scaffold diversity found in natural products, it is, therefore, logical to group them into the fourth category of diverse libraries. Interestingly, looking at the scaffold diversity of the same libraries considering only those scaffolds populated by at least 25 compounds leads to identical clusters with a simple shift of PC50C toward higher values (Figure 7D). Simple rules based on the size (number of classified druglike and unique compounds) and on PC50C values (all classes, classes with more than 25 compounds) of 18 collections are provided (Table 5) as a guide to classify libraries not investigated herein.

Setting Up a Library of Nonredundant Classes. To set up a single data set for registering all commercially available scaffolds, all classes (except those arising from the reference MDDR database) depicted by the previous analysis were merged into a single library. Redundancy of the scaffolds was removed by working with InChI codes, which enable the detection of duplicates and tautomers. The resulting SBI (*S*caffold of the *BI*oinformatics Group of the CNRS) collection contains a total of 21 393 unique classes, out of which a surprisingly high number (16 583) are exclusively found at one supplier (Table 7). Interestingly, compounds contained in the classification represent 811 375 compounds, out of which 556 107 have a unique InChI representation. Although MDDR-derived scaffolds have not been incorporated into the SBI scaffold library, it is interesting to note that 88% of the MDDR scaffolds are nonoverlapping with those arising from vendors (Table 7). About 3000 scaffolds are necessary to cover all previously investigated biological activities by the MDDR. If a target space orthogonal to that addressed by the MDDR has to be investigated, we therefore suggest screening any of the exclusive SBI scaffolds.

A more restrictive data set of 2498 classes comprises scaffolds with a density of at least 25 compounds (Table 8). Of these, 921 classes have only one supplier as their source. A total of 329 (1.5%) scaffolds are discarded when the compounds contained in a class are checked for uniqueness by InChI. An R-group decomposition of all classes into Markush structures indicates a distribution of substituents

Table 8. Number of Scaffolds Which Are At Least/Exactly in *n* Screening Collections (DBs)

	InChI		InChI and at least 25 compounds	
# of DBs *n*	# of scaffolds in at least *n* DBs	# of scaffolds in exactly *n* DBs	# of scaffolds in at least *n* DBs	# of scaffolds in exactly *n* DBs
1	21393	16583	2498	921
2	4810	2532	1577	431
3	2278	1009	1146	106
4	1269	501	853	251
5	768	300	602	194
6	468	179	408	133
7	289	97	275	84
8	192	71	191	70
9	121	39	121	39
10	82	25	82	25
11	57	20	57	20
12	37	19	37	19
13	18	3	18	3
14	15	6	15	6
15	9	7	9	7
16	2	1	2	1
17	1	1	1	1

SCAFFOLD DIVERSITY OF SCREENING LIBRARIES

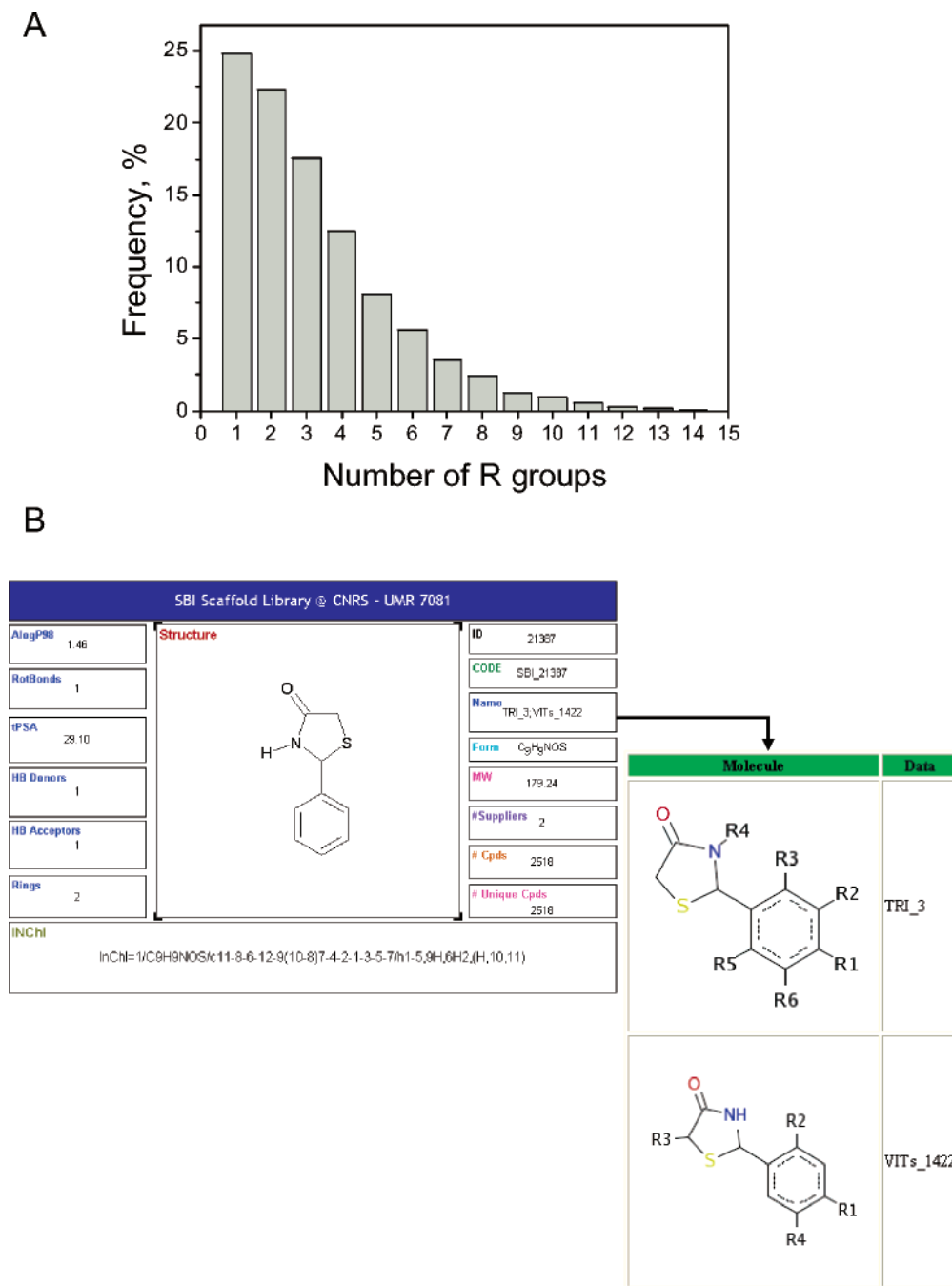*J. Chem. Inf. Model., Vol. 46, No. 2, 2006* **521**

Figure 8. The SBI scaffold library. (A) Distribution of the number of R groups for each scaffold. (B) Browsing the library. For each scaffold, molecular descriptors (AlogP98, number of rotatable bonds, topological polar surface area, number of H-bond donors and acceptors, number of rings, molecular weight, and number of unique compounds), vendor information (identity and number of suppliers), and a unique SBI code enable an easy navigation in the chemistry space covered by commercial scaffolds. Selecting a particular scaffold (e.g., 2-phenylthiazolidin-4-one) returns the corresponding classes indexed by commercial sources (TRI_3, VITs_1422; see a list of indexes in Table 1) and the related Markush structures.

following a monoexponential decay (Figure 8A). A total of 75% of the stored scaffolds offer at least two substituents and, thus, real diversity. The scaffold library can be easily browsed by substructure, physicochemical properties, or suppliers of the corresponding compounds (Figure 8B). A unique code for each scaffold refers to the individual suppliers and the corresponding Markush structures, thereby enabling the comparison of commercial sources for a particular scaffold (Figure 8B).

A molecular complexity of the SBI scaffold library was investigated as recently proposed by Selzer et al.[34] by computing circular FCFP_4 fingerprints and extracting

FCFP_4 sizes and densities (Figure 9). FCP_4 density calculated for all scaffolds of the SBI library indicate that a large majority of scaffolds are complex enough (FCFP_4 density > 1) to ensure biological activity. A putative application of the SBI library could then be the selection of low-molecular-weight fragments for NMR screening.[35−37] Because of their small size, the scaffolds selected herein present a relatively high average self-similarity (average Tanimoto coefficient of 0.74 using FCFP_4 fingerprints). Customizing a fragment library out of the SBI data set would, therefore, require the selection of the "least-substituted" compounds for a subset of dissimilar molecular scaffolds.
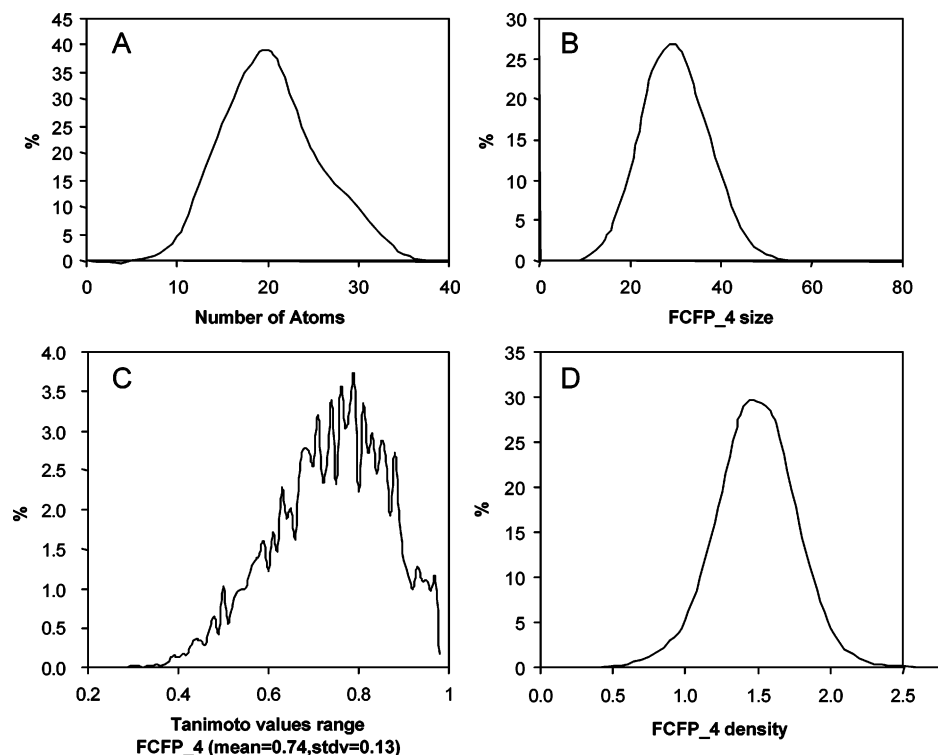
Figure 9. Analyzing the molecular complexity of the scaffold library. (A) Number of heavy atoms. (B) FCFP_4 size: number of bits set in the SciTegic functional connectivity fingerprints[48] using a fragment diameter up to four bonds. (C) Self-similarity plot using FCFP_4 fingerprints and Tanimoto coefficient. (D) FCFP_4 density:[34] FCFP_4 size/number of heavy atoms.

It should be noted that several scaffold-based libraries have already been reported in the past. Agrafiotis et al.[38] described a probe library based on 50 representative scaffolds comprising 300 000 druglike compounds dedicated to primary screening. Another design of a scaffold library was recently reported by Card et al.,[39] where 275 555 compounds (starting with 1 994 133 molecules from 17 vendors, then filtered by MW range) have been clustered according to their constituent fragments (segmented at rotatable bonds) and similar compounds were grouped (Tanimoto index > 0.85). This resulted in 20 360 small molecular-weight fragments covering approximately 80% of the scaffold component space. Our library presents the advantage of covering most commercially available compounds and archiving scaffolds as a medicinal chemist would do by intuition, thus enabling an easy navigation in scaffold space and the selection of the most relevant compounds according to simple user-defined queries.

On the Use of ClassPharmer Scaffolds. In the current study, we have considered a scaffold from the ClassPharmer definition: a "chemically aware" MCS taking into account its chemical environment (e.g., a MCS substituted by an aliphatic carbon chain will be different from the same one substituted by an aromatic ring). We are aware that many different scaffold definitions are possible (recall Figure 1) and that the partitioning of compounds will be dependent upon the selected scaffold definition. There are both advantages and drawbacks in utilizing ClassPharmer for computing molecular scaffolds out of large libraries. A first true advantage is the ad hoc detection of MCS, which enables a classification of all compounds of the library. Alternative strategies based on the storage of precomputed chemical features[23] do not guarantee this exhaustiveness. Second, the fuzziness of ring closure and atom match definitions can be customized depending on the purpose. Here, we chose exact definitions of the latter parameters to ensure a chemically unique definition of each scaffold. Although tolerating nonexact atom matches would enable taking into account bioisosterism in the scaffold definition and, thus, significantly decrease the number of scaffolds, fuzzy ring closure is clearly not suited for archiving scaffolds as it would allow the definition of inadequate substructures (e.g., three connected carbon atoms of a phenyl ring) as classes. A postclassification analysis of scaffolds present in our database is still possible, notably by comparing their nonbonded interaction potentials[40] or molecular shapes.[41]

Third, ClassPharmer MCSs describe not only the minimum common substructure but also its chemical environment, which enables a classification mirroring, for the most part, the intuition of a medicinal chemist. Hence, many scaffolds already identified by vendors within their collections[42] can be recovered in the SBI scaffold library. The ClassPharmer MCS presents the advantage of being of various sizes (from a ring scaffold to a superstructure, recall Table 1) and, thus, reconciles multiple definitions of a scaffold. Proposed classifications are easier to interpret (notably for a medicinal chemist) than those arising from more complex hierachical descriptions.[13,43−45] Last, importing compounds from a new collection into an existing classification is straightforward and allows the quick evaluation of the scaffold overlap of different collections.

A clear drawback of our approach is its low speed. When a standard PC with 1 GB of RAM is used, only collections with less than 150 000 compounds can be classified within 48 CPU hours. The regular upgrade of the scaffold library is, thus, considerably penalized. Meanwhile alternative classification approaches using hiercharchial descriptions[13,43−45]

SCAFFOLD DIVERSITY OF SCREENING LIBRARIES

J. Chem. Inf. Model., Vol. 46, No. 2, 2006 **523**

or combining fingerprints and MCS methods[46] have been developed and might be considered under the conditions that (i) it also produces chemically meaningful classes and (ii) a significant increase in performance can be observed for the same initial (huge) library size.

A limitation, for the purpose of scaffold archiving, is the redundancy observed in the clustering (e.g., a particular compound is often found in multiple classes). Although class redundancy is not necessarily a problem in mining HTS data, as it exactly reflects the point of view of a medicinal chemist, it was a real hurdle in our study to quantify the population covered by each class. To overcome this problem, we developed a very simple approach which selects the most "central scaffold" of each compound. It should be stated that our protocol still generates a significant number of singletons. Because of the overall low speed of the classification procedure, we have not considered merging all singletons and reclassifying this subset to populate existing classes or to generate additional clusters. Likewise, reclustering singletons by similarity to existing cluster substructures[46] is another interesting alternative to reduce the number of singletons.

It is clear that different scaffold definition and clustering methods will lead to quite different outcomes for a single library. The herein-described statistics are, therefore, likely to be very sensitive to the archiving protocol.

CONCLUSIONS

The molecular diversity of 17 commercially available screening collections covering 2.4 million compounds was evaluated by computing graph-based maximum common substructures for each library. Two metrics (NC50C and PC50C) were developed to facilitate the comparison of libraries of various sizes. The herein-analyzed commercial collections could be grouped into four categories depending on their size and PC50C value (percentage of scaffolds accounting for 50% of the classified compounds). Our classification reflects the history of each collection and the way it had been compiled (combinatorial libraries and medicinal chemistry libraries). Merging all classes led to a library of nonredundant scaffolds that can easily be browsed for different purposes such as (i) defining a scaffold-focused library[47] starting from an existing hit and, thus, quickly generating structure–activity relationships, (ii) defining a general purpose library where a few copies of user-selected diverse scaffolds are cherry picked,[39] and (iii) setting up a collection of small molecular-weight fragments for structural biology screening[37] (X-ray and NMR) by selecting the least substituted compound(s) for user-defined classes.

ACKNOWLEDGMENT

Supporting Information Available: Chart A, filtering rules in OpenEye Filter program; Chart B, queries to analyze the overlap of the screening collections; Scheme A, database scheme with the central table structures containing the essential information relative to 21 393 scaffolds; Table A, number of classified compounds overlapping pairwise; Table B, percentage of overlap of classified compounds of a database A with database B; Table C, number of overlapping classes by pairwise comparison; and Table D, percentage of overlapping classes by pairwise comparison of a database A with database B. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) National Chemical Library, National Center for Scientific Research (CNRS). http://chimiotheque.ujf-grenoble.fr/induk.html (December 2005).

(2) Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. NIH Molecular Libraries Initiative. *Science* **2004**, *306*, 1138−1139.

(3) Webb, T. R. Current directions in the evolution of compound libraries. *Curr. Opin. Drug Discovery Dev.* **2005**, *8*, 303−308.

(4) Bocker, A.; Derksen, S.; Schmidt, E.; Teckentrup, A.; Schneider, G. A hierarchical clustering approach for large compound libraries. *J. Chem. Inf. Model.* **2005**, *45*, 807−815.

(5) Martin, Y. C. Diverse viewpoints on computational aspects of molecular diversity. *J. Comb. Chem.* **2001**, *3*, 231−250.

(6) Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.; Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greaney, P.; Morley, D.; Hubbard, R. E. Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 643−651.

(7) Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular diversity in chemical databases: comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750−763.

(8) Shemetulskis, N. E.; Dunbar, J. B., Jr.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the diversity of a corporate database using chemical database clustering and analysis. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 407−416.

(9) Sirois, S.; Hatzakis, G.; Wei, D.; Du, Q.; Chou, K. C. Assessment of chemical libraries for their druggability. *Comput. Biol. Chem.* **2005**, *29*, 55−67.

(10) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: Using the MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.

(11) Voigt, J. H.; Bienfait, B.; Wang, S.; Nicklaus, M. C. Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702−712.

(12) Nilakantan, R.; Bauman, N.; Haraki, K. S. Database diversity assessment: new ideas, concepts, and tools. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 447−452.

(13) McFayden, I.; Walker, G.; Alvarez, J. Enhancing hit quality and diversity within assay throughput constraints. In *Chemoinformatics in Drug Discovery*; Oprea, T., Ed.; Wiley-VCH: Weinheim, Germany, 2004; pp 143−173.

(14) McGregor, J. J.; Willett, P. Use of a maximum common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 137−140.

(15) Su, A. I.; Lorber, D. M.; Weston, G. S.; Baase, W. A.; Matthews, B. W.; Shoichet, B. K. Docking molecules by families to increase the diversity of hits in database screens: computational strategy and experimental evaluation. *Proteins* **2001**, *42*, 279−293.

(16) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(17) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsions: A new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.

(18) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP-retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.

(19) *ClassPharmer Suite*, version 3.2−3.5; Bioreason, Inc.: Santa Fe, NM.

(20) Charifson, P. S.; Walters, W. P. Filtering databases and chemical libraries. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 311−323.

(21) *Filter 1.0*; OpenEye Scientific Software, Inc.: Santa Fe, NM.

(22) *Cliff, 1.23*; Molecular Networks GmbH, D-91052 Erlangen, Germany.

(23) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E., Jr. LeadScope: Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302−1314.

(24) *OEChem*, version 1.3; OpenEye Scientific Software, Inc.: Santa Fe, NM.

(25) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(26) Willett, P. An algorithm for chemical superstructure searching. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 114−116.

(27) Brown, R. D.; Downs, G. M.; Willett, P.; Cook, A. P. F. A hyperstructure model for chemical structure handling: Generation and atom-by-atom searching of hyperstructures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 522−531.

(28) *InChI (IUPAC International Chemical Identifier)*, version 1.0; IUPAC: Research Triangle Park, NC, 2005.

(29) *JChemBase*; ChemAxon Ltd.: Budapest, Hungary.

(30) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.

(31) Nilakantan, R.; Nunn, D. S. A fresh look at pharmaceutical screening library design. *Drug Discovery Today* **2003**, *8*, 668−672.

(32) Xue, L.; Bajorath, J. Distribution of molecular scaffolds and R-groups isolated from large compound databases. *J. Mol. Model.* **1999**, *5*, 97−102.

(33) Feher, M.; Schmidt, J. M. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218−227.

(34) Selzer, P.; Roth, H. J.; Ertl, P.; Schuffenhauer, A. Complex molecules: do they add value? *Curr. Opin. Chem. Biol.* **2005**, *9*, 310−316.

(35) Baurin, N.; Aboul-Ela, F.; Barril, X.; Davis, B.; Drysdale, M.; Dymock, B.; Finch, H.; Fromont, C.; Richardson, C.; Simmonite, H.; Hubbard, R. E. Design and characterization of libraries of molecular fragments for use in NMR screening against protein targets. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2157−2166.

(36) Schuffenhauer, A.; Ruedisser, S.; Marzinzik, A. L.; Jahnke, W.; Blommers, M.; Selzer, P.; Jacoby, E. Library design for fragment based screening. *Curr. Top. Med. Chem.* **2005**, *5*, 751−762.

(37) Zartler, E. R.; Shapiro, M. J. Fragonomics: fragment-based drug discovery. *Curr. Opin. Chem. Biol.* **2005**, *9*, 366−370.

(38) Agrafiotis, D. K.; Lobanov, V. S.; Salemme, F. R. Combinatorial informatics in the post-genomics era. *Nat. Rev. Drug Discovery* **2002**, *1*, 337−346.

(39) Card, G. L.; Blasdel, L.; England, B. P.; Zhang, C.; Suzuki, Y.; Gillette, S.; Fong, D.; Ibrahim, P. N.; Artis, D. R.; Bollag, G.; Milburn, M. V.; Kim, S. H.; Schlessinger, J.; Zhang, K. Y. A family of phosphodiesterase inhibitors discovered by cocrystallography and scaffold-based drug design. *Nat. Biotechnol.* **2005**, *23*, 201−207.

(40) Watson, P.; Willett, P.; Gillet, V. J.; Verdonk, M. L. Calculating the knowledge-based similarity of functional groups using crystallographic data. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 835−857.

(41) Sauer, W. H.; Schwarz, M. K. Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 987−1003.

(42) De Laet, A.; Hehenkamp, J. J. J.; Wife, R. L. Finding drug candidates in virtual and lost/emerging chemistry. *J. Heterocycl. Chem.* **2000**, 669−674.

(43) Cases, M.; Garcia-Serna, R.; Hettne, K.; Weeber, M.; van der Lei, J.; Boyer, S.; Mestres, J. Chemical and biological profiling of an annotated compound library directed to the nuclear receptor family. *Curr. Top. Med. Chem.* **2005**, *5*, 763−772.

(44) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical scaffold clustering using topological chemical graphs. *J. Med. Chem.* **2005**, *48*, 3182−3193.

(45) Xu, Y. J.; Johnson, M. Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 912−926.

(46) Stahl, M.; Mauser, H. Database clustering with a combination of fingerprint and maximum common substructure methods. *J. Chem. Inf. Model.* **2005**, *45*, 542−548.

(47) Krier, M.; Araujo-Junior, J. X.; Schmitt, M.; Duranton, J.; Justiano-Basaran, H.; Lugnier, C.; Bourguignon, J. J.; Rognan, D. Design of small-sized libraries by combinatorial assembly of linkers and functional groups to a given scaffold: application to the structure-based optimization of a phosphodiesterase 4 inhibitor. *J. Med. Chem.* **2005**, *48*, 3816−3822.

(48) *Pipeline Pilot*, version 4.2; SciTegic Inc.: San Diego, CA.

CI050352V