N° 20 • 1er Décembre 2005

La Gazette du CINES

Centre Informatique National de l'Enseignement Supérieur

BioinfoDB : un inventaire de molécules commercialement disponibles à des fins de criblage biologique

Didier Rognan
Institut Gilbert Laustriat, CNRS UMR7175-LC1, Illkirch
Laboratoire de Bioinformatique du Médicament
Email : didier.rognan@pharma.u-strasbg.fr

Introduction

A cours des dix dernières années, la miniaturisation des essais biologiques a considérablement accéléré leur débit. Il est relativement courant, pour la majorité des grandes compagnies pharmaceutiques, de cribler quelques centaines de milliers de molécules afin d'identifier les molécules les plus prometteuses dans un programme de découverte de nouveaux médicaments [1]. Pour des raisons essentiellement économiques, le secteur académique ne peut suivre cette marche en avant malgré des tentatives louables, notamment en France, de constituer une collection de molécules (chimiothèque) dite « patrimoine » en provenance de divers laboratoires académiques français [2].

Il est par contre relativement aisé d'archiver, sous forme de base de données, la très grande majorité des 3 millions de molécules commercialement disponibles auprès de divers fournisseurs connus, et de sélectionner au moyen de filtres chémoinformatiques (criblage in silico) les molécules les plus appropriées pour un projet de recherche particulier. Diverses bases de données (ACD-Screen, ChemNavigator) existent déjà pour subvenir à ce besoin. Elles présentent toutefois le désavantage d'êtres très onéreuses, et de donner un simple accès

SOMMAIRE

a des données brutes non traitées. Nous avons donc entrepris, il y a maintenant 4 ans, de développer et de maintenir une base de données de molécules « candidats-médicaments » commercialement disponibles (BioinfoDB).

Développement de BioinfoDB

La mise au point d'une telle base de données a répondu à un cahier des charges bien défini :

- elle doit être la plus exhaustive possible et décrire la totalité des molécules commercialement disponibles en poudre et achetables à l'unité ;
- elle doit prendre en compte la redondance (une même molécule présente chez de multiples fournisseurs) et la disponibilité afin de faciliter la gestion du stock;
- elle doit être suffisamment bien annotée (fournisseurs, références commerciales, descripteurs physicochimiques) afin d'en affiner les critères de sélection;
- elle doit prendre en compte des données chimiques majeures (ionisation, protonation, tautomérie, stéréochimie) contrôlant la nature même des molécules archivées ;
- elle doit éviter autant que possible de multiples interconversions de formats électroniques (diverses représentations moléculaires : 1-D, 2-D, 3-D);
- elle doit être archivée sous une forme permettant un stockage et une interrogation aisées (tables MySQL interrogeable par une interface Web) ;
- son actualisation doit être périodique (idéalement trimestrielle) afin de refléter fidèlement les stocks fournisseurs réellement disponibles, ce qui nécessite la mise au point d'un protocole automatisé de constitution/actualisation.

Nous avons choisi le logiciel PipelinePilot [3] afin d'établir un protocole entièrement automatisé (Figure 1) allant de la saisie des données brutes jusqu'au stockage des molécules « candidats-médicaments » correctement annotées dans des tables MySQL.

Après avoir lu l'ensemble des catalogues électroniques (Figure 1 : Etapes 1 et 2) couvrant 23 fournisseurs, 38 collections de criblages et 3 millions de molécules, un identifiant unique est donné a chaque composé et une première étape de filtration est appliquée afin d'éliminer les contre-ions, les molécules chimiquement incorrectes (structures erronées) et trop compliquées. Un jeu de 162 règles [4] est par la suite appliqué afin de ne garder que des « candidats-médicaments » potentiels dans un fichier unique (Etape 3). La quatrième étape de notre protocole consiste à détecter les duplicats en comparant chaque molécule une à une (une représentation simplifié dite « en ligne » de type SMILES canonique est utilisée a cet effet) et à conserver pour chaque molécule présente en plusieurs exemplaires, la totalité des sources commerciales. Les deux étapes suivantes (Etapes 5 et 6, Figure 1) consistent à convertir la structure de chaque molécule en une représentation 3-D adéquate, puis à les ioniser correctement à pH physiologiques. Un certain nombre de descripteurs physicochimiques sont alors calculés (Etape 7, Figure 1) et rajoutés dans des champs séparés, afin d'annoter la base de données dans une représentation 2-D facilement interrogeable (Etape 8; Figure 1). Pour chaque molécule archivée dans la base finale (1,860,000 composés au total), les diverses représentations (1-D, 2-D, 3-D) sont sauvegardées afin de répondre à

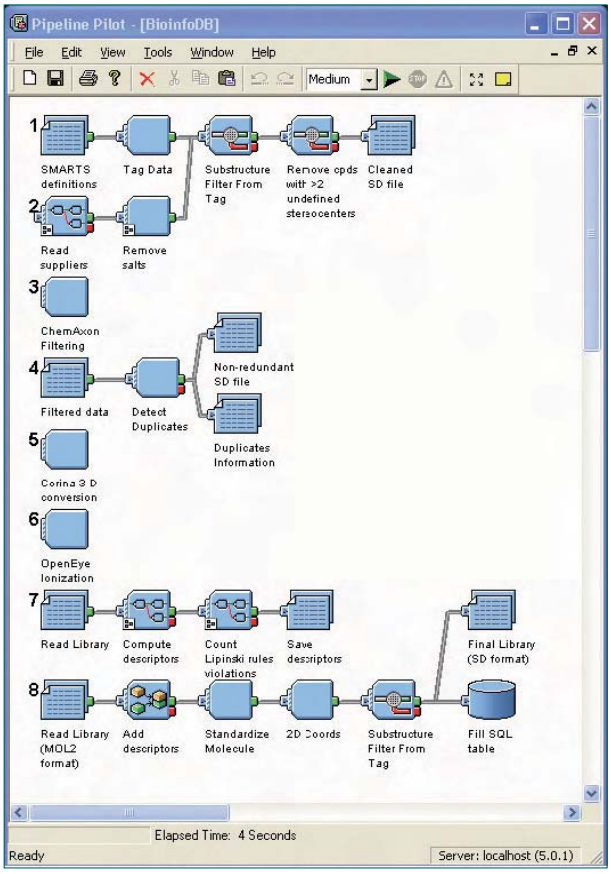


Figure 1 : Protocole automatisé de constitution de la chimiothèque BioinfoDB

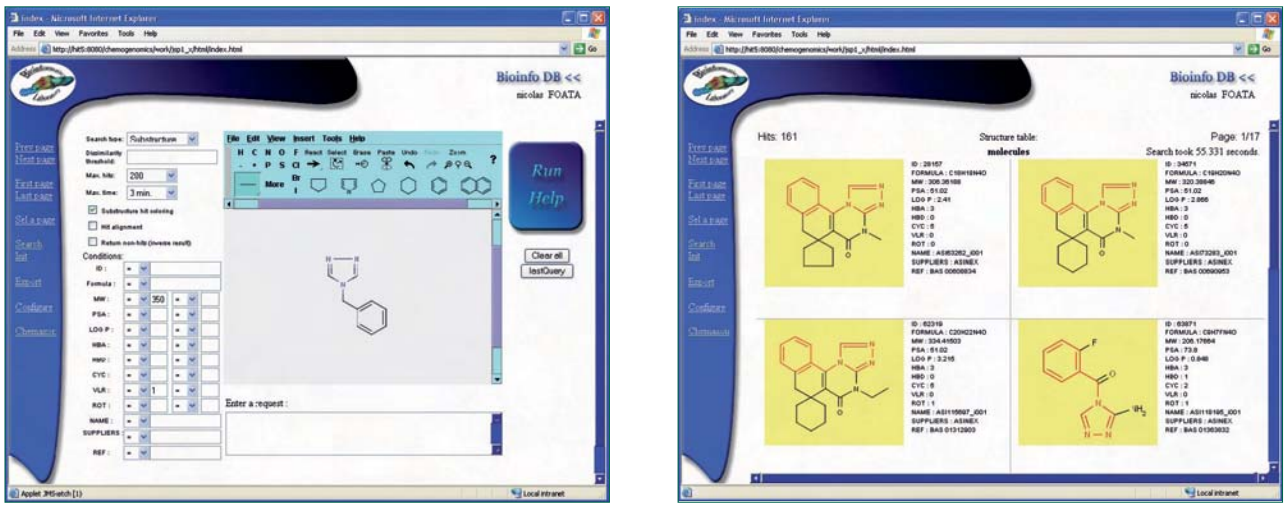


Figure 2 : Navigation dans BioinfoDB (gauche: interface de navigation pour une requête simple; droite: résultats de la requête)

des besoins différents (ex : calculs d'empreintes chimiques binaires en 1-D, navigation en 2-D, modélisation moléculaire en 3-D). La dernière opération consiste à stocker l'ensemble de l'information dans des tables MySQL afin d'en faciliter l'interrogation. L'ensemble du protocole est entièrement automatisé et prend environ 2-3 jours sur un PC standard.

Navigation dans BioinfoDB

Afin d'assurer une navigation aisée dans une interface conviviale et ne nécessitant pas l'installation d'une application particulière sur le client, nous avons choisi d'archiver les structures chimiques au moyen du logiciel JChemBase [5] sur notre serveur (PC/Windows), et d'interroger les tables MySQL au moyen de JSP (Java Server pages).

L'interface d'interrogation (Figure 2) permet de sélectionner n'importe laquelle des 1,860,000 molécule de notre chimiothèque par des requêtes simples formulée soit par structure/sous-structure chimique (un éditeur moléculaire est utilisé pour dessiner des structures chimiques en 2-D), soit similarité à une structure de référence, soit par respect de valeurs seuil sur des descripteurs moléculaires (poids moléculaire, surface polaire, logP, donneurs et accepteurs de liaisons hydrogènes, nombre de cycles et de liaisons de rotations, nombre de violations des règles de Lipinski), soit par interrogation de sources commerciales (fournisseur, référence catalogue). Dans la mesure ou la totalité des informations structurales (sous forme d'empreintes binaires) sont stockées en mémoire cache, l'interrogation de la base est très rapide, la durée d'une requête allant de quelques dixièmes de seconde (requête simple basée sur une structure) à un maximum d'une minute (requête complexe alliant recherche structurale à descripteurs moléculaires).

Une fois les molécules d'intérêt identifiées, il est très facile d'exporter le niveau d'information choisi par l'utilisateur (structures et/ou liste de références fournisseurs et/ou descripteurs) dans un fichier unique qui pourra être utilisé a diverses fins :

- constitution d'une liste de molécules à l'achat auprès des fournisseurs correspondants (si plusieurs fournisseurs existent pour une molécule, la totalité des fournisseurs et des références catalogues correspondantes est fournie à l'utilisateur) ;
- constitution d'une chimiothèque focalisée sur une besoin bien précis et pouvant faire l'objet par exemple d'un criblage in silico (recherche de similarité, recherche de pharmacophore, docking). Il est à noter que l'ensemble des 1,860,000 structures (2-D, 3-D) de BioinfoDB est accessible au CINES (SGI Origin3800) et que l'utilisateur peut ainsi facilement constituer sa chimiothèque 2-D ou 3-D en concaténant les molécules sortant d'une liste issue de notre navigateur.

Perspectives

BioinfoDB n'est pour le moment consultable qu'en Intranet au Laboratoire de Bioinformatique du Médicament (CNRS UMR7175-LC1). Afin d'en assurer une distribution plus large, nous avons toutefois décidé de la rendre accessible à tout utilisa-

teur du CINES. Cette Gazette en fera l'écho au moment où la base sera consultable en ligne (probablement fin Décembre). D'ici là, toute personne intéressée à consulter BioinfoDB à des fins académiques peut me contacter directement.

Remerciements

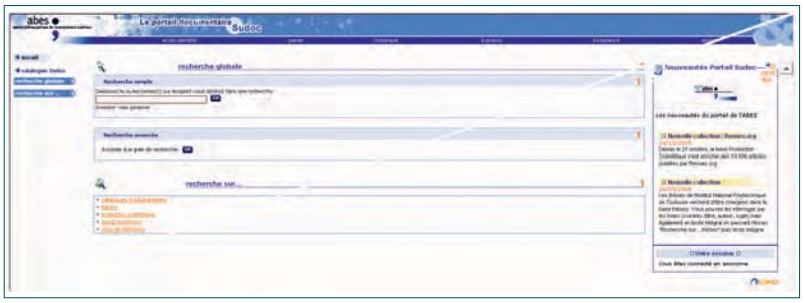
Je remercie particulièrement l'ensemble de mes collaborateurs ayant contribué à l'existence de cette base de données et particulièrement Guillaume Bret, Mireille Krier et Nicolas Foata.

Références

1. **Beyond uHTS: ridiculously HTS?** Mander, T., *Drug Discov. Today ;* 5:223-225., 2000.
2. **Chimiothèque Nationale,** Centre National de la Recherche Scientifique (CNRS), http://chimiotheque.ujf-grenoble.fr/induk.html.
3. SciTegic Inc., San Diego, CA 92123-1365, USA.
4. **Filtering databases and chemical libraries** Charifson, P. S. and W. P. Walters, *J Comput Aided Mol Des; 16: 311-23, 2002.*
5. **JChemBase,** ChemAxon Ltd., Budapest, 1037 Hongrie.

Le portail documentaire Sudoc

www.portail-sudoc.abes.fr
Pour l'équipe Portail Documentaire Sudoc
Marianne Giloux - ABES



Pourquoi le portail documentaire Sudoc ?

Après avoir mis en place le réseau et le catalogue Sudoc[1] (cf. La Gazette n°15-16), l'ABES[2] se devait de travailler sur l'accès aux documents en ligne, service très demandé par les utilisateurs. En effet, le catalogue ayant mis en place les outils nécessaires à la localisation des documents possédés par les établissements de l'enseignement supérieur et de la recherche, il s'agissait de faciliter l'identification et l'accès aux documents « en ligne ».

A l'issue d'une étude préalable et d'un appel d'offre effectué début 2003, la société Archimed[3] a été choisie pour mettre en place une solution de portail documentaire dont les objectifs énoncés étaient les suivants :

- répondre aux attentes des utilisateurs en permettant l'accès simultané à différents catalogues, bibliographies et bases de données, mais également à une somme de ressources en ligne (et ce, en fonction des droits d'accès de chacun.)
- enrichir les outils de recherche, faciliter et diversifier les accès en mettant en place un méta-moteur de recherche et d'indexation pour tout type de ressources.

[1] *Système Universitaire de Documentation : http://www.sudoc.abes.fr*
[2] *Agence Bibliographique de l'Enseignement Supérieur : http://www.abes.fr*
[3] *http://www.archimed.fr*