BIOMEDICAL DATA JOURNAL
datajournals.eu

Zhang Y. *Biomed Data J.* 2015; 1(1): 39-41
http://dx.doi.org/10.11610/bmdj.01107

ISSN 2367-5322

***datajournals.eu:* The European Open Data portal**

Policy Paper

Open-access and Structured Data in Drug Discovery

Yixin Zhang

B CUBE, Center for Molecular Bioengineering, Technische Universität Dresden, Dresden, Germany
http://www.bcube-dresden.de/

ARTICLE INFO:

RECEIVED: 24 Sep 2014
REVISED: 15 Nov 2014
ACCEPTED: 17 Dec 2014
ONLINE: 06 Jan 2015

KEYWORDS:

Open Access
Drug Screening
Drug Discovery
Image Processing
DNA Library
Datasets as Topic

ABSTRACT

A data journal in the biomedical field is an innovative task with potential benefits to the scientific society, the pharmaceutical and biotechnology industries, as well as the medical institutions and authorities. It can serve as a source to stimulate, using computational modeling and bioinformatics, the bridging of biomedicine and basic biology researches and connecting pharmaceutical and biotechnology industries with academy. In the era of various high throughput technologies and big data, it could become a powerful driving force to combine expertise to understand the complexity of life and to catalyze new innovative therapeutics and diagnosis.

The developments of various high-throughput and/or high-content drug-screening techniques are aiming not only to probe a large number of chemical compounds, but also to obtain deep insights into the molecule/molecule interaction associated with the diversity of chemical space, the structure-activity relationship, as well as the pharmacological mechanism. Moreover, the cell-based drug-screening approaches can shed new light on the dynamics and regulation of proteins and genes associated with various physiological and pathological states.

Recently, the European Union-funded OSL (OpenScienceLink) project has launched an open-access journal to provide the urgently needed platform, through which the scientists can publish and share not only their results, but also original data. While many emerging interdisciplinary research topics in biosciences are also a part of the "big data" concept (omics), the abundant information which resulted from various technological advances has given researchers unprecedented access to resources. However, they also pose many daunting challenges. For example, very strong assumptions are often made about mathematical properties that may not at all reflect what is really going on at the level of physicochemical and biological micro-processes. "Big data," no matter how comprehensive or well analyzed, needs to be complemented by "big judgment."[1]

The "omics" methods are aiming to obtain deep insights into the science that governs the biological processes in pathophysiological conditions as well as the therapeutic effects of biomedical treatments through an all-encompassing approach. However, such promises could only be fulfilled when we can interpret the empirical observation not only in a meaningful biomedical context, but objectively, thus preventing biased deduction. The big data collected using various tech-

✉ Corresponding Author: TUD, B Cube, Arnoldstrasse 18, 01307 Dresden; E-mail: yixin.zhang@bcube-dresden.de

niques in the field of drug discovery is basically similar to the other fields of biosciences, involving e.g. physics that operate the instrumentation and analysis; chemistry to shed light on the structure-activity relationship; structural biology to understand the protein functions; medicine to interpret the biological and therapeutic effects; and moreover, bioinformatics for data processing and mining; and mathematics for statistical analysis and modelling. Since we are still at the dawn of such complex and large amount of information, methods to analyse the data comprehensively and objectively are often lacking.

To share data represents an attractive way to tackle the challenges; sharing allows researchers with different backgrounds to give their insights, especially people not directly involved in the drug discovery. On the one hand, physicists and engineers, and on the other hand bio-informaticians and mathematicians can analyse data through bottom-up and top-down approaches, respectively. The question is no longer whether original datasets should be shared, but how to establish an integral platform for scientists across many different fields and, using datasets collected in different experiments and instrumentations, how to make this a common and well-rewarded part of the research culture.

Traditionally, the datasets involved in drug discovery campaigns are mainly related to the sizes of chemical libraries as well as the number of target proteins when the target-based screening approach is used. Therefore, the amount of data is determined by the number of chemical compounds (typically in the range of thousands to millions of compounds), the number of target proteins (typically in the range of 10 to 100 proteins), and the number of concentrations and replicates. The amount of information and its complexity could be further increased when structural biology and bioinformatics are applied to shed light on the structure-activity relationship. In recent years, there has been a big boom of data size in drug discovery research (as discussed in the following paragraphs). While these technological advances are often achieved through the synergy of distantly related disciplines, procession of such complex datasets in many cases remains a mathematical and computational challenge. To share the information, anyone can download the data themselves, generate models and carry out analysis using their own methods. Such an open-access platform could have unexpected impact on the development of science and technology, especially in the fields involving interdisciplinary, complex and big data. This is analogical to what open source software has done for evolving information technology, or free-access free-content internet encyclopaedia (e.g. Wikipedia) has contributed to general reference work. In the following paragraphs, two examples will be highlighted to illustrate developments, as well as the respective challenges.

Image-based high-content screening as an emerging phenotypic screening technology has been quickly developed in the past few years.[2] It aims to address some intrinsic disadvantages of target-based high-throughput screening. The developments in many seemingly unrelated fields, ranging from fluorescence microscopic technology to computation and image analysing algorithms have made such an approach possible. It leads to datasets much larger than what we have used to experience in high throughput screening experiments. However, although more and more drug screening campaigns have been carried out using this approach, translation of big data into high-content and medically relevant information lags far behind. The challenges include practical difficulties handling large high-content screening data as well as designing disease related model systems. The latter challenge has led to a current trend to replace imaging of cells in 2D culture with imaging of cells in 3D bio-matrices which has further increased the data size and complexity for imaging, data processing and analysis.

Another important emerging technology in the field of drug discovery is DNA-encoded chemical library.[3] As a chemical analogue to peptide and antibody display technology, it can be used to construct libraries of billions of compounds. The selection procedure coupled with a deep sequencing method allows the selection against one protein performed in a single well and selections against many different targets to be performed in parallel. Using genetic code as information storage for chemical structures has resulted in chemical libraries of unprecedented size, as well as the large datasets of many different formats, including raw sequencing data and billions of chemical structures. Although recently many chemical libraries have been synthesized and many selection experiments have been carried out using this approach, the strength of this method in drug discovery remains to be demonstrated. The challenges include establishing a statistical model and algorithm for hit identification, designing selection experiments under biochemically relevant conditions, and increasing the structural diversity of the chemical libraries. The binding of protein to DNA can be already directly visualized during the sequencing process.[4] If this technology can be ap-

plied to a DNA-encoded chemical library, it will lead to a direct on-chip hit identification with biophysical binding profiles. The feasibility of this approach remains to be demonstrated, because the real time imaging process to acquire the binding profiles will lead to further increase of data size and complexity, as compared with the current methods for DNA-encoded chemical library.

Both high-content screening and DNA-encoded chemical library are examples of many emerging techniques in the field of drug discovery which involve large and complex datasets. The establishment of an open-access and structured data journal will provide the opportunity to allow people from different fields to tackle the challenges in the era of big data. We will welcome contributions on datasets for, but not limited to:

- Using bioinformatics and computational methods in drug screening
- Simulation of cellular networks in pharmacological treatment
- Cell-based high-content drug screening and image analysis
- Drug screening using chemical library to probe protein-ligand binding or enzyme inhibition
- DNA and RNA-array in drug screening
- DNA-encoded chemical library
- Structural biology dataset in drug screening
- Combining bioinformatics and computational methods with experimental methods to establish biochemical and biological models.

Of particular interest are original works on:

- Bioinformatics and computational methods helping the development of general and open source public databases
- Computational methods and algorithms for analysing large datasets in drug screening
- Computational methods and algorithms for inter-database data analysis and mining.

Acknowledgement

The preparation of this manuscript was supported by the EU-funded project OpenScienceLink (Grant agreement 318652).

References

[1] Shah S, Horne A, Capellá J. Good Data Won't Guarantee Good Decisions. Harvard Business Review. 2012 Apr;90(4). Available from: https://hbr.org/2012/04/good-data-wont-guarantee-good-decisions.

[2] Singh S, Carpenter AE, Genovesio A. Increasing the Content of High-Content Screening: An overview. J Biomol Screen. 2014 Apr 7;19(5):640-650. DOI: 10.1177/1087057114528537.

[3] Nutiu R, Friedman RC, Luo S, Khrebtukova I, Silva D, Li R, Zhang L, Schroth GP, Burge CB. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nat Biotechnol. 2011 Jun 26;29(7):659-64. DOI: 10.1038/nbt.1882.

[4] Goodnow, RA, editor. A Handbook for DNA-Encoded Chemistry: Theory and Applications for Exploring Chemical Space and Drug Discovery. Hoboken, NJ: Wiley; 2014.