Open Data, open doors

Clore Laboratory's Mike Cawthorne outlines how the barriers surrounding drug discovery could be overcome if Open Data can be properly utilised

Historically, most research on the discovery and development of new medicines has occurred in commercial settings – pharmaceutical and biotechnology companies. Despite many successes, the low productivity of drug discovery research is well documented, as is the search for new business models in which companies are increasingly looking to collaborate with academic groups to maintain their clinical development pipelines.

Finding new bioactive molecules

In recent decades, most of the pharma industry has used empirical high-throughput screening (HTS) to identify starting points for drug discovery programmes. This approach requires the assembly of millions of physical compound screening samples, the development of miniaturised robotic assays and very large expenditure to set up and maintain. Because of the high costs, there are only a few examples of HTS facilities in an academic setting. Additionally, effective drug development requires knowledge and understanding of the relationship between chemical structure and biological activity. Large bodies of such knowledge have been generated over the years inside companies, but the information is proprietary. However, new technologies and initiatives are making more effective academic drug discovery possible.

Open Access data initiatives

Large-scale databases of structure-activity relationships are becoming available to researchers through a variety of publicly or charitably funded initiatives. The following examples give an indication of what is becoming available. ChEMBL is a manually curated database of molecules, their biological activities and their drug-like properties, maintained by the European Bioinformatics Institute (EBI) at the Wellcome Trust Genome Campus, Hinxton, UK. The current version (19) contains over 1.5 million compound records, with almost 13 million assay results across 10,500 drug targets. ToxBank is making toxicogenomics data available, a data type previously very poorly represented in the public domain. Combined with longstanding public databases such as the Protein DataBank of the three-dimensional structures of drug target proteins, an invaluable resource for drug discovery is being created.

Chemoinformatics

The information in these repositories is only useful if it can be analysed and mined to guide discovery projects. In parallel, chemoinformatics tools are being developed that can interrogate these databases and build sophisticated models that can be used to design new compounds. 'Virtual screening' is analogous to experimental HTS but occurs in a computer and involves, for example, the search for database compounds that can fit into a protein binding site, or with shape similarity to a known active compound. Only the best scoring compounds are selected for experimental testing. These methods are so computationally efficient that they can be performed not only on models of existing compounds, but also on much larger numbers of compounds that could be made, but which do not yet exist. This allows exploration of 'chemical space' that is orders of magnitude larger than that available to HTS; potentially billions of compounds.

Another application is to predict new activities of existing compounds, for example, by using the structure-activity relationships that are available in ChEMBL. A molecular similarity query identifies 'hit' compounds that have chemical similarity to a query compound. In many cases, the known biological activities of these hits will be the same as those of the query compound, but it is also possible that additional biological activities are observed for the hits. This suggests that the query compound may also be active against these other biological targets. Drug repurposing, prediction of side effects and prediction of the mode of action of a drug are applications of this type of approach.

Impact on academic drug discovery

The initiatives described above reduce or eliminate many of the barriers that previously limited drug discovery outside large commercial organisations. The ability of virtual screening methods to focus experimental resources on a relatively small number of compounds obviates the need to develop very expensive HTS facilities or the time involved in developing robust automated assays – in any case, many of the novel drug targets emerging from biology research may not be amenable to an HTS approach.

Once a lead has been identified, the wealth of data now publicly available can be used to guide its optimisation, not just for compound efficacy, but also for pharmacokinetic properties such as compound half-life and metabolism – an area about which academic researchers have historically been less aware.

Taken together, these new technologies combine with the traditional academic strength in understanding of disease biology to provide an exciting future for drug discovery in academia.

Professor Paul Finn, at the Buckingham Institute of Translational Medicine, is a leading expert in chemoinformatics, and his team are open to collaborations and joint grant applications.

THE UNIVERSITY OF BUCKINGHAM

Mike Cawthorne
Professor of Metabolic Diseases
Clore Laboratory
University of Buckingham

tel: +44 (0)1280 820309

mike.cawthorne@buckingham.ac.uk
www.buckingham.ac.uk/bitm