Virtual Hydrocarbon and Combinatorial Databases for Use with CAVEAT

Yongliang Yang,[†] Dmitri V. Nesterenko,[†] Ryan P. Trump,[‡,§] Ken Yamaguchi,[‡,||] Paul A. Bartlett,[‡] and Dale G. Drueckhammer*[,†]

Department of Chemistry, Stony Brook University, Stony Brook, New York 11794, and Center for New Directions in Organic Synthesis, Department of Chemistry, University of California, Berkeley, California 94720

Received July 6, 2005

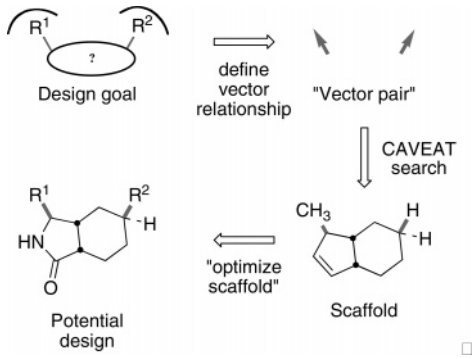
Three new virtual databases have been developed for use with the bond-orientation-based database searching program CAVEAT. These consist of a database of trisubstituted monocyclic hydrocarbons having ethyl, vinyl, and phenyl substituents; a database of unsubstituted bicyclic hydrocarbons; and a database of core structures from established combinatorial synthetic methods having hydrogen, ethyl, vinyl, and phenyl substituents at the readily varied positions. Each collection of molecules was subjected to a batch conformational search, minimization, and conversion to a vector database for use with CAVEAT.

INTRODUCTION

As a computational tool in molecular design, CAVEAT is unique.[1−3] By searching three-dimensional databases, the program identifies molecular frameworks that can serve as templates to position functional groups in a specific relative orientation. In contrast to search algorithms that specify geometric relationships by interatomic distances, CAVEAT characterizes the three-dimensional search query by a vector relationship among bonds, as illustrated in Scheme 1. Thus, a typical search query would be defined by the relative orientation of bonds leading to substituents in a desired relationship; CAVEAT could then identify, from a structural database, molecules with bonds that project in the same relative orientation. Grafting the substituents onto these frameworks, with further design optimization, provides potential targets for synthesis. CAVEAT is not limited in the number of bond vectors defined in the search query, and applications based on matching three vectors have been demonstrated. CAVEAT was originally conceived for the design of conformationally constrained peptides and peptidomimetics,[1,3−6] but it has also been employed in drug design,[7,8] in the design of small molecule receptors[9] and of chiral ligands for asymmetric catalysis.[10,11]

Database searching methods have also been widely used in virtual screening for drug design based on the docking of small molecules into a binding site of an enzyme or other protein receptor.[12−14] However, these applications require databases with features that are quite different from those desired for CAVEAT searches. While drug design efforts rely largely on functional groups that interact directly with a target protein through hydrogen bonds and other polar or hydrophobic interactions, CAVEAT is only concerned with the relative three-dimensional arrangement of bonds, since the functionality will be introduced later in the design process. As a consequence, the identity of the functional groups on the molecules in the 3D database, indeed, whether the substitutable positions are identified as C−C, C−H, or N−C bonds, and so forth, is immaterial. Databases of relatively simple, unadorned structures, therefore, provide as much structural diversity as databases of complex, functionally diverse compounds.

Scheme 1



Although CAVEAT is a powerful tool for many structure-based design applications, its utility only extends as far as the databases of three-dimensional molecular frameworks that are available for searching. The Cambridge Structural Database (CSD)[15] is the premier source of molecular structures, but it is only a sampling of three-dimensional structure space and has many molecules that are impractical as templates. As complementary databases, the Berkeley group developed the TRIAD and ILIAD databases, consisting of collections of computer-generated molecules.[1,2] TRIAD is a database of more than 400 000 tricyclic hydrocarbons containing three- to six-member rings, all possible patterns of unsaturation including one exocyclic methylene group on each ring, and all possible stereoisomers (excluding enantiomers, since the CAVEAT search itself considers both the native and enantiomeric versions of a 3D structure). ILIAD is a database of more flexible structures assembled from

* Corresponding author phone: (631) 632−7923; email: dale.drueckhammer@sunysb.edu.
[†] Stony Brook University.
[‡] University of California.
[§] Current address: Discovery Research, GlaxoSmithKline, 5 Moore Drive, Research Triangle Park, North Carolina 27707.
[||] Current address: Systems Biology Department, Berlex Biosciences, 2600 Hilltop Drive, Richmond, California 94804.

10.1021/ci050277o CCC: $30.25
Published on Web 11/05/2005

combinations of five of the simple building blocks $-CH_2-$; $-CH=CH-$; $CH_2=C<$; $-C\equiv C-$; $-S-$; $-S-S-$; *o*-, *m*-, and *p*-phenylene; *o*- and *m*-disubstituted cyclopentadienyl; and trimethylenemethane.

Our work on the application of CAVEAT to receptor design for small molecule recognition and on peptidomimetics stimulated our desire for additional databases to complement those already available, namely, with cores of less complexity and greater synthetic accessibility than that represented by the CSD or TRIAD/ILIAD. We now describe methods for the rapid construction of the following databases of virtual scaffolds for use with CAVEAT: (1) a library of trisubstituted monocyclic hydrocarbon structures, (2) a database of simple unsubstituted bicyclic hydrocarbon structures, and (3) a library of scaffolds from reported combinatorial chemistries.

METHODS

General. The two hydrocarbon libraries were constructed on a Silicon Graphics Octane computer with a 190 MHz processor running IRIX v.6.5 OS. Cerius2 version 4.0 and Catalyst version 4.6 from Accelrys were used for library generation and conformer generation. Structural minimization was performed with Macromodel v.6.2 on a 64 knot cluster in the structural biology computing center at Stony Brook. The Gemini database was constructed and structural minimizations were performed on a Silicon Graphics Octane computer with dual 250 MHz processors running IRIX v. 6.5 OS. Vector databases (indices to the 3D databases) were generated using the CAVEAT program itself.

Generation of Trisubstituted Hydrocarbon Core Structures. Core structures were built in the 3D sketcher module of Cerius2. The positions for substitution on the core structure were selected and the hydrogen, ethyl, vinyl, and phenyl substituents were defined in the Builders2-Analog Builder module. Virtual libraries were generated while writing structures directly to an sd file. For the library file generated from each core structure, the UNIX script "find_hydrogen_sd" was used to generate a new file containing only the trisubstituted structures (all UNIX scripts are available in the Supporting Information).

Construction of Bicyclic Hydrocarbon Structures. Structures corresponding to all possible patterns of unsaturation of seven bicyclic ring systems were entered manually into the Cerius2 interface.

Construction of Combinatorially Accessible Structures—Gemini. Structures corresponding to selected literature and internally developed combinatorial syntheses[16−23] were entered manually into the Cerius2 interface with the hydrogen, ethyl, vinyl, and phenyl substituents at points of variation.

Conformer Generation and Minimization and Vector Database Construction. The script "fix-msi-sd" (Supporting Information) was used to remove any duplicate molecules in the sd file, to insert the term "-ISIS-" in the header information for each molecule so that Catalyst would recognize the stereochemistry, and to remove the charge notation. Multiple conformations were generated using the Catalyst program, and all conformers within 20 kcal/mol of the global minimum were retained. The output .cpd files were exported to .mmod files, which are the input files for Macromodel. Each resulting file was run through the batch minimization protocol of Macromodel using a modified script to make it K shell compatible.[12] All conformers within 5 kcal/mol of the minimum for each structure were retained. All of the output conformer files from the Macromodel BatchMin process have the same header information "CONF", which does not distinguish them from each other. The script "fix-header-out" was written in awk to insert the filename and conformer numbers in the header information of output files from the Macromodel minimization process so that CAVEAT would recognize them as different conformers. These minimized conformer files were grabbed by a script "grab-minimized-source", written in K shell, to put them into CAVEAT source databases. The source database and vector database were constructed using the CAVEAT utilities.

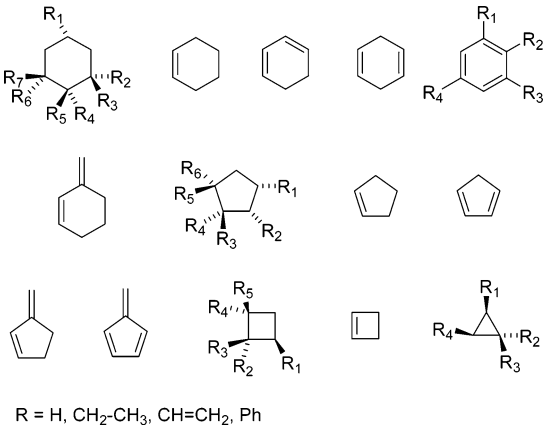


Figure 1. Base Structures in the TRISUB Database

RESULTS AND DISCUSSION

Cyclohexane was chosen as the initial core structure for diversification with combinations of ethyl, vinyl, and phenyl substituents. These three substituents were chosen as representatives of the common geometries of sp^3, sp^2, and aryl functional groups. Variation of all 12 positions of cyclohexane to form 4^{12} structures would create an extremely large file, perhaps too large to even be recorded on our largest available memory drive, and would have a great deal of redundancy. We, thus, decided to develop a database with only three positions of substitution in each structure. Trisubstituted structures were chosen because a three-vector match is sometimes desired and because the third substituent may provide some conformational influence on the ring system, especially with more flexible rings, and thus provide more diversity. For generation of all of the possible trisubstituted cyclohexanes, it was necessary to vary only 7 of the 12 positions (Figure 1). The analogue builder feature of the Cerius2 program suite (Accelrys) was used to generate the library of structures in which the seven selected positions of cyclohexane were varied with hydrogen, ethyl, vinyl, and phenyl substituents. This process generated a large collection of structures ($4^7 = 16\,384$) that still contained a great deal of redundancy, as each possible arrangement of three substituents is represented in a large number of structures with variable substitution at the other four positions. The next step in the procedure was to minimize the size of the database while maintaining all possible three-vector combinations by selecting only the trisubstituted structures.

This goal was achieved by viewing the order of generation of structures as a base 4 numbering system, with each

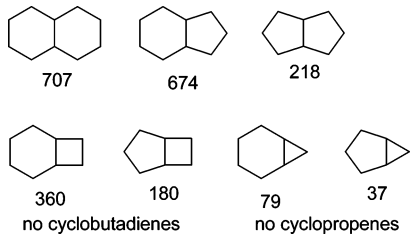


Figure 2. Core Structures in the BIAD Database.

structure corresponding to a seven-digit base 4 number. In this analysis, each digit refers to one of the seven R groups and the value of each digit indicates the substituent at that position (0 = H, 1 = ethyl, 2 = vinyl, and 3 = phenyl). Unix scripts were written to (a) generate the set of base 4 numbers in which all but three of the digits are 0 (H atoms), (b) convert them to base 10 numbers, (c) add 1 to each (because the structures are numbered starting from 1 rather than 0), (d) select the structures corresponding to these numbers, and (e) save them to a new file.

This file, thus, contains only the desired cyclohexane structures in which exactly three of the positions are substituted by ethyl, vinyl, or phenyl groups. The procedure for the generation of trisubstituted cyclohexanes was then applied to the 13 additional ring systems shown in Figure 1. The R groups represent the positions that were varied for each structure in order to generate all possible trisubstituted structures. For those structures in which R groups are not shown, all positions were varied, including the exocyclic methylene hydrogens.

The individual structures were processed further to generate the energetically favorable conformations of each molecule. Specifically, the CATALYST program (Accelrys) was used to generate all of the reasonable conformations of each structure. Each conformer was minimized using the Batch-Min routine of MacroModel, and all unique conformations within 5 kcal/mol of the minimum energy conformation were retained. The groups of elaborated ring systems generated by this approach were combined and processed into the vector database for use with CAVEAT. A total of about 757 000 structures was obtained, including all conformational isomers.

An additional database of unsubstituted bicyclic hydrocarbons was developed based on seven ring systems shown in Figure 2. For each ring system, structures having every possible pattern of unsaturation were entered manually into the Cerius2 interface. Structures containing one or two exocyclic double bonds (methylene groups) in addition to endocyclic double bonds were also included. The total number of structures based on each ring system, including stereoisomers, is also given in Figure 2. In total, 2255 structures were obtained. These structures were subjected to a conformational search, energy minimization, and selection of structures within 5 kcal/mol of the lowest energy structure using methods described above. This new database has been tentatively named BIAD. In principle, these bicyclic "core" structures could be identified in CAVEAT searches of the TRIAD database. However, the conformation of the bicyclic system may be influenced by the third ring of the tricyclic structures, an issue that is avoided with this bicyclic database (moreover, no conformational search was performed in generating TRIAD).

The above hydrocarbon databases sample a large area of conformational space, and compounds are often found with the desired vector relationship in CAVEAT searches. However, a facile synthetic route to the substituted scaffold compound may not be readily apparent. Indeed, a shortcoming of the TRIAD and ILIAD databases, as well as the above databases, is their "generic" nature. That is, as hydrocarbon

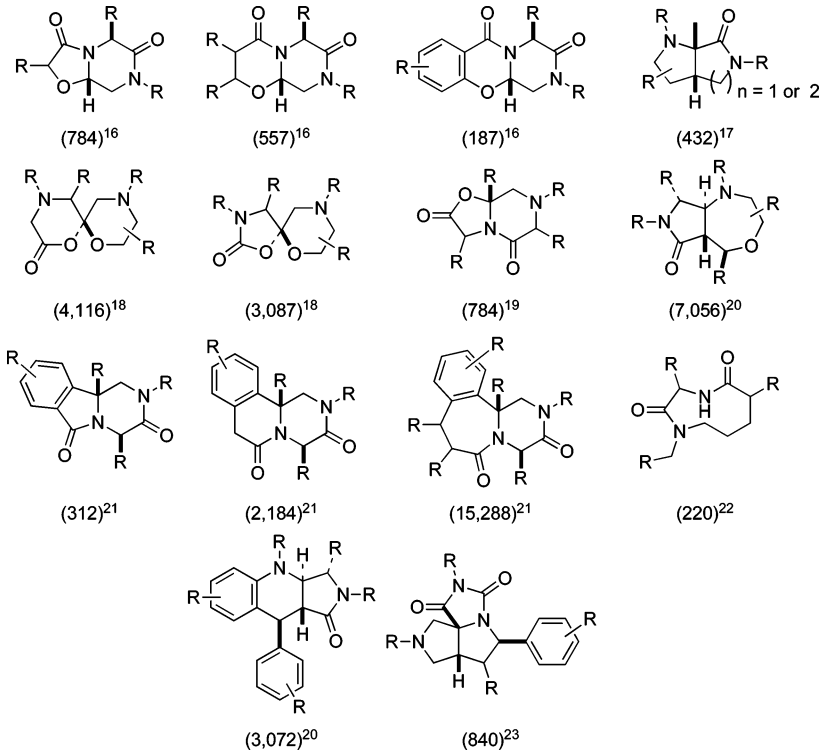


Figure 3. Core Structures in the GEMINI Database. The number of structures represented by each chemotype and reference to the synthesis are indicated.

scaffolds, the core structures may be challenging synthetically, especially with the desired substituents. Thus, search hits are not likely to be used per se but point to frameworks that, with appropriate heteroatom substitution, might be useful as a starting point for design. In this light, a diverse 3D database of readily accessible molecular frameworks would be highly desirable.

The large body of literature on combinatorial chemistry provides many densely substituted core scaffolds for which facile and robust synthetic routes are available. Thus, an additional database, dubbed GEMINI, was constructed that sampled scaffolds from chemistries developed by the Berkeley group and from published combinatorial syntheses. For each scaffold, an approach similar to that described above was used to build a virtual array that placed hydrogen, ethyl, vinyl, and phenyl groups at the positions of the scaffold where the validated chemistry suggested that substitution would be facile. GEMINI contains the 14 cores in Figure 3, represented by 38 219 molecules in 323 533 different conformations. Because of the low computational requirements for constructing and processing the database, additional cores from the literature could be added quickly.

The TRISUB, BIAD, and GEMINI databases are now being used routinely, in addition to the other available databases, in CAVEAT applications. The TRISUB database has provided interesting new leads in molecular design that are currently being developed at Stony Brook. Although the smaller and less-diverse BIAD database does not consistently provide hits in CAVEAT searches, it is likely to be valuable in certain applications. GEMINI continues to be attractive as a source of readily accessed frameworks and as a model for expandable or proprietary CAVEAT databases, as illustrated above. In addition to the value of these databases themselves, this work provides general methods that may be valuable in the development of additional databases for use with CAVEAT and for other applications based on the variable substitution of simple core structures.[24]

ACKNOWLEDGMENT

The work at Stony Brook was supported by National Science Foundation Grant CHE0213457; we also thank Victor Hornak (Stony Brook Center for Structural Biology) and Zhenkai Liang (Stony Brook Department of Computer Science) for assistance. At Berkeley, K.Y. was supported by a fellowship from the National Science Foundation. The Center for New Directions in Organic Synthesis is supported by Bristol-Myers Squibb as a sponsoring member and Novartis Pharma as a supporting member.

Supporting Information Available: UNIX scripts used in database construction and processing. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Lauri, G.; Bartlett, P. A. CAVEAT: A program to facilitate the design of organic molecules, *J. Comput.-Aided Mol. Des.* **1994**, *8*, 51−66.
(2) (a) CAVEAT: A Program to Facilitate the Design of Organic Molecules. http://www.cchem.berkeley.edu/∼pabgrp/Data/caveat.html. (b) TRIAD and ILIAD: Comprehensive 3-Dimensional Databases of Computed Structures. http://www.cchem.berkeley.edu/∼pabgrp/Data/TandI.html
(3) Bartlett, P. A.; Etzkorn, F. A.; Guo, T.; Lauri, G.; Liu, K.; Lipton, M.; Morgan, B. P.; Shea, G. T. Intuitive- and Computer-Assisted Approaches to the Design of Conformationally Restrained Peptides and Their Mimics. In *Chemistry at the Frontiers of Medicine, Proceedings of the Robert A. Welch Foundation Conference on Chemical Research XXXV*; Houston, TX, 1992; pp 45−68.
(4) Weiss, G. A.; Collins, E. J.; Garboczi, D. N.; Wiley, D. C.; Schreiber, S. L. A Tricyclic Ring-System Replaces the Variable Regions of Peptides Presented by 3 Alleles of Human MHC Class-I Molecules. *Chem. Biol.* **1995**, *2*, 401−407.
(5) Sefler, A. M.; Kozlowski, M. C.; Guo, T.; Bartlett, P. A. Design, Synthesis, and Evaluation of a Depsipeptide Mimic of Tendamistat. *J. Org. Chem.* **1997**, *62*, 93−102.
(6) Smith, W. W.; Bartlett, P. A. Macrocyclic Inhibitors of Penicillopepsin. 3. Design, Synthesis, and Evaluation of an Inhibitor Bridged Between P2 and P1′. *J. Am. Chem. Soc.* **1998**, *120*, 4622−4628.
(7) Artis, D. R.; Brotherton-Pleiss, C.; Pease, J. H. B.; Lin, C. J.; Ferla, S. W.; Newman, S. R.; Bhakta, S.; Ostrlich, H.; Jarnagin, K. Structure-Based Design of Six Novel Classes of Nonpeptide Antagonists of the Bradykinin B_2 Receptor. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 2421−2425.
(8) Takano, Y.; Koizumi, M.; Takarada, R.; Kamimura, M. T.; Czerminski, R.; Koike, T. Computer-aided Design of a Factor Xa Inhibitor by Using MCSS Functionality Maps and a CAVEAT Linker Search. *J. Mol. Graphics Modell.* **2003**, *22*, 105−114.
(9) Yang, W.; He, H.; Drueckhammer, D. G. Computer-Guided Design in Molecular Recognition: Design and Synthesis of a Glucopyranose Receptor. *Angew. Chem., Int. Ed.* **2001**, *40*, 1714−1718.
(10) Kozlowski, M. C.; Panda, M. Computer-aided design of chiral ligands Part I. Database search methods to identify chiral ligand types for asymmetric reactions. *J. Mol. Graphics Modell.* **2002**, *20*, 399−409.
(11) Kozlowski, M. C.; Waters, S. P.; Skudlarek, J. W.; Evans, C. A. Computer-Aided Design of Chiral Ligands. Part III. A Novel Ligand for Asymmetric Allylation Designed Using Computational Techniques. *Org. Lett.* **2002**, *4*, 4391−4393.
(12) Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **2002**, *6*, 439−446.
(13) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002**, *7*, 1047−1055.
(14) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813−1818.
(15) Cambridge Structural Database (CSD). http://www.ccdc.cam.ac.uk/products/csd/.
(16) Smith, L. R.; Bartlett, P. A. Novel, amino-acid derived heterobicycles: scaffolds for β-turn mimics and targets for combinatorial synthesis. *Mol. Online* **1998**, *2*, 58−62.
(17) Marx, M. A.; Grillot, A.; Louer, C. T.; Beaver, K. A.; Bartlett, P. A. Synthetic design for combinatorial chemistry. Solution and polymer-supported synthesis of polycyclic lactams by intramolecular cyclization of azomethine ylides. *J. Am. Chem. Soc.* **1997**, *119*, 6153−6167.
(18) Trump, R. P.; Bartlett, P. A. Amino acid-derived heterocycles as combinatorial library targets: Spirocyclic ketal lactones. *J. Comb. Chem.* **2003**, *5*, 285−291.
(19) Lewis, J. G.; Bartlett, P. A. Amino acid-derived heterocycles as combinatorial library targets: Bicyclic aminal lactones. *J. Comb. Chem.* **2003**, *5*, 278−284.
(20) Spaller, M. R.; Thielemann, W.; Brennan, P. E.; Bartlett, P. A. Combinatorial Synthetic Design. Solution and Polymer-Supported Synthesis of Heterocycles via Intramolecular Aza-Diels−Alder and Iminoalcohol Cyclizations. *J. Comb. Chem.* **2002**, *4*, 516−522.
(21) Todd, M. H.; Ndubaku, C.; Bartlett, P. A. Amino acid derived heterocycles: Lewis acid catalyzed and radical cyclizations from peptide acetals. *J. Org. Chem.* **2002**, *67*, 3985−3988.
(22) Virgilio, A. A.; Bray, A. A.; Zhang, W.; Trinh, L.; Snyder, M.; Morrissey, M. M.; Ellman, J. A. Synthesis and evaluation of a library of peptidomimetics based upon the β-turn. *Tetrahedon* **1997**, *53*, 6635−6644.
(23) Peng, G.; Sohn, A.; Gallop, M. A. Stereoselective solid-phase synthesis of a triaza tricyclic ring system: A new chemotype for lead discovery. *J. Org. Chem.* **1999**, *64*, 8342−8349.
(24) For further information on the databases described here, contact Dale Drueckhammer for TRISUB or BIAD or Paul Bartlett (paul_bartlett@berkeley.edu) for GEMINI. For TRIAD and ILIAD, see ref 2.

CI050277O