Articles

The Properties of Known Drugs. 1. Molecular Frameworks

Guy W. Bemis* and Mark A. Murcko

Vertex Pharmaceuticals, 130 Waverly Street, Cambridge, Massachusetts 02139-4242

Received April 19, 1996[⊗]

In order to better understand the common features present in drug molecules, we use shape description methods to analyze a database of commercially available drugs and prepare a list of common drug shapes. A useful way of organizing this structural data is to group the atoms of each drug molecule into ring, linker, framework, and side chain atoms. On the basis of the two-dimensional molecular structures (without regard to atom type, hybridization, and bond order), there are 1179 different frameworks among the 5120 compounds analyzed. However, the shapes of half of the drugs in the database are described by the 32 most frequently occurring frameworks. This suggests that the diversity of shapes in the set of known drugs is extremely low. In our second method of analysis, in which atom type, hybridization, and bond order are considered, more diversity is seen; there are 2506 different frameworks among the 5120 compounds in the database, and the most frequently occurring 42 frameworks account for only one-fourth of the drugs. We discuss the possible interpretations of these findings and the way they may be used to guide future drug discovery research.

Introduction

The drug design process is largely driven by the instincts, intuition, and experiences of pharmaceutical research scientists. It is often instructive to attempt to "capture" these experiences by analyzing the historical record, i.e., successful drug design projects of the past. The inferences drawn from this analysis can play an important role in shaping our thinking on current and future projects. For this reason, we would like to analyze the structures of a large number of drugs—the ultimate product of a successful drug design effort. There is a wealth of information implicitly encoded in the two-dimensional and three-dimensional structures of molecules that are currently sold as drugs. This includes toxicity, stability (both chemical and metabolic), synthetic accessibility, starting material costs, and the like. Our goal for this paper is to begin to deconvolute this information in order to apply it to the design of new drugs.

There are several computational tools available for this analysis: substructure searching using one of several commercially available software packages (e.g. Merlin, ISIS, Unity),[1−3] automated ring searching using one of several published algorithms,[4−8] and shape descriptor methods.[9−12] We use shape descriptor methods because they are easily implemented and are flexible enough to allow the analysis to be performed in an automated way.

We analyze the Comprehensive Medicinal Chemistry (CMC) database[13] which contains two-dimensional and predicted three-dimensional structures and important biochemical properties for known drugs. The CMC database has been developed from Pergammon's Comprehensive Medicinal Chemistry series.[14]

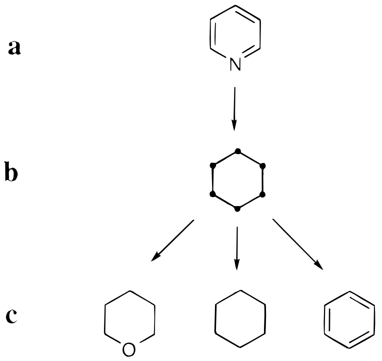


Figure 1. Graph representation of molecules.

Methods

The current version of the CMC database (v. 94.1) includes more than 6700 compounds. However, many of these do not meet our criteria for various reasons, e.g., imaging agents, dental resins, and veterinary compounds. Thus, our first task was to identify and remove these compounds. We eliminated all compounds for which no therapeutic activity class was given, as well as compounds which fell into any of the following classes: radiopaque agents, contrast agents, solvents, anesthetics, disinfectants, topicals, local agents, spermicides, wetting agents, flavoring agents, pharmaceutical aids, surgical aids, dental, surfactants, sunscreens, ultraviolet screens, emetics, preservatives, aerosol propellants, chelators, keratolytics, insecticides, astringents, herbicides, laxatives, sweeteners, dental caries prophylactics, adhesives, dentistry, pharmaceutic aids, veterinary, buffers, scabicides, and ectoparasiticides. After this process, the CMC database had 5120 remaining entries.[15]

Our analysis of the structures in the CMC database has been carried out on two levels, using atomic properties and graph properties. Atomic properties include such information as element type, atomic hybridization, and atomic charge. Graph properties of molecules are the connectivity properties of the atoms representing a molecule, that is, the information that may be derived from a molecular structure by considering each atom to be a vertex and each bond to be an edge on a graph.[16] The graph for a particular molecule may be considered an archetype for each instance of that molecular shape.

[*] To whom correspondence should be addressed. E-mail: bemis@vpharm.com.

[⊗] Abstract published in *Advance ACS Abstracts,* July 1, 1996.

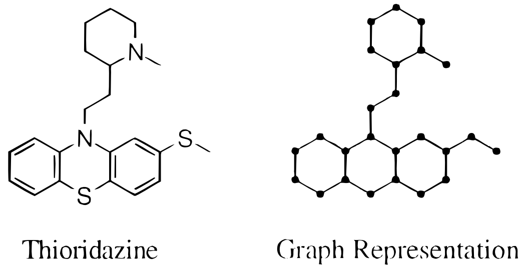


Figure 2. Graph representation of a typical drug molecule.

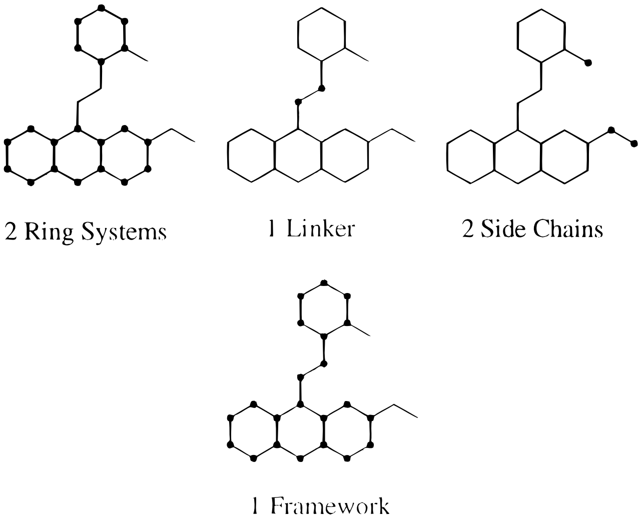


Figure 3. Distinguishing between ring systems, linkers, and side chains.

That is, for a molecule such as pyridine (Figure 1a), the molecular graph or archetype is the graph with six vertices (Figure 1b). The same archetype represents molecules such as benzene, cyclohexane, and pyran, among many others (Figure 1c). Thus the structures of molecules can be readily analyzed in terms of a hierarchy in which molecular archetypes are at the top, and individual molecules are at the bottom (Figure 1).

When analyzing drug molecules, one is faced with a slightly more complicated set of graphs than in the simple example shown in Figure 1. To demonstrate this point, we might consider the antidepressant thioridazine, which is shown along with its graph representation or archetype in Figure 2. We can now pick out structural elements which can be used to further order groups of atoms within a molecular graph. We may dissect any molecule into four units: ring, framework, linker, and side chains. We adopted the following definitions to aid our analysis.

Ring Systems. We define ring systems to be cycles within the graph representation of molecules and cycles sharing an edge (a connection between two atoms or a bond). For example, benzene, naphthalene, and anthracene are all single ring systems. Treating cycles this way makes sense from a chemical structural point of view. As an approximation, the cycles and fused cycles in a molecule represent rigid units in which many degrees of freedom are removed from a collection of atoms.

Linker Atoms. Atoms that are on the direct path connecting two ring systems are defined as linker atoms. As can be seen in Figure 3, thioridazine has a two-atom linker connecting the two ring systems. Molecules such as biphenyl have a zero atom linker—the six-membered rings are different ring systems.

Side Chain Atoms. Any nonring, nonlinker atoms are defined as side chain atoms. Figure 3 shows that thioridazine has two side chains: a single-atom side chain attached to the six-ring and a two-atom side chain attached to the fused tricyclic ring system.

Framework. The framework is defined as the union of ring systems and linkers in a molecule. As shown in Figure 3, the thioridazine molecule consists of two ring systems: a six-ring and three linearly fused six-rings. Together these rings and linkers define the framework of this molecule. The concept of a framework is central to our paper, and provides an important distinction between our present work and work done previously.[6]

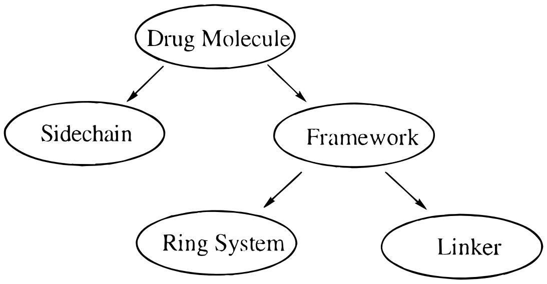


Figure 4. Hierarchical description of molecules.

We can now classify molecules and their constituent atom groupings into a hierarchy as shown in Figure 4. This classification scheme is very useful for analyzing the structures of drug molecules for several reasons. First, well-represented frameworks can be identified, and emphasis can be placed on these for new drug discovery. Second, linkers and ring systems can be identified for potential use in a combinatorial-type approach to compound library generation. Third, compound libraries may be evaluated for their relationship to the shapes of known drugs. In other words, we can evaluate how well the diversity space of a library overlaps with our representation of drug-space.

We begin our analysis by identifying side chain atoms, which is done as follows. Each atom bonded to only one other atom is identified as a side chain atom and removed from the molecule. This process is repeated until either the molecule disappears (acyclic molecules) or until each atom is bonded to at least two other atoms. The remaining atoms are identified as the framework atoms. The next step in our analysis is the identification of atoms within the framework that are in rings (or cycles in the graph) using a depth-first search.[17] Any atom not part of a ring is identified as a linker atom. This process follows the hierarchy shown in Figure 4.

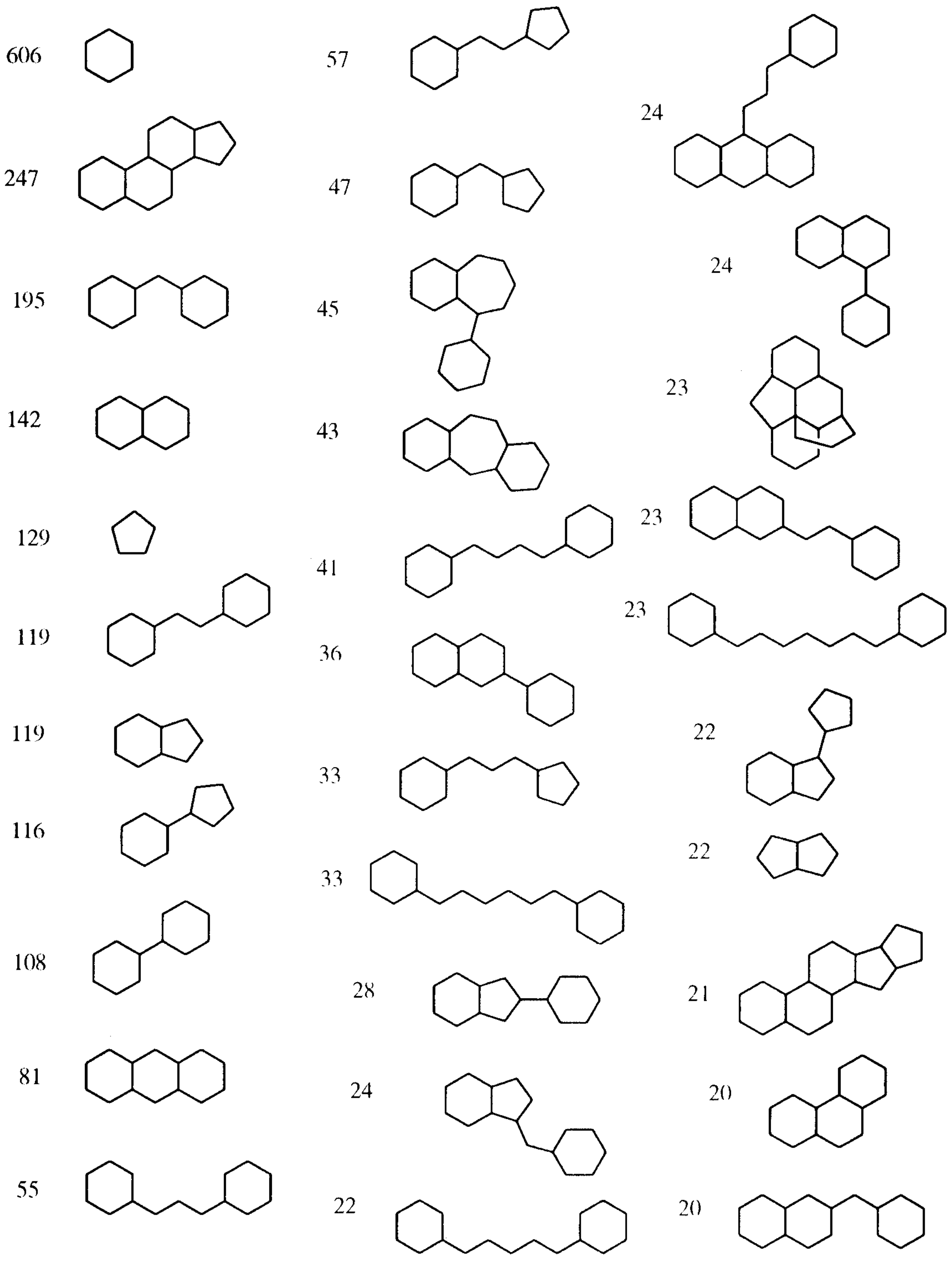
The molecular frameworks obtained in this manner were grouped into clusters of identical shape description. Our analysis has been carried out in two ways: we have conducted both a purely graph theoretical analysis and an analysis which also considers atomic properties. Both methods follow essentially the same formal procedure with the only difference being the shape descriptor used. For the graph analysis we used two-dimensional triangle shape descriptors[12] and for the analysis including atomic properties we used topological torsions.[11] For computation of topological torsions, we found it necessary to retain the π electrons associated with framework atoms when side chains were removed. For example, cyclohexanone would have the sp^2 oxygen tagged as a side chain atom, and the sp^2 carbon tagged as having two associated pi electrons. On the basis of the topological torsion representation, the cyclohexanone framework would therefore have a different shape description than the cyclohexane framework. The cyclohexanone framework is therefore represented with two dots next to the sp^2 carbon to indicate the associated electrons. We have used this notation in Charts 2 and 3.

Results

First we summarize the results of the graph theory (archetype) analysis and then the atomic property (instance) analysis. Finally, we discuss the relationship between the two kinds of analysis.

From the graph theory analysis, there are 1179 different frameworks among the 5120 compounds analyzed. Of these frameworks, 783 (66%) are unique, i.e.,

Chart 1. Graph Frameworks for Compounds in the CMC Database as Classified by Connectivity Triangles (Numbers Indicate Frequency of Occurrence)



they are found in only one drug molecule. Chart 1 shows graph frameworks for compounds in the CMC database as classified by connectivity triangles. We have shown only frameworks that exist in at least 20 drugs. This set of 32 frameworks accounts for 50% of the 5120 total drug molecules. Clearly the six-ring is the most commonly used framework for these drugs. Acyclic molecules (those with no framework) account for 306 (6%) of the molecules we examined.

Our second method of analysis uses topological torsions[11] for classification. Several atom properties (atom type, hybridization, and bond order) are considered. Somewhat more diversity is seen; there are 2506 different frameworks among the 5120 compounds in the database. Again, a large majority of these frameworks (1908, or 76%) are unique. Chart 2 shows atomic property-based drug frameworks (drug instances) that occur in the CMC at least 10 times. Naturally, because this classification scheme accounts for hybridization and bond order, one would expect a more diverse set of frameworks to be required to represent the drug database. Even so, this set of 41 frameworks accounts for

Chart 2. Atomic Frameworks for Compounds in the CMC Database as Classified by Topological Torsions (Numbers Indicate Frequency of Occurrence)

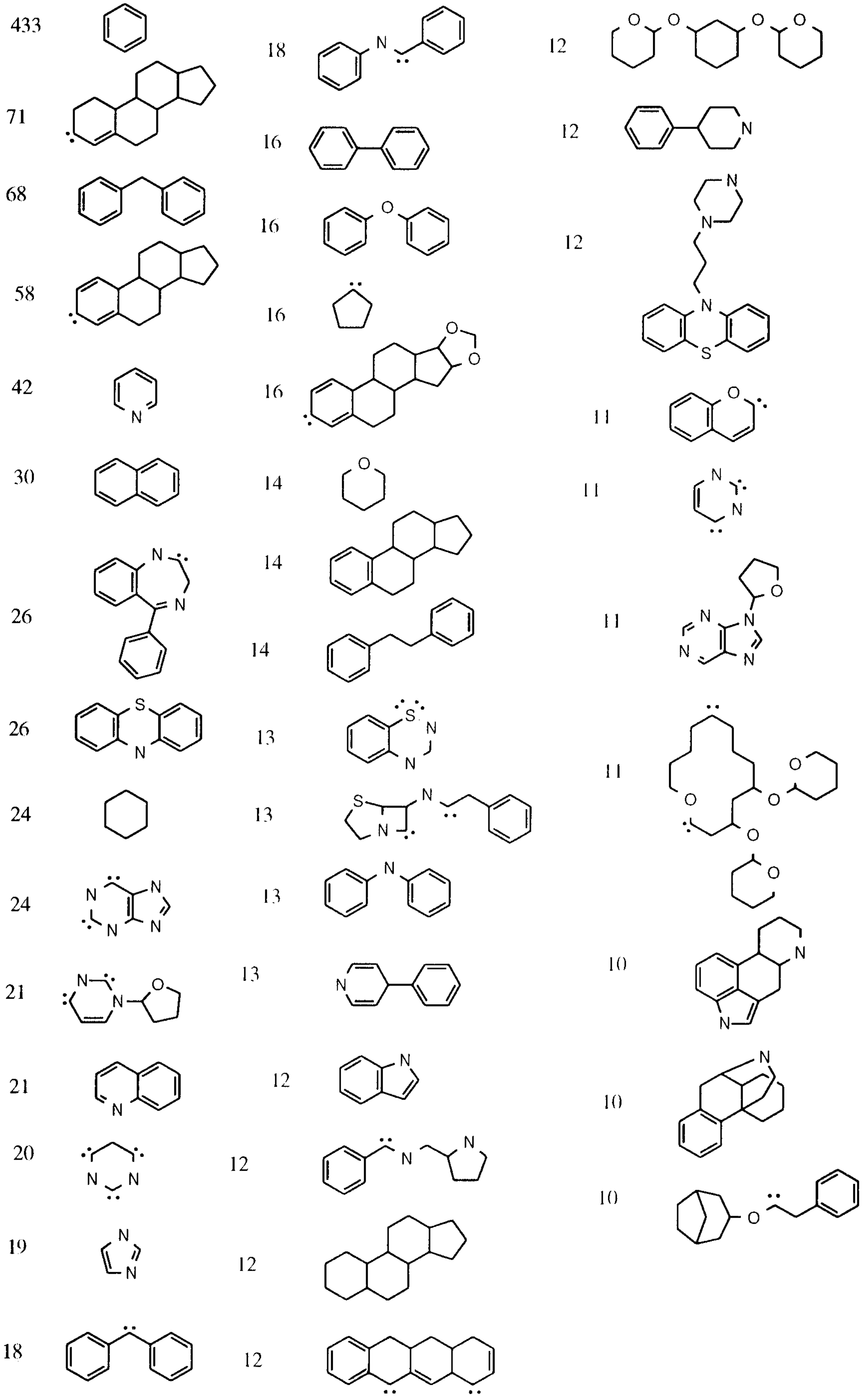
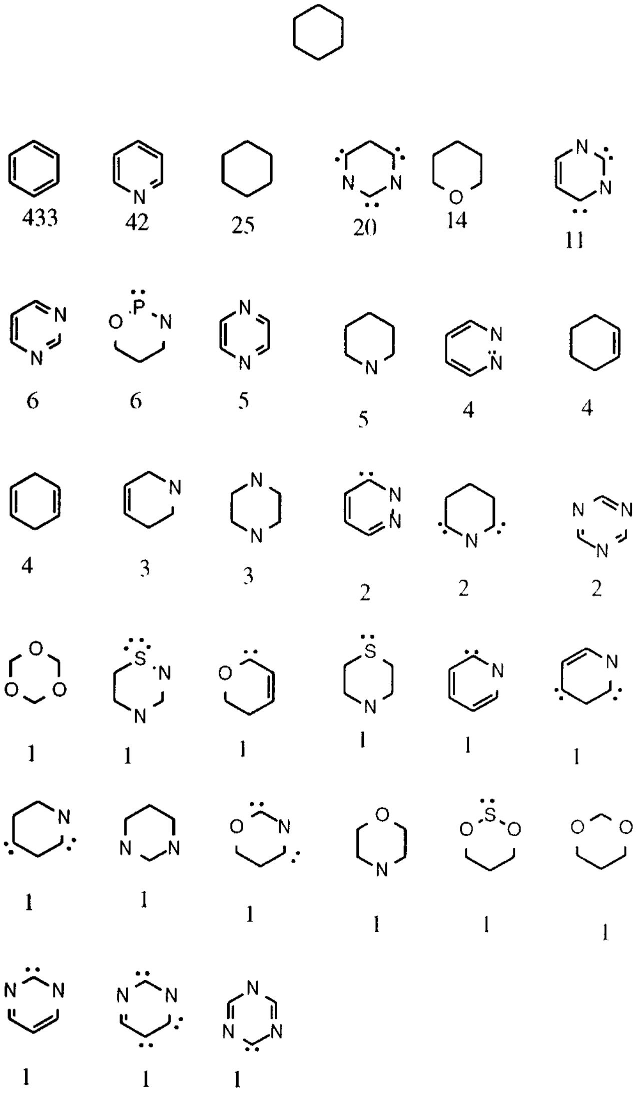


Chart 3. All Six-Membered Rings Found in the CMC Database (Numbers Indicate Frequency of Occurrence)

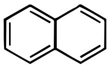


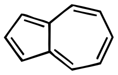
1235 (24%) of the 5120 molecules we examined. Clearly benzene is the most commonly used framework for these drugs.

It is instructive to understand the relationship between the graph theory frameworks, which can be viewed as providing a “high-level” or “generic” classification scheme, and the atom property-based frameworks, which further subdivide classes of frameworks based on their chemical properties. As an example, we may consider the atomic property based framework for the most popular graph theory based framework—the six-ring. Chart 3 shows the set of six-ring atomic frameworks that accounts for the 606 six-ring scaffolds found in our filtered version of the CMC database. Overwhelmingly, the most common six-ring atomic framework is benzene. Of the drug molecules we considered, 8.5% (433 out of 5120) have benzene as their molecular framework.

Chart 1 can be further broken down (by inspection) into rings and linkers. The linkers present are chains with zero to seven nodes shown in Chart 4—rings and linkers. Rings have a dashed line showing points where linkers can potentially be attached. By using this set of 14 rings (with eight potential attachment points) and eight linkers, we can derive the molecular frameworks for over half of known drugs (as defined by our subset of the CMC database).

A problem sometimes encountered when using molecular shape descriptors is multiple representations, cases where different shapes are represented by identical shape descriptions. There are a number of ways to deal with this problem, such as adding more detail to the shape descriptor or using multiple shape descriptors.[6] For small data sets such as the CMC, perhaps the simplest solution is to look through groups of molecules with identical shape descriptions and pick out cases of multiple representation. This is the method we used. An example of multiple representation is found in the topological torsion shape description of these two molecules:


A


B

We found two examples of the **B** molecular framework grouped with 30 examples of the type **A** framework so we assigned them to separate clusters.

Finally, we should note that as a control, a partial analysis was performed also on the complete CMC database (approximately 6700 compounds), and the results were substantially the same.

Discussion

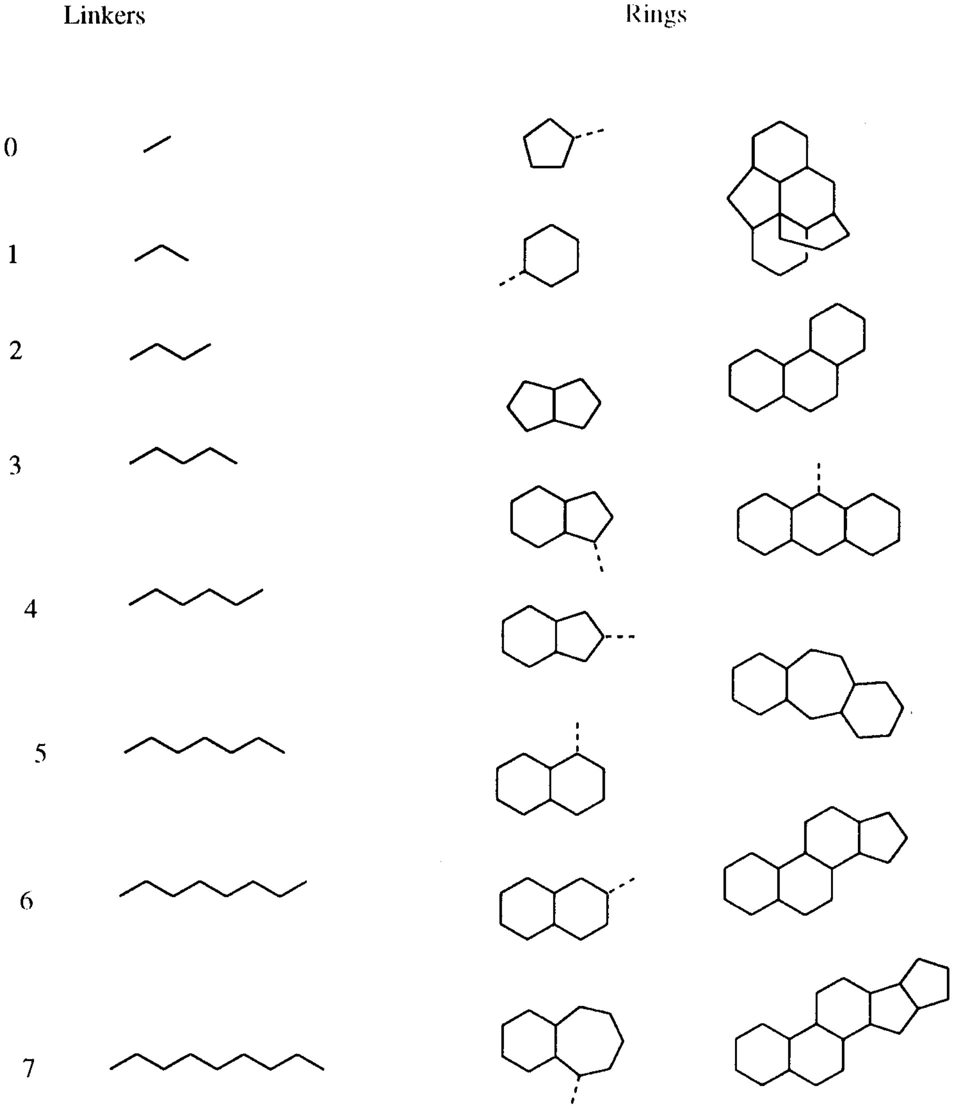
This is our first attempt at classifying the shapes of drug molecules, and our goal is to provide a “high-level overview” of the gross structural features of these molecules. Accordingly, for purposes of this research, we have deliberately defined “shape” in simple terms. The first classification scheme ignores such important features as the details of substituents on rings, chain branching, bond order, atom types, stereochemistry, and three-dimensional conformation. The second classification method does account for bond order and atom types.

There is no reason to believe that the set of 5120 molecules in our database represents all the possible shapes that a drug may take. However, it is instructive to examine the universe of known drugs to see what patterns may exist. Once these patterns have been deduced, the drug designer may apply them in various ways. For example, one might attempt to bias a *de novo* design program or a combinatorial chemistry effort to produce a set of molecules which either contains or does not contain these patterns.

The reader must bear in mind that “shape” in this work refers to the two-dimensional topological graph of the molecules. While three-dimensional shape is partially encoded in the two-dimensional graph of a molecule, we expect that the three-dimensional conformations of drugs with the same topological shape will not all be similar, although certain conformations would be expected to appear more frequently than others.

Of course, the preferences we have identified for certain shapes do not necessarily reveal some fundamental truth about drugs, receptors, metabolism, or toxicity. Instead, it may reflect the constraints imposed by the scientists who have produced these drugs. Constraints due to synthetic or patent considerations, cost, or a general conservatism (i.e., a tendency to make

Chart 4. Graph Representations of the Rings and Linkers for the Most Common Drug Frameworks Found in Chart 1[a]



[a] Linkers are depicted with open valences on each end; the number of nodes in each linker is given to the left. Rings are depicted with dashed lines indicating possible points of attachment for linkers.

new compounds which are structurally similar to known compounds) all may be reflected in these findings.

However, *half of the known drugs fall into only 32 shape categories.* The drugs which possess these topological shapes (Chart 1) are quite different in polarity, conformation, hydrogen-bonding potential, and other properties; they bind to different classes of receptor; and they serve different pharmacological needs. And yet, they all have the same topological shape.

In part, the results in Chart 1 stem from the simplicity of our classification scheme, but it also may reflect some of the properties which are beneficial for producing drugs. For example, if we consider the set of 32 frameworks in Chart 1, we see that most (23) contain at least two six-rings linked or fused together. We also see that only three of these frameworks have more than five rotatable bonds.

A "pharmacological promiscuity" parameter could be provided for each of our frameworks. This was suggested to us by one external reviewer and several internal reviewers. This parameter would be defined by the ratio of targets to frameworks, that is, the number of pharmacological targets acted upon by drugs composed of a particular framework divided by the number of drugs made from that framework.

As an example, the biphenyl molecular framework (Chart 2) constitutes 16 drugs in our database. The CMC lists the following distinct therapeutic classes for these drugs: antiamebic, antifungal, antiinfective, antihypercholesteremic, antihyperlipoproteinemic, fasciolicide, antirheumatic, analgesic, anti-inflammatory, antithrombotic, uricosuric, and antiarrhythmic. The pharmacological promiscuity parameter for this molecular framework is therefore 12/16 or 0.75.

This parameter would be extremely useful for several purposes such as choosing a scaffold upon which to begin a combinatorial design effort. Unfortunately, the exact pharmacological target for each drug is not known, and often multiple therapeutic categories are listed for drugs, so this analysis would require either dealing with a very restricted subset of drugs or grouping together similar low-level pharmacological targets.

It is intriguing to consider ways in which our analysis might be used to direct a *de novo* design effort. For example, on the basis of the above-mentioned observation that two six-membered rings are a common motif, one might begin a *de novo* exercise by docking two benzene rings into the active site using shape-based methods that ignore electrostatics.[18,19] Next, one could link or fuse these rings into a single ligand using one of several algorithms,[20−23] placing special emphasis on scaffolds found in Chart 1 (directed linking). Finally, one could assign atom types for the ligand based on electrostatic complementarity with the active site,[19] placing special emphasis on the atomic distributions found for the scaffolds found in Charts 2 and 3 (directed atom assignment). Some minimization would likely be needed as different atomic hybridizations are overlaid on the initial benzene fragments.

Many other approaches also are possible. For example, one might attempt to utilize the frameworks found in Chart 2. These could be used as seed structures for *de novo* structure generation by random combination of fragments[24,25] and linkers such as those in the ILIAD database.[23] Finally, our collection of "rings and linkers" in Chart 4 might be used in conjunction with fragment perception algorithms[26] and similarity methods[27] to select compounds for synthesis and testing from a combinatorial library or compound collection database.

Future research in the area of "drug database mining" will focus on other properties of known drugs including their flexibility, log *P*, solubility, and a more detailed shape description that includes such features as charge and hydrogen bonding potential.

Acknowledgment. We would like to thank Ajay, Chris Baker, Joshua Boger, Chris Lepre, Roger Tung, Pat Walters, Keith Wilson, and Bob Zelle for their suggestions and their careful reading of the manuscript. We thank Scott Thomas for help with processing the CMC database.

References

(1) Available from DAYLIGHT Chemical Chemical Information Systems, Inc., Irvine, CA.

(2) Available from MDL Information Systems, Inc., San Leandro, CA.

(3) Available from Tripos, Inc., St. Louis, MO.

(4) Klingebiel, U.; Specht, K. Automatic Generation of the Chemical Ringcode from a Connectivity Chart, *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 113−116.

(5) Randic, M. Ring ID Numbers, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 142−147.

(6) Nilakantan, R.; Bauman, N.; Haraki, K.; Venkataraghavan, R. A Ring-Based Chemical Structural Query System: Use of a Novel Ring-Complexity Heuristic. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 65−68.

(7) Domokos, L. Beilstein Ring Search System. 1. General Design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 663−667.

(8) Fan, B. T.; Panaye, A.; Doucet, J.-P.; Barbu, A. Ring Perception. A New Algorithm for Directly Finding the Smallest Set of Smallest Rings from a Connection Table. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 657−662.

(9) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; JohnWiley & Sons, Inc.: New York, 1990.

(10) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 82−85.

(11) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.*, **1987**, *27*, 82−85.

(12) Bemis, G. W.; Kuntz, I. D. A fast and efficient method for 2D and 3D molecular shape description. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 607−628.

(13) Comprehensive Medicinal Chemistry (CMC-3D) Release 94.1 is available from MDL Information Systems Inc., San Leandro, CA.

(14) *Comprehensive Medicinal Chemistry, Vol. 6*; Hansch, C., Sammes, P. G., J. B., Taylor, Series Eds.; Pergamon: Oxford, 1990.

(15) A similar process of removing compounds from the CMC has been carried out as part of an analysis of the molecular weights of known drugs: Kim, E. E.; Baker, C. T.; Dwyer, M. D.; Murcko, M. A.; Rao, B. G.; Tung, R. D.; Navia, M. A. Crystal Structure of HIV-1 Protease in Complex with VX-478, a Potent and Orally Bioavailable Inhibitor of the Enzyme. *J. Am. Chem. Soc.* **1995**, *117*, 1181−1182.

(16) For a good introduction to molecules as graphs, see: Hansen, P. J.; Jurs, P. C. Chemical Applications of Graph Theory. *J. Chem. Ed.* **1988**, *65*, 574−580.

(17) Cormen, T. H.; Leiserson, C. E.; Rivest, R. L. *Introduction to Algorithms*; MIT Press: Cambridge, 1990; pp 477−485.

(18) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

(19) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **1992**, *13*, 505−524.

(20) Roe, D. C.; Kuntz, I. D. BUILDER v.2: Improving the chemistry of a de novo design strategy. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 269−282.

(21) Lewis, R. A.; Roe, D. C.; Huang, C.; Ferrin, T. E.; Langridge, R.; Kuntz, I. D. Automated site-directed drug design using molecular lattices. *J. Mol. Graph.* **1992**, *10*, 66−78.

(22) Lauri, G.; Bartlett, P. A. CAVEAT: a Program to Facilitate the Design of Organic Molecules. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 51−66.

(23) Gillet, V. J.; Newell, W.; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A. P. SPROUT: Recent Developments in the De Novo Design of Molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 207−217.

(24) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. A Method for Automatic Generation of Novel Chemical Structures and Its Potential Applications to Drug Discovery. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 527−530.

(25) Pearlman, D. A.; Murcko, M. A. CONCERTS: Dynamic connection of fragments as an approach to de novo ligand design. *J. Med. Chem.* **1996**, *39*, 1651−1663.

(26) Barakat, M. T.; Dean, P. M. The Atom Assignment Problem in Automated De Novo Drug Design. 2. A Method for Molecular Graph and Fragment Perception. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 351−358.

(27) Willett, P. Algorithms for the Calculation of Similarity in Chemical Structure Databases. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; JohnWiley & Sons, Inc.: New York, 1990; pp 43−64.

JM9602928