Original Article

ScafBank: a public comprehensive Scaffold database to support molecular hopping

Bi-bo YAN[1,#], Meng-zhu XUE[2,#], Bing XIONG[2,*], Ke LIU[1], Ding-yu HU[2], Jing-kang SHEN[2,*]

[1]Electronics and Information College, Yangtze University, Jingzhou 434023, China; [2]State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 201203, China

Aim: The search for molecules whose bioactivities are similar to those of given compounds or to optimize the initial lead compounds from high throughput screening has attracted increasing interest in recent years. Our goal is to provide a publically searchable database of scaffolds out from a large collection of existing chemical molecules.
Results: Although a number of *in silico* methods have emerged to facilitate this process, which has become known as ″scaffold hopping″ or "molecular hopping", there is an urgent need for a database system to provide such valuable data in the drug design field. Here we have systematically analyzed a collection of commercially available small molecule databases and a bioactive compound database to identify unique scaffolds and we have built apublically searchable database. The analysis of approximately 4 800 000 of these compounds identified 241 824 unique scaffolds, which are stored in a relational database (http://202.127.30.184:8080/db.html). Each entry in the database is associated with a molecular occurrence and includes its distribution of molecular properties, such as molecular weight, logP, hydrogen bond acceptor number, hydrogen bond donor number, rotatable bond number and ring number. More importantly, for scaffolds derived from the bioactive compounds database, it also contains the original compounds and their target information.
Conclusion: This Web-based database system could help researchers in the fields of medicinal and organic chemistry to design novel molecules with properties similar to the original compounds, but built on novel scaffolds.

Keywords: chemoinformatics; privileged structure; scaffold database; molecular hopping
Acta Pharmacologica Sinica (2009) 30: 251–258; doi 10.1038/aps.2008.22; published online 19th January 2009

Introduction

"Scaffold hopping" or "molecular hopping" is a recent concept in the field of medicinal chemistry that refers to the search for compounds with bioactivities similar to given original structures[1,2]. Many computational approaches have been used to generate heterogeneous structures with bioactivity similar to that of a given structure of interest. These include *de novo* molecular design, virtual screening, pharmacophore search, topology similarity search and shape similarity search[3–8]. Some approaches require the three-dimensional structure of the target, so that the binding site information can be taken into account in the search for molecules with favorable interactions. Other approaches require known active compounds at hand, so that information about the ligand's properties, such as molecular hydrophobicity and hydrophilicity, charge, and shape features, can be used to constrain database mining. The scaffold hopping method has been proved very effective in molecular design, as shown in many review papers. For example, in work conducted at Abbott Laboratories, Zhao *et al* performed scaffold modification on a screening hit and identified potent and selective growth hormone secretagogue receptor (GHS-R) antagonists[9]. The scaffold of the hit antagonist was changed from a phenylisoxazole ring to a tetralin carboxamide, dramatically improving its binding affinity and other physiochemical properties. Also, several pharmaceutical companies have modified the hydrophobic ring part of the classic statin-class of HMG-CoA reductase inhibitors and thus invented new entities that they have been able to introduce into the drug market.

There are several methods of molecule fragmentation that produce meaningful fragments, such as scaffolds and functional groups. This analysis usually involves three steps:

[#] These authors contributed equally to this work.
* Correspondence to Prof Bing XIONG and Prof Jing-kang SHEN.
E-mail bxiong@mail.shcnc.ac.cn and jkshen@mail.shcnc.ac.cn
Received 2008-11-03 Accepted 2008-12-08

1. divide the molecules into fragments based on some rules to produce substructures;

2. obtain a unique list of the identified substructures;

3. process the substructures to assess their importance and to identify the most interesting molecules.

The first fragmentation step is crucial, because it determines what kinds of substructures will be produced. The method proposed by Bemis and Murcko[10] is to fragment the whole molecule as rings, linkers and chains. The linker is a minimum atom path connecting the ring parts. The rings are familiar to chemists and obvious from the chemistry of the molecule; all remaining atoms belong to the chains. The framework of the molecule consists of all rings and linker fragments. This method has been implemented by several research groups with slight modifications. Xu[11] modified the ring definition to include the unsaturated ring-bonded atoms in order to maintain the charge and geometric properties of the ring system. By doing this, he was able to program a system to classify the compounds based on their molecular framework. Another important approach invented by Lewell *et al* is to fragment the molecule based on 11 simple chemical reaction types[12]. The resulting fragments are ready for use in *in silico* synthesis to form a virtual library. These two basic, yet very important method complement each other in some situations. The former tends to produce a large scaffold, with no clues as to how to synthesize it by means of organic chemistry, whereas the latter tends to cut the meaningful scaffold into several atomic building blocks.

Although scaffold hopping methods have empowered researchers to optimize their lead compounds, to the best of our knowledge, there is no publicly accessible database containing this invaluable scaffold information. Many public small molecule databases are focused on the whole molecule level, such as the ZINC[13] and PubChem databases[14]. These databases are used to search for entire molecules by similarity or substructures, rather than to identify interesting substitution fragments within individual molecules. To support Web-based molecular hopping, we have constructed a comprehensive database of unique scaffold structures by systematically fragmenting the ZINC molecular database, a large database (derived from several commercially available molecular libraries) containing more than 4.6 million compounds. We have also performed the same operations on the small molecular ligand dataset of the DrugBank database and the MDL Drug Data Report (MDDR)[15] database in order to derive additional scaffolds. These fragment structures are associated with the properties of the original compounds from which the scaffold was derived. All this information, as well as the 2D structures of the scaffolds, is stored in a Web-based database system called ScafBank, which also implements substructure- and fingerprint-based similarity searches to enable researchers to quickly find feasible scaffolds. We believe that this valuable scaffold database will support medicinal chemists by allowing them to search with their own input fragments, facilitating molecular hopping studies.

Materials and methods

Molecular databases

The commercially available small molecular libraries used for high throughput screening were retrieved from the ZINC Web site (http://blaster.docking.org/zinc/). After molecules with similarities greater than 0.9 based on the fingerprint comparison method were removed, only 819 061 of the 4 600 000 small molecules in the ZINC database remained. We downloaded these datasets for scaffold and functional group analysis. The physicochemical properties of these molecules were calculated using JChem software and saved for later use. In order to compare these data with bioactive compounds, 1030 approved drugs contained in DrugBank[16] and approximately 160 000 compounds from the MDDR database were collected and analyzed.

Scaffold analysis

To identify the scaffold structures hidden in the molecules, the recursive scaffold analysis method was adopted and implemented on the basis of the open source C^{++} programming library OpenBabel2.0[17]. As described by Bemis and Murcko[10], the algorithm works by first going through the molecule graph to trim off chain atoms. This is done by continually removing the atoms bonded to only one heavy atom until no more such atoms can be found. The recursive scaffold analysis implemented here is based on the HierS system[18]. Contiguous fragment searching was conducted to find side chain trimmed molecules, as illustrated in Figure 1. For every fragment, the basic ring scaffolds were found and deleted one by one to produce a new molecule (which may contain several fragments). The new molecules were subjected to a further recursive scaffold analysis. To identify the basic ring system, the smallest set of smallest rings (SSSR) method was used to indicate the ring atoms[19]. If two rings are connected by one or more atoms, then they are associated together as one ring scaffold. This process is continued until no ring can be associated to any others. The recursive scaffold process was performed until only one ring system remained. Each of the resulting fragments was added to the final scaffold list and written out as a molecule.

The RECAP method was used to produce functional

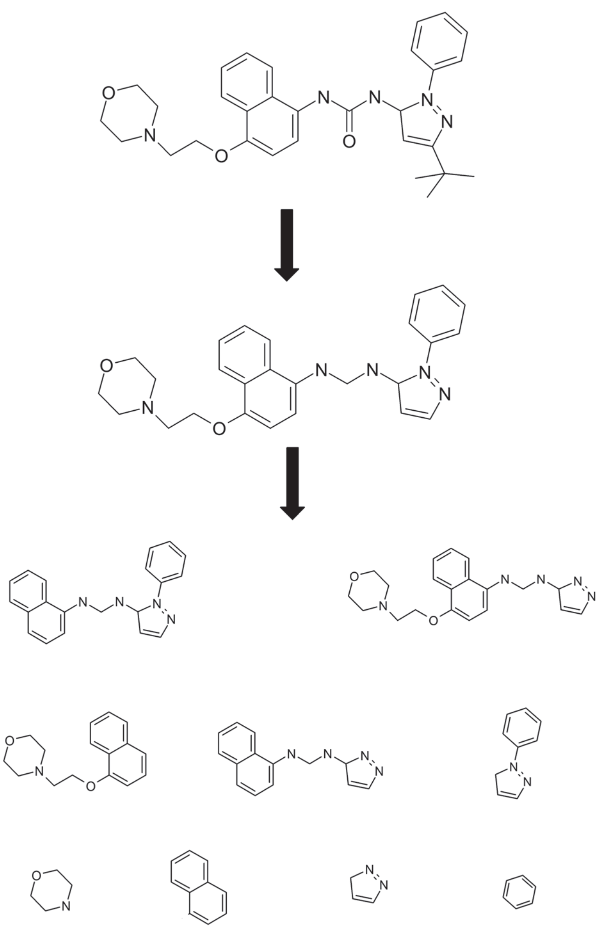


Figure 1. Illustration of the recursive scaffold analysis procedure, which fragments the molecule into scaffolds.

groups from the molecular database for analysis. This method was chosen because it is a reaction-based method, and the functional groups are ready to be used in virtual library construction. The Fragment program in the JChem package was utilized to accomplish this task[20]. The reverse reactions were represented with the SMIRK language and written into an XML file for fragmentation, giving the program the ability to recognize the reaction center and to produce the final functional groups.

To remove duplicate scaffolds and functional groups, the Java programming library Chemical Development Kit (CDK) was used as the basis for implementing a canonical SMILES representation of the chemical structures[21]. The canonical scaffolds and functional groups were processed to obtain a unique list of molecular fragments. Python scripts were then written to collect information about the original molecules (logP, molecular weight, hydrogen bond donor number, hydrogen bond acceptor number and ring number). The distributions of these properties were calculated and stored in a MySQL database for analysis[22].

In order to compare the results from these commercially available datasets with those from drug-like or marketed drug molecules, the 2D structures of bioactive molecules in the MDDR and DrugBank databases were subjected to the same analyses. To facilitate identification of the privileged structures, the frequencies of the scaffolds in the MDDR database and the drug target information associated with their original compounds were analyzed. This target information was used to retrieve the scaffold's target information Shannon entropy using the following equations:

$$STE=-\sum_{i}(N_i/N_{all})*\log_2(N_i/N_{all})$$

$$NSTE=STE/-\log_2(1/N_{all})$$

where *STE* is the scaffold target entropy, N_i is the number of molecules associated with the ith target class, and N_{all} is the total number of molecules associated with this scaffold. To normalize this entropy, the value was scaled by the entropy of all molecules.

Database system

The widely used MySQL database management system was selected to build the scaffold database[22]. All scaffolds were manipulated using the JChem database management system due to its efficiency at structure searching. The two-dimensional structures were imported into the database by the jcman program from the JChem package and then stored in a structure table. Other relevant information, such as the molecular weight and logP distributions of the original molecules associated with this scaffold, was stored in another MySQL table. Substructure and fingerprint-based similarity searching was implemented to facilitate Web-based searching. When querying using substructure and structure similarity, the JChem database was searched, and the IDs of the resulting molecules were collected and further used in querying other information tables. These results were then combined together and shown in a Web page format.

Results and Discussion

Analysis of scaffolds derived from the ZINC database

A recursive scaffold analysis similar to the HierS method[18] was adopted to analyze the ZINC database and store the unique scaffolds in a MySQL database. To provide an overview of the database, the scaffold occurrence numbers in the original ZINC database[13] were calculated and are shown in Figure 2. Most of the scaffolds are unique in the ZINC database (only one molecule contains that scaffold).

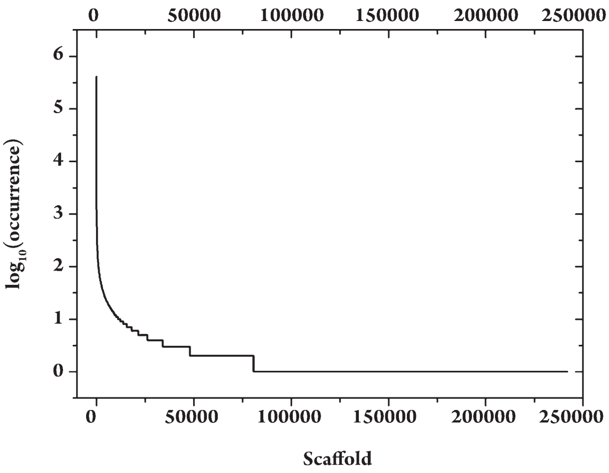


Figure 2. Molecular occurrences in the ZINC database of scaffolds in ScafBank. The occurrence number was scaled by log10 for clarity.

The most frequent scaffold is benzene, which is contained in approximately half of the ZINC database molecules. About 80 000 scaffolds have a molecular occurrence greater than 2, and about 10 000 scaffolds have an occurrence greater than 10. The scaffold database was also analyzed based on some common properties, such as molecular weight, ring number, aromatic atom number, and aliphatic atom number. This analysis shows that the molecular weight distribution has a typical Gaussian distribution shape and a mean of 280 Dalton, similar to the distribution in the ZINC database. Scaffolds are composed of ring and linker atoms. An analysis of the ring numbers in the scaffolds indicates that about 100 000 scaffolds contain three ring systems and two hetero-ring systems. This gives the database a large number of ring combinations to support scaffold hopping, which benefits researchers seeking substitutions for their query scaffolds. From the distribution of aromatic/aliphatic atoms, the mean number of aromatic atoms in the scaffold database was found to be approximately eight, and for aliphatic atoms, it was about twelve. For comparison, these properties were also calculated from the original ZINC database. Most of the properties were found to have distributions similar to those in the scaffold database.

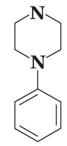
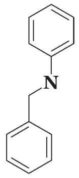
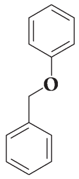
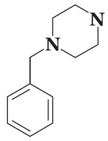
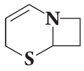
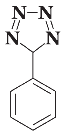
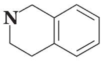
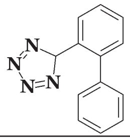
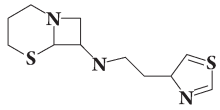
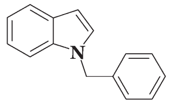
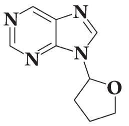
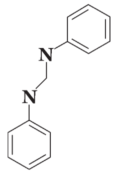
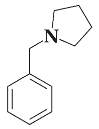
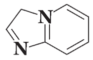
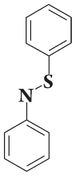
We also analyzed the functional groups in the ZINC database. To do this, the default RECAP method[12] was used to fragment the molecules in the ZINC database to get reaction-based building blocks. Surprisingly, this resulted in only 11 958 unique fragments. This represents a small portion of the 819 061 molecules in the ZINC database. This may indicate that the molecules in the ZINC database are made from a relatively small number of building blocks with simple chemical reactions. The same RECAP method was used to analyze the 1 030 approved drugs in DrugBank[16], where it yielded a total of 1 599 unique fragments. These results demonstrate that the ZINC database, a collection of commercially available molecular databases, tends to contain common functional groups and may only represent a small fraction of the total functional group space.

Scaffold analysis of MDDR database

As many researchers have demonstrated, bioactive molecular databases contain high-hanging fruit information about target family-related privileged structures. These privileged structures could be utilized to design focused libraries that target specific protein families. To extract this information, a scaffold analysis of the MDDR database was conducted using the same procedure described above. After the identification of unique scaffolds, the target Shannon entropy was calculated for each scaffold in order to identify interesting fragments. Larger normalized scores indicate that a scaffold is found in more targets and can be considered a privileged structure. Some of these are listed in Table 1. This is consistent with findings indicating that most of these valuable scaffolds are in the modulators of the GPCR protein family. Compared with previous investigations, this method provides a systematic way to extract privileged structures and also gives a ranking for these scaffolds, enabling researchers to check more easily for the interesting ones. Further library design using methods of combinatorial chemistry is underway and will be reported elsewhere.

To further assess the uniqueness of these scaffolds, we compared the scaffolds derived from the ZINC database with those derived from the DrugBank and MDDR databases[15]. As shown in Table 2, there is only a small overlap between the ZINC and MDDR scaffolds (12 946 scaffolds in common), indicating that the two scaffold databases complement each other to cover a larger scaffold space. When comparing the ZINC and MDDR database scaffolds (by removing the market drug compounds) with the DrugBank scaffolds alone, it was shown that the ZINC scaffolds cover about 53.1% of the scaffolds found in DrugBank, whereas the MDDR database covers about 78.6% of the DrugBank scaffolds. This is consistent with the fact that the MDDR database is more drug-like than the ZINC database. A combination of the ZINC and MDDR scaffolds covers about 83.44% of the DrugBank scaffold space, indicating that researchers will find a suitable scaffold for their projects in most cases.

Table 1. Examples of interesting scaffolds with large entropy scores.

Scaffold	Occurrence number[A]	STE[B]	NSTE[B]
	2749	0.562	0.033
	1880	0.555	0.039
	1757	0.650	0.052
	1620	0.596	0.053
	1599	0.041	0.009
	1280	0.115	0.029
	1279	0.65	0.077
	969	0.067	0.024
	923	0.011	0.005
	636	0.578	0.073
	596	0.253	0.094
	553	0.548	0.086
	538	0.561	0.160
	515	0.547	0.160
	511	0.506	0.134

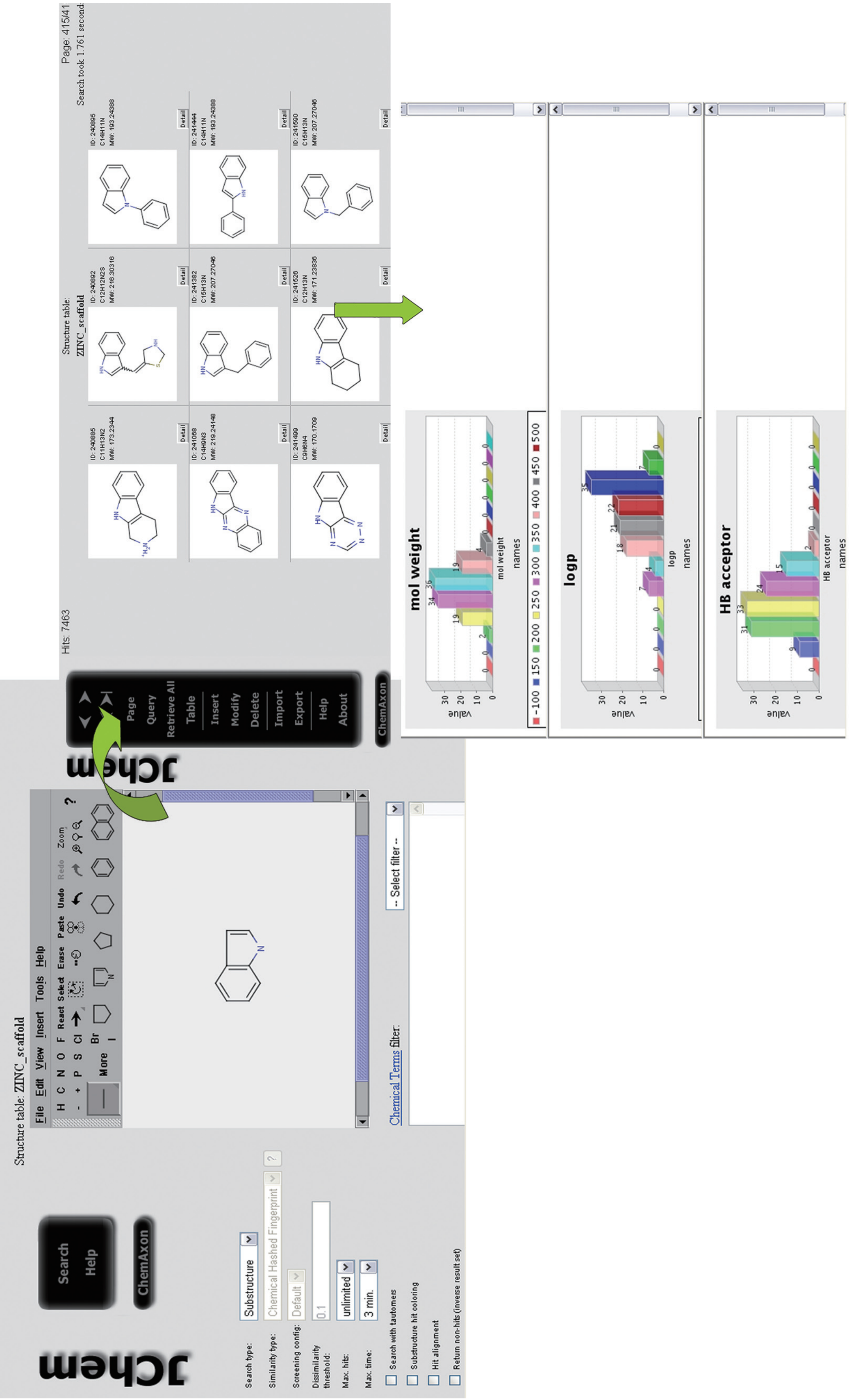
A, occurrence number is the number of molecules in the MDDR database contain this scaffold. B, STE, and NSTE are calculated according to the equation in Materials and Methods.

Table 2. Comparison of scaffolds across the three datasets derived from the ZINC, MDDR and DrugBank databases. The value in each cell is the number of common scaffolds found in the datasets and the number in parentheses is the percentage of common scaffolds.

Common scaffolds	ZINC	MDDR	DrugBank
ZINC	241 824 (100%)	12 946 (9.00%)	661 (53.14%)
MDDR	12 946 (5.35%)	143 780 (100%)	978 (78.62%)
DrugBank	661 (0.273%)	978 (6.80%)	1 244 (100%)

Web interface and searching options

To facilitate the use of the database by researchers, we constructed a Web site called ScafBank (http://202.127.30.184:8080/db.html) to host these analyzed data. Through the Web site, users can browse the unique scaffolds, as well as the associated information. In addition, they can search the database using substructure- or fingerprint-based similarity measures. As shown in Figure 3, users can draw the 2D structure online with the program Marvin or they can upload a molecule into Marvin. A database search can then be con-



Search type: Substructure
Similarity type: Chemical Hashed Fingerprint
Screening config: Default
Dissimilarity threshold: 0.1
Max. hits: unlimited
Max. time: 3 min.
□ Search with tautomers
□ Substructure hit coloring
□ Hit alignment
□ Return non-hits (inverse result set)

Structure table: ZINC_scaffold

Chemical Terms filter: -- Select filter --

Structure table: ZINC_scaffold
Hits: 7463
Page: 415/41
Search took 1.761 second

ID: 240985 C11H13N2 MW: 173.2344
ID: 240992 C12H12N2S MW: 216.30316
ID: 240995 C14H11N MW: 193.24388
ID: 241068 C14H9N3 MW: 219.24148
ID: 241382 C15H13N MW: 207.27046
ID: 241444 C14H11N MW: 193.24388
ID: 241499 C9H6N4 MW: 170.1709
ID: 241526 C12H13N MW: 171.23836
ID: 241590 C15H13N MW: 207.27046

Figure 3. The Web interface of the ScafBank database.

ducted to find similar scaffolds in the database. The interface provides the option of specifying filtering rules, such as how many molecules to output or how many hydrogen bonds the resulting molecules should contain. This gives the researcher the flexibility to retrieve scaffolds based on their own scaffold hopping research. After the database is searched, the molecules retrieved are depicted on Web pages. Each scaffold is associated with the molecular property distribution of its original molecules, which may be useful for combinatorial library design. Also, users can double click on the scaffold and open a new Marvin window, in which they can calculate additional properties of the scaffolds, such as conformation and charge.

To further demonstrate the capability of this scaffold database, we queried the ScafBank with a two-ring scaffold (Figure 4). Similarity searching at a similarity level 0.7, returned a total of 97 hits by the MDDR scaffold database. We collected the results and found some interesting scaffolds that could be used as substitutes in the query, some of which are listed in Figure 4. These resulting scaffolds are reasonable from the viewpoint of medicinal chemists. Further real applications of ScafBank through combinatorial library design and synthesis are in progress and will be reported elsewhere.

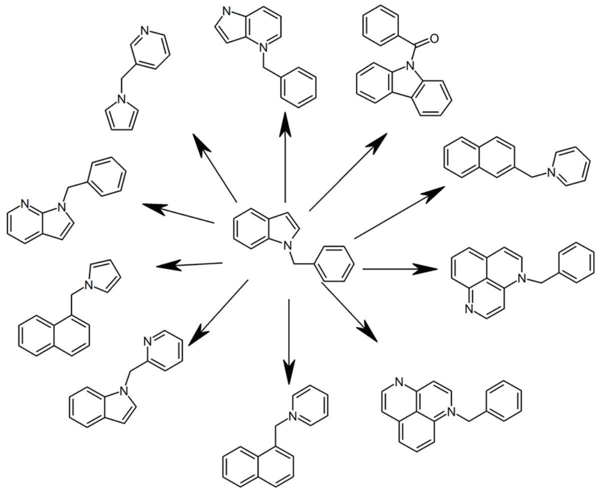


Figure 4. A case study. The query scaffold is in the middle and some of the results are listed around the query structure.

Discussion

Scaffold hopping is an active research field in chemoinformatics, and many computational methods are being devised to help medicinal chemists develop novel ideas for the hit-to-lead optimization and improve the druggability for these bioactive compounds. Here we conducted scaffold analysis on three common databases and compiled these scaffolds into a relational database, which will enable researchers to perform scaffold substitution query studies. The original databases used to extract the scaffolds include most of the commercially available compounds. Using the canonical SMILES representation, we collected unique scaffolds into a relational database. Which removes the redundancies in this database and simplifies the post-analysis of the query results.

As demonstrated by numerous medicinal chemistry studies, scaffold hopping is a more general application of bioisosteric design, a process in which a target scaffold is replaced by another scaffold, which is sometimes considerably different in structure but still has similar properties. We hope our ScafBank database may be used in scaffold hopping to obtain molecules with better bioavailability or selectivity. Another straightforward application of the scaffold database is the identification of important scaffolds and further subjects for combinatorial chemistry library design. This "privileged structure" approach has already demonstrated its potential in developing GPCR modulators[2]. The bioactivity-related entropy score in the ScafBank could be used to prioritize the scaffolds, which helps researchers judge the importance of the scaffolds and select the most interesting scaffolds with which to construct a combinatorial library to increase the chance of finding hit or lead compounds.

Although scaffold substitution is a useful method in medicinal chemistry, as reviewed by Babaoglu and Shoichet[23], molecules are composed of various fragments. The bioactivity is not just simply summarize the contributions of these fragments. Sometimes the molecule act in an integrated way. In the case of scaffold hopping, changing one part of the ligand may also affect other parts of the molecule because of variations in subtle torsion angle change, in the orientations of other groups connected to these scaffolds, and in the physicochemical properties of the molecules substituted in the scaffold. It should be noted that in our database system, only chemical 2D similarity is considered for scaffold hopping. Users should, therefore, not think the scaffolds returned from database search is the final decision to use for substitution. Instead, it is just a starting point from which to further determine the feasibility of scaffold hopping.

Conclusion

In summary, a comprehensive, Web-accessible scaffold database was built by recursive scaffold analysis of the ZINC, DrugBank and MDDR databases. By comparing these unique scaffolds with the scaffolds derived from approved drugs in DrugBank, it was found that the scaffolds covered approximately 83% of DrugBank scaffold space. To our knowledge, this is the first public database specifically constructed for scaffolds. This database may assist researchers in pharmaceutical fields in conducting scaffold hopping to design novel molecules with higher potency or pharmaceutical potential.

Acknowledgements

This project was supported by the National Natural Science Foundation of China (Grant No 30600784) and the Knowledge Innovation Program of the Chinese Academy of Sciences, SIMM0709QN-18 (Grant No O7G602R012)

We thank Dr David L BURK and Dr Prof Albert M BERGHUIS (McGill University, Canada) for reading the manuscript and for helpful discussions. We thank ChemAxon Inc for providing the JAVA software packages JChem and Marvin.

Author contribution

Bing XIONG and Jing-kang SHEN designed the project; Bi-bo YAN, Meng-zhu XUE, Ke LIU and Bing XIONG performed the research; Ding-yu HU did the case study and analyzed the data; Bing XIONG and Jing-kang SHEN wrote the article.

References

1 Brown N, Jacoby E. On scaffolds hopping in medicinal chemistry. Mini Rev Med Chem 2006; 6: 1217–29.

2 Zhao H. Scaffold selection and scaffold hopping in lead generation: a medicinal chemistry perspective. Drug Discov Today 2007; 12: 149–55.

3 Bajorath J. Integration of virtual and high-throughput screening. Nat Rev Drug Disc 2002; 1: 882–94.

4 Barnum D, Greene J, Smellie A, Sprague P. Identification of common functional configurations among molecules. J Chem Inf Comput Sci 1996; 36, 563–71.

5 Eckert H, Bajorath J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. Drug Discov Today 2007; 12: 225–33.

6 Klebe G. Virtual ligand screening: strategies, perspectives and limitations. Drug Discov Today 2006; 11: 580–94.

7 Khedkar SA, Malde AK, Coutinho EC, Srivastava S. Pharmacophore modeling in drug discovery and development: an overview. Med Chem 2007; 3: 187–97.

8 Schneider G, Fechner U. Computer-based *de novo* design of drug-like molecules. Nat Rev Drug Discov 2005; 4: 649–63.

9 Zhao H, Xin Z, Liu G, Schaefer VG, Falls HD, Kaszubska W, *et al.* Discovery of tetralin carboxamide growth hormone secretagogue receptor antagonists via scaffold manipulation. J Med Chem 2004; 47: 6655–7.

10 Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. J Med Chem 1996; 39: 2887–93.

11 Xu J. A new approach to finding natural chemical structure classes. J Med Chem 2002; 45: 5311–20

12 Lewell XQ, Judd DB, Watson SP, Hann MM. 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 –22.

13 Irwin JJ, Shoichet BK. ZINC–a free database of commercially available compounds for virtual screening. J Chem Inf Model 2005; 45: 177–82.

14 PubChem database. Avaiable from http://pubchem.ncbi.nlm.nih.gov/

15 MDL Drug Data Report database (MDDR). Avaible from http://www.mdli.com/

16 Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, *et al.* Woolsey DrugBank: a knowledgebase for drugs, drug actions and drug targets. J Nucleic Acids Res 2006; 34: D668–72

17 Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, *et al.* The Blue Obelisk– interoperability in chemical informatics. J Chem Inf Model 2006; 46: 991–8. Avaiable from http://openbabel.sourceforge.net/

18 Wilkens SJ, Janes J, Su AI. HierS: hierarchical scaffold clustering using topological chemical graphs. J Med Chem 2005; 48: 3182–93.

19 Figueras J. Ring perception using Breadth-first search. J Chem Inf Comput Sci 1996; 36: 986–91.

20 ChemAxon. Avaiable from http://www.chemaxon.com/jchem/

21 Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The chemistry development kit (CDK): an open-source Java library for chemo- and bioinformatics. J Chem Inf Comput Sci 2003; 43: 493–500.

22 MySQL database system. Avaible from http://www.mysql.com/

23 Babaoglu K, Shoichet BK. Deconstructing fragment-based inhibitor discovery. Nat Chem Biol 2006; 2: 720–3.