Structure-Based Design and In Silico Virtual Screening of Combinatorial Libraries. A Combined Chemical– Computational Project

1 AUTHOR:

Jerome Baudry
University of Tennessee
60 PUBLICATIONS   **1,731** CITATIONS

Structure-Based Design and In Silico Virtual Screening of Combinatorial Libraries

A Combined Chemical–Computational Project

W

Jerome Baudry*

School of Chemical Sciences, University of Illinois at Urbana–Champaign, Urbana, IL 61801; *jerome@scs.uiuc.edu

Paul J. Hergenrother

Department of Chemistry, University of Illinois at Urbana–Champaign, Urbana, IL 61801

Combinatorial chemistry is a fast-expanding field of chemistry, and laboratory experiments directed toward chemistry students have been recently reported in the literature. For example, liquid-phase combinatorial synthesis of compounds in the organic chemistry laboratory have been described, coupled with simple and elegant activity assays through visual inspection of insect repellency (1), in vitro antibiotic activity (2), or identification of the product's odor (3). In addition, a solid-phase synthesis laboratory experiment for the combinatorial preparation of oligopeptides has also been recently described (4). These laboratory experiments have introduced advanced students to the important aspects of industrial and academic combinatorial chemical synthesis. For obvious reasons of practicability, these previous experiments were aimed at generating a low number of product compounds created following a predefined synthesis strategy that had been optimized by the instructors.

A particularly successful and interesting application of combinatorial chemistry is in the field of drug design. Medicinal chemists often design and synthesize combinatorial libraries of several hundreds of compounds that are focused towards a particular property. Examples include similarity with (or diversity from) a known active structure or therapeutic class of compounds or searching for binding to a particular biological target. In order to screen large, sometimes massive, libraries and assess their potential affinity for a particular target, in silico (computational) high-throughput approaches are employed to dock and rank virtual libraries of compounds. Several successful applications of this computational approach have been described, and these allow the analysis of a combinatorial design prior to synthesis. This in silico screening has obvious advantages in terms of time gain and cost reduction (see ref 5 for an application of this approach and refs 6 and 7 for reviews).

The present article describes a laboratory project, part of the advanced undergraduate and graduate chemistry course "Combinatorial Chemistry" at the University of Illinois at Urbana–Champaign. The project aims to introduce students to several aspects of combinatorial chemistry through a combined and multidisciplinary chemical–computational approach. The goal of the project is to: (i) have the students design combinatorial libraries biased toward binding a specific protein target, (ii) have the students propose a specific synthetic pathway for their library, and (iii) have the students enumerate the hundreds of library members and evaluate them for their binding affinity to their protein target.

Description of the Class and the Role of the Computational Project

Combinatorial Chemistry is a one-semester course designed to familiarize students with aspects of modern combinatorial chemistry and combinatorial biology. Both advanced undergraduate and graduate students can take the course; in practice, the majority of students in the course are first- or second-year chemistry graduate students. This course is taught twice-a-week for 80 minutes each session for a total of 31 class periods. Although textbooks are recommended for the students (8, 9), most of the material is taken from the primary literature or topical reviews, and the examples are updated yearly. The course begins with a brief history of drug discovery, highlighting the events that led to the widespread usage of combinatorial chemistry in the 1990s. A series of lectures then covers the basic aspects of combinatorial chemistry, which includes: solid-phase synthesis, beads and linkers, split pool versus parallel synthesis, solution-phase libraries, focused versus diverse libraries, encoding a combinatorial library, and dynamic combinatorial chemistry. Next, screening of combinatorial libraries is overviewed and then presented in the context of tools for studying protein function. This segment covers classic genetics, RNA interference, and chemical genetics, where the compounds are created through combinatorial chemistry. At this point in the course there is an examination.

The second half of the course begins with various combinatorial approaches to catalyst discovery and development, and then the class segues into combinatorial biology. The combinatorial biology portion of the course begins with a discussion of peptide and protein display techniques, with an emphasis on phage display. Catalytic antibodies, catalytic nucleic acids, and RNA aptamers are then discussed, followed by a lecture on the yeast two- and three-hybrid systems, with a special emphasis on their applicability to compound screening. The final segment of the combinatorial biology portion describes various in vitro and in vivo techniques for the evolution of enzymatic activity and protein function. The class then finishes with lectures on polymeric combinatorial libraries and applications of combinatorial chemistry to materials, inorganic chemistry, and molecular recognition. During finals week there is a three-hour comprehensive examination.

In addition to the two examinations, students are evaluated based on their performance on two projects. The first project entails writing a report on a combinatorial library that

has been published in the literature. The library is chosen by the students and approved by the instructor; libraries appropriate for this project incorporate design, synthesis, and evaluation elements. The second project is described in this manuscript: the design, in silico synthesis, and screening of a combinatorial library against a protein target. The target is chosen by the students, and they are limited only by the presence of a protein in the Protein Data Base (PDB). By the time of the project (~2/3 of the way through the course), course lectures had covered several examples of proteins that could be potential targets for small molecule binding. In addition, many students are already working in a research lab and could choose a protein target relevant to their research labs' programs. Although the course instructors could propose possible protein targets to students who could not choose their own target, this possibility did not arise. This project was utilized in both the fall of 2002 and the fall of 2003 course offerings. Between these two classes, 40 students took part in the project: 39 graduate students and 1 undergraduate student. In general, their level of computational knowledge was low going into the course. Thirty-seven of the graduate students where members of the chemistry department, whereas the other 2 graduate students were enrolled in the biophysics and computational biology program.

Flowchart of the Project

The flow of the project is described in Figure 1. In the initial part of the project, which takes place outside of the computer laboratory, the students select a biological target of interest, perform a literature search on the biological or therapeutic role of the target, and identify in the literature (or from the 3D structure of the target, see below) potential compounds on which to base the initial library design. The students are instructed to propose possible synthetic pathways based on knowledge gained from in-class discussions of combinatorial chemistry. After this preliminary work, the students perform the computational part of the project: they use computational chemistry software to enumerate their library, dock it in their target, analyze the results to identify the most interesting hits, and look for trends that indicate the presence of particularly active R groups. These points are described in details below.

Project Protocol

Initial Demonstration

A two-hour demonstration of the programs and of the protocol is given to the students in the computer lab.[1] In this session the instructors, using a library designed for docking in a particular target, operate the software and project their computer's monitor on a large, wall-mounted screen for all the students to follow. The process is detailed in a step-by-step fashion, with comments on every aspect of the process. Fully annotated hard copies of the protocol are provided to the students with each step clearly outlined. The students can access the computer lab at all hours during the following weeks to work on their project, and class hours are set up to solve potential technical problems or answer questions about usage of the software. The students are given six weeks
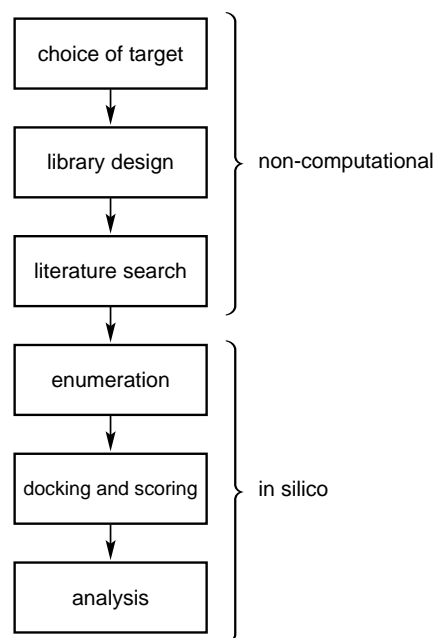

Figure 1. General flowchart of the project.

to complete the project. In our experience, the first few students working on their projects require frequent help from the instructors. After the first couple of weeks and the completion of their projects, the first few students were willing and able to help their fellow students. Little intervention from the instructors was required.

Choice of Protein Target, Literature Search, and Library Design

The choice of a protein target is left entirely to the students. The only restriction is that a 3-dimensional structure of the target exists and is deposited in the PDB (6). Ideally, a known ligand (inhibitor or activator) has also been described, either in the form of a co-crystallized molecule in the PDB structure or from the literature. The students are asked to justify why they choose a particular target, both in terms of medicinal interest and the appropriateness of a combinatorial optimization strategy.

The design of the combinatorial library by the students is based on the analysis of the functional groups of known ligand(s), as well as on mutagenesis studies described in the literature, when available. The students are asked to describe the synthetic scheme they propose for completion of the library, based on the classroom lectures. To adapt to the available computer power and time frame given to complete the project, it was recommended that the number of final compounds in the designed library not exceed 1000 to 2000 compounds (5 seconds/ligand: ~3 hours of calculations for the docking of a ~2000 compounds library).

Preparation of the Protein Target

In the first stage of the computational part of the project, the students access the PDB and download the 3D structure

of the protein. They use a standard Web browser to access and search the PDB and save the coordinates of the target on their computer's hard drive.

Often, PDB structures contain more than one copy of the target in the crystal unit cell. The students used the program MOE *(10)* to prepare the protein structure for subsequent screening. This requires the deletion of all but one of the copies of the protein, extraction of the ligand from the protein, removal of ions or water molecules that may compete with compound binding and addition of hydrogen atoms to the protein and the ligand. The final protonated structure of the protein target is saved in PDB format from MOE on a computer disk.

Enumeration

The program MOE is used to create and enumerate the library in silico. The students draw their scaffold (or scaffolds if they choose to investigate more than one), and define the reagents' attachment points on the scaffold(s). The structure and the attachment points of the scaffold(s) are saved in the MOE proprietary database. The students draw in MOE the structure of each of the R groups that will be attached to the scaffold and define the attachment point on the R groups. As for the scaffold(s), each list of R groups is stored in a MOE database that contains all the individual compounds and their attachment points. The program MOE automatically enumerates the library (i.e., attach the R groups to the scaffold in a combinatorial fashion) and produces a log file that the students are invited to examine to verify the successful completion of the enumeration. A library database is created that contains the enumerated products. A short energy-minimization of all the products is performed to obtain 3D structures without steric clashes. The final library is saved on disk as a file with the "SD" format, containing the minimized 3D structures of all enumerated compounds. The students, in addition to the designed library, were instructed to include in their docking a set of 244 "neutral" compounds that are low molecular weight (less that 220 g/mol), very hydrophobic, and not optimized for interaction with any particular receptor's functional group. This neutral library will serve as a benchmark against which the designed library will be evaluated: if the proposed library is indeed correctly designed to optimize binding in the ligand binding pocket, the compounds of the neutral library will score poorer relative to the designed library. In addition to these two libraries (designed and neutral), the students are also asked to dock the known ligand found in the crystal structure or other relevant ligands that they find in the literature; referred to as the natural ligand(s). This allows the assessment of the designed library relative to the known ligand(s), as the known ligand is expected to score better than most of the compounds in the designed library. Any compound of the designed library that is found to score better than or close to the score of the natural ligand suggests a particularly interesting molecule and set of R groups. Conversely, the use of neutral library and natural compounds also helps the students to identify problems in the process. For instance, a poor-scoring natural ligand or high-scoring neutral library suggests flawed computational parameters, allowing the students to identify the possible

problem(s) and discuss it in their report, as well as suggest possible remedies.

Docking and Scoring

The docking and scoring of the enumerated library is performed using the LigandFit *(11)* module within the Cerius2 program *(12)*. LigandFit is a widely used program in industrial and academic research groups. The program employs a cavity-detection algorithm combined with a fast ligand-conformational search engine that allows the optimization of the bond structure of a ligand in a protein cavity. The speed of the process allows for fast processing of each ligand, typically 5 seconds per ligand on the SGI octane workstations used in the project (limiting the number of docked conformations to 1000 per molecule). Using this program, large libraries can be screened during a few hours of computer time. For each ligand, a binding score (ligscore) is calculated that correlates with the interaction energy between each ligand and its protein environment. In practice, the students define the binding site from the co-crystallized ligand or from important protein residues found in the literature. The docking program processes the file containing the structures of the designed library, of the neutral library, and of the natural ligand(s). The best-scoring binding mode for each of the docked molecules is saved to disk. Each molecule has an index number that allows for its rapid identification in the analysis process.

Analysis of the Results

The compounds are sorted according to their calculated ligscore values. This ranks the compounds so that a rank of one corresponds to the best-scoring compound, a rank of two is the next best-ranking compound, and so forth. The analysis of the results of the virtual docking is performed by plotting the rank versus the calculated ligscore for all of the docked compounds. As described above, the designed library is considered interesting if some of its compounds are ranking better than the natural ligand, and if most of its compounds are ranking better than the neutral library. A plot of score versus rank for a typical interesting library, reproduced from a student's report is shown in Figure 2 (see full report in Supplementary Material[W]). In Figure 2, 39 compounds out of a 1024-compound library targeted against a cyclin-dependant kinase protein (CDK 2) were found to rank better than a known inhibitor for this target, the natural ligand. Most of the neutral library compounds were ranked between 350 and 900. Several compounds of the designed library did not rank as well as the neutral library compounds. This is often due to steric repulsions between the ligands and the protein in relatively small binding sites: when large, high molecular weight compounds from the designed library are too large to fit into the cavity, more and more steric clashes are found and penalize the docking score of these compounds (the protein's structure is kept rigid in the docking experiment, and therefore the binding site's geometry cannot adapt itself to compounds larger that the binding cavity). This effect is illustrated in Figure 3, reproduced from the same student's report. Figure 3 plots the rank versus the molecular weight of the docked compounds (a measure of how large

and bulky they are). While the best-ranking compounds do not show any particular correlation between the rank and the weight, compounds from the designed library that ranked below ~850 display a correlation between their rank and their molecular weight, indicating that compounds that are larger obtain less and less favorable binding scores and ranks. The compounds of the neutral library, all being of small molecular weight, show no correlation whatsoever between their rank and their molecular weight. Figure 3 also shows that best-ranking compounds from the designed library have a molecular weight that is significantly higher than compounds in the neutral library, showing that best ranking compounds are scoring well not just because they are small and avoid steric clashes, but because they indeed feature the chemical characteristics needed to achieve a good binding with their target.

From the results presented in Figure 2, an R group analysis of the best ranking compounds is performed. Students identify the R groups that are consistently present in the best-ranking compounds. Students conclude with the proposal for a smaller, focused library optimized for binding.

Discussion of the Project and Future Directions

Overall, the students successfully completed the project. They were able to perform the computational tasks and critically analyzed their final results. Comments by the students indicated that the multiple software manipulations were the weak link of the process: several students realized during the final analysis that there were problems with their calculations and had to start the process over to achieve successful docking and analysis. However, this multi-programs approach allowed the students to gain experience on more than one software platform, similar to what they will experience in a academic or industrial research setting.

The most common problems, often successfully identified by the students themselves, were either: (i) problems with the preparation of the protein (charges not attributed, presence of ions or crystallographic water molecules in the binding site competing with ligand binding), or (ii) incorrectly defined attachment points in their R group lists (resulting in non-correct enumeration of the corresponding compounds in the library). Most of these problems could have been identified prior to the time-consuming enumeration–minimization–docking processes by examining carefully the different log files created by the programs. While the students were invited during the initial two-hour demonstration to pay attention to these log files, not all did so when performing their own first screening work. In the future, the students will be asked to report explicitly on the content of these log files, minimizing the risk of potential enumeration problems. A few students chose to design and screen combinatorial libraries of several thousands compounds, which led to long calculation times and made it impossible to repeat the experiment in the event of a problem. In the future, the students will be asked to calculate the predicted time needed to dock their library before they perform the actual docking calculations, which should provide for an "early warning" if the proposed libraries are too large.

Conclusion

This combined chemical–computational project has introduced the students to the computational tools and approaches used in the design of combinatorial libraries in drug-discovery research. The students successfully manipulated several software programs and critically analyzed the results of the computations. They were able to identify the most promising R groups in their design and propose focused combinatorial libraries optimized for binding.

ᵂSupplemental Material

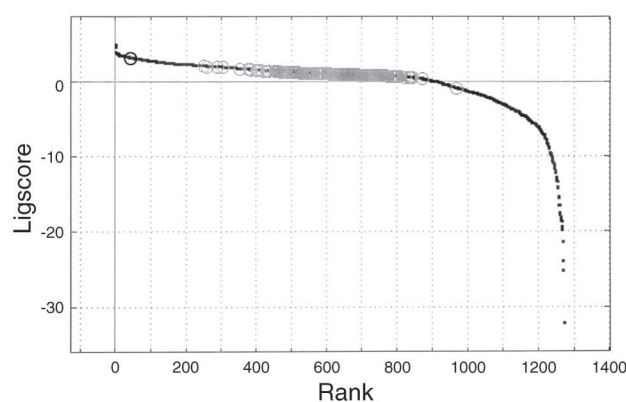A student's report is available in this issue of *JCE Online*.

Figure 2. Plot of the compound number versus the ligscore. The compound numbers are assigned based on the ligscore rank, with lower numbers indicating higher ligscore. The compound with the black circle indicates the known inhibitor and the compounds with gray circles correspond to compounds of the neutral library. Lower numbers correspond to better-docking compounds with higher ligscores.
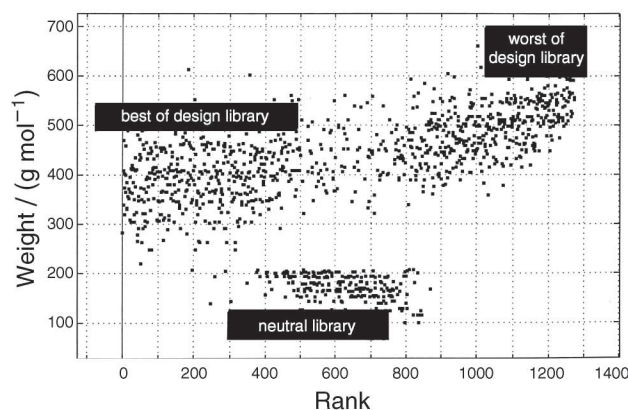

Figure 3. Plot of the compound number versus the molecular weight of the docked ligands.

ation facility.In the Classroom

Note

1. See the Computing Home Page, *http://vizlab.scs.uiuc.edu* (accessed Mar 2005), for a description of the visualization facility.

Literature Cited

1. Miles, W. H.; Gelato, K. A.; Pompizzi, K. M.; Scarbinsky, A. M.; Albrecht, B. K.; Reynolds, E. R. *J. Chem. Educ.* **2001,** *78,* 540–542.
2. Wolkenberg, S. E.; Su, A. I. *J. Chem. Educ.* **2001,** *78,* 784–785.
3. Birney, D. M.; Starnes, S. D. *J. Chem. Educ.* **1999,** *76,* 1560–1561.
4. Truran, G. A.; Aiken, K. S.; Fleming, T. R.; Webb, T. R.; Markgraf, J. H. *J. Chem. Educ.* **2002,** *79,* 85–86.
5. Aronov, A. M.; Munagala, N. R.; Kuntz, I. D.; Wang, C. C. *Antimicrob. Agents Chemother.* **2001,** *45,* 2571–2576.
6. Joseph-McCarthy, D. *Current Drug Discovery* **2002,** *March,* 20–23. *http://www.currentdrugdiscovery.com/2002/march.html* (accessed Feb 2005).
7. Bajorath, J. *Current Drug Discovery* **2002,** *March,* 24–28. *http://www.currentdrugdiscovery.com/2002/march.html* (accessed Feb 2005).
8. Seneci, P. *Solid-Phase Synthesis and Combinatorial Technologies,* 1st ed.; John Wiley & Sons: New York, 2000.
9. Beck-Sickinger, A.; Weber, P. *Combinatorial Strategies in Biology and Chemistry,* 1st ed.; John Wiley & Sons: New York, 2001.
10. MOE (The Molecular Operating Environment), Version 2003.02. Chemical Computing Group Inc., Montréal, Canada.
11. Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. *J. Mol. Graphics. Model.* **2003,** *21,* 289–307.
12. Cerius2. Version 4.9. Accelrys, Inc., San Diego, CA. *http://www.accelrys.com/cerius2/* (accessed Apr 2005).

894 Journal of Chemical Education • Vol. 82 No. 6 June 2005 • www.JCE.DivCHED.org