Drug-like Index: A New Approach To Measure Drug-like Compounds and Their Diversity

Jun Xu* and James Stevenson

Research & Development Center, Boehringer Ingelheim Pharmaceuticals, Inc., 900 Ridgebury Road,
Ridgefield, Connecticut 06877-0368

Received March 23, 2000

Combinatorial organic synthesis (combinatorial chemistry or CC) and ultrahigh-throughput screening (UHTS) are speeding up drug discovery by increasing capacity for making and screening large numbers of compounds. However, a key problem is to select the smaller set of “representative” compounds from a virtual library to make or screen. Our approach is to select drug-like as well as structurally diverse compounds. The compounds, which are not very drug-like, are less taken into account or excluded even if they contribute to the diversity of the collection. Hence, the first step in the compound selection is to rank compounds in drug-like “degree”. To quantify the drug-like “degree”, drug-like index (DLI) is introduced in this paper. A compound’s DLI is calculated based upon the knowledge derived from known drugs selected from Comprehensive Medicinal Chemistry (CMC) database. The paper describes the way of this knowledge base is formed and the procedure for selecting drug-like compounds.

INTRODUCTION

High-throughput screening (HTS) and combinatorial chemistry (CC)[1−3] bring increased capacity for making and screening a large number of compounds in a relatively short time. However, we can become lost in the large chemical space built by ourselves if we are not guided to a right direction (i.e. drug-like compound space). HTS and CC need tools to select compounds from a large chemical compound space (such as a virtual library) to avoid screening millions compounds which may yield few genuine hits, redundant information or lead to structurally unattractive hits that cannot be readily modified to yield a viable drug candidate. This is why several approaches have been developed to select compounds.[4−12] Common thoughts for the compound selection are summarized as follows:

(1) The entire chemical diversity space is not accessible, but its subset can be derived, ideally, this subset represents the structural diversity of the entire space.

(2) The chemical structure diversity space can be represented by a structural descriptor (such as, molecular topological indices, fingerprints, or pharmacophore) space for computations. Mathematical methods, such as, principal component analysis (PCA) are used to reduce the number of the descriptors, before they are used to form the diversity space.

(3) The structural descriptor space can be partitioned into subspaces by a variety of statistical methods, such as hierarchical or nonhierarchical clustering, artificial neural networks, etc.

(4) Picking “representative” compounds from every partitioned subspace can form the representative diversity space.

What is missed in the common thoughts is that structure diversity space is a relative space (there is no absolute original point), which has to be referred to a reference point. For drug discovery, a meaningful reference point is a drug-like cluster center. Because nondrug-like compounds should not be arbitrarily included no matter how diverse they are. Including such compounds increases the chances of spending effort on hits that have structural liabilities that may not be removable in the lead optimization process. Also nondrug-like compounds may be more susceptible to producing artifactual activity in screening assays, e.g. through detergent-like properties.

Our drug-like cluster center consists of a set of the distributions of selected structural descriptors; the distributions are derived from a database of existing drugs. Actually, there are many different types of drugs. In this work, we study only non-peptide orally used drugs. However, the method and program can be applied in studying other types of drugs. There are many published structural descriptors, some of them may be used for ranking drug-like compounds, and some of them may be not. We selected 25 structural descriptors by discussing with medicinal chemists and SAS analysis. To keep this paper in proper length, we will discuss the methods for selecting structural descriptors in other paper. Structural descriptors can correlate to each other; some of them may be redundant; however, if they have different and significant distribution in the considered drug class, they can be used for drug-knowledge extraction. The structure descriptors, which are not bioactively relevant, will generate noise for the drug-like cluster center.[13] On the other hand, we make sure that the descriptors maintain their identity and clearly interpretable structural significance throughout the process. For example, principal component analysis (PCA) is used in other approaches to reduce the dimensions of chemical space. That is, it projects a higher dimensional chemical space to a lower dimensional chemical space without losing too much information concerning the relationship between members of the chemical space. However, the

10.1021/ci000026+ CCC: $19.00
Published on Web 08/02/2000

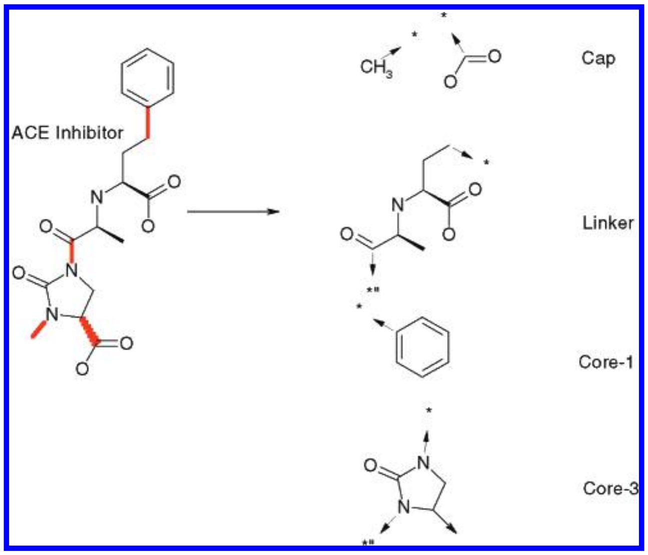


Figure 1. Building blocks and joint bonds.

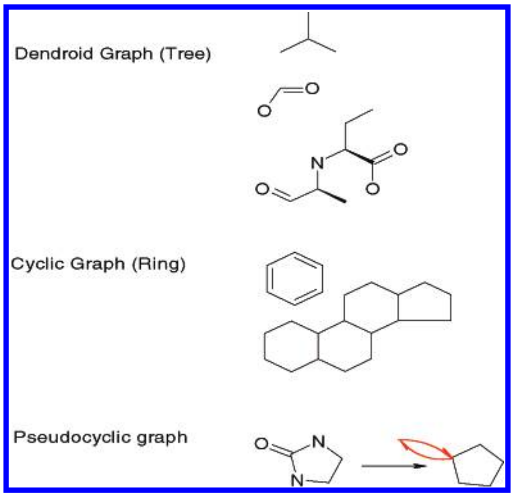


Figure 2. Building block classifications.

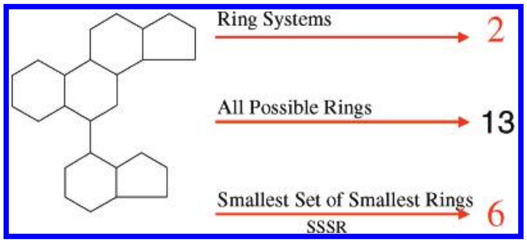


Figure 3. Three ways for ring counting.

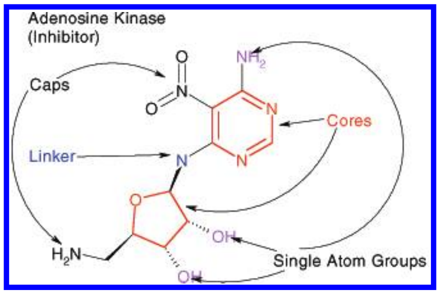


Figure 4. The relations among different types of building blocks.

new dimensions in the lower dimensional chemical space will lose their physical meaning, and it is not easy to determine which descriptors are related to biological activity.

This approach has been tested on following five libraries: 1. the CMC database (MDL product, most of marketed drug structures are included), 2. the MDDR database (MDL product, medicinal chemistry related compounds), 3. the ACD database (MDL product, general available chemicals), 4. the AsInEx catalog (AsInEx, Inc., 62,971 small molecules), and 5. one of BI PI compound collections.

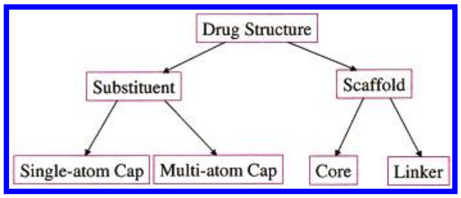


Figure 5. Hierarchy of drug structure.

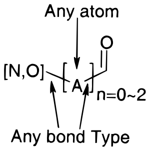


Figure 6. N-level bonding patterns.

Lipinski's "rule of 5"[14] is the simplest description of the drug-like cluster center, although it reflects a limited amount of information. Their work indicates that molecular weight, number of H-bond donors, number of H-bond acceptors, and clogP, are drug related structural descriptors. To make our drug-like cluster center more robust, more structural descriptors are introduced and generated as shown later. There have been number of efforts to distinguish drug-like compounds from nondrug-like compounds. Ajay, Walters, and Murcko[15] used a Bayesian neural network to recognize drugs and nondrugs. Bemis and Murcko studied side chain distributions in CMC database,[21] and Wang and Ramnarayan published their work on "drug-likeness".[22] Ghose, Viswanadhan, and Wendoloski have reported their knowledge-based approach.[16]

In this paper, we start with concepts of structural diversity, select and introduce a set of drug relevant topological descriptors, and compute the distributions of the descriptors of drug structures (oral drugs in CMC database). Based upon the distributions, a drug-like compound cluster center is formed. The cluster center is the knowledge base (KB), which is used to rank compounds in any library in term of their "drug-like" indices (DLI). The curve of relative population versus the DLI value is a simple mean to view the structure diversity of a compound library. There are no criteria to absolutely distinguish drugs from nondrugs, hence, we do not intend to use DLI approach to do "drug structure pattern recognition". We use DLI to rank compounds in a library and let medicinal chemists to decide DLI threshold to select drug-like compounds.

METHOD

Chemically, a set of compounds made from greater number of structural "building blocks" is structurally more diverse than the set made from smaller number of "building blocks". If this assumption is agreed, the consequent question will be that how the building blocks are defined. Naturally, chemists believe structural building blocks are functional groups and core structures. Bemis and Murcko[16] have similar idea about the building blocks. They characterized the hierarchy of drug structures in terms of rings, links, and molecular frameworks.

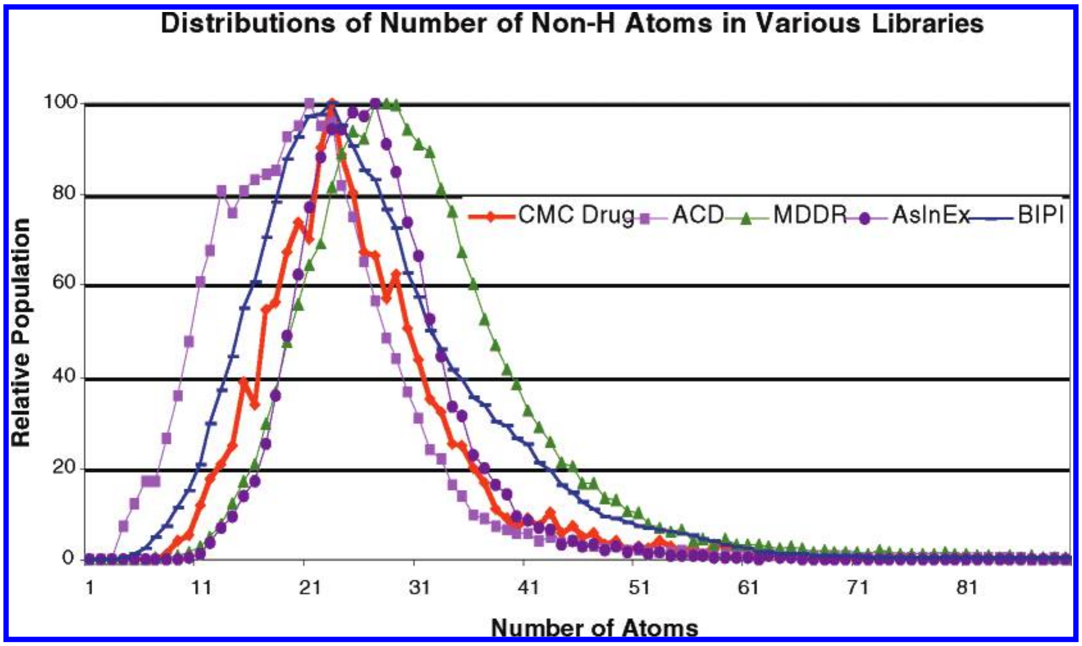


Figure 7. Distribution of the number of non-H atoms of drugs (in CMC database) verses the distributions for other compound collections. CMC, ACD, and MDDR are all MDL database products, AsInEx is the compound catalog of AsInEx, Inc., and BIPI is our in-house compound database.

Table 1. The Structural Diversity of CMC Database

total building blocks	2538
structural diversity	52.5%
side chains	1405
rings	1133
heterorings	969

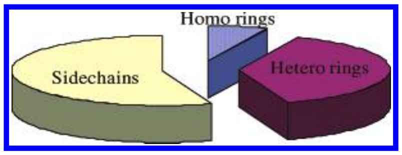


Figure 8. The side chain, ring distributions in drug structures.

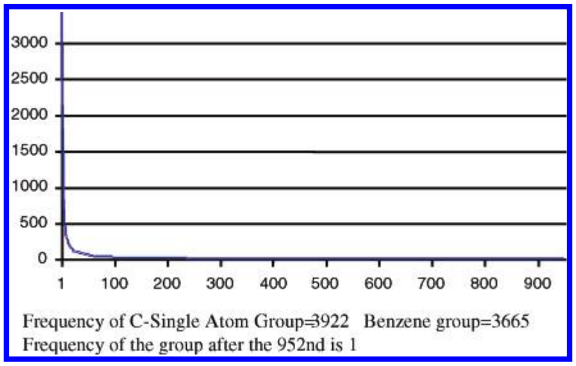


Figure 9. The population distribution for all building blocks.

To study structure diversity and similarity quantitatively, the building block should be defined rigorously. Therefore, we use following rules to define and derive the building blocks:

1. A cap (substituent, or side chain) is an acyclic substructure with one attachment connecting to other structure building blocks.

2. A linker is an acyclic substructure with more than one attachment connecting to other structure building blocks.

3. A core is a cyclic substructure without linker or cap.

4. A drug structure consists of a scaffold with or without substituents (or caps).

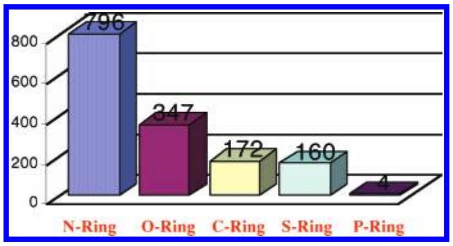


Figure 10. Distribution of heterorings.

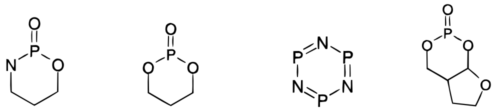


Figure 11. Four P-heterocyclic rings.

5. A scaffold has at least one core. If it has more than one core, linkers should connect the cores.

6. A nonring bond containing at least one ring-atom is called as "joint bond". By breaking joint bond(s), a structure will fall into a set of building blocks, i.e., cores, linkers, and caps.

7. An unsaturated bond is treated as a pseudo-ring (a ring with only two vertexes, see Figure 2) if it is attached to a cyclic system.

Figure 1 shows how building blocks are derived from a structure.

The building blocks are classified into dendroid graph (tree) and cyclic graph. As indicated above, the unsaturated bonds attached to cyclic system are considered as pseudo-cycles. Any building block has following properties: (1) number and location of attachment(s), (2) topologic attribute (acyclic or cyclic), and (3) ring topologic attributes (simple ring, fused ring, heteroring, saturated/unsaturated ring, pseudo-ring).

Figure 2 shows the examples of the classifications.

There are several ways to count rings in a structure. The counting methods are as follows:

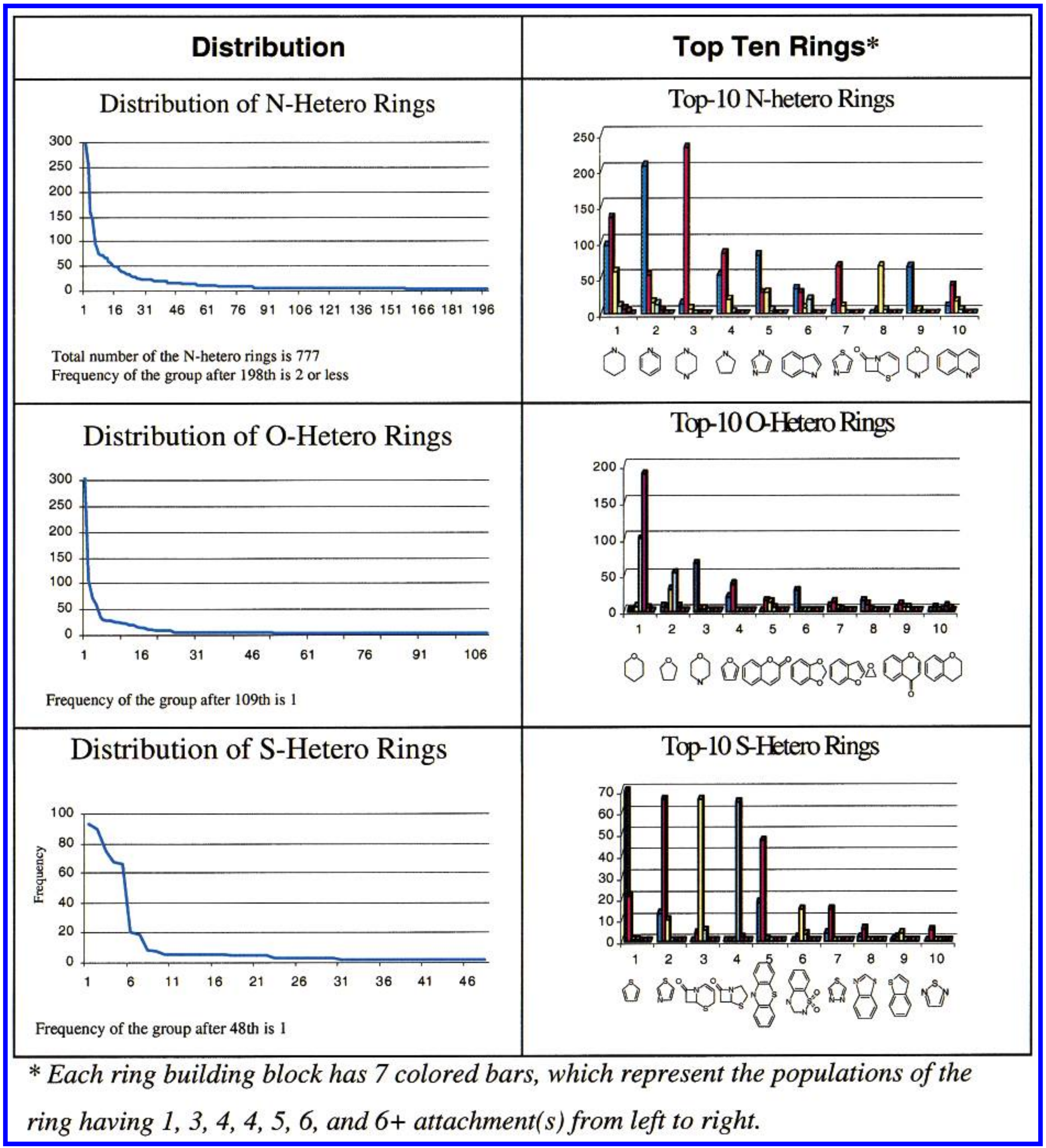


Figure 12. Heterocyclic building blocks. * Each ring building block has 7 colored bars, which represent the populations of the ring having 1, 3, 4, 4, 5, 6, and 6+ attachment(s) from left to right.

1. Ring System: A ring system is the ring having no joint bond to connect to any other ring.

2. The Smallest Set of Smallest Rings (SSSR): SSSR is the smallest ring building blocks necessary to form other ring systems.

3. All Rings: All possible rings are all possible ring paths can be found in every ring system within a structure.

Figure 3 illustrates these concepts.

Fused rings can have different activity behaviors if they are not fused together. To characterize this feature, “fusion degree” is introduced. The fusion degree can be calculated by

$$Ring\_\ fusion\_\mathrm{deg}\ ree = \frac{Number\_\mathrm{of}\_SSSR}{Number\_of\_ring\_systems} \qquad (1)$$

These building block types and their correlated functions are as follows:

1. Single atom group: single non-H atom attached directly to a cyclic system, which may be a component of a pharmacophore, such as OH, $NH_{(1-2)}$.

2. Core: a cyclic substructure without linker or cap, which makes drug molecule in special shape and selectable to a drug target.

3. Linker: acyclic structure linking cores to form a scaffold, which makes drug molecule flexible to bound to a drug target.

4. Scaffold: a set of cores linked together by a link or linkers, which is the common part (or signature) of a drug family.

Figure 4 explains the relations among different types of building blocks.

Based on the above-mentioned considerations, a typical drug molecule has following structural hierarchy (Figure 5):

When a building block is derived from a drug structure, the connection information is also prereserved. This information includes the number of attachments, the location(s) of the attachment(s), and the times of each attachment being connected.

To measure "drug-like" degree for a structure, two steps are taken:

1. "Drug-like" cluster center, i.e., knowledge base (KB), is computed from a drug database.

2. Based on the "building block" features of a given compound, calculate the "drug-like" index (DLI) by means of comparing the features against the "drug-like" cluster center.

Before "drug-like" cluster center is extracted, every drug structure is normalized by calculating smallest set of smallest rings, canonizing all types of aromatic rings, marking ring bonds, linker bonds and nonlinker chains, separating all building blocks, and recording the distributions of all types of building blocks.

The drug-like cluster center consists of the distributions of 25 selected structural descriptors. The descriptors are as follows: 1. number of non-H atoms: to describe the molecular size (Note: we do not use molecular weight, because a small molecule with heavier atoms may be consider as a bigger molecule for its heavier molecular weight.), 2. The number of SSSR: to describe molecular shape, 3. the molecular cyclized degree (MCD)

$$MCD = 100 - 100 \times \frac{Number\_of\_atoms}{Number\_of\_bonds + Number\_of\_moieties} \quad (2)$$

4. the number of non-H rotating bonds: to describe molecular flexibility, 5. the number of non-H polar bonds: describe molecular polarity, 6. the number of carbon atoms as the terminal group, to describe molecular lipophilicity, 7. the number of N atoms with at least one hydrogen atom, 8. the number of hydroxyl group, 9. the number of H-bond donors, 10. the number of H-bond acceptors, 11. the number of N atoms and O atoms, 12. the number of 2-degree chain atoms (the acyclic atom connected with 2 non-H atoms is a 2-degree chain atom), 13. the number of 2-degree cyclic atoms (the cyclic atom without non-H atom substitution is a 2-degree cyclic atom), 14. the number of 3-degree chain atoms (the acyclic atom connected with 3 non-H atoms is a 3-degree chain atom), 15. the number of 3-degree cyclic atoms (the cyclic atom with one non-H atom substitution is a 3-degree cyclic atom), 16. the number 1-level bonding pattern, 17. the number 2-level bonding pattern, 18. the number 3-level bonding pattern (The bonding pattern is defined by Figure 6), 19. the number of building blocks, 20. the number of aromatic systems, 21. the number of cyclic building blocks, 22. the number of linkers, 23. the number of caps,

24. the maximum size of SSSR, and 25. the maximum cap size.

These 25 descriptors are selected by analyzing their distributions among various libraries. These analyses were done to ensure that the distributions of the 25 descriptors of drugs are significantly different from other compound collections. Figure 7 shows the distributions of one of the descriptors (number of non-H atoms in a structure) from five different compound databases.

Figure 7 indicates these compound collections have different molecular size distributions. Comparing with drugs in CMC, ACD compounds shift to the left, MDDR compounds shift to the right, BIPI compounds cover all the size range, and is a greater compound collection. From the significant analysis, we are convinced that drug-like compounds do have a cluster center, which is formed in the distributions of the selected 25 structural descriptors.

The drug-like cluster center formation algorithm is described by following pseudo codes:

```
Algorithm: drug_cluster_center_generation(CMC Database)
{
    filter out non-oral bioactive drugs; // such as radiopaque agents, solvents, etc.
    exclude peptides, mixtures, etc.
    For ( i=0; i<total number of considered drugs; i++)
    {
        extract_building_blocks(structure(i));
        calculate all 25 descriptors;
        record populations for the 25 descriptors;
    }
    For ( i=0; i<25; i++)
    {
        normalize(descriptor(i)); //the score range: 0% ~ 100%
    }
    save the cluster center;
}
```

For a given structure, the same algorithm as described above calculates the 25-descriptor values. The values are then mapped onto the cluster center, and the "drug-like" index (DLI) is calculated in following formula:

$$DLI = \sqrt[n]{\prod_{i=1}^{n} Score(descriptor(i))} \quad (3)$$

It should be noted that (3) is not an Euclidean distance. Logically, the Euclidean distance means that all descriptors are independent, i.e., their relations are "OR", in other words, if one of the descriptors produces good score, DLI will have good score. We believe their logic relations are "AND", that is a good DLI value requires every descriptor has to have good score. If one of the score is zero, the DLI will be zero.

PROFILES OF DRUG STRUCTURES

The data set used to compute "drug-like" cluster center is MDL Comprehensive Medicinal Chemistry database release 98.1 (CMC 98.1). This release contains total 7497 records. 2233 records are excluded for being nonoral drugs, such as radiopaque agents, imaging agents, dental resins, veterinary compounds, and peptides or proteins. After preprocessing, 5264 compounds are remained. 428 compounds are excluded from the remaining compounds for being small simple compounds, inorganic compounds, heavy metallic-organic compounds, and mixture etc. Only 4836 structures are actually accepted for building the cluster center.

The calculation is carried out on a SGI workstation. It took about 15 min. 2538 building blocks are found. If structural

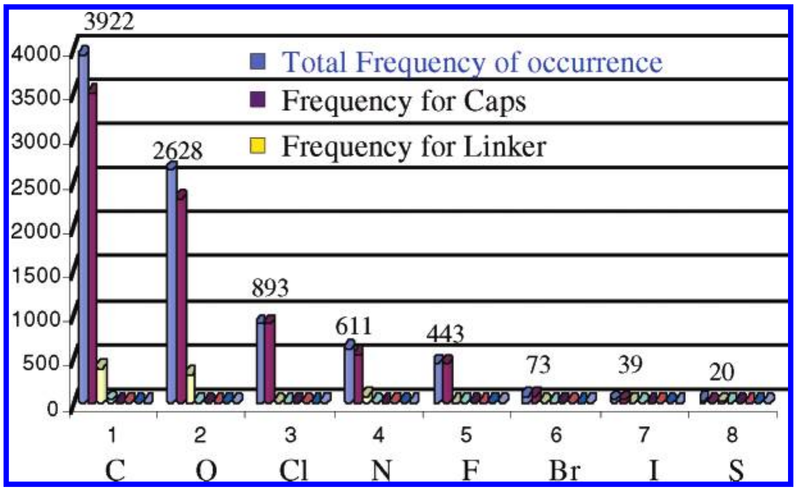


Figure 13. Distribution of single atom group.

Table 2. The 25 Rules for Drug-like Compounds

descriptor	best	min.	max.
no. of non-H atoms (molecular size)	22	10	50
no. of SSSRs	3	1	6
cyclized degree	10%	4%	18%
no. of non-H rotatable bonds	9	3	35
no. of polar bonds	8	1	25
no. of methyl terminal groups	0	0	7
no. of amino H-bond donors	0	0	3
no. of hydroxyl H-bond donors	0	0	4
no. of H-bond donors	1	0	5
no. of H-bond acceptors	3	0	10
no. of O atoms and N atoms	4	0	15
no. of 2-degree acyclic atoms	1	0	12
no. of unsubstituted cyclic atoms	8	1	19
no. of 3-degree acyclic atoms	1	0	5
no. of substituted cyclic atoms	6	1	15
no. of 1-level bonding pattern	0	0	5
no. of 2-level bonding pattern	0	0	5
no. of 3-level bonding pattern	0	0	4
no. of building blocks	1	4	11
no. of aromatic systems	1	0	3
no. of cyclic building blocks	2	1	5
no. of linkers	1	0	3
no. of caps	2	0	8
max. SSSR size	6	3	13 (vary)
max. cap size (in number of atoms)	1	0	12

diversity is measured in the ratio of the number of building blocks to the number of the number of input structures, then the CMC's structural diversity = 2538/4836 = 52.5%. The profiles of the drug structures in CMC collection are listed in Table 1.

The distribution of building blocks is depicted in Figure 6.

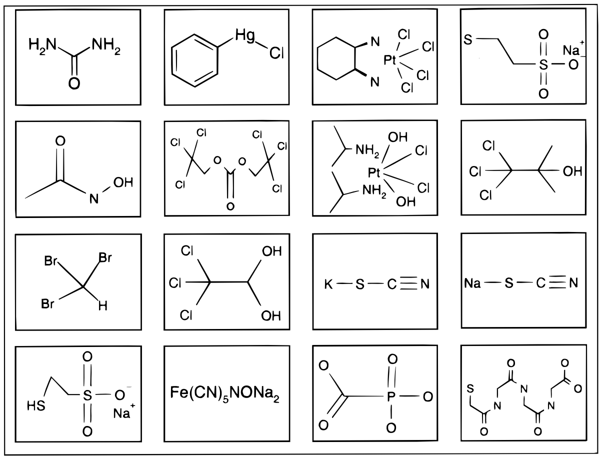
The pie chart shows that the most of “drug-like” structures have heterorings. The most of the heterorings are nitrogen-heterorings (Figure 10).

As shown in Figure 9, only 3.9% building blocks are popular. In other words, only about 100 building blocks are statically significant in CMC database.

As expected, $CH_{(1\sim3)}$ group and benzene ring are most popular building blocks.

The distribution of different types of rings is shown in Figure 10. Nitro-heterocyclic rings are most popular. Often, oxygen-heterocyclic or sulfur-heterocyclic rings have nitrogen atom in the ring system.

Table 3. Some Low DLI Value Compounds from CMC Database



There is no pure phosphorus-heterocyclic ring in CMC database. The only 4 phosphorus-heterocyclic rings are listed in Figure 11.

The distributions of poplar heterorings and their top-10 structures are shown in Figure 12.

Only 8 single atom groups are found in CMC drug database. As expected, the methyl group is the most popular group (see Figure 13). There is some population of methylene group. Most of the sulfur atoms in drugs are in ring systems or chains, not as a single atom group. Br and I are rare.

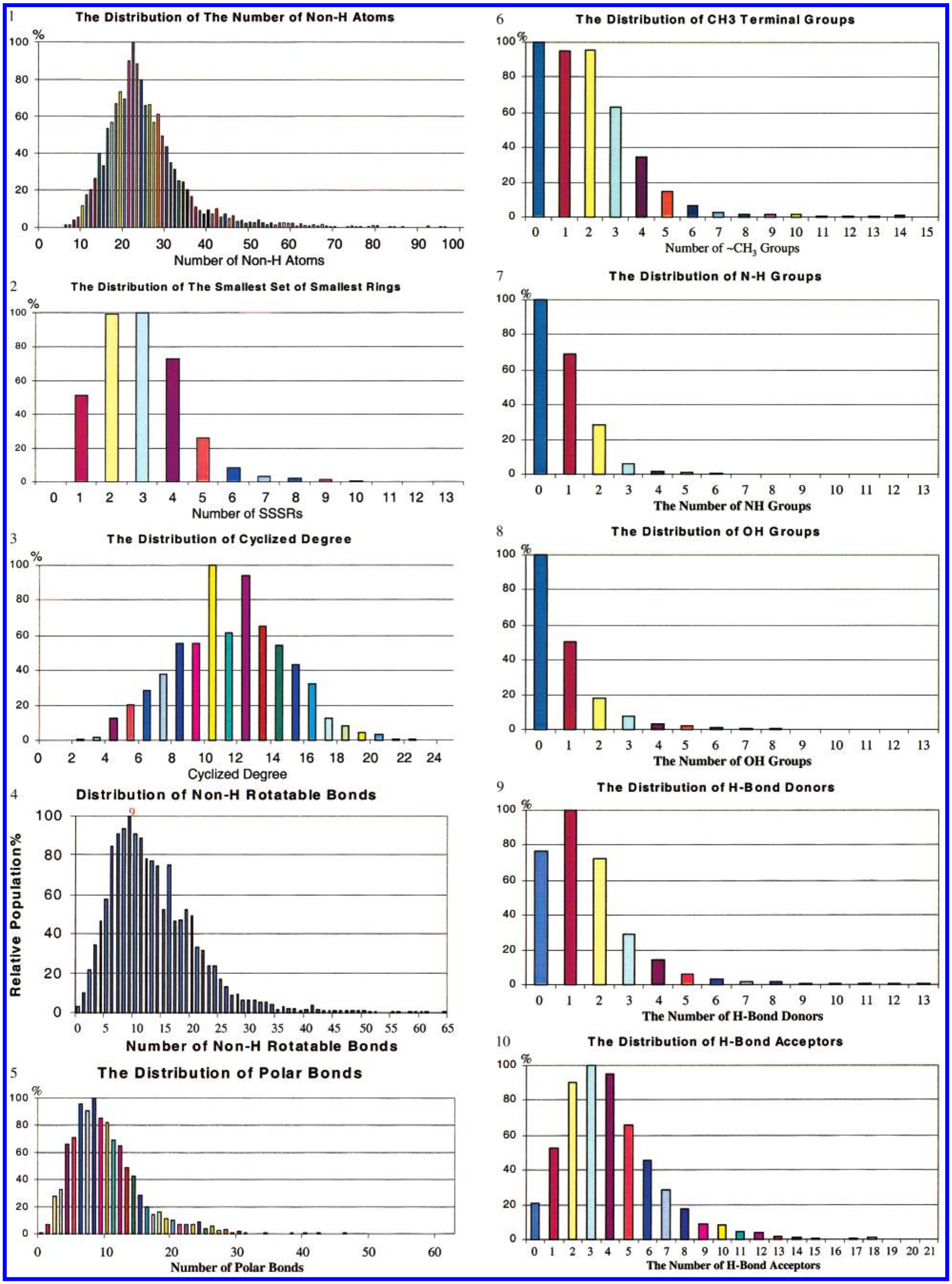
The distributions of 25 structure descriptors for CMC drugs are listed in Figure 14.

Based on these distributions, the 25 rules for drug-like compounds are summarized in Table 2.

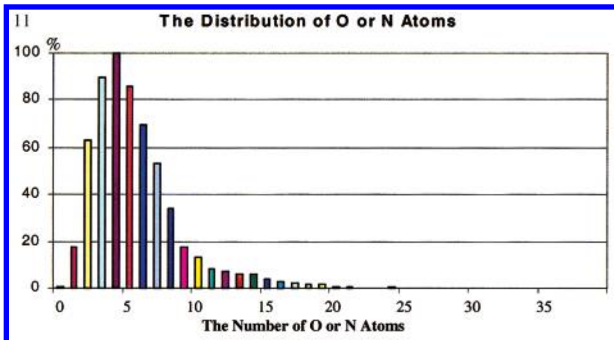
These rules are considered as the super set of Lipinski's rules. It is noted that the upper boundaries for the numbers of H-bond donors and H-bond acceptors are consistent with Lipinski's rules, but the molecular size boundary is higher. Again, these rules show the best values and the ranges for drugs. These rules become useful guidelines in drug design and combinatorial library design.

DRUG-LIKE INDEX (DLI) AND DIVERSITY

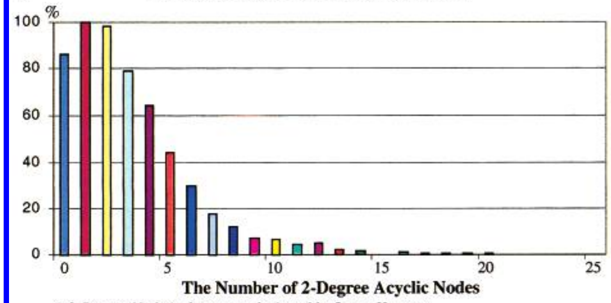
As mentioned before, the drug-like cluster center is built to be a practical reference point to measure drug-biased structure space. That is, the DLI is used to rank the


1
The Distribution of The Number of Non-H Atoms
%
100
80
60
40
20
0
0 10 20 30 40 50 60 70 80 90 100
Number of Non-H Atoms
2
The Distribution of The Smallest Set of Smallest Rings
%
0 1 2 3 4 5 6 7 8 9 10 11 12 13
Number of SSSRs
3
The Distribution of Cyclized Degree
%
0 2 4 6 8 10 12 14 16 18 20 22 24
Cyclized Degree
4
Distribution of Non-H Rotatable Bonds
9
Relative Population%
0 5 10 15 20 25 30 35 40 45 50 55 60 65
Number of Non-H Rotatable Bonds
5
The Distribution of Polar Bonds
%
0 10 20 30 40 50 60
Number of Polar Bonds
6
The Distribution of CH3 Terminal Groups
%
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Number of ~CH3 Groups
7
The Distribution of N-H Groups
%
0 1 2 3 4 5 6 7 8 9 10 11 12 13
The Number of NH Groups
8
The Distribution of OH Groups
%
0 1 2 3 4 5 6 7 8 9 10 11 12 13
The Number of OH Groups
9
The Distribution of H-Bond Donors
%
0 1 2 3 4 5 6 7 8 9 10 11 12 13
The Number of H-Bond Donors
10
The Distribution of H-Bond Acceptors
%
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
The Number of H-Bond Acceptors

11 The Distribution of O or N Atoms

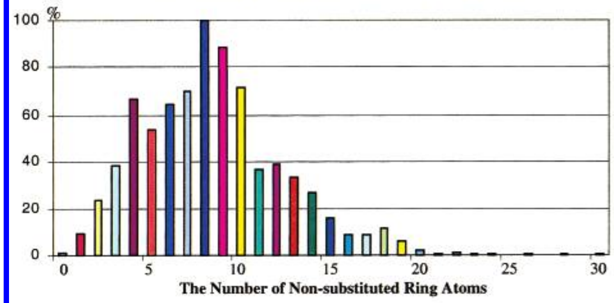


12 The Distribution of 2-Degree Nodes

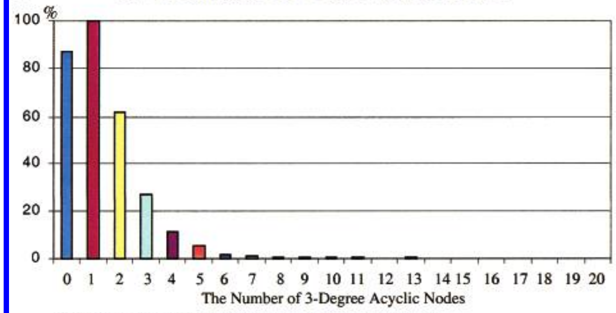


** 2-Degree Node is the atom substituted by 2 non-H atoms*

13 The Distribution of Non-substituted Ring Atoms

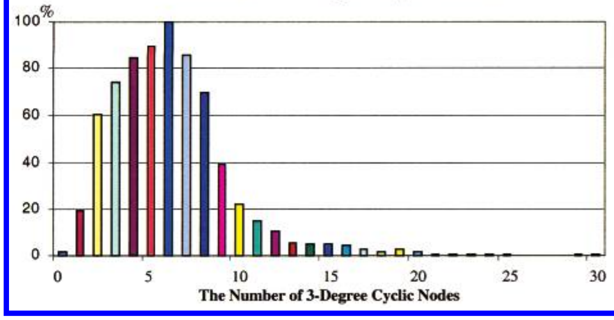


14 The Distribution of 3-Degree Acyclic Nodes

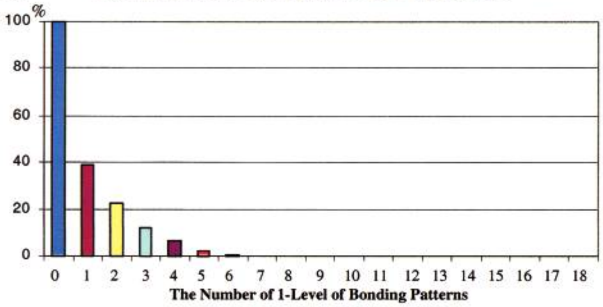


** 3-Degree Node is the atom substituted by 3 non-H atoms*

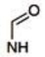
15 The Distribution of 3-Degree Cyclic Nodes



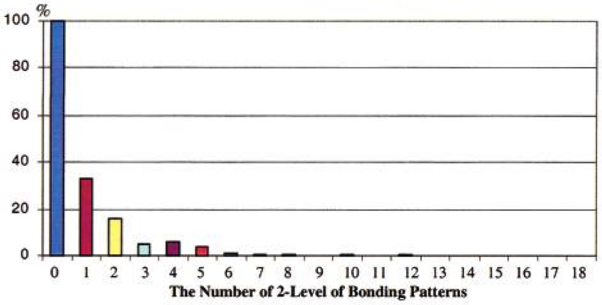
16 The Distribution of 1-Level Bonding Patterns



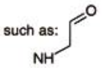
1-Level Bonding Pattern is a substructure, such as:



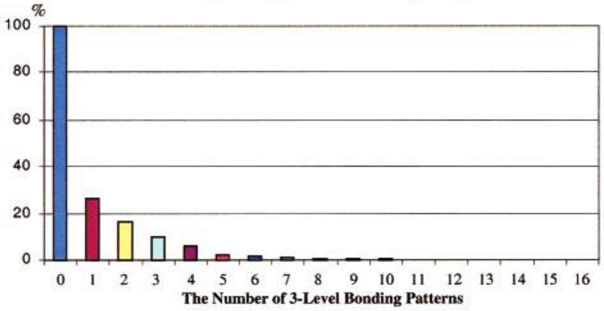
17 The Distribution of 2-Level Bonding Patterns



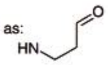
2-Level Bonding Pattern is a substructure, such as:



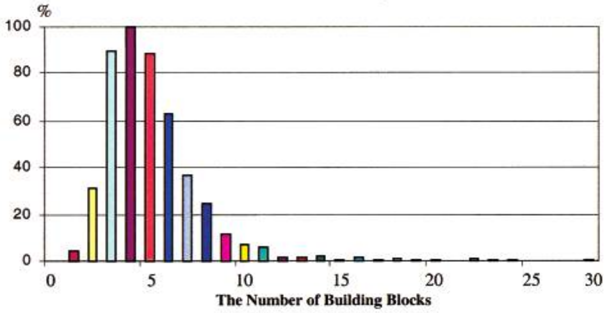
18 The Distribution of 3-Level Bonding Patterns



3-Level Bonding Pattern is a substructure, such as:



19 The Distribution of Building Blocks



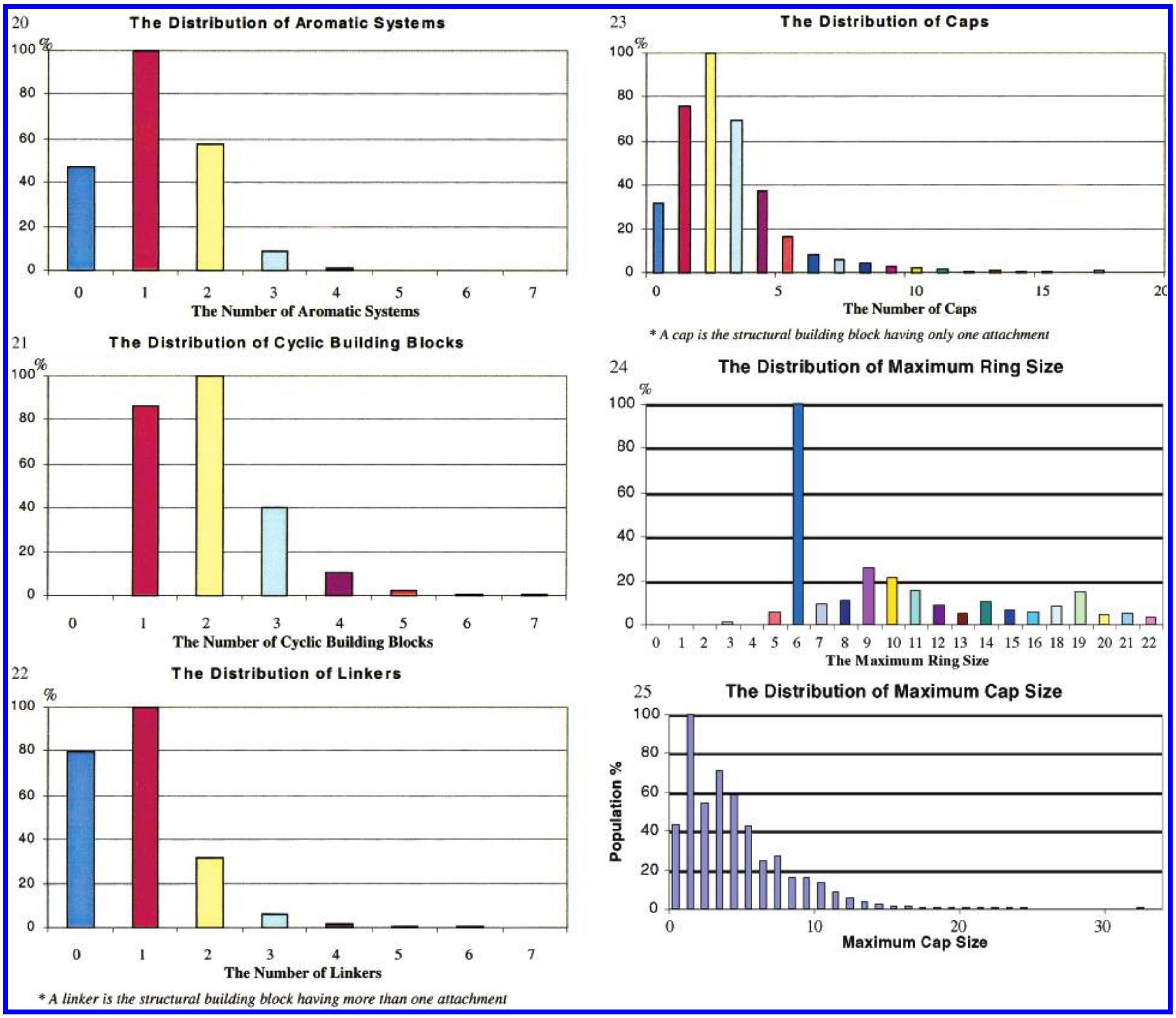


Figure 14. Drug cluster center.

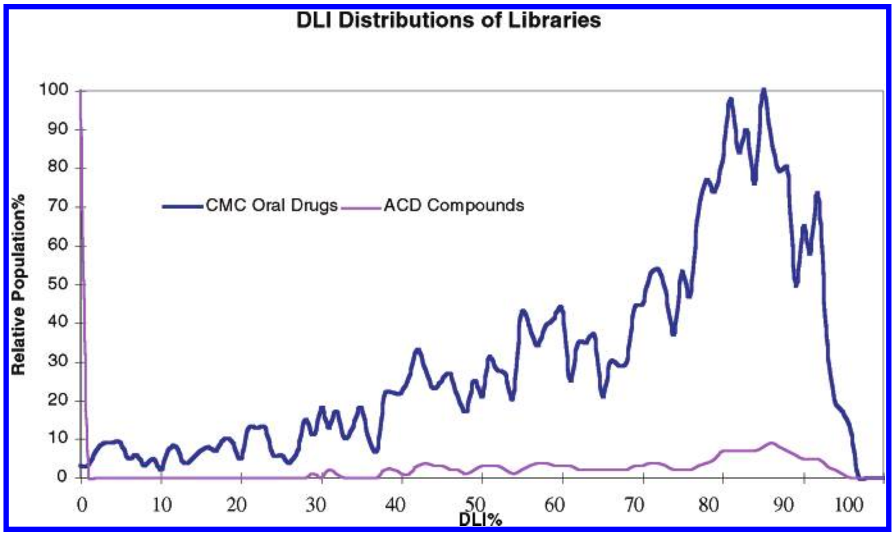


Figure 15. The diversity comparison of CMC oral drugs and ACD compounds.

compounds in a library. When the compounds are ranked with DLI, the structural diversity of the library can be viewed in a DLI distribution graph. In the DLI distribution, the *X*-axis represents DLI of the compounds, and the *Y*-axis represents the relative population of the compounds, which have a given DLI value bin, in this paper, the bin width is 1%. The DLI distribution of CMC drugs is shown in Figure 15.

To compare the structural diversities between two libraries, their DLI distributions are overlaid so chemists can have a clear picture of the diversity. Figures 15 and 16 show the diversity difference among CMC oral drug, ACD compounds, and one of our proposed combinatorial libraries by

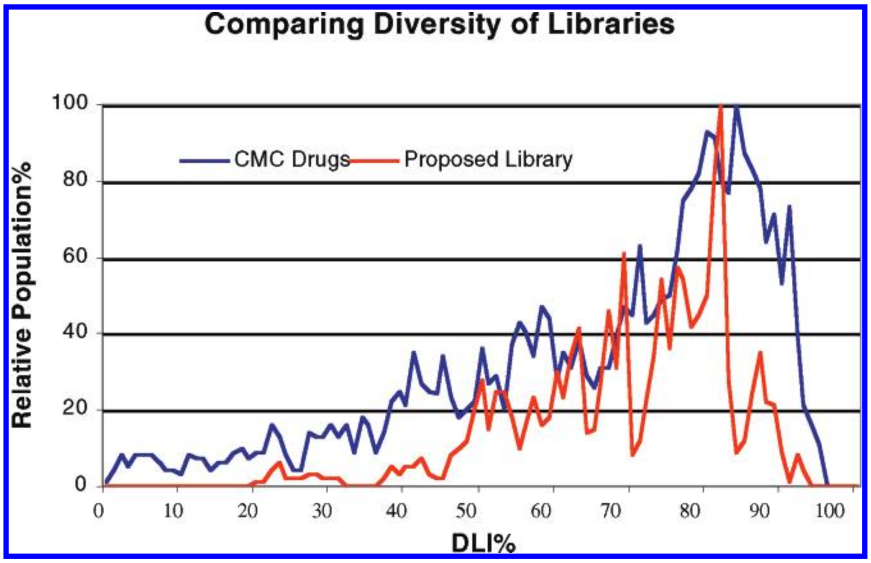


Figure 16. The diversity comparison of a BI combinatorial library and CMC drugs.

Table 4.

algorithm	function	remark
atom-by-atom match	to count unique structural building blocks	refs 17−18
finding SSSR	to find smallest set of smallest rings, normalize aromatic ring systems, find all types of ring building blocks	ref 17
marking ring bonds, chain bonds and linking bonds	to distinguish different types of building blocks	new algorithm based on ref 17
extracting building blocks	to characterize the structural composition (building blocks) of drug compounds	new algorithm
25 structural descriptors calculations	to construct drug like cluster center	some descriptors have been proposed by other authors

means of overlaid the DLI distributions. There are 70,939 compounds out of 250,282 ACD compounds having DLI% less than 1%. In other words, 28.3% ACD compounds are not drug-like. On the other hand, there are 60% ACD compounds are >50% drug-like. By examining the nondrug-like ACD compounds, we find that they are normally simple small molecules, small reagents, peptides, toxic compounds, etc. This analysis shows that CMC oral drugs and ACD compounds do have very different DLI distributions.

This comparison also gives ideas that where is the location of a proposed combinatorial library in drug-like chemical space. By de-convoluting the values of DLI, we can understand why a library has lower/higher average DLI value: such as the number of smallest set of smallest rings is too great, and the number of hydrogen bond donors is too many, etc. The narrow range of DLI distribution of a library indicates the library's diversity is not very good. For a compound acquisition, DLI technology is used as a flexible filter to exclude unwanted compounds. Although there is no ubiquitous threshold, our experience shows that any compound with DLI < 15% is not a good candidate for HTS. Table 3 shows some "bad" compounds (still drugs) filtered by DLI from CMC database.

PROGRAMS AND ALGORITHMS

This DLI project consists of two programs: Knowledge Base Extractor (KBE) and DLI Calculator (DLIC). KBE reads drug structures from a SD file and outputs "drug-like" cluster center. DLIC loads the cluster center and the structures of a combinatorial library (in SD file format) and then outputs to a new SD file with DLI values and the DLI distribution. Both programs are written in C language and running on UNIX, NT and Windows systems.

Table 4 lists the main algorithms used by the two programs; some algorithms, such as structure and substructure match algorithms, SSSR algorithm, are Xu's previous work.

CONCLUSIONS

DLI is a way to rank a set of compounds to make it easier for us to design a diversified yet drug-like combinatorial library. Two compounds having similar DLI values may have similar or different topological scaffolds. Since DLI is based on existing drugs, any marketed drug should have passed many research and development barriers before it arrives the market. This means that DLI is a global measurement for drug-like compounds. A HTS hit might have a lower DLI value, and its DLI may be improved after lead optimization. On the other hand, the drug-like cluster center should be a moving target, because diseases are evolutive, and new drugs are released from time to time. It has been reported that new drugs are larger, more lipophilic, and less permeable, having more internal H-bonds. HTS hits are more lipophilic than phase-2 or marketed drugs etc.[20] To keep in step with the new development of drugs, we can extract the new drug-like cluster center from an updated drug database.

ACKNOWLEDGMENT

Dr. Peter Farina and Dr. Ronald Magolda are appreciated for their valuable comments and continuously encouragement.

REFERENCES AND NOTES

(1) Moos, W. Introduction: Combinatorial Chemistry Approaches the Next Millennium. In *A Practical Guide to Combinatorial Chemistry*; Czarnik, A. W., DeWitt, S. H., Eds.; ACS: Washington, DC, 1997; pp 1−16.

(2) Craig, F. F. Screening of Combinatorial Libraries. In *A Practical Guide to Combinatorial Chemistry*; Czarnik A. W., DeWitt, S. H., Eds.; ACS: Washington, DC, 1997; pp 399−412.

(3) Terrett, N. K.; Gardner, M.; Gordon, D. W.; Kobylecki, R. J.; Steele, J. Combinatorial Synthesis − The Design of Compound Libraries and their Application to Drug Discovery. *Tetrahedron* **1995**, *51*, No. 30, 8135−8173.

(4) Willett, P. Using Computational Tools to Analyze Molecular Diversity. In *A Practical Guide to Combinatorial Chemistry*; Czarnik, A. W., DeWitt, S. H., Eds.; ACS: Washington, DC, 1997; pp 17−47.

(5) Clark, R. D. OptiSim: An Extended Dissimilarity Method for Finding Diverse Representative Subsets. *J. Chem. Inf. Comput. Sci.* **1997**; *37*, No. 6, 1187−1188.

(6) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, No. 3, 572−584.

(7) Barnard, J. M. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, No. 6, 644−649.

(8) Martin, E. J.; Blaney, J. M.; Siani, M. A. et al. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, No. 9, 1431−1436.

(9) Davies, K. Using Pharmacophore Diversity to Select Molecules to Test from Commercial Catalogues. In *Molecular Diversity and Combinatorial Chemistry*; Chaiken, I. M., Janda, K. D., Eds.; ACS: Washington, DC, 1996; pp 309−316.

(10) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, No. 1, 28−35.

(11) Agrafiotis, D. K.; Lobanov, V. S. An Efficient Implementation of Distance-Based Diversity Measures Based on *k*−*d* Trees. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, No. 1, 51−58.

(12) Gillet, V. J.; Willett, P.; Bradshaw, J.; Green, D. V. S. Selecting Combinatorial Libraries to Optimize Diversity and Properties. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, No. 1, 169−177.

(13) Bayada, D. M.; Hamersma, H.; van Geerestein, V. J. Molecular Diversity and Representativity in Chemical Databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, No. 1, 1−10.

(14) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Advanced Drug Delivery Reviews* **1997**, *23*, 3−25.

(15) Ajay, W.; Walters, W.; Murcko, M. A. *J. Med. Chem.* **1998**, *41*, 3314−3324.

(16) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55−68.

(17) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(18) Xu, J. GMA: A Generic Match Algorithm for structural Homorphism, Isomorphism, Maximal Common Substructure Match and Its Applications. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 25−34.

(19) Xu, J.; Zhang, M. HBA: A new algorithm for structural match and applications. *Tetrahedron Computer Methodology* **1989**, *2*, No. 2, 349−356.

(20) Lipinski, C. A. Solubility Screening In An Early Drug Discovery Setting, Proceeding of Early ADME and Toxicology in Drug Discovery: Techniques for Accelerating and Optimizing Drug Candidate Selection; November 12−13, 1998, Berkeley, CA.

(21) Bemis, G. W.; Murcko, M. A. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095−5099.

(22) Wang and Ramnarayan *J. Comb. Chem.* **1999**, *1*, 524−533.

CI000026+