Open-access chemistry databases evolving slowly but not surely

Publicly available initiatives still have a long way to go to catch up with proprietary counterparts.

Monya Baker

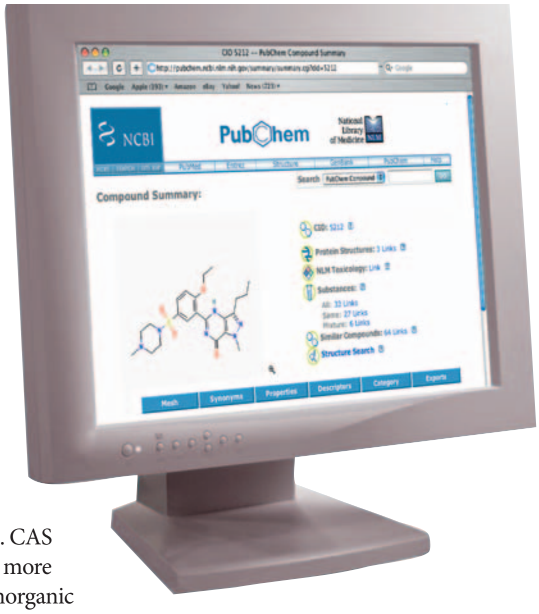
In terms of freely available resources chemists have long been green with envy with what is available to biologists.

For over two decades, open-access databases like GenBank and the Protein Data Bank (PDB) have been instrumental in helping biologists translate gene and protein sequences into biological relevance. Chemists, particularly those in academia, would dearly love free access to databases that can identify and predict biological activity from chemical structures. Despite recent efforts to satisfy these needs, many say there is still a long way to go before this dream is realised.

Several open-access chemistry resources have sprung up in the past few years (TABLE 1). Perhaps the best known is PubChem, launched in September 2004 by the US National Institutes of Health. PubChem stores assay results from designated screening centres on a collection of about 100,000 small molecules, plus other contributed data, all searchable by structure.

"Until PubChem came on the scene, the state of chemoinformatics compared to bioinformatics was 20 years behind," says Christopher Lipinski, who formulated the eponymous rule-of-five criteria for drug bioavailability.

Open-access initiatives are a welcome arrival, but the chemical data still pale in comparison to what already exists in other databases and the published literature, says Andrew Hopkins, Associate Research Fellow in Knowledge Discovery at Pfizer. "PubChem and ChemBank are positive initiatives, but are not yet on a comparable level to the proprietary databases or commercial screening databases."



PubChem's director, Stephen Bryant, says he lacks the staff and mandate to collect data from published literature and patents. The powerhouse of the chemistry database field, Chemical Abstracts Service (CAS), a division of the American Chemical Society (ACS), employs hundreds of scientists to help abstract, index, check and curate the data from some 50 patent authorities and 9,500 journals. CAS currently contains records for more than 29 million organic and inorganic substances. PubChem, by comparison, has 13 million entries representing about 8 million chemical structures (as of August 2006).

PubChem has the right attitude but lacks the resources, says Brian Shoichet, Professor of Pharmaceutical Science at University of California, San Francisco. The screening data are less rigorous than those in peer-reviewed articles, and contain many false positives. Deposited data aren't curated, and so mistakes in structures, units and other characteristics can and do occur. Shoichet worries that chemists who use PubChem to identify compounds used to perturb a specific biological process (so-called tool compounds) will find molecules that send them on a wild goose chase.

Part of the reason is that building the chemistry equivalent of open-access biology databases is no easy task. Biological molecules have straightforward sequences that work easily with databases. Making chemical structures readable by machines can be tricky, particularly for varying properties like ionization states. Worse, since structural similarity doesn't guarantee similar bioactivity, structures without other data aren't always useful.

Bryant argues that caveats are posted on the site, and says that a confounding factor is that many chemists have not mastered the skills necessary to search effectively. Searchers might, for example, be able to identify false positives by checking whether molecules or structurally similar ones are active across many assays. Aware of the problem, the NIH began offering PubChem classes in August this year.

Such errors need not be a fatal flaw, says Stuart Schreiber, Director of

Doubt over HIV vaccines
p710

New paradigm for treating depression
p711

Double victory for generics companies
p712

Bob Ruffolo discusses the merits of early clinical development centres
p714

Gleevec alternative approved
p717

Chemical Biology at the Broad Institute. "What will happen is what happened in genomics. People will take the suitable data and bring them into an analysis environment." Schreiber oversees ChemBank, a high-throughput screening project that aims to understand cellular pathways using small molecules, and which deposits all its data in PubChem.

Despite flaws, the very existence of a common repository could help change the way researchers analyse chemical data. If pharmaceutical companies could use a common public place to deposit data, this would allow academics to develop better software to help identify compounds with similar activity.

Staff and sponsors at PubChem are developing a suite of software tools, including a chemical equivalent of the BLAST software commonly used to identify homologous genes and proteins. Companies such as Bristol-Myers Squibb, GlaxoSmithKline and Merck have already met with the NIH about donating crystal structures of proteins with bound ligands in the hope of creating more predictive computational tools. Even proprietary databases are contributing some data. Elsevier, for example, has contributed several million structures to PubChem. Commercial contributors see their entries as an experiment in advertising, says Bryant. "It's a way of saying 'I have further information about this structure'."

And that could be the very information a researcher needs. Some users have already downloaded the entire PubChem database, presumably to integrate it with proprietary information. A mix of public and proprietary databases can be used to ask different questions about molecules, such as whether they are commercially available, what literature and patent references exist or how to make them. Workplaces offer varying levels of access, and scientists have a roster of tricks to find the information they need. For instance, Jean-Claude Bradley, a chemistry professor at Drexel University, has started publishing structures from his research on a blog to make more information available freely.

Another solution to this fragmented, proprietary system was launched by three entrepreneurs in November 2005. Modelled on Google, eMolecules lets chemists search by structure and substructure and pulls up information from several government databases, plus catalogue data (minus prices) from about 150 chemical vendors. Right now, searchers use eMolecules mainly to find what compounds are available and from which suppliers. Projects are underway to allow the search engine to pull up structures from published patents and peer-reviewed literature.

Table 1 | **Several databases are now freely available to chemists**

Database	Information	Sponsor
ChEBI (Chemical Entities of Biological Interest): http://www.ebi.ac.uk/chebi/	Links to relevant protein data; nomenclature	European Bioinformatics Institute
ChemBank: http://chembank.broad.harvard.edu/	Curated bioactivity data; links to PubChem	Broad Institute (funded by US NCI)
ChemID Plus: http://chem.sis.nlm.nih.gov/chemidplus/	Chemical name, formula and structure; physical and toxicological properties	US National Library of Medicine
*CHEMnetBASE: http://www.chemnetbase.com/	Chemical names; physical properties; references	Chapman & Hall/CRC Press LLC
DrugBank: http://redpoll.pharmacy.ualberta.ca/drugbank/cgi-bin/molSearch.cgi	Chemical, pharmacological, pharmaceutical and target data for approved drugs	University of Alberta (funded by Genome Alberta/Genome Canada)
eMolecules (formerly Chmoogle): http://www.emolecules.com/	Commercial availibility; database results such as DrugBank, PubChem, NIST webbook	eMolecules
PubChem: http://pubchem.ncbi.nlm.nih.gov/	Bioactivity data; protein structure links; published literature	National Institutes of Health
QueryChem: http://llama.med.harvard.edu/~jklekota/QueryChem.html	Searches public databases and the web using a combination of text and structure	Broad Institute/Harvard University
ZINC: http://blaster.docking.org/zinc/	Commercially available compounds rendered in ready-to-dock 3-D formats	University of California, San Francisco

Source: Jean-Claude Bradley/*Nature Reviews Drug Discovery*. A comprehensive list of specialized chemistry databases can be found at http://library.caltech.edu/collections/chemistry.htm#DATABASESP. *Not entirely free, full information by subscription only. NCI, National Cancer Institute.

Existing algorithms that turn chemical structures into machine-readable text strings could go a long way to collating information from different sources onto one screen, but even if technical hurdles were resolved, open-access and proprietary information don't mix. "As soon as you have some sort of access control or entry fee, that represents a huge barrier for bridging different resources," says Warren DeLano, Head of DeLano Scientific, which provides open-source molecular visualization systems. "The worldwide web wouldn't work if you had to put a coin in every time you wanted to follow a link."

The problem gets even worse for chemoinformaticians who want to use data from thousands of compounds to generate predictive rules. Though the amount of publicly available data is increasing, older, more-extensive and better-validated data are largely inaccessible. Linking structures to the appropriate data within published documents is very difficult to automate, says eMolecules CEO Klaus Gubernator. "Proximity does not guarantee that data close to a structure is associated with the structure." Another problem with extrapolating data from published documents is that with the exception of *Nature Chemical Biology* and Prous's *Drugs of the Future*, journals have yet to make depositing chemical structure information a condition of publishing.

An ideal collaborative resource would be designed for large-scale data mining, contain curated historical data, and have data standards and deposition tools that could constantly bring in data from the published literature. Hopkins and Shoichet think that a governmental or philanthropic organization could purchase a quality curated proprietary database, make it freely available, and set up policies and standards so that data are funnelled into the database as they are published. Although proprietary databases would still be in demand for specialized services and many journals would not opt in, such efforts would mean an explosion in useful information. "It could happen in a few months," Hopkins says, "if you bring the right people to the party." Unfortunately, no-one seems to be willing to take up the challenge.

More likely in the near future is that publicly available data will grow as the screening projects advance, that some journals can be convinced that their articles should provide machine-readable descriptions of the structures discussed, and that the expertise and tools for using available data will grow. In other words, the party might take longer to get started than hoped for, but it should be worth the wait.