Wellcome boost for open-access chemistry

Philanthropic acquisition gives the academic chemogenomics community invaluable access to well-curated proprietary data.

Sarah Houlton, London, UK

The Wellcome Trust has recently awarded a 5-year, UK£4.7 million grant to transfer well-structured chemogenomics data from the publicly listed company Galapagos to the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI). The data will be incorporated into the Institute's collection of open-access data resources for biomedical research, and maintained by a team that is now being recruited.

EMBL-EBI, based in Hinxton, Cambridge, UK, had already identified the strategic need for a chemogenomics data resource to help translate insights from the Human Genome Project into medical advances. Janet Thornton, Director of EMBL-EMI, says: "Chemogenomics data are an essential component in future drug discovery efforts, but the value of this is only practically realized when such data are effectively integrated against genome databases and functional-genomics data."

Public databases of chemogenomics data have been established in recent years, the largest of which is PubChem (http://pubchem.ncbi.nlm.nih.gov), hosted by the US National Institutes of Health. However, lack of curation of publicly deposited data is a significant limitation to its utility (*Nature Rev. Drug Discov.* 7, 632–633; 2008). Also, as yet, ▸

the nature of the data — particularly data that could be valuable for drug discovery efforts — is not yet comparable with that available in typical pharmaceutical company databases (*Nature Rev. Drug Discov.* 5, 707–708; 2006).

The Wellcome Trust's acquisition of Galapagos's data is set to change this. "I think there will be a number of immediate wins for the biological community, such as gaining rapid access to defined lists of compounds that are likely to modulate specific genes, gene families or pathways," says John Overington, Senior Director of Discovery Informatics at Galapagos, who was closely involved in negotiating the transfer.

The Galapagos databases started life at Inpharmatica, a informatics spin-out company of the University College London, UK, acquired by Galapagos in 2006. Inpharmatica began in-house drug discovery projects in 2000 to exploit its technology platform through identifying 'good' targets. "We had a good idea of what made a druggable gene, and started to link this to chemical databases," says Overington. "Nothing was already available that allowed us to make this leap between a protein sequence and a small molecule, peptide or protein therapeutic, which had some defined effect on a target."

So they developed a large-scale structure–activity relationship (SAR) database, StARlite (SARs in the literature), which links targets, functional assay results, and absorption, distribution, metabolism, elimination and toxicity (ADMET) properties to compound structure and target sequence or structure. It currently contains 450,000 distinct compounds — 35% of which meet all of the 'rule of five' criteria for oral drug bioavailability — 2 million bioactivities and nearly 4,000 molecular targets.

"A key feature of our database is that we record the biological effects of making changes to a molecule's structure. If it is made hydrophobic or bigger, does it affect, say, cell penetration or activity?" Overington explains. "These differences are often key to converting an *in vitro* tool compound into a useful drug."

Another major informatics challenge is the difficulty in searching across public and proprietary databases because of different data structures. An integration layer, SARfari, was therefore developed by Inpharmatica to enable companies to incorporate their in-house data into the database. "First we developed a G-protein-coupled receptor platform and then a kinase system," says Overington. "Future plans include another SARfari for antibacterials, to support integrated chemistry-led and biology-led target selection."

In addition, a related database called CandiStore contains the structures, targets and latest development stage of clinical development candidates, including both small molecules and other classes of therapeutics. "The aim is to track drug failures and understand why they failed… it's a work in progress; some areas are well-populated and others will be built using the grant," says Overington.

> "There will be a number of immediate wins for the biological community."

In the long term, Overington explains that they may introduce a deposition mechanism for new data, but they do not want to replicate PubChem, as one of the strengths of StARlite is the curated, consistent nature of the data. This is always going to be a drawback with large repository-like databases such as PubChem, explains its Director Steve Bryant. "Our informatics challenge is how to compare chemicals effectively when different chemists draw them differently," he says. "There are also biological challenges — how to describe what was done. We've asked depositors to provide a bottom-line summary — what are the true positives, the most active chemicals in their primary screens, and so on. These are very informative to non-experts when we have them."

This consistency of data is the big difference between the new EMBL-EBI database and other publicly accessible databases, says Andrew Hopkins, professor of medicinal informatics at the University of Dundee, UK. "It's been normalized for searching and designed to be mined," he says. "Having to download everything and reformat it before it can be searched effectively puts people off, and there's no guarantee that the search would succeed."

Paul Clemons, Director of Computational Chemical Biology Research at the Broad Institute — which hosts ChemBank (http://chembank.broad.harvard.edu), another large curated database of small molecules and biological screens that is freely available — thinks that public access to well-curated chemogenomics databases is crucial. "They let creative academics have access to the sort of data that each pharma company has separately had for many years," he says. "We occasionally hear criticism from industry colleagues that some of the data-mining activities we're doing have been done before. But we're only now getting access to these data sets, so I'm sure academics have sometimes redeveloped analysis methods that were developed secretly in pharma before."

As well as the new opportunities for academic groups to develop chemoinformatics approaches, there is growing excitement at the prospect of free access to the EMBL-EBI data for public drug discovery projects in areas such as neglected diseases. "I want to be one of the first adopters — there's a wealth of ideas we want to try," says Hopkins. "A good example is that we can now improve upon the original druggability analysis for neglected disease pathogens [see http://TDRtargets.org] by linking predicted druggability to sets of chemical tools for screening. In the previous iteration, the chemistries couldn't be disclosed."

"Another example," he continues, "is that we can use this large public data set to build large-scale virtual assay banks using machine learning processes that learn from the underlying data to predict new biological activities of compounds."

Clemons also highlights the value for chemical biology in general. "I think that the biggest benefit in the end will be the ability to make new and more specific tool compounds for cell biological research that leverage the information about what's been made before, and what it did when it was exposed to biological assays."

It is anticipated that scientists will be able to start using the database in early 2009. The funding will first ensure that the data will be available as a complete downloadable database for local installation. User-friendly web-based front-ends and programmable web services are expected to follow.

After 5 years, EMBL-EBI will need to find further funding for maintenance and curation of the database. "The future depends on how it develops," says Thornton. "In the longer term, we are hoping it will be part of Elixir, the large infrastructure project for biological data, but that is a long way off. EMBL-EBI was very keen to acquire these data and there is a lot of scope in the future for developing resources around them."