

Optimization of combinatorial library design

Merck Frosst researchers develop and test new algorithm for fast and effective filtering of reagent sets

“This program reduces or eliminates the time a medicinal chemist spends examining reagents which a priori cannot be part of a ‘good’ library.”

Christopher I. Bayly (left) and Jean-François Truchon (right) in the library at Merck Frosst Canada & Co. in Kirkland, Québec

MDL[®] Available Chemicals Directory (MDL ACD) is an indispensable, easy-to-access, structure-searchable database for sourcing chemicals. Like all MDL content, it supports a variety of research activities. For example, ACD has played an important role in an innovative approach to optimizing the virtual screening of reagents in combinatorial library design.

Jean-François Truchon and Christopher I. Bayly of Merck Frosst Canada & Co. have recently published a paper¹ describing a novel, robust computer algorithm for optimizing reagent lists used in the design of combinatorial libraries for matrix synthesis—and the database they queried in selecting the reagents for their program was MDL Available Chemicals Directory.

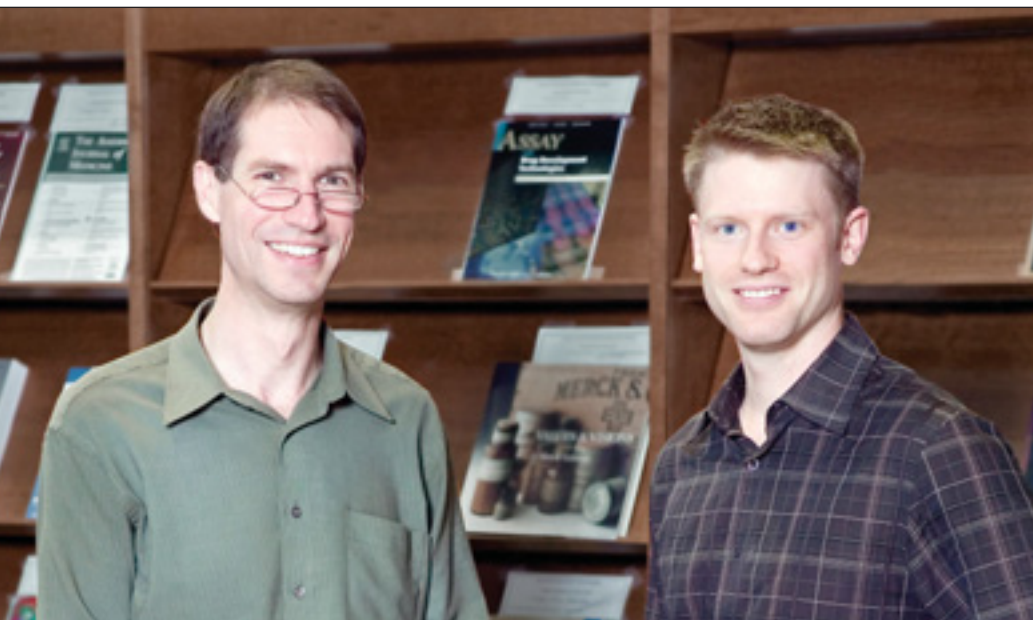
According to Truchon and Bayly, high-throughput synthesis and screening of large numbers of compounds have opened up unprecedented possibilities for drug development. In the introduction to their

paper, they state: “Although thousands of compounds can be synthesized rapidly with automation technology, scientists have soon realized that it is not sufficient to blindly generate many compounds for screening. Rather, a rational evaluation of the combinatorial library is desirable in order to maximize the outcome of an expensive synthesis campaign. This can be achieved by paying careful attention to both the design of the synthetic scheme and the reagent selection. In this latter task, a great amount of choice is offered: for example, the MDL Available Chemicals Directory provides chemists with a way to find a supplier for any of more than 510,000 chemical substances representing over 1.4 million individual chemical products from 685 suppliers. If a combinatorial library is made with two or more reagent functionalities, between 100,000 and many billions of different products are accessible. The ACD offers from 145 to about 6,000 reactants per functional group, whereas around 20 need to be chosen in a typical case to carry forward the synthesis. Therefore, between 86% and 99.7% of the reagents available would need to be pruned from the original sets!”

The challenge grows even more daunting when one considers the importance of product properties in the pruning process. “The idea that reagent selection is dependent upon the library template and should therefore be done according to the products seems intuitive. With millions or billions of candidate products, optimal selection becomes intractable by

photographer: Kevin Clark

¹ Truchon, Jean-François; Bayly, Christopher I.; “GLARE: A new approach for filtering large reagent lists in combinatorial library design using product properties.” *J. Chem. Inf. Model.* 2006, 46 (4), pp. 1536-1548.



humans; computer-assisted selection becomes essential," say Truchon and Bayly.

In support of this approach, their paper presents "a novel computer algorithm, called GLARE (Global Library Assessment of REagents), that addresses the issue of optimal reagent selection in combinatorial library design. This program reduces or eliminates the time a medicinal chemist spends examining reagents which *a priori* cannot be part of a 'good' library. Our approach takes the large reagent sets returned by standard chemical database queries and produces often considerably reduced reagent sets that are well-behaved with respect to a specific template. The pruning enforces 'goodness' constraints such as the Lipinski rule-of-five on the product properties such that any reagent selection from the resulting sets produces only 'good' products. The algorithm we implemented has three important features:

1. As opposed to genetic algorithms or other stochastic algorithms, GLARE uses a deterministic greedy procedure that smoothly filters out non-viable reagents;
2. The pruning method can be biased to produce reagent sets with balanced size, conserving proportionally more reagents in smaller sets;
3. For very large combinatorial libraries, a partitioning scheme allows libraries as large as 10^{12} to be evaluated in 0.25 s on an IBM AMD Opteron processor."

Truchon and Bayly validated and optimized their algorithm on a diverse set of 12 combinatorial libraries. They extracted the reagents for all the libraries from MDL

ACD using an in-house Web tool called the Virtual Library ToolKit (VLTK) that applied standard filters to eliminate metals and chemically incompatible functional groups. Overall, the libraries created for the study offered diverse scaffolds consisting of four cyclic and eight linear structures with 15 different kinds of reagents: alcohol, aldehydes, primary and secondary amines, sulfonyl chlorides, carboxylic acids, α -haloketones, boronic acids, aryl bromides, halides, amino acids, isocyanates, thioureas, 2-aminoethanol and isocyanides. "We believe that this set of combinatorial libraries and the structures are of general interest for algorithmic optimization validation and comparisons for the scientific community," says Truchon. "...and we are grateful to Elsevier MDL for allowing us to publish the molecular structures of the reagents in the supplementary material accompanying our paper."

The data sets

Truchon's and Bayly's paper begins by describing the datasets used to study and validate the GLARE algorithm. The authors used experimentally synthesized combinatorial libraries that had been documented in the recent literature, primarily in organic synthesis and molecular modeling journals. They also designed one hypothetical tetrapeptide library to test the algorithm on a very large dataset. The 12 libraries ranged from 10^5 to 10^{12} potential products and were built from two to four reagent basis sets each. "We believe the libraries represent real-life problems to which the GLARE algorithm can be applied," says Truchon.

"The principal objective of the algorithm was to provide a combinatorial set of virtual products that satisfied user-defined filtering rules. Another objective was to maximize the number of reagents to give the synthetic chemist as many choices as possible and to enable subsequent filtering steps."

Jean-François Truchon &
Christopher I. Bayly
Merck Frosst Canada & Co

Elsevier MDL collaboration with Matrix Scientific

Elsevier MDL and Matrix Scientific are collaborating to help researchers in pharmaceutical, agrochemical, polymer, coatings and academic research locate and acquire chemical materials with improved ease and reliability. Under the partnership, Matrix Scientific provides Elsevier MDL with regular electronic updates of product and pricing information from Matrix Scientific chemical catalogs, which facilitates timely updates to the MDL Available Chemicals Directory database.



As part of the agreement, Matrix Scientific is including MDL numbers in their printed catalogs, which Hiram S. Allen, president of Matrix Scientific, says will be a huge benefit to customers. "We are very pleased with the results we have achieved over many years of inclusion in MDL Available Chemicals Directory," says Allen. "At a recent American Chemical Society national exposition, we surveyed people about where they looked when searching for research chemicals. The number-one response was MDL Available Chemicals Directory. We look forward to a long and beneficial association with Elsevier MDL."



Hiram S. Allen, President of Matrix Scientific

The property calculations

Because they wanted to address the “high-throughput” end of the spectrum, the authors validated the algorithm with properties such as the number of hydrogen bond acceptors (HBA), the number of hydrogen bond donors (HBD), the number of non-hydrogen atoms (nonH) and the calculated log of the octanol-water partition (logP). Using the nonH property instead of the molecular weight was advantageous in being more tolerant of molecules containing bromine, chlorine or sulfur atoms which are frequently part of drugs and development candidates. The fast property evaluations obtained with GLARE were the direct result of treating many product properties in the combinatorial libraries as the sum of the reagent property plus an overall correction for the scaffold and connection to it. The paper shows that this is true for HBA, HBD, nonH and logP, but the authors state the same strategy could be applied to evaluate product prices, polar surface area and other useful properties.

The core algorithm and filtering techniques

The principal objective of the algorithm was to provide a combinatorial set of virtual products that satisfied user-defined filtering rules. Another objective was to maximize the number of reagents to give the synthetic chemist as many choices as possible and to enable subsequent filtering steps. It was felt that optimally fulfilling these two objectives would lead to a maximally effective and maximally “good” library.

The optimization itself worked through a greedy iterative algorithm that was shown to converge to good solutions. The idea was to score each reagent based on its likelihood of leading to “good” products. The worst reagents were then eliminated before the next iteration. The authors found that a goodness of 95% was largely sufficient, as it kept just a few reagents on the border of what was considered viable. However, it was also possible to reduce the goodness threshold with the aim of obtaining more reagents. The authors considered the effectiveness of the pruned library, i.e., the proportion of the initial reagent list to keep. They found it important to progressively

eliminate fewer reagents as the calculation evolved. They developed an equation relating the optimal number of reagents to keep at first iteration to the initial goodness of the library and the goodness threshold. Using this equation minimized the computational time while maintaining an optimum effectiveness. However, they also found that this pruning strategy overly penalized small reagent sets when reagent lists of large and small size were optimized together. To deal with this problem, they developed a scaled pruning scheme that eliminated reagents proportional to the list size.

The partitioning technique

The fast property calculation, iterative algorithm and pruning scheme resulted in the rapid, efficient processing of reagents, but very large virtual libraries remained intractable without the use of a partitioned sampling technique. In this approach, all the reagents from all dimensions were scored at one time through the partitioning of the base sets. Instead of performing a full combination of the reagents, the authors did a partial combination using partitions of the reagent basis sets from each dimension and attributing scores to the monomers. In contrast to random sampling, partitioning ensures that all remaining reagents are examined at each iteration, while the system performs an even lower number of property evaluations. The authors found that small partitions reduced the computational burden and still gave good libraries, but also reduced the overall effectiveness of the system. To balance this, they recommend a minimum partition size of 16 reagents.

Truchon and Bayly state that the performance of the GLARE algorithm is better than that previously obtained by other groups using similar algorithms. For example, GLARE can optimize large four dimensional libraries using multi-objective rules in under one second on affordable computers. They propose this method for general use in fast and effective filtering of reagent sets that need to conform to a given set of product properties.

Truchon and Bayly propose the GLARE algorithm for general use in fast and effective filtering of reagent sets that need to conform to a given set of product properties.



Elsevier MDL
2440 Camino Ramon,
Suite 300
San Ramon, CA 94583
Tel: +1 (925) 543-5400
Fax: +1 (925) 543-5401
www.mdl.com

MDL and Available Chemicals Directory are registered trademarks or trademarks of MDL Information Systems, Inc. ('Elsevier MDL') in the United States and/or other countries. All other product and company names mentioned herein may be trademarks or registered trademarks of their respective holders.

Copyright © 2006 Elsevier MDL. All rights reserved