Published in final edited form as:

Lancet Oncol. 2016 March ; 17(3): 286. doi:10.1016/S1470-2045(16)00095-4.

The canSAR data hub for drug discovery

Cindy H Chau, Barry R O'Keefe, and William D Figg

The recent explosion of complex biological data has generated an unprecedented wealth of information in the arena of drug discovery. The ability to integrate these datasets and learn how to analyse the complexity of the data as a whole across different scientific disciplines will determine the translational relevance derived from large-scale omics data. This need has given rise to a growing number of software tools and bioinformatics portals designed to efficiently access and use large amounts of data to accelerate drug discovery. However, navigating across several databases can be time consuming and cause equivocal and disjointed outcomes.

The demand for a resource that brings together diverse data from various scientific domains resulted in the creation of canSAR by the Cancer Research UK Cancer Therapeutics Unit (The Institute of Cancer Research, London, UK). Launched in 2011, canSAR is a publicly accessible database that functions as a one-stop shop approach to everything related to cancer-drug discovery. The range of multidisciplinary data spans from genomics, pharmacological, drug, and chemical data, to protein networks, cancer cell lines, and clinical data. In its latest update, canSAR (3.0) continues to expand its comprehensive approach with the addition of a 3D-structure-based druggability-assessment feature for almost 3 million cavities on the surface of nearly 110 000 macromolecules. The overall goal is to enable rational target selection and validation, and to help precision-guided translational research to ultimately improve the drug discovery process. For example, researchers can query the known characteristics of a protein or target, examine expression data, determine interacting or binding partners, and identify any potential small-molecule inhibitors through in-silico screening. Additionally, extended tools such as bioactivity profiling help to categorise drug activities, which is in turn helpful for compound profiling. The end result is a portal for rapid access to quality-controlled information and researcher-friendly tools with which to integrate disparate data streams. One particularly useful feature is the readily accessible documentation of the methods and step-by-step outputs of the data-processing tools available.

Such an intuitive interface is an invaluable asset to cancer-translational researchers. The importance and usefulness of a data-aggregating resource is directly related to its ongoing stewardship and governance of the data. In its current iteration, canSAR has undergone 5 years of refinement and improvement to ensure that relevant data is linked to the original sources and updated monthly. The site also enables easy access to data sources and the most recent date that each of these sources has been updated. Properly curated and annotated data is essential, because the advent of large omic datasets derived from increasingly precise technologies has led to an exponential rate of scientific development and progress without a concurrent increase in the capacity of individual researchers to accurately assess the resulting

data. This reality needs increased transparency and data sharing, which could save time and costs for all.

Although the canSAR site makes single-stop assessments of data possible, comparative benchmarking should still be done with other integrated databases such as the Centre for Therapeutic Target Validation (CTTV) platform, which was recently launched and developed through a collaborative effort by GlaxoSmithKline (Middlesex, UK), the Wellcome Trust Sanger Institute (Hinxton, UK), and the European Bioinformatics Institute (Saffron Walden, UK). CTTV also draws from public-data resources to collate large sets of information about the associations between potential drug targets and diseases from various types of data, and provides the means to build data-analysis pathways tailored to individual research interests. Although both canSAR and CTTV function as powerful hubs for effective target validation and use many of the same basic sources of data, canSAR leverages on its cancer-focused annotation of clinical and biological pathway data to good effect with expansive integrated links. As the database continues to become more sophisticated, the maintenance of a navigable web interface remains a priority; canSAR relies upon feedback from end users and the willingness to regularly upgrade capabilities and the user experience to stay current. Keeping active users engaged will help to improve the performance metrics of the site and result in continuous improvement as a reliable and functional platform.

Although these databases look promising, they will need to deliver tangible benefits in terms of improving the efficiency of drug discovery to justify their methods. What remains to be seen is how researchers can successfully interpret the collective data for innovative target selection, and how well these data can be integrated into widely used, user-modifiable interfaces such as Pipeline Pilot and D36O. With such tools in hand, researchers should at least have a strong sense of how much information has already been published and where both the large data sets and published literature coalesce around tangible questions that require answering. It is often in these gaps of knowledge that the missing pieces of the drug-discovery puzzle are found.