ELSEVIER

Mining of existing biological data makes it possible to transfer knowledge between originally independently pursued drug discovery projects, thereby enhancing our understanding of biological mechanisms.

The opportunities of mining historical and collective data in drug discovery

Q2 **Anne Mai Wassermann, Eugen Lounkine, John W. Davies, Meir Glick and L. Miguel Camargo**

In Silico Lead Discovery, Novartis Institutes for Biomedical Research, 250 Massachusetts Avenue, Cambridge, MA 02139, USA

Vast amounts of bioactivity data have been generated for small molecules across public and corporate domains. Biological signatures, either derived from systematic profiling efforts or from existing historical assay data, have been successfully employed for small molecule mechanism-of-action elucidation, drug repositioning, hit expansion and screening subset design. This article reviews different types of biological descriptors and applications, and we demonstrate how biological data can outlive the original purpose or project for which it was generated. By comparing 150 HTS campaigns run at Novartis over the past decade on the basis of their active and inactive chemical matter, we highlight the opportunities and challenges associated with cross-project learning in drug discovery.

Introduction

Advances in screening technologies and an increasing awareness of the value of biological data utilization have led to the accumulation of diverse small molecule bioactivity data in public and corporate repositories [1,2]. The wealth of information ranges from quantitative (such as gene expression, cellular phenotypes and protein phosphorylation) to categorical (such as clinical adverse events) data. The availability of collective, large-scale datasets in the public domain and institutional, historical data generated over time by companies has led to the recognition that experiments have utility beyond their initial intended purpose. Therefore the design, execution and interpretation of an experiment, in particular HTS, must be done in the context of collective and institutional knowledge.

It has been increasingly recognized that historical and/or multidimensional datasets can be used to characterize and compare small molecules (Fig. 1a). The generation of biological descriptors from such datasets has enabled mechanism-of-action (MoA) hypotheses such as molecular target identification, prediction of phenotypic drug response (e.g. preclinical toxicity and clinical adverse events) and repositioning of existing drugs. Furthermore, biological descriptors have been applied to virtual screening, potency prediction and subset design.

Whereas small molecules can be characterized by their outcomes in different biological assays, it is much less realized that small molecules can also characterize assays. In a similar manner to

Corresponding authors: Wassermann, A.M. (annemaiwassermann@gmail.com, Camargo, L.M. miguel.camargo@novartis.com)

Anne Mai Wassermann
obtained her Masters in molecular biomedicine and her PhD in computational life sciences from the University of Bonn, Germany. During her PhD, she worked on the computational analysis of structure–activity relationships. In 2012, she joined the *In Silico* Lead Discovery group at the Novartis Institutes for Biomedical Research in Cambridge, MA, as a Presidential Postdoctoral Fellow. At Novartis, she investigated the use of bioactivity profiles for small molecule subset design, virtual screening and mechanism-of-action prediction. In October 2014, she finished her postdoctoral training to take on a position in the Molecular Informatics group at Pfizer, Cambridge, MA.

Eugen Lounkine
is a research investigator at Novartis Institutes for Biomedical Research in the *In Silico* Lead Discovery Group. Having studied molecular biomedicine at the University of Bonn, Germany, from 2003 to 2007, Dr Lounkine has since been focusing on computational approaches to life sciences. In 2009, he graduated in computational life sciences at the University of Bonn under Professor Bajorath. In 2010, Eugen joined Novartis Institutes for Biomedical Research (NIBR) in Cambridge, MA, as a postdoctoral fellow focusing on off-target prediction and linking *in silico* and *in vitro* activity profiles to clinical phenotypes and adverse drug reactions.

John W. Davies
is the global head of *In Silico* Lead Discovery at the Novartis Institutes for Biomedical Research in Cambridge, MA. John has a PhD in chemistry from University College London, UK.

Meir Glick
is the head of the *In Silico* Lead Discovery group at the Novartis Institutes for Biomedical Research in Cambridge, MA. His current research focus is the application of chemometric and systems chemical biology approaches to bridge the phenotype–target and target–lead gaps in drug discovery. Previously, he was a trained as a postdoc in the area of computational chemistry at the University of Oxford, UK.

L. Miguel Camargo
is a senior investigator at Novartis Institutes for Biomedical Research in Cambridge, MA. He has several years of drug discovery experience working as a computational biologist across different stages of drug development. His current research areas include the application of systems biology methods to lead discovery as well as how to use public and institutional knowledge for hypothesis generation and validation. Before Novartis, Miguel was with Merck & Co. for 12 years. Miguel has a PhD from the University of Cambridge, UK.

1359-6446/06/$ - see front matter © 2014 Published by Elsevier Ltd. http://dx.doi.org/10.1016/j.drudis.2014.11.004

Please cite this article in press as: Wassermann, A.M. et al. The opportunities of mining historical and collective data in drug discovery, Drug Discov Today (2014), http://dx.doi.org/10.1016/j.drudis.2014.11.004

GLOSSARY

Biomarker an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention that can be objectively measured and evaluated
Drug repositioning administration of a known drug for a new indication
High-content screening a phenotypic screen where the phenotype is morphological changes in a cell. The screening technology mostly uses digital microscopy and subsequent digital image processing to capture multiple readouts simultaneously
Microarray 2D array on a solid surface where each spot contains biological material. For example, in DNA microarrays small amounts of a specific DNA sequence are attached to each spot. They are subsequently hybridized to the cDNA in a sample. The hybridization is quantified and gene expression levels in the sample are determined
Neural network more precisely artificial neural network: a machine-learning method inspired by biological nervous systems. Highly interconnected nodes process information in parallel to solve a problem. As with the human brain, neural networks learn by example
Next-generation sequencing (NGS) high-throughput sequencing that produces a high number of sequences in parallel and where development was pushed by the demand for low-cost sequencing. NGS is used for genome sequencing, transcriptome profiling, DNA–protein interactions and epigenome characterization
On- or off-target an on-target is the intended, primary target of a drug. Through this target it achieves its desired therapeutic effect. An off-target is a secondary target that is modulated in addition to the primary target. In contrast to the on-target, it does not contribute to therapeutic efficacy but its modulation can cause undesired side-effects
Pharmacodynamics study of how a drug affects the body (at different concentrations)
Pharmacophore a set of molecular features present in a ligand in a specific geometric arrangement. This feature set is required for molecular recognition by the biological target
Phenotypic screen a screen that aims at identifying substances that modulate the phenotype of a cell or an organism in a defined manner. In contrast to target-based screens, the molecular target of an active agent is not known a priori (i.e. the desired phenotype can result from different mechanisms-of-action)
Polypharmacology the activity of a compound at multiple targets
Primary data high-throughput screening activity data. Typically, activities are based on measurements at a single compound concentration
Resource description framework a method for conceptual description of information using defined syntax and subject-predicate-object (triplet) expressions. Example: drug (subject) binds to (predicate) protein (object)
Scaffold hopping the process of finding a small molecule that has the same desired bioactivity but a different chemotype (i.e. a substantially different chemical structure)
Semantic web a movement to make information in the web readily interpretable by machines that can be used to search, share and integrate information more easily. This requires a better structuring of documents and common data formats in the web
Support vector machine a supervised machine-learning approach to separate two different classes. The method projects the input training data into a higher dimensional descriptor space and finds a hyperplane that best separates the two classes in that space. A test object is projected into the same space and classified based on which side of the hyperplane it is located
Two-dimensional polyacrylamide gel electrophoresis (2D-Page) a gel electrophoresis technique to separate proteins in mixtures. The proteins are separated by two different properties in two dimensions (i.e. in two directions that are 90° apart) on 2D gels

how we assign properties to compounds from biological descriptors (i.e. different assay endpoints) and learn from those similarities, we can use biological information on screened compounds to improve our understanding of the properties and relationships of the assays themselves (Fig. 1b). For example, phenotypic screening is currently experiencing a renaissance in drug discovery [3] and, in many cases, the same phenotypic assay readout can result from different modulations of a biological network. Similar small molecule chemotypes active in multiple assays can indicate that similar biological mechanisms are being perturbed. Therefore, by comparing different screening campaigns through the activity of their compounds, relationships between the biology of initially independent assays can be revealed. It also goes, therefore, that relationships between drug discovery projects (e.g. modulations of the same or similar pathways) can enable knowledge transfer between projects. Additionally, such comparisons can also help determine whether there are other factors (e.g. assay technology or chemical composition bias) that could link to more-appropriate assays or compound libraries.

In this review, we tackle both viewpoints on the small-molecule-profile and assay-profile landscapes and highlight the value of using institutional and collective knowledge. First, we give an overview of applications of different biological descriptors that have been introduced to characterize small molecules and highlight successes that have been achieved with these approaches so far. Second, we describe the lessons learnt from a systematic analysis of more than 150 HTS assays at Novartis. Caveats such as assay similarities resulting from technology biases, rather than common biological mechanisms of active compounds, are outlined, along with a fair assessment of the opportunities and challenges associated with cross-project learning. Our review aims to highlight the fact that broad small molecule profiling offers the opportunity to increase our understanding of biology further, as well as that of the compounds and the technologies we use in addition to simply identifying chemical matter to modulate an intended target or phenotype.

Biological signatures of small molecules and their applications

Compounds have been described by a variety of biological signatures, ranging from biochemical profiles to gene expression and clinical adverse events. Although the methods to derive and compare them vary substantially, there are four generic applications of bioactivity profiles. (i) MoA elucidation – aims at identifying the protein or an otherwise well-defined, chemically

Please cite this article in press as: Wassermann, A.M. et al. The opportunities of mining historical and collective data in drug discovery, Drug Discov Today (2014), http://dx.doi.org/10.1016/j.drudis.2014.11.004