Inter*net* Electronic Journal of
Molecular Design

March 2004, Volume 3, Number 3, Pages 143–149

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Nenad Trinajstić on the occasion of the 65th birthday
Part 9

Guest Editor: Douglas J. Klein

New Diversity Criterion and Database Compression Method

Bing Liu,[1] Aijun Lu,[1] Lei Zhang,[1] Haibo Liu,[1] Zhenming Liu,[2] and Jiaju Zhou[1]

[1] Laboratory of Computer Chemistry, Institute of Process Engineering, Chinese Academy of Sciences, P.O. Box 353, Beijing 100080, China
[2] The Chemistry Department of Peking University, China

Received: August 18, 2003; Revised: October 29, 2003; Accepted: December 29, 2003; Published: March 31, 2004

Citation of the article:
B. Liu, A. Lu, L. Zhang, H. Liu, Z. Liu, and J. Zhou, New Diversity Criterion and Database Compression Method, *Internet Electron. J. Mol. Des.* **2004**, *3*, 143–149, http://www.biochempress.com.

Internet Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

New Diversity Criterion and Database Compression Method[#]

Bing Liu,[1] Aijun Lu,[1] Lei Zhang,[1] Haibo Liu,[1] Zhenming Liu,[2] and Jiaju Zhou[1,]*

[1] Laboratory of Computer Chemistry, Institute of Process Engineering, Chinese Academy of Sciences, P.O. Box 353, Beijing 100080, China
[2] The Chemistry Department of Peking University, China

Received: August 18, 2003; Revised: October 29, 2003; Accepted: December 29, 2003; Published: March 31, 2004

Internet Electron. J. Mol. Des. **2004**, *3* (3), 143–149

Abstract

Based on the topological scaffold classification approach to cluster a structural database, we propose a new criterion to evaluate the diversity of a chemical structural database. This criterion is defined as the ratio of scaffold number to total structure number in the database. Six databases have been evaluated by this criterion. To reduce the size of a database under the minimum losing structural diversity, a novel effective database compression method has been developed. The number of selected structures in each compounds group with common scaffold is determined by empirical K–4–5 rules, and the selected structures are those with the higher drug–like value (DLV). A validity test has been made by adding 200 new random nonoverlapping natural products into NCI3D, and the losing percent of that test data set is about 10.5%. Results show that NCI3D, MNPD (marine natural products database) and TCMD (traditional Chinese medical database) have 68.7%, 60.3%, 54.4% size reduced respectively by this method.

Keywords. Structure scaffold diversity; database compression; topological scaffold; database diversity; chemical structural database.

Abbreviations and notations

SCA, topological scaffold–based classification approach	CANCER, cancer screen database
SSD, structure scaffold diversity	AIDS, AIDS antiviral screen database
DCM, database compression method	TCMD, traditional Chinese medical database
DLV, drug–like value	MNPD, marine natural products database

1 INTRODUCTION

With the rapid development of high throughput screening and combinatorial chemistry, large chemical structural database has become a necessary tool for drug design in pharmaceutical industry [1–3]. Many new chemical structural databases with thousands of compounds have been on the market in recent years [4,5]. But these large data sets consume too much time in leading drug screening, and make the chemists very confused in selecting potent compounds. Many algorithms have been used to resolve this problem [6–10]. The common target of these algorithms is to select a

[#] Dedicated to Professor Nenad Trinajstić on the occasion of the 65th birthday.

* Correspondence author; phone: 0086–10–62626703; fax: 0086–10–62561822; E–mail: jjzhou@lcc.icm.ac.cn.

database subset with minimizing diverse losing.

Many typical diverse analysing methods are based on structural descriptors, which could match the special project well, but they cannot give a universal solution for all structural databases. Such as Pearlman and Smith had developed cell–based diversity arithmetic for a given receptor using structure and activity data [9]. Two algorithms, Nilakantan's ring–scaffold based [11] and Xu's topological scaffold based [6], give a description to universal structural database classification.

In this paper, we propose a new criterion SSD, the abbreviation of structure scaffold diversity, to evaluate the diversity of a universal structural database based on the topological scaffold. Also, an effective method to compact a chemical structural database is proposed.

2 DATABASES AND METHODS

First of all, we classified the compounds by their topological scaffolds. Then we compute and compare their drug–like values and give a diverse evaluation. At last, we compact the structural database according to the former steps. The algorithm used to classify the structures is the topological scaffold–based classification approach (SCA). The databases tried out our tests are NCI3D, AIDS, CANCER, MDDR–3D, TCMD and MNPD.

2.1 Databases

NCI3D (1994 Released), AIDS (October 1999 Released) [12], CANCER (August 1999 Released) [13], are public databases released by the National Cancer Institute (USA) which comprise 126705, 42390, and 32443 compounds respectively. All compounds have the 3D structure, CAS registry number, and NSC number. NCI3D is the largest freely available public domain chemical structural database. Several prior publications have described its history and related projects [14–21]. Various version of NCI3D can be downloaded from many websites [22].

MDDR–3D is a commercial structural database developed by MDL Information Systems, Inc, and contains 132726 3D models [23].

TCMD, a new commercial database, is developed by our Laboratory of Computer Chemistry (LCC) and has 9127 entries. A typical entry includes detailed 3D molecular structures, English names and synonyms, physical properties, natural sources and references information. Bioactivity data are available for 3000 of the entries. There are 3922 traditional Chinese medical plant species including standard expression on effect.

MNPD, a new marine natural products subject database, is also developed by LCC. It contains 8078 marine natural products with the compounds names, 3D structures, bioactivities for 39.31% of them, CAS Registry Numbers for 15.02% of them, property data for 46.17% of them, sources and references information.

Those 3D structural databases have some molecules without coordinates, thus they can’t describe a molecule completely in SDF connect table. So before computing the scaffolds, we delete these molecules using a batch program. The databases and their structure numbers are listed in Table 1.

Table 1. Topological scaffold number and SSD value of these compared databases

Database	No. of structures	Scaffold Number	SSD
NCI3D	126089	23776	18.86%
CANCER	32440	13458	41.49%
AIDS	42389	17986	42.43%
MDDR–3D	132726	51124	38.52%
TCMD	9126	3183	34.88%
MNPD	8078	2729	33.78%

2.2 Topological Scaffold–Based Classification Approach (SCA)

Xu has presented a new concept to classify molecular structures, the topological scaffold [6]. The step to define the topological scaffold can be described as follows: (1) Define ring bonds. A ring bond is a bond with both its atoms in the same ring. (2) Define linker bonds. A linker bond is not a ring bond, but both of its atoms are connect to ring directly or indirectly. (3) Define chain bonds. A chain bond is a bond neither a ring bond nor a linker bond. (4) Define the topological scaffold. A topological scaffold is a structure that contains ring bond and linker bond but no chain bond. Figure 1 gives the illustration of the definition of a topological scaffold.

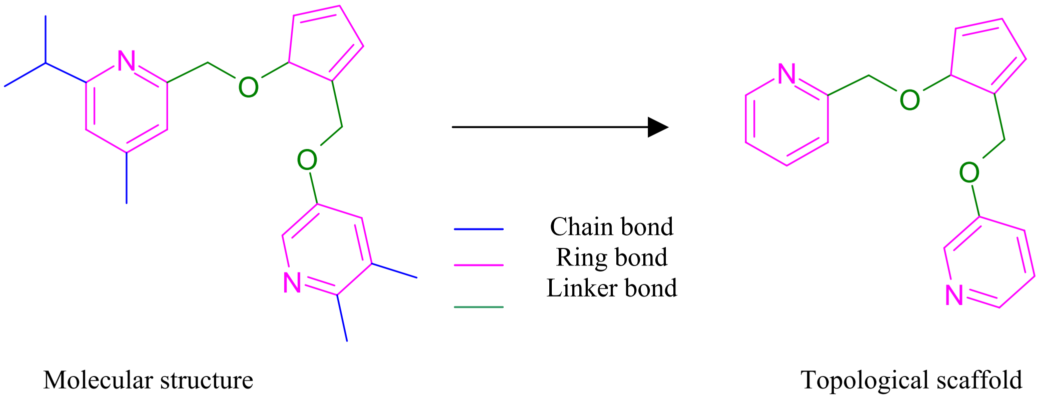


Figure 1. The definition of topological scaffold

Unlike conventional approaches in majority for clustering chemical compounds, SCA based on topological scaffold not based structural descriptors. The main idea is that if the structures have the same topological scaffold they will be in the same group. The SCA has been implemented as a grogram developed by C++ language on Windows 95/ NT and UNIX platforms.

2.3 Database Compression Method

Database diversity evaluation targets finding an efficient data subset. The first thing of all is that how many compounds will be reserved better in compacted database. We must emphasize that the best compressed database size is not just equal to its scaffold number. We present a simple example to explain it. The substitution of a methyl group at the proper site can convert an active drug into an inactive compound, which could make the database have a lower SSD value. But fortunately, it only takes up very small proportion, otherwise it could contradict the 2D–QSAR model principle. We consider one scaffold for 3–5 compounds could assure both the higher diversity and a lower information decrease. Secondly, we must resolve which compounds will be selected in a database.

The criterion for number selecting is K–4–5 rule (See Table 2). If the number of compounds with common scaffold is equal to or lower than four in a database, these compounds will be reserved all in new compacted database. If the number is larger than four but smaller than one hundred, they will be reserved four of them, while if the number is larger than one hundred, they will be reserved five.

Table 2. The K–4–5 criterion for number selecting

Number of compounds shared one scaffold	Reserve number
$N \leq 4$	keep
$4 < N \leq 100$	4
$N > 100$	5

Which compounds should be selected? Generally, typical whole–molecule descriptors, such as LogP, can express the molecule character well. But it could only express very limited information about the details of molecular sub–structural differences [9]. While some other structural descriptors, such as ring number, number of H–bond donors, number of H–bond acceptors, number of rotating bonds, can complement that defect.

According to the Lipinski's rule of 5, drug–like compounds should have appropriate molecular weight, H–bond donors, H–bond acceptors, and LogP value [24]. Larger molecular weight could make the molecule difficult to move through the cell membrane. Solubility of drug will be very small in water if the LogP value is large enough, which make it very difficult to been transported in body. On the contrary, if the LogP value is very small, it can hardly pass through the cell membrane.

Using statistic method, Xu has limited the number of smallest set of smallest rings, the number of rotating bonds and average electro–negativity etc., for drug design [6]. And he added these descriptors to complement Lipinski's rule of 5.

Those descriptors' contribution to drug design are not due to individual action but as a community. We integrate those descriptors as a new parameter, named Drug–Like Value (DLV), to express the potential possibility of a compound becoming drug. The parameters and their statistic drug–like value are listed in Table 3.

Table 3. Different molecular descriptors and their statistic drug–like value

Cluster parameters	Definition	Drug like value
HD	Number of H–bond donors	≤ 5
HA	Number of H–bond acceptors	0–8
AB	Number of aromatic bonds	0–28
SSSRS	Number of smallest set of smallest rings	1–9
AZ	Average atomic numbers	6–10
RB	Number of rotating bonds	0–14
AE	Average electro–negativity	2.55–3.02
MW	Molecular weight	≤ 500
LogP	Octanol–water partition coefficient	≤ 5

Because compounds in the same group have uniform ring scaffold, in other words have equal SSSRS, we will discard this descriptor’s contribution. The definition of DLV can be described as follows: (1) Set the initial DLV value to zero. (2) Compute the parameters listed in Table 2 for every molecule in database and give them all equal weight exponent. The LogP value is computed by XLogP program [25]. (3) If the value of a parameter is within the field of drug–like value, the molecular DLV increases with one. Otherwise, the DLV will remain constant.

$$DLV = \sum_{i=HD}^{\log P} 1 \; or \; 0 \qquad (1)$$

We compact the larger database into smaller based on the K–4–5 criterion and the DLVs. The higher DLV molecule has the priority of being selected. If many molecules have same DLV, they could be selected randomly. This database compression method has been implemented as a program (DCM) in C++ language on Windows platforms.

3 RESULTS AND DISCUSSION

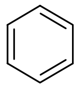
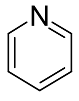
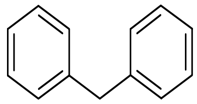
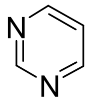
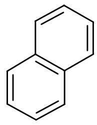
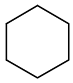
Structure Scaffolds Diversity (SSD), which by definition is the ratio of topological scaffolds number to total structures number, will be used as a database diverse evaluating criterion. The topological scaffolds number and their SSD values of those six databases are listed in Table 1.

From the value of SSD, AIDS has the highest structure diversity, and others are CANCER, MDDR–3D, TCMD, MNPD and NCI3D respectively. They have the SSD value between 30% and 45% except NCI3D.

In fact, most scaffolds in those databases, including NCI3D, have one or two compounds each. Only a few small scaffolds contain many compounds. The possibility of these compounds turning into drugs is almost zero. So we can discard most of them bounteously. The scaffold structures, shared by more than 1000 compounds in NCI3D, are listed in Table 4.

The compacted database of NCI3D (SNCI3D) is composed of 39470 structures computed by DCM. We recalculate the topological scaffolds of the SNCI3D by SCA. There are 23776 scaffolds total and the SSD is 60.24%. The same method is implemented on TCMD and MNPD. And the SSD of those two compacted databases is 76.42% and 57.04% respectively. In this way, we achieve the goal of keeping the diversity and reduce the database size notably.

Table 4. The scaffolds structures shared by more than 1 000 compounds in NCI3D

scaffold	number	scaffold	number	scaffold	number
	19186		1627		1376
	1344		1241		1176

The validation of structure diversity is extremely difficult, perhaps impossible [9]. It only can test what extent the useful information can be conserved. In order to test the validity of DCM, we add random non–overlapping 200 natural products structures into NCI3D and compute this new mixed database using SCA and DCM. The results are listed in Table 5.

Table 5. The results of validity test

Item	data
Total number of mixed database	126289 (MNP 200 plus NCI3D 126089)
Computed scaffolds	23931
SSD	18.95%
Total number of reduced size database	39644
The ratio of MNP losing	10.5%

There are 10.5% MNP have been lost because the compacting process is hard to avoid random selection when many compounds have the same DLV.

4 CONCLUSIONS

The ratio of scaffold number to total structure number in a structural database is a useful and simple criterion to evaluate the diversity of a universal chemical structure database. The database compression method proposed in this paper is an efficient approach to both reduce the database size and keep the information of structural diversity and utility.

Acknowledgment

The authors acknowledge Jun Xu [6] for providing the SCA program.

5 REFERENCES

[1] E. Vangrevelinghe, K. Zimmermann, J. Schoepfer, R. Portmann, D. Fabbro, and P. Furet, Discovery of a Potent and Selective Protein Kinase CK2 Inhibitor by High-Throughput Docking, *J. Med. Chem*. **2003**, *46*, 2656-2662.

[2] E. Perola, K. Xu, T. M. Kollmeyer, S. H. Kaufmann, F. G. Prendergast, and Y. P. Pang, Successful Virtual Screening of a Chemical Database for Farnesyltransferase Inhibitor Leads, *J. Med. Chem*. **2000**, *43*, 401-408.

[3] J. V. de Julian–Ortiz, J. Galvez, C. Munoz–Collado, R. Garcia–Domenech, and C. Gimeno–Cardona, Virtual Combinatorial Syntheses and Computational Screening of New Potential Anti–Herpes Compounds, *J. Med. Chem*. **1999**, *42*, 3308–3314.

[4] J. Lei and J. J. Zhou, A Marine Natural Product Database, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 742–748.

[5] M. He, X. J. Yan, J. J. Zhou, and G. R. Xie, Traditional Chinese Medicine Database and Application on the Web, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 273–277.

[6] J. Xu, A New Approach to Finding Natural Chemical Structure Classes, *J. Med. Chem.* **2002**, *45*, 5311-5320.

[7] R. Nilakantan, N. Bauman, and K. S. Haraki, Database diversity assessment: New ideas, concepts, and tools, *J. Comp.–Aided Mol. Design*. **1997**, *11*, 447–452.

[8] R. S. Pearlman and K. M. Smith, Novel Software Tools for Chemical Diversity, *Perspectives Drug Discovery Design*. **1998**, *9*, 339–353.

[9] R. S. Pearlman and K. M. Smith, Metric Validation and the Receptor–Relevant Subspace Concept, *J. Chem. Inf. Comput. Sci.* **1999**, *9*, 28–35.

[10] D. M. Bayada, H. Hamersma, and V. J. van Geerestein, Molecular Diversity and Representativity in Chemical Databases, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1–10.

[11] R. Nilakantan, N. Bauman, K. S. Haraki, and R. Venkataraghavan, A Ring Based Chemical Structural Query System: Use of a novel Ring–Complexity heuristic, *J. Chem. Inf. Comput. Sci.* **1990**, *30,* 65–68.

[12] http://dtp.nci.nih.gov/docs/aids/aids_data.html.

[13] http://dtp.nci.nih.gov/docs/cancer/cancer_data.html.

[14] J. H. Voigt, B. Bienfait, S. Wang, and M. C. Nicklaus, Comparison of the NCI Open Database with Seven Large Chemical Structural Databases, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702–712.

[15] M. T. Zehnacker, R. H. Brennan, G. W. A. Milne, J. A. Miller, and M. J. Hammel, The NCI Drug Information System. 6. System maintenance, *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 193–197.

[16] G. W. A. Milne, M. C. Nicklaus, J. S. Driscoll, S. Wang, and D. Zaharevitz, National Cancer Institute Drug Information System 3D Database, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1219–1224.

[17] G. W. A. Milne and J. A. Miller, The NCI Drug Information System. 1. System overview, *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 154–159.

[18] G. W. A. Milne, A. Feldman, J. A. Miller, G. P. Daly, and M. J. Hammel, The NCI Drug Information System. 2. DIS pre–registry, *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 159–168.

[19] G. W. A. Milne, M. C. Nicklaus, J. S. Driscoll, S. Wang, and D. Zaharevitz, National Cancer Institute Drug Information System 3D Database, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1219–1224.

[20] G. W. A. Milne, J. A. Miller, and J. R. Hoover, The NCI Drug Information System. 4. Inventory and shipping modules, *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 179–185.

[21] G. W. A. Milne, A. Feldman, J. A. Miller, and G. P. Daly, The NCI Drug Information System. 3. The DIS chemistry module, *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 168–179.

[22] http://cactus.nci.nih.gov/ncidb2/download.html.

[23] http://www.mdli.com/pdfs/MDDRds.pdf.

[24] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv. Drug Delivery Rev.* **1997,** *23*, 3–25.

[25] R. X. Wang, Y. Fu, and L. H. Lai, A New Atom–Additive Method for Calculating Partition Coefficients, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615–621.