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

Received August 30, 2005

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” or the NIH Molecular Libraries Initiative, 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 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, is the way of measuring the diversity of a compound library. Very informative testimonials about this key aim were shared with the community elsewhere. Nevertheless, molecular diversity is heavily dependent on the descriptors, metrics, and multivariate methods used to assess it. Most studies on commercially available compound libraries have traditionally used physicochemical and topological descriptors, summed up into a score or encoded into fingerprints or hash codes 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), the largest rigid fragment or rings, molecular frameworks or Murcko’s scaffolds with or without descriptors (e.g., topological torsions), and molecular fragments as generated by the RECAP method. 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 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.
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. 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 implemented in OpenEye’s Filter program were then used to select the most suitable compounds 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 (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 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.

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. 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:

\[
\text{CSS} = \frac{\text{MW}_{\text{compound}} - \text{MW}_{\text{scaffold}}}{N_R}
\]

where \(\text{MW}_{\text{compound}}\) is the molecular weight of a compound, \(\text{MW}_{\text{scaffold}}\) 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 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 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, and is outlined as follows:

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

As for the removal of interclass redundancy, the R-group decomposition was implemented in Python on the basis of OpenEye OEChem.

**Setting Up a Scaffold Library.** All classes (excluding the singletons) were assembled from the generated classifications to form a scaffold library. Computing InChI 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 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 D₂ antagonists (Figure 4) randomly chosen from the Hert data set 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),
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., phenylpiperazine) but with a different chemical environment (the benzofuranyl-piperazine 2 is not classified with the phenylpiperazines 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

Figure 4. Structure of 20 true dopamine D-2 antagonists randomly extracted from the Hert data set.30
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 drug-like 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 ($250 < \text{MW} < 500$) 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 (ASIg, 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)
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.\textsuperscript{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.\textsuperscript{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 4,390 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 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 3. Classification Results

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

\* Classification defined as containing at least two unique compounds. \*\* Classification defined as containing at least 25 unique compounds. \*\*\* Percentage of interclass redundancy (percentage of compounds present in multiple classes). \*\*\*\* Number of classes accounting for 50\% of classified compounds. \*\*\*\*\* Percentage of classes accounting for 50\% of classified compounds.

### Table 4. Example of Interclass Redundancy

<table>
<thead>
<tr>
<th>Compound</th>
<th>MW(scaffold)</th>
<th>NR</th>
<th>CSS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\* The compound exemplified here (CD 05668, Maybridge) has a molecular weight of 413.32 and three proposed scaffolds highlighted in bold. \* NR: number of R groups. \* CSS (Central scaffold score) = (MW\textsubscript{compound} - MW\textsubscript{scaffold})/NR.
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, 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 5.** Classification of Collections According to Their Size and Relative Scaffold Diversity (PC50C)

<table>
<thead>
<tr>
<th>Category</th>
<th>libraries</th>
<th>size$^b$</th>
<th>PC50C$^c$</th>
<th>PC50C$_{25}$$^d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>large combinatorial libraries</td>
<td>CBG, CDIc, IBSs</td>
<td>&gt;100 K</td>
<td>&lt;3</td>
<td>&lt;13</td>
</tr>
<tr>
<td>medium combinatorial libraries</td>
<td>ASIg, ASIp, SPE, TRI, VITs</td>
<td>50–100 K</td>
<td>&lt;3</td>
<td>&lt;13</td>
</tr>
<tr>
<td>diverse libraries</td>
<td>CDIi, CST, IBSn, TIMs</td>
<td>&lt;50 K</td>
<td>&lt;4</td>
<td>10–15</td>
</tr>
<tr>
<td>highly diverse libraries</td>
<td>CNR, MAY, MDDR, NET, TIMn, VITt</td>
<td>&lt;50 K</td>
<td>&lt;4</td>
<td>&gt;15</td>
</tr>
</tbody>
</table>

$a$ 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.

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

<table>
<thead>
<tr>
<th>Category</th>
<th>Scaffold</th>
<th>Identifier</th>
<th>Suppliers</th>
<th>Uniques compounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large combinatorial Libraries</td>
<td><img src="large_combinatorial.png" alt="Image" /></td>
<td>SBI_4853</td>
<td>15</td>
<td>22 988</td>
</tr>
<tr>
<td>Medium combinatorial Libraries</td>
<td><img src="medium_combinatorial.png" alt="Image" /></td>
<td>SBI_2909</td>
<td>10</td>
<td>8 592</td>
</tr>
<tr>
<td>Diverse Libraries</td>
<td><img src="diverse.png" alt="Image" /></td>
<td>SBI_2654</td>
<td>1</td>
<td>322</td>
</tr>
<tr>
<td>Highly Diverse Libraries</td>
<td><img src="highly_diverse.png" alt="Image" /></td>
<td>SBI_21089</td>
<td>1</td>
<td>106</td>
</tr>
</tbody>
</table>
Figure 7. Scaffold diversity of screening collections.
collections labeled to 22 academic laboratories, each of them with a different alphabetical order. For duplicate classes, a single copy has been conserved corresponding to the first encountered library sorted by alphabetical order. For example, the French National Chemical library classes are really diverse in terms of scaffold architecture and generally present a larger choice of proprietary low-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 category of highly diverse libraries (CNR, MAY, NET, ASIp, SPE, TRI, and VITs). In a third group are found 5 and 6) of medium-sized combinatorial libraries (ASIp, ASIIp, SPE, TRI, and VITt). 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) is a repository of compounds collected at 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. 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 (Scaffold of the Boinformatics 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 7. Distribution of Classes for the SBI Scaffold Library

<table>
<thead>
<tr>
<th>library</th>
<th>classes</th>
<th>exclusive classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASIIg</td>
<td>3485</td>
<td>1240 (36%)</td>
</tr>
<tr>
<td>ASIIp</td>
<td>1964</td>
<td>1729 (88%)</td>
</tr>
<tr>
<td>CBG</td>
<td>3194</td>
<td>1213 (38%)</td>
</tr>
<tr>
<td>CDIIc</td>
<td>3425</td>
<td>2325 (68%)</td>
</tr>
<tr>
<td>CDIIi</td>
<td>2306</td>
<td>974 (42%)</td>
</tr>
<tr>
<td>CNR</td>
<td>391</td>
<td>299 (76%)</td>
</tr>
<tr>
<td>CST</td>
<td>1010</td>
<td>307 (30%)</td>
</tr>
<tr>
<td>IBSn</td>
<td>756</td>
<td>504 (67%)</td>
</tr>
<tr>
<td>IBSs</td>
<td>3484</td>
<td>1845 (57%)</td>
</tr>
<tr>
<td>MAY</td>
<td>1543</td>
<td>1052 (68%)</td>
</tr>
<tr>
<td>NET</td>
<td>941</td>
<td>722 (77%)</td>
</tr>
<tr>
<td>SPE</td>
<td>3260</td>
<td>1304 (46%)</td>
</tr>
<tr>
<td>TIMn</td>
<td>162</td>
<td>48 (30%)</td>
</tr>
<tr>
<td>TIMs</td>
<td>1954</td>
<td>700 (36%)</td>
</tr>
<tr>
<td>TRI</td>
<td>1338</td>
<td>1098 (82%)</td>
</tr>
<tr>
<td>VITs</td>
<td>2149</td>
<td>759 (35%)</td>
</tr>
<tr>
<td>VITt</td>
<td>402</td>
<td>264 (66%)</td>
</tr>
<tr>
<td>MDDR</td>
<td>3057</td>
<td>2696 (88%)</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th># of DBs</th>
<th># of scaffolds in at least n DBs</th>
<th># of scaffolds in exactly n DBs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21393</td>
<td>16583</td>
</tr>
<tr>
<td>2</td>
<td>4810</td>
<td>2532</td>
</tr>
<tr>
<td>3</td>
<td>2278</td>
<td>1009</td>
</tr>
<tr>
<td>4</td>
<td>1269</td>
<td>501</td>
</tr>
<tr>
<td>5</td>
<td>768</td>
<td>300</td>
</tr>
<tr>
<td>6</td>
<td>468</td>
<td>179</td>
</tr>
<tr>
<td>7</td>
<td>289</td>
<td>97</td>
</tr>
<tr>
<td>8</td>
<td>192</td>
<td>71</td>
</tr>
<tr>
<td>9</td>
<td>121</td>
<td>39</td>
</tr>
<tr>
<td>10</td>
<td>82</td>
<td>25</td>
</tr>
<tr>
<td>11</td>
<td>57</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>37</td>
<td>19</td>
</tr>
<tr>
<td>13</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th># of DBs</th>
<th># of scaffolds in at least n DBs</th>
<th># of scaffolds in exactly n DBs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2498</td>
<td>921</td>
</tr>
<tr>
<td>2</td>
<td>1577</td>
<td>431</td>
</tr>
<tr>
<td>3</td>
<td>1146</td>
<td>106</td>
</tr>
<tr>
<td>4</td>
<td>853</td>
<td>251</td>
</tr>
<tr>
<td>5</td>
<td>602</td>
<td>194</td>
</tr>
<tr>
<td>6</td>
<td>408</td>
<td>133</td>
</tr>
<tr>
<td>7</td>
<td>275</td>
<td>84</td>
</tr>
<tr>
<td>8</td>
<td>191</td>
<td>70</td>
</tr>
<tr>
<td>9</td>
<td>121</td>
<td>39</td>
</tr>
<tr>
<td>10</td>
<td>82</td>
<td>25</td>
</tr>
<tr>
<td>11</td>
<td>57</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>37</td>
<td>19</td>
</tr>
<tr>
<td>13</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Statistics are only given for comparison purposes.
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. 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. 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 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.
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 features23 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 potentials40 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 collections42 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 hierarchical 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 hierarchical descriptions13,43–45

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 fingerprints46 using a fragment diameter up to four bonds. (C) Self-similarity plot using FCFP_4 fingerprints and Tanimoto coefficient. (D) FCFP_4 density:44 FCFP_4 size/number of heavy atoms.
Scaffold Diversity of Screening Libraries

or combining fingerprints and MCS methods 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 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 on the Internet at http://pubs.acs.org.

ACKNOWLEDGMENT

M.K. is the recipient of a CIFRE Grant (No. 738/2002) provided by the “Association Nationale pour la Recherche Technique” and Idéalp Pharma (Villeurbanne, France). We thank Pat Bacha and Vincent Vivien (Bioreason Inc.) for their support throughout the course of this work.

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


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


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

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


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