

Screening Assistant

Un Logiciel de Gestion de Chimiothèques

Aurélien Monge, Alban Arrault, Christophe Marot et 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

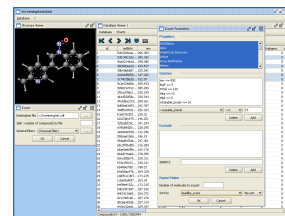
<http://www.univ-orleans.fr/icoa/screeningassistant/>



INTRODUCTION

La réalisation de tests de criblages virtuels nécessite d'être effectuée sur un ensemble le plus complet et le plus représentatif possible de molécules. Un nombre très important de structures chimiques peut facilement être obtenu auprès des fournisseurs de composés chimiques. La gestion de ces structures nécessite cependant de prendre en compte plusieurs paramètres. Il faut tout d'abord éliminer les doublons, qui feraient perdre non seulement du temps lors d'un criblage virtuel, mais aussi de l'argent lors d'un criblage réel. En fonction des projets, on ne gardera que les molécules drug-like, lead-like, ou bien qui répondent à des critères spécifiques. Une fois toutes ces étapes réalisées, il peut s'avérer que le nombre de composés restant soit encore trop important. On sélectionne alors un sous-ensemble en utilisant des techniques basées sur la diversité des structures.

Nous développons un logiciel qui permet de traiter ces problèmes. Nous illustrons l'utilisation de ce logiciel par un exemple concret : choisir 1000 molécules lead-like à acheter parmi les bases de 4 fournisseurs.



1 CREATION DE LA CHIMIOTHEQUE

Pour présenter les fonctionnalités du logiciel nous prendrons comme exemple une base constituée de quatre fournisseurs. Les quatre fichiers SDF correspondant à ces bases sont insérés (figure 1). Lors de cette insertion plusieurs prétraitements sont effectués :

- suppression des contre-ions
- addition des hydrogènes
- correction des états de protonations
- calcul d'un code unique : le code InChI [1]
- identification des doublons par ce code unique
- calcul de scores drug-like et lead-like
- détection des « frequent hitters »

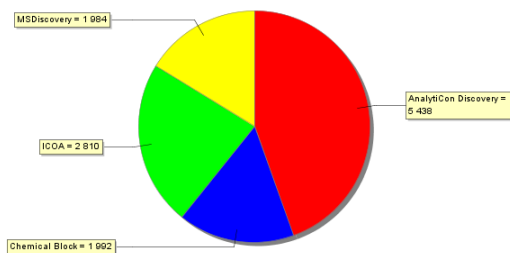
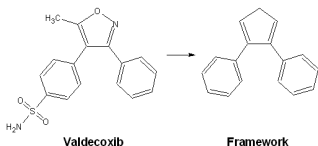


Figure 1 : répartition des composés de la base globale entre les différents fournisseurs.

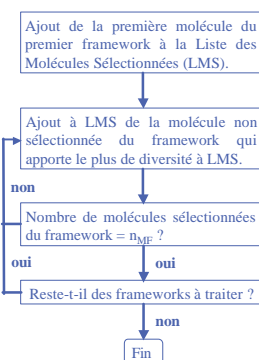
3 CHOIX DES COMPOSES

Une fois le fournisseur choisi, il faut sélectionner les molécules destinées au criblage. Il s'agit ici aussi de travailler avec la diversité, mais nous allons également utiliser la notion de frameworks introduite par Bemis [3]. Nous avons apporté une modification à cette notion de frameworks afin de garder l'information sur l'aromaticité de la molécule.



L'algorithme de diversité développé conserve la répartition des molécules par frameworks. En effet il est important de favoriser la notion de diversité des "squelettes" des molécules car il est beaucoup plus intéressant d'avoir des hits avec des squelettes différents.

L'algorithme sélectionne 1023 molécules parmi les 1578 composés lead-like de la base Chemical Block.



$$n_{\text{fin}} = n_{\text{nb de framework}} \times \frac{\text{nb de molécules souhaitées}}{\text{nb de molécules de la base}}$$

Figure 4 : algorithme de diversité conservant la répartition des frameworks de la base initiale.

2 CHOIX DU MEILLEUR FOURNISSEUR

Il est beaucoup plus avantageux du point de vue financier de commander toutes les molécules chez un même fournisseur. Cela nécessite de choisir le fournisseur qui possède les composés les plus représentatifs. Il faut pour cela explorer l'espace chimique de la base entière et choisir le fournisseur qui recouvre cet espace au maximum. Dans notre exemple nous considérerons l'espace chimique lead-like. L'étude des figures 2 et 3 nous permet d'éliminer la base Analyticon Discovery pour ce projet. La base Chemical Block est la plus intéressante des 4 bases en terme de choix de composés lead-like.

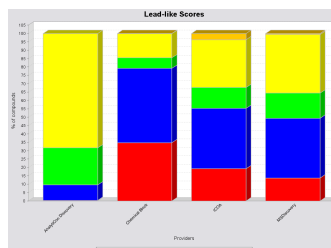


Figure 2 : la répartition des composés en fonction de leurs scores lead-like pour chaque fournisseur est représentée. Les composés avec un score = 1 (en rouge) sont considérés comme lead-like. La base Chemical Block est celle qui a le plus grand pourcentage de composés lead-like.

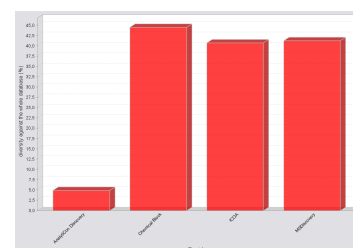
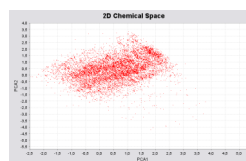


Figure 3 : diversité estimée de la base de chacun des fournisseurs. Cette diversité est estimée par l'étape de dissimilarité du Stochastic Clustering Algorithm [2]. On utilise comme descripteurs les fingerprints SSKey3DS. La base Chemical Block est la plus représentative de la base globale.

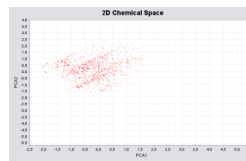
4 VISUALISATION DE LA SÉLECTION

La visualisation de l'espace chimique permet d'effectuer un contrôle visuel de la diversité des molécules sélectionnées. Screening Assistant permet la visualisation de l'espace chimique. Cet espace chimique est étalonné sur 18 000 molécules sélectionnées par diversité parmi 2,6 millions. Une Analyse en Composantes Principales a été réalisée sur ces 18 000 molécules en utilisant comme descripteurs la masse moléculaire, le logP et les fingerprints SSKey3DS. Nous observons que les molécules sélectionnées couvrent l'espace lead-like de la base de départ.

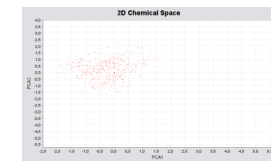


Étape 1 : base globale.

Étape 2 : composés lead-like.



Étape 2 : composés lead-like de Chemical Block.



Étape 3 : Sélection finale.

CONCLUSION

Nous développons un logiciel, Screening Assistant, implémentant les étapes de la gestion d'une chimiothèque et de la création d'un sous-ensemble de test. Ce logiciel programmé en Java stocke les données dans un serveur MySQL. Il est conçu pour mettre à jour automatiquement les bases de données à partir des fichiers SDF des fournisseurs. Les structures sont identifiées de manière unique en utilisant le code InChI. Les descripteurs relatifs aux propriétés drug-like des composés sont calculés. Les caractères drug-like et lead-like des molécules sont également évalués. Le logiciel permet également l'extraction de molécule par diversité. Ce système est particulièrement adapté pour la création de chimiothèques destinées au criblage virtuel ou réel.

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

2 - Reynolds, C. H.; Druker, R.; Pfahle, L. B. Lead Discovery Using Stochastic Cluster Analysis (SCA): A New Method for Clustering Structurally Similar Compounds. J. Chem. Inf. Comput. Sci. 1998, 38, 305-312

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