DATABASES

Compound bioactivities go public

http://www.ebi.ac.uk/chembl/

There was a time, in the not-too-distant past, where pharmaceutical companies often had the edge over academic research in the drug discovery field, in part for a simple but powerful reason: vast repositories of compound bioactivity data. In any given year, a typical large pharmaceutical company would have the capacity to run hundreds of high-throughput screens of on the order of about a million compounds, creating primary actives in the order of millions and dose-response curves that number in the tens of thousands. In contrast to data obtained in the bioinformatics world—where genomic sequences were publicly deposited in GenBank, protein sequences found their home in SwissProt and protein crystal structures could be accessed by everyone through the Protein Data Bank—there were scarce opportunities for academic researchers to access and mine large amounts of bioactivity data. This factor (among others) led to a disconnect between industrial and academic drug discovery efforts.

Fortunately, in the last decade the rules of the game have shifted considerably, with certain types of compound bioactivity data entering the public domain: PubChem, probably the best known and one of the largest efforts of this kind so far, is a public repository of predominantly US National Institutes of Health Molecular Libraries Roadmap screening data, ChemBank publicizes screening results from the Broad Institute of Harvard and MIT and DrugBank compiles more narrow but also much more in-depth information on a set of thousands of US Food and Drug Administration–approved drugs, and so on. However, 18 January, with the official release of the ChEMBL database (ChEMBLdb, version Chembl_02, http://www.ebi.ac.uk/chembldb/index.php) at the European Bioinformatics Institute, marked an important addition to the public domain: a database containing large-scale, manually curated (and thus usually higher quality) data covering a large amount of chemical space with largely standardized data taken from dose-response experiments and with annotated literature references.

Funded by a Strategic Award from the Wellcome Trust for the next 5 years, the ChEMBLdb and related tools are being developed by the ChEMBL group, which is headed by John Overington at the European Bioinformatics Institute. The initial database, originally called StARlite, was acquired from Galapagos in 2008. The current release of the database contains about 2.4 million activities of 622,824 compounds, measured against 7,192 targets, 4,364 of which are protein targets (the others are cell lines or organisms). This release of ChEMBLdb also includes 24,000 natural products, which is particularly noteworthy because databases linking natural products to protein targets have been notoriously difficult to find in the past. Data are abstracted from a total of nearly 34,000 publications, taken from 12 prominent medicinal chemistry journals such as the *Journal of Medicinal Chemistry*. Hence, coverage of public bioactivity data, while not complete, is by far the most comprehensive ever seen in a public database.

Although this is a major step forward indeed, equivalent commercial databases may contain up to ten times as many structures (according to the companies). In particular, many commercial databases contain compound data from patents, which is not represented in ChEMBLdb at the current stage. One hopes that with the increasing influence of automated text mining (and, in particular, mining of chemical structures) in the life science domain, this information can be incorporated into future versions of the database.

What is particularly useful about ChEMBLdb is the many items of data provided for every compound in a format amenable to computerized data mining: gene names are provided for targets, and measured activities are translated into a 'standard value', while at the same time the database provides the numerical information given in the original publication. More than half of the data in ChEMBL are K_i, IC_{50} or EC_{50} values, thereby helping to automate selection of compounds with a given bioactivity cutoff. In addition to providing on-target activity data, targets are also annotated with protein ontologies, enabling the user to follow chemogenomics approaches such as those used in target deorphanization projects.

Although ChEMBLdb can be accessed through a web interface as well as downloaded by FTP for local data mining efforts (and also integration into in-house databases), sophisticated data mining tools are being developed in the ChEMBL group based on target families relevant in drug discovery. At present the most advanced of those tools is Kinase SARfari (http://www.sarfari.org/kinasesarfari/), which, in addition to screening data, provides sequence, alignment and structural information on the kinase protein family. The present release of Kinase SARfari (version 1.18) links ~17,000 compounds to 959 kinase domains comprising nearly 69,000 bioactivity data points. The web interface enables searching for related kinases both from the ligand or bioactivity side and from the three-dimensional structure of the protein target. A version for G protein–coupled receptor ligands, termed GPCR SARfari, is being developed, and tools for other target classes would certainly be useful to researchers in each respective area. In addition, a database of known drugs (termed DrugStore and containing in excess of 1,500 entries) as well as a database of clinical compounds (named CandiStore and containing more than 12,000 entries in various stages of clinical development) will also be released shortly.

Overall, the databases the ChEMBL group now provides are enriching scientific research in the drug discovery field tremendously, allowing both academics and smaller start-up companies to perform knowledge-based drug discovery projects on an unprecedented scale. The current grant of the Wellcome Trust will support development and updates of ChEMBLdb through 2013, and the impact that these data will have on drug discovery research will be profound—something that will surely still be true after that date. Several analyses based on ChEMBLdb data are on the way in academia and the data is being integrated into data repositories in pharmaceutical companies. Hence, the value of these data has been recognized by academia as well as industry—and it is hoped that this, as well as related public bioactivity repositories, will continue to grow and thrive in the future. ■

Competing financial interests
The author declares no competing financial interests.

REVIEWED BY ANDREAS BENDER

Andreas Bender is with the Leiden / Amsterdam Center for Drug Research, Leiden University, Leiden, The Netherlands.
e-mail: andreas.bender@cantab.net