

Conception of a Huge Chemical Database and Test Set Design for Virtual and High Throughput Screening

Aurélien Monge, Alban Arrault, Christophe Marot and Luc Morin-Allory
Institut de Chimie Organique et Analytique, UMR CNRS 6005,
Université d'Orléans BP 6759, 45067 ORLEANS Cedex 2, France.



aurelien.monge@univ-orleans.fr

<http://www.univ-orleans.fr/icoa/screeningassistant/>



INTRODUCTION

Designing a screening set of molecule is one of the first steps in drug-discovery projects. This task needs to start with an important number of molecules and to deal with both druglikeness and chemical diversity. We are developing a software for the management of databases of millions of compounds. This software is conceived to automatically update databases from chemical providers SD files. Duplicates structures are identified using the InChI unique code. Drug-like related properties are computed and reactive functions are identified. The druglikeness and the leadlikeness of the compounds are evaluated with in-house scores. Analysis functionalities are also provided, allowing charting various properties of the compounds in the database.

This software has been used to create a virtual database of millions of compounds in our laboratory. This database is used in several drug discovery projects, and a screening set design for a COX2 docking model is shown as an application.

SOFTWARE

Our Java software *Screening Assistant* is based on both *JOELib* for computational chemistry and *MySQL* for the database management. Databases are automatically updated with new providers' SDF. As only new structures must be added, this operation requires a reliable unique code for the structures. First the counter-ions of the structures are deleted and the structures are reprotated, then InChI code is computed. This project developed by IUPAC in collaboration with NIST have been launched with the aim to become a free standard unique code [1].

After duplicates identification, the next step is to select compounds of interest. In drug-discovery projects the most general criterion for compounds is good oral bioavailability, also called drug-like criterion. We have developed a score based on the widely used Lipinski rules, previous studies in our laboratory and properties repartitions in drug databases [2]. This score gives a real number representing the penalty for the compounds. Compounds with scores = 1 are drug-like, and compounds with scores > 2 are predicted to have bad oral bioavailability. A similar score have been developed for leadlikeness estimation.

Obviously druglikeness and lead-likeness are general concept, an some project can need more specialised rules. Many drug-related properties are computed by *Screening Assistant*, and each of these properties can be used to create specific filters (fig. 1).

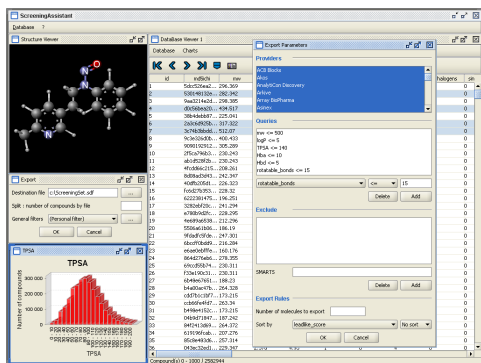


Figure 1: Screening Assistant showing chart and choice of criteria for creating a screening set.

DIVERSITY AND FRAMEWORKS

It is interesting to study the increase of the diversity and of the number of frameworks [3] with the number of compounds in databases (fig. 2).

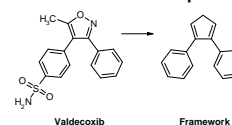


Figure 2: Valdecixib and his framework.

We instinctively expect that diversity and frameworks will increase slower as the size of databases increases. Actually the increase of the number of frameworks is roughly linear with $R^2 = 0.89$ (fig. 3). The diversity increase is less linear than frameworks $R^2 = 0.71$ and shows a trend to increase slightly slower for bigger databases.

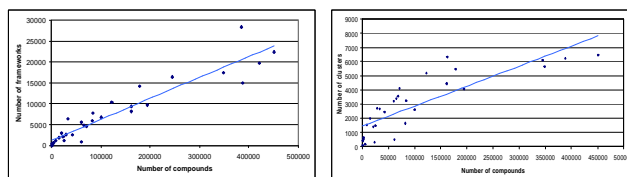


Figure 3: increase of frameworks (left) and diversity (right) with the size of the chemical databases from 32 providers.

With this observations we can conclude that it does not exist an optimal chemical databases size and that it is important to consider as many molecules as possible in the first steps of drug-discovery programs.

VIRTUAL DATABASE

Screening Assistant was used to combine the databases of 32 chemical providers resulting in a virtual database of 3.8 millions of compounds (fig. 4). 2.6 millions of structures are identified as unique. Among this compounds 1.9 millions are drug-like and 900 000 are lead-like according to our in-house drug-like and lead-like scores. 87 000 unique scaffolds have been identified.

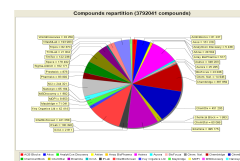


Figure 4: Distribution of the providers in the global database.

SCREENING SET DESIGN

We used *Screening Assistant* and our virtual database to generate a screening set for a COX2 Pharmacophore / Docking model. As testing 2.6 millions of molecules is very time consuming, a subset of molecules is tested. First we have studied properties of the set of active molecules used in the model and remove all compounds with properties very different of active molecules' properties. This results in a subset of 1.1 millions of compounds. Next, diversity is used to reduce this number to approximately 800 000 compounds. The diversity is evaluated using fingerprints based descriptors and Tanimoto metric. Moreover scaffolds are taken into account and the distribution of the scaffolds in the subset of 1.1 millions of structures is kept.

CONCLUSION

We have a virtual database of 2.6 millions of unique compounds with precomputed druglike related properties, fingerprints and frameworks. This allows to extract quickly a screening set using chosen criteria, diversity, or framework distribution. Although we use this database to generate screening sets for virtual models in our laboratory, we have also applied these algorithms to design databases for high throughput screening. It allows to design databases without duplicates, with desired properties and with maximal diversity. In addition to that, our system is very helpful for updating our database, adding only new compounds from providers' databases.

1 - The IUPAC Chemical Identifier Project : <http://www.iupac.org/projects/2000/2000-025-1-800.html>

2 - Mozziconacci, J. C. ; Arnoult, E. ; Baurin, N. ; Marot, C. ; Morin-Allory, L. Preparation of a molecular database from a set of 2 million compounds for virtual screening applications : gathering, structural analysis and filtering. 9th Electronic Computational Chemistry Conference (ECCC9).

3 - Bemis, G.W. ; Murcko, M.A. The Properties of Known Drugs. 1. Molecular Frameworks. *J.Med.Chem* 1996, 39, 2887-2893.