

FULL PAPER

## Distribution of Molecular Scaffolds and R-Groups Isolated from Large Compound Databases

Ling Xue<sup>1,†</sup> and Jürgen Bajorath<sup>1,2</sup>

<sup>1</sup>MDS Panlabs, Computational Chemistry & Informatics, 11804 North Creek Pkwy. S., Bothell, WA, 98011-8805, USA. Tel: +1-425-487-8297; Fax: +1-425-487-8276; E-mail: jbajorath@panlabs.com

<sup>2</sup>Department of Biological Structure, University of Washington, Seattle, WA 98195, USA.

Received: 16 March 1999/ Accepted: 28 April 1999/ Published: 28 May 1999

**Abstract** We describe an approach to isolate molecular scaffolds and R-groups from known chemical compounds in order to generate scaffold and R-group databases from two large compound collections, Optiverse™ and Maybridge™. The distributions of molecular scaffolds and R-groups in the parent databases were analysed and compared. We find that a limited number of scaffolds and R-groups account for the majority of database compounds and that most of the scaffolds occur only once or twice in the compound databases. Diversity analysis suggests that the compound and scaffold databases have similar molecular diversity. Implications for library design are discussed.

**Keywords** Compound libraries, Diversity, Molecular scaffolds, R-groups, Statistical distribution

### Introduction

Combinatorial chemistry and high throughput screening methods have modified approaches to drug discovery in a significant way [1,2]. Many thousands of compounds can now be prepared by combinatorial means and screened against biological targets. This scenario is changing the way computational chemistry supports drug discovery programs. The efficient design of combinatorial libraries and the analysis of large amounts of screening data have become major challenges for computational chemists. Molecular diversity

and similarity are important aspects of combinatorial library design [3]. Libraries may be designed reaction-based, i.e., by specifying starting materials and chemical transformations for different reactions, or product-based, i.e., by generating combinations of defined molecular frameworks, or scaffolds, and R-groups [4,5]. A major challenge in library design is to find a reasonable balance between molecular diversity, however defined, and synthetic feasibility of computed compounds [5]. Product-based design strategies depend on the availability of molecular scaffolds. If scaffold libraries are available, synthetically accessible scaffolds can be pre-selected before compound libraries are computed. This provides an opportunity, among others, to consider synthetic feasibility of compounds early in the design process. To aid in product-based library design, we generate and analyse databases of molecular scaffolds from known compounds. Here we describe a computational method to isolate scaffolds and R-groups and apply this technique to two

Correspondence to: J. Bajorath

<sup>†</sup>Present address: Thetagen Inc., 600 Broadway, Seattle, WA 98122, USA.

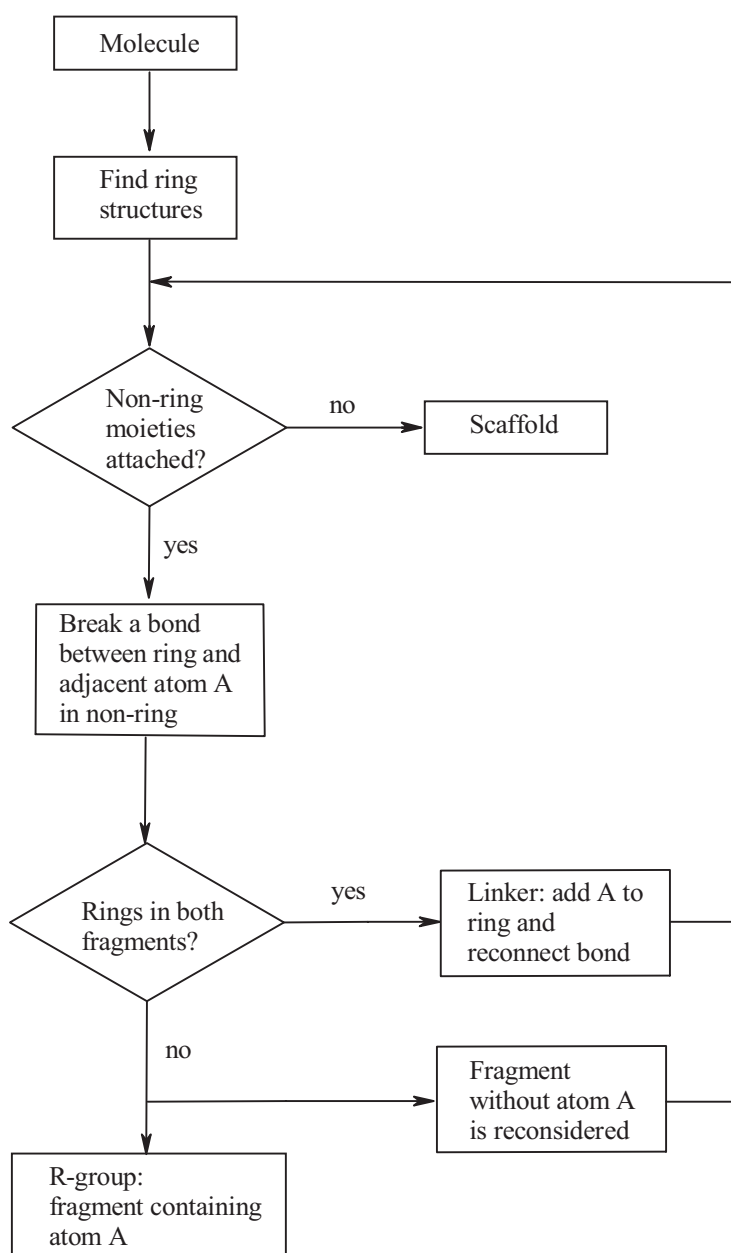
conceptually different databases. Analysis of scaffold and R-group distributions shows that a limited number of recurrent molecular fragments dominate the composition of compounds in the evaluated databases.

## Materials and methods

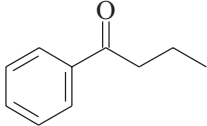

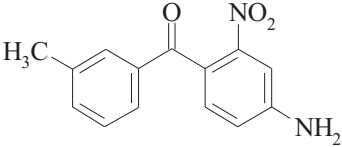
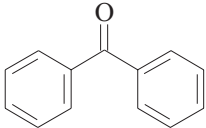
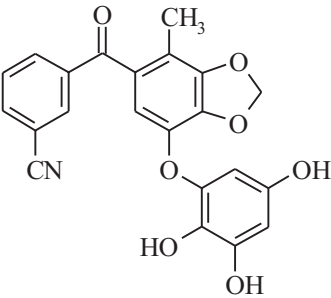
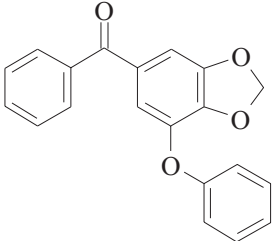
For our analysis, two different compound databases, Optiverse<sup>TM</sup> (OV) [6] and Maybridge<sup>TM</sup> (MB) [7] were used as examples. OV is a screening library, based on diversity design [8], and contains 117,976 compounds. MB contains 58,239 compounds and intermediates commonly used in

medicinal chemistry. Only a very limited number of compounds in these databases (1,214 in OV and 1,060 in MB) contain no ring structures. These compounds were removed from the databases prior to our analysis. All computations were carried out with MOE [9]. As described in the Results section, an algorithm was developed to isolate scaffolds and R-groups from known compounds. This algorithm was implemented in MOE using SVL code [10]. Average Tanimoto coefficients ( $T_c$ ) [11] were calculated using a 2D fingerprint [12]. Partitioning of compound and scaffold databases into unique classes of molecules was carried out using the QuaSAR-Cluster function [13] of MOE using 57 MDL SSKey-type structural fragments [14,15], the number of aro-

**Figure 1** The diagram illustrates the approach applied in this study to isolate scaffolds and R-groups from compound databases. For each compound, we identify ring structures, break bonds between rings and the rest of the molecule, and determine whether the resulting fragments contain R-groups. If not, the broken bond is part of a linker and the reconnected fragments represent a possible scaffold. The process is repeated until no new connection points and R-groups are found



**Table 1** Representative scaffolds and R-groups isolated following the approach described in Figure 1

Molecule [a]	Scaffold	R-groups
		COCH <sub>2</sub> CH <sub>2</sub> CH <sub>3</sub>
		NO <sub>2</sub> , CH <sub>3</sub> , NH <sub>2</sub>
		CH <sub>3</sub> , CN, OH

[a] prototypic compounds

matic bonds, hydrogen bond acceptors, and fraction of rotatable bonds per molecule as descriptors [16].

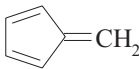


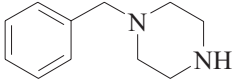

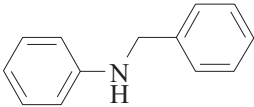

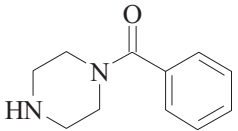
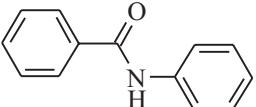
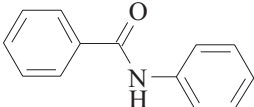

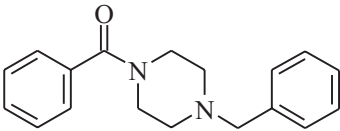
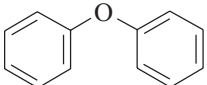

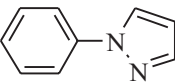
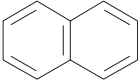


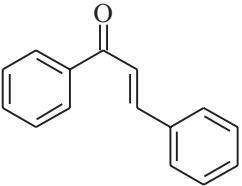
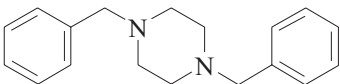
## Results

We use a hierarchical description of molecules, akin to, for example, Bemis & Murcko [17], and define a "scaffold" as a molecular fragment without R-groups, and an "R-group" as any functional group or (non-ring) side chain with only one connection point to the rest of the molecule. R-groups are distinct from linkers that connect ring structures and that are part of scaffolds. Table 1 shows prototypic examples of compounds, scaffolds, and R-groups. An algorithm to isolate scaffolds and R-groups is described in Figure 1. This algorithm has been implemented in MOE using SVL (the program is provided as supplementary material). Using this approach, 52,529 unique scaffolds and 4,486 R-groups were isolated from OV compounds and 15,690 scaffolds and 2,851 R-groups from MB. Only 2,945 scaffolds and 407 R-groups were identical in OV and MB. Thus, combined databases contain 65,274 scaffolds and 6,930 R-groups.

The ratio of the number of original compounds to scaffolds is 2.2 for OV and 3.7 for MB. Thus, on average, we isolate a unique scaffold from every two OV and from every four MB compounds. However, the distribution of scaffolds in the databases is far from average. For both OV and MB, a small number of scaffolds represent a large percentage of the compounds. Table 2 lists some of the scaffolds that dominate the composition of MB and OV molecules. In OV, 4,205 scaffolds (8% of all scaffolds) account for 50% of all compounds. Thus, 48,324 scaffolds (92%) occur in 58,988 compounds (compound/scaffold ratio of 1.2). In MB, 767 scaffolds (5%) account for 50% of the compounds. Thus, ~14,923 scaffolds (95%) occur in 29,120 compounds (compound/scaffold ratio of 1.9). Thus, in both databases, more than 90% of the scaffolds occur only once or twice. Table 2 shows that the most frequently observed scaffolds are small aromatic structures or heterocycles and that OV and MB have three identical scaffolds in their top ten lists.

Significant trends are also detected when R-group distributions are analysed. Small numbers of R-groups are found in the majority of compounds. Table 3 lists the prevalent R-groups in OV and MB. In both cases, the top ten R-groups account for almost 75% of R-groups in the databases, and

**Table 2** Top ten scaffolds in the Optiverse and Maybridge databases

No.	Maybridge Scaffold	Percent [a]	Optiverse Scaffold	Percent [a]
1		7.76		4.46
2		3.14		0.70
3		1.47		0.50
4		1.23		0.44
5		0.60		0.39
6		0.49		0.37
7		0.48		0.36
8		0.41		0.35
9		0.38		0.32
10		0.37		0.29

[a] fraction of database compounds that share this scaffold

**Table 3** Top ten R-groups in the Optiverse and Maybridge databases

R-group	Rank	Percent	Accumulated %
<b>Optiverse:</b>			
-CH <sub>3</sub>	1	26.5	26.5
-OCH <sub>3</sub>	2	11.8	38.3
-OH	3	9.0	47.3
-Cl	4	8.9	56.2
-NO <sub>2</sub>	5	5.6	61.7
-F	6	3.5	65.3
-CF <sub>3</sub>	7	2.9	68.2
-Br	8	2.5	70.7
-CH <sub>2</sub> CH <sub>3</sub>	9	1.8	72.5
-COOCH <sub>2</sub> CH <sub>3</sub>	10	1.5	73.9
<b>Maybridge:</b>			
-Cl	1	20.1	20.1
-CH <sub>3</sub>	2	19.7	39.7
-CF <sub>3</sub>	3	7.1	46.9
-OCH <sub>3</sub>	4	6.0	52.9
-F	5	4.9	57.8
-NO <sub>2</sub>	6	4.8	62.6
-OH	7	3.7	66.3
-CN	8	3.1	69.5
-NH <sub>3</sub>	9	2.8	72.3
-Br	10	2.5	74.8

"Rank" is based on the occurrence of the R-group in the database compounds and given in "Percent". "Accumulated %" means the total percentage of the R-group in the database together with higher ranked R-groups (e.g., -CH<sub>3</sub>, -OCH<sub>3</sub>, and -OH account for 47.3 % of all R-groups in the Optiverse database).

the majority of R-groups occur only once. Common organic functional groups, including halogen substituents, nitro- and hydroxyl-groups are among the top ten R-groups. The methyl group alone accounts for 25% of the R-group occurrences in OV and 20% in MB.

Table 4 summarises results of diversity analysis on the compound and scaffold databases. Similar average  $T_c$  values between 0.3 and 0.4 were obtained for both OV and MB and the resulting scaffold databases. Since average  $T_c$  calculations can only detect significant differences in overall diversity, we have also partitioned the compound and scaffold databases to determine the number of unique compound classes in each database [13]. Although the compound databases are much larger than the scaffold databases, the number of unique classes and unique compounds are similar in each case. Thus, on the basis of the calculations summarised in Table 4, the molecular diversity of isolated scaffolds is similar to the diversity of the compound databases.

## Discussion

A variety of methods have been developed to identify biologically active molecules by database analysis [18,19], screen databases for molecules with desired properties [20], and identify drug-like molecules [21,22]. By contrast, our approach was primarily developed to sample molecular scaffolds for product-based library design. It is important to note that our algorithm follows a hierarchical description of molecules and is not reaction-based. Thus, it is conceptually different from techniques that divide molecules on the basis of chemical reaction information such as RECAP [23]. The method described here is in part similar to the molecular framework analysis of Bemis and Murcko [17] who analysed 5,120 compounds in the Comprehensive Medicinal Chemistry database [14]. It was found that these 5,120 molecules include 1,179 distinct frameworks and that 32 of these frameworks account for the shape of 50% of these molecules, when 2D shape descriptors were applied [17].

Our algorithm was designed to analyse compounds with ring structures, and non-ring compounds were therefore deleted from OV and MB prior to application of the method. Nevertheless, we were able to analyse the OV and MB databases, since only 2,274 of a total of 176,215 compounds in these databases (~1.3 %) did not contain ring structures. OV and MB were selected as examples because they represent different types of compound databases, a diverse combinatorial screening library (OV) and a compound collection focused on molecules commonly used in medicinal chemistry (MB). Our analysis revealed that a limited number of scaffolds and R-groups dominate the composition of compounds in both databases, and that more than 90% of identified molecular scaffolds occur only once or twice in these large compound collections. In cluster analysis studies on a variety of compound databases using  $T_c$  calculations with a 75% similarity cutoff, a 40% overlap between a small subset of OV and MB but no undistinguishable compounds ( $T_c = 1$ ) were detected [15]. Our findings provide a molecular explanation for these earlier observations.

The compound and scaffold databases display a very similar degree of diversity, suggesting that R-groups, as defined herein, do not significantly increase molecular diversity in these databases. This suggestion may seem counter-intuitive but is consistent with two of our findings: The vast majority of scaffolds occur only in one compound and, secondly, a small number of R-groups dominate the compounds in both databases. Thus, scaffolds largely capture the diversity of database compounds, when isolated as described here.

Taken together, our findings have several implications for product-based library design. The relatively small sets of preferentially used molecular scaffolds in these databases provide preferred pathways for compound synthesis. However, a less biased distribution of scaffolds in designed libraries may be desirable. The diversity encoded in these molecular scaffolds can only be significantly increased if scaffolds are decorated with many different combinations of R-groups. Thus, design strategies that explore combinations of care-

**Table 4** Diversity analysis. The number of unique classes and the number of singletons (unique molecules), obtained by partitioning of the database, are reported for each compound and scaffold database

	Maybridge				Optiverse			
	Size	Unique classes	Singletons	T <sub>c</sub> [a]	Size	Unique classes	Singletons	T <sub>c</sub> [a]
Compounds	58,239	337	18	0.33	117,976	402	8	0.36
Scaffolds	15,690	358	19	0.36	52,529	390	10	0.37

[a] average Tanimoto coefficient

fully selected molecular scaffolds and large R-group libraries are thought to provide a balance between molecular diversity and synthetic feasibility of computed compounds.

**Supplementary material available** The SVL code to isolate scaffolds and R-groups from compound databases is provided to accompany the manuscript.

**Acknowledgment** We wish to thank Jeff Godden for many helpful discussions and suggestions.

## References

- Kauvar, L. M.; Laborde, E. *Curr. Opin. Drug Discovery and Development* **1998**, *1*, 66.
- Kubinyi, H. *Curr. Opin. Drug Discovery and Development* **1998**, *1*, 16.
- Bures, M. G.; Martin, Y. C. *Curr. Opin. Chemical Biology* **1998**, *2*, 376.
- Gillet, V. J.; Willet, P.; Bradshaw, J. J. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731.
- Ferguson, A. M.; Patterson, D. E.; Garr, C. D., Underiner, T. L. *J. Biomol. Screen.* **1996**, *1*, 65.
- Garr, C. D.; Peterson, J. R.; Schultz, L.; Oliver, A. R.; Underiner, T. L.; Cramer, R. D.; Ferguson, A. M.; Lawless, M. S.; Patterson, D. E. *J. Biomol. Screen.* **1996**, *1*, 179.
- Maybridge Chemical Company LTD, Trevillet, Tintagel, Cornwall PL34 OHW, UK.
- Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. *J. Med. Chem.* **1996**, *39*, 3049.
- MOE, Molecular Operating Environment, Chemical Computing Group, Montreal, Canada; <http://www.chemcomp.com/feature/deploy.htm>
- SVL, Scientific Vector Language; <http://www.chemcomp.com/feature/svl.htm>
- Flower, D. R. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379.
- Sheridan, R. P.; Bush, B. L. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756.
- QuaSAR-Cluster; <http://www.chemcomp.com/article/cluster.htm>
- MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA.
- McGregor, M. J.; Pallai, P. V. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443.
- Xue, L.; Godden, J.; Gao, H.; Bajorath, J. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, in press.
- Bemis, G. W.; Murcko, M. A. *J. Med. Chem.* **1996**, *39*, 2887.
- Gillet, V. J.; Willet, P.; Bradshaw, J. J. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165.
- Gao, H.; Williams, C. I.; Labute, P.; Bajorath, J. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 164.
- Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. *J. Adv. Drug Deliv. Rev.* **1997**, *12*, 3.
- Ajay, W. P.W.; Murcko, M. A. *J. Med. Chem.* **1998**, *41*, 3314.
- Sadowski, J.; Kubinyi, H. *J. Med. Chem.* **1998**, *41*, 3325.
- Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. W. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511.