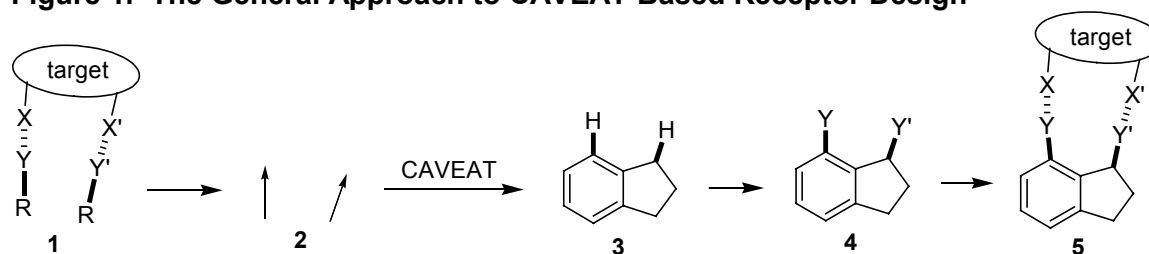


Quantum Mechanical Optimization of Virtual Molecular Databases

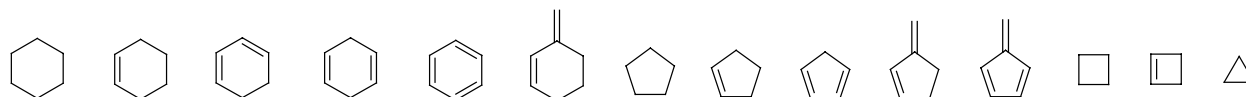
The primary area of research in my group is focused on the use of the computer program CAVEAT¹⁻⁴ in the design of specific receptors for molecules and ions by the general approach outlined in figure 1. Computational methods are used to model a complex **1**, in which Y and Y' represent functional groups that bind to complementary functional groups X and X' of the target molecule. The bonds in bold are then used to define a pair of vectors **2**, with the starting point of each vector corresponding to the atoms R and R' and the direction of each vector corresponding to the direction of the bond to Y or Y'. CAVEAT is used to search a database to identify structures having a pair of bonds matching the vector pair. For the vector pair **2**, structure **3** might be identified from a database with the C-H bonds in bold matching the vector pair. The structure **4** incorporating the Y and Y' groups into the bonds that match the vector pair thus has the groups in the proper position and orientation to form the complex **5** with the target molecule. This design approach was first demonstrated in the development of a selective receptor and sensor for glucose.⁵

Figure 1. The General Approach to CAVEAT-Based Receptor Design



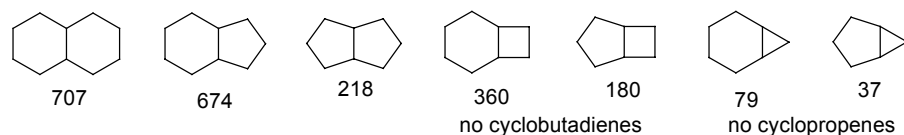
The early stages of this project used almost exclusively the TRIAD database of computer generated tricyclic hydrocarbons developed by Bartlett and coworkers specifically for use with CAVEAT.^{1,2} More recently, two new databases for use with CAVEAT have been developed in this group.⁶ Initially a library was created of trisubstituted monocyclic hydrocarbons having three of the hydrogens replaced with ethyl, vinyl, or phenyl substituents. The substituents provide additional bonds for matching the vectors defined in a CAVEAT search. The hydrocarbon core structures used are shown in figure 2. The analog builder feature of the Cerius2 program suite (Accelrys) was used to generate libraries of multisubstituted structures and a Unix script was written to select the structures in which only three of the positions were replaced with non-hydrogen substituents. Conformational searching and molecular mechanics minimization was conducted and all unique conformations within 5 kcal/mol of the minimum energy conformation of each compound were retained. The groups of elaborated ring systems generated by this approach were combined and processed into the vector database for use with CAVEAT. This database named TRISUB contains a total of about 757,000 structures and has become the primary database used in this project.

Figure 2. Core Structures in the TRISUB Database



An additional database of unsubstituted bicyclic hydrocarbons was developed based on seven ring systems shown in figure 3.⁶ For each ring system, structures having every possible pattern of unsaturation were entered manually into the Cerius2 interface. Structures were included containing one or two exocyclic double bonds (methylene groups) in addition to

Figure 3. Core Structures in the BIAD Database.



intracyclic double bonds. The total number of structures based on each ring system, including stereoisomers, are also given in figure 3. In total, 2,255 structures were obtained. These structures were also subjected to a conformational search, energy minimization, and selection of structures within 5 kcal/mol of the lowest energy structure. This new database has been tentatively named BIAD. The development of these two databases along with a database GEMINI developed by Paul Bartlett's group, our collaborator at UC Berkeley, was recently published.⁶

These newly developed databases have contributed greatly to this project such that the accessibility of reasonably simple and synthetically accessible structures using CAVEAT is no longer the primary limitation of this project. These databases have been the source of structures in more recent design efforts from this group.⁷⁻¹⁰ The greater database limitation is now the limited accuracy of the 3-dimensional structures, which have all been calculated using molecular mechanics in MacroModel. While molecular mechanics might be expected to give reasonably accurate structures for the simple hydrocarbons these databases and this is generally true, there are limitations. One observation in our work has been the clear inaccuracy of conformations about single bonds between sp^2 and sp^3 hybridized carbon atoms, which are very prevalent in our databases. Furthermore, while a 0.3 Å error in distance between a pair of structurally distant atoms may seem fairly small, such error may be very significant in design efforts. As a result of such errors, many initial hits turn out to not be good hits upon further scrutiny by higher-level computational modeling. Perhaps more importantly, many potentially good hits may not be identified at all in the CAVEAT search if the structural error is greater than the error range defined in the search.

This proposal for use of the Seawulf cluster aims to overcome the database limitations described above by reoptimizing the structures in the BIAD and TRISUB databases using quantum calculations. While semiempirical calculations would offer some improvement, the initial plan is to perform ab initio calculations at the HF/3-21G level. In several selected examples, this level gives results essentially indistinguishable from structures calculated at the higher B3LYP/6-31G(d) level but with significantly less computer time. Ab initio calculations on these hundreds of thousands of structures would clearly not be feasible on the single or dual-processor machines used in the development of these databases but seem feasible using the Seawulf cluster. Initially the relatively small BIAD database (2,255 distinct compounds with multiple conformations of many) will be processed at the HF/3-21G level. This is expected to require roughly 1,000 CPU hours, as representative structures have been optimized in <30 minutes on a 2.0 GHz PC. The output file from MacroModel in the development of this database will serve as the source of the structural files.⁶ The MaxCycle option in Gaussian or equivalent command in the program employed will be used to assure that a single problematic structure does not become a bottleneck. The BIAD refinement will serve as a test case for developing the process.

Once the BIAD refinement has been completed and the results properly evaluated, the refinement of the TRISUB database will be pursued. A decision will first be made of whether this much larger database can indeed be processed at the HF/3-21G level given the amount of computer time available. It is likely that the time required will be in excess of 100,000 CPU hours to process the entire database, though the optimization of the comparatively small BIAD database should give a better indication. If the HF/3-21G optimization of TRISUB is deemed to

not be feasible, a decision will be made of a semiempirical method by comparing the accuracy of several representative structures calculated using AM1, PM3, and other available semiempirical methods. As necessary, the work will be performed in installments by optimizing a portion of the structures using increments of allotted processor time. Even if it is only possible to initially optimize about 10% of the total structures, the database resulting from this subset of the larger databases could in itself be very useful for further work using CAVEAT. A sampling of structures will be evaluated after each installment to evaluate the results, though relative to most computational projects this work will require minimal data analysis along the way.

A revised application for renewal of a previous grant from the National Science Foundation supporting our overall CAVEAT project is due Jan. 12, 2007. The database refinement proposed here will be a part of that application. Thus, I would like to have approval of at least some initial allotment of computer time before this deadline so that documentation of this approval can be included with the NSF application. The grant would not start until at least July 2007, though we may be able to begin the computational work in the Spring if computer time is available. I have extensive experience in computational chemistry performed on a PC, while my students and I have some experience processing large molecular databases in UNIX systems as demonstrated by our database development work described in reference 6. Still, this work does not represent my group's primary expertise and we will probably need to solicit some assistance either from the support staff of the Seawulf cluster or from other connections that we have used in the past.

References

1. Lauri, G.; Bartlett, P. A. "CAVEAT: A program to facilitate the design of organic molecules" *J. Computer-Aided Molecular Design* 1994, 8, 51-66.
2. <http://www.cchem.berkeley.edu/%7Epabgrp/Data/caveat.html>
3. Bartlett, P. A.; Shea, G. T.; Telfer, S. J.; Waterman, S. "CAVEAT: A Program to Facilitate the Structure-derived Design of Biologically Active Molecules" In Roberts, S. M. (Ed.) *Molecular Recognition: Chemical and Biological Problems*, Royal Society of Chemistry, London, 1989, pp. 182-196.
4. 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.
5. 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.
6. Yang, Y.; Nesterenko, D.; Trump, R. P.; Yamaguchi, K.; Bartlett, P. A.; Drueckhammer, D. G. "Virtual Hydrocarbon and Combinatorial Databases for Use with CAVEAT", *J. Chem. Information Modelling*, 2005, 45, 1820-1823.
7. Huang, H.; Drueckhammer, D. G. "A Computer-Designed Macrocyclic Zinc Receptor" *Chem. Commun.*, 2005, 5196-5198.
8. Zhu, Y.; Drueckhammer, D. G. "Transition State Modeling and Catalyst Design for Hydrogen Bond-Stabilized Enolate Formation" *J. Org. Chem.*, 2005, 70, 7755-7760.
9. Huang, H.; Drueckhammer, D. G. "A Modular Molecular Tweezer Designed using CAVEAT" *Chem. Commun.*, 2006, 2995-2997.
10. Lin, C.; Drueckhammer, D. G. "Computer-guided design of a Cu(II) receptor and sensor" *New. J. Chem.*, 2006, 30, 1725-1730.