Published in final edited form as:

Drug Discov Today Technol. 2015 July ; 14: 31–36. doi:10.1016/j.ddtec.2015.03.004.

Reporting biological assay screening results for maximum impact

Evan Bolton

National Center for Biotechnology Information, Bldg. 38A/8S810, National Library of Medicine, U.S. National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA.

Abstract

A very large corpus of biological assay screening results exist in the public domain. The ability to compare and analyze this data is hampered due to missing details and lack of a commonly used terminology to describe assay protocols and assay endpoints. Minimum reporting guidelines exist that, if followed, would greatly enhance the utility of biological assay screening data so it may be independently reproduced, readily integrated, effectively compared, and rapidly analyzed.

Graphical Abstract

Introduction

The ability to perform biological assay screening is ubiquitous. Many universities have both the appropriate equipment and access to large chemical substance libraries necessary to produce vast quantities of bioactivity data. For example, the U.S. National Institutes of

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of Interest

The author declares no conflict of interest.

Health (NIH) Molecular Libraries Program (MLP) project (1) unleashed a torrent of publically accessible biological assay screening results over its ten year lifespan. Most of these MLP screening centers were located at universities. Given the public availability of assay screening data, attention has turned to comparison and analysis.

MLP funded the creation of the PubChem resource (2–4) in 2004 at the National Library of Medicine (NLM, part of NIH) to archive and host its output, a sizeable +200 million biological assay screening endpoints resulting from thousands of biological high throughput screening (HTS) assays, involving thousands of biological targets of keen scientific interest, performed on hundreds of thousands of small molecule chemicals. The emergence of this unprecedented access to public domain biological assay screening data was enhanced a few years later at the European Bioinformatics Institute (EBI) by the ChEMBL project (5), a free resource providing bioactivity data for small molecules manually abstracted from tens of thousands of journal articles found in key medicinal chemistry journals. As data systems containing large quantities of bioactivity screening data, PubChem and ChEMBL were not new. The novelty was the depth and breadth of biological assay screening information they provided for scientists (worldwide) to freely use, including coverage of biological targets of acute therapeutic interest. These projects provided a venue and way to disseminate new contributions of biological assay screening data for the public.

In a relatively short period of time the availability and accessibility of open screening data went from near nothing to a deluge. Resources like PubChem and ChEMBL added substantial value to this information by integrating it together and with other scientific resources; however, harnessing this treasure trove involves difficulties that continue to the present day. In the case of PubChem, many details about an assay are available only in non-structured text (making it difficult to compare assays) or are not present at all (requiring contact with the data contributor for missing details). The lack of enforced standards and the lack of expert manual curation in PubChem means that the same biological assay reported by different labs (or even the same lab) may appear dissimilar, with variations in the assay description, readouts reported, target definition, and approaches to determining bioactivities, as it depends on the individual data contributor to decide how best to annotate their data. In the case of ChEMBL, despite expert manual curation of data from publications, many biological assay protocol details are not abstracted, preventing direct comparison between assays without reading the publications. Furthermore, a lack of consistent bioactivity data reporting between journals (or within the same journal) means some important details about biological assay screening results may be absent, requiring contacting authors for further details. The inadequacies and inconsistencies of bioactivity data reporting limits the extent the data can be integrated, compared, and analyzed.

The pharmaceutical industry has developed best practices, including terminologies and informatics platforms, to help normalize and analyze biological assay screen data within their organizations (6–10). Unfortunately, these tend to be proprietary and closed off from the open data space. A positive sign that these best practices may become more generally accessible includes the “Assay Guidance Manual” eBook (11) developed in collaboration between Eli Lilly & Company and the National Center for Advancing Translational Sciences (NCATS, part of NIH), that seeks to help investigators identify probes that

modulate the activity of biological targets, pathways, and cellular phenotypes. Designed to include an open submission and review process, it may help to encourage further contributions of useful terminologies and approaches to handling and analyzing biological assay screening data known within proprietary data spaces.

When PubChem and ChEMBL began, vocabularies, ontologies, and minimum reporting standards for bioassay screening data were not commonly available. Today, this is no longer the case. Biological assay screen and bioactivity reporting standards (12), guidelines (13), and terminologies (14–16) are available and evolving as are their applications for annotation and analysis purposes (17–20). Reviewed here are approaches to minimum reporting standards for biological assay screening results with an emphasis towards important considerations to maximize the utility of published data.

Emerging standards for minimum data reporting

Minimum assay HTS reporting guidelines

A lack of reporting standards for biological assay HTS data prompted investigators from three MLP screening centers to suggest guidelines in 2007 on the key information that should be provided for every HTS assay (13). Five core areas were emphasized: *Assay*, *Library*, *HTS Process*, *Post-HTS Analysis*, and *Results*. The *Assay* covers the nature, strategy, reagents, and protocol of the screen. The key aspects of the *Assay* include a description of the assay logic, which would include adequate description of positive and negative controls, sensitivity to types of assay interference, sources of all reagents, and a clear summary of the protocol. The *Library* describes the constituency of the samples (such as the type of chemicals and core scaffolds), how the samples are presented to the assay, sample source (vendor/synthesized), and quality control procedures. The *HTS Process* covers relevant description of aspects of assay controls (and their arrangement), the number of assay plates, assay duration, dispensing systems, detectors, data outputs, correction/normalization procedures, and assay performance metrics (e.g., Z factor, Z', etc.). The *Post-HTS Analysis* describes how actives were selected, how they were retested, how the sample identity was confirmed, and whether the actives were purified or resynthesized. The *Results* give and describe the outcomes of the HTS assay, including confirmed actives, how initial actives were later disproved, and (relative activity) ranking strategies of screened samples.

Minimum Information About a Bioactive Entity (MIABE)

Considering the quantity of biological activity data published in the literature, and a clear lack of consistency in how they are reported, the Minimum Information About a Bioactive Entity (MIABE) guidelines were published in 2011 (12). Created by a diverse set of representatives from pharmaceutical companies, universities, and bioactivity data resource providers, recommendations were established to delineate the key data necessary to maximize the benefit of published bioactivity results. A primary aim of MIABE is to help provide standardization of reporting and collection of data in an effort to improve data quality and availability. MIABE emphasizes the use of controlled vocabularies to describe bioactivity information. It recommends the availability of publication data in an easy to exchange file format. Complementary to the earlier HTS bioassay guidelines, MIABE

emphasizes three core areas: *Contact*, *Compound*, and *Assay*. The *Contact* includes a stable primary contact (person and/or institution) responsible for the bioactivity result. The *Compound* description includes three main parts to identify pertinent details on the entity whose bioactivity is being measured, including *molecule properties*, *molecule production*, and *physicochemical properties*. The *molecule properties* emphasize the primary name, molecule type, IUPAC chemical systematic name, IUPAC InChI, chemical structure, salt and the (final) bioactive prodrug/metabolite form, as known. The *molecule production* provides details about the purity and how the sample was acquired, including applicable details on the synthetic route, isolation procedure, or manufacturer (including product number). The *physicochemical properties* include molecular weight (and whether the weight includes waters of hydration and salt), experimentally determined properties like water solubility and Log P (also, Log D, when appropriate), and computed properties (including the program used and version). The *Assay* description includes separate sets of guidance by assay type, including *in vitro* (cell-free), cellular, whole organism, pharmacokinetic, and toxicological studies. The guidance emphasizes key details and parameters about the assay should be provided so the biological assay results can be reproduced. For *in vitro* assays, this includes aspects like primary target, assay details, assay parameters, delivery systems, assay results, and secondary gene targets. Cellular assays should include details like cell type, culture conditions, agonism/antagonism indications, assay results, secondary cellular assays, and toxicological observations. Whole organism studies should include details including organism specific information, disease model, dosing route, results, toxicological observations, and drug-drug interactions. Pharmacokinetic studies should include details like absorption, protein binding, dosing route, dosing schedule, half-life, $V_{\max}$, distribution volume, bioavailability, metabolites and excretion information.

Ontologies for biological assay screening

Terminologies and ontologies to describe biological assay screening data were in their infancy when PubChem was first launched in 2004. Since this time two key ontologies have been created: Open Biomedical Investigations (OBI) (14) and BioAssay Ontology (BAO) (16). Complementary in nature, OBI is more general purpose to describe experiments, while BAO is specific to the biological assay screening domain.

Open Biomedical Investigations (OBI)

OBI (14) provides terminology to represent biomedical investigations which it considers as a process with several parts such as study design, execution, and results. Each study part may have many subparts. OBI's comprehensive nature is both good and bad. While it provides extensive means to describe an experiment, it is somewhat expansive and can be intimidating for the casual user to wield. It is easy to envision two scientists, regardless of proficiency, documenting the same experiment in different (yet equivalent?) representations with OBI. Training and further community guidance on best practices when using OBI to describe biological assay screening experiments may be necessary to ensure community wide consistency in its use. In addition, if LIMS and ELN system providers were to harness OBI in a consistent fashion, it is not difficult to imagine that all necessary OBI annotation describing an experiment could be generated automatically as a report for inclusion with a

publication. Considering MIABE strongly recommends use of standardized vocabulary and key details about an experiment to be available in a format for ready data exchange, a combination of LIMS/ELN providers and OBI could be a powerful combination to help improve data sharing and experiment cross comparisons, including biological assay screening.

BioAssay Ontology (BAO)

BAO (16) was created by researchers within one of the former MLP screening centers to help standardize, organize, and semantically describe biological HTS assays like those found in PubChem. In its latest form, BAO 2.0 (15) emphasizes six major parts of an assay: *Bioassay Component*, *Format*, *Method*, *Biological Component*, *Screened Entity Component*, and *Endpoint Component*. It includes other constructs that help to define the organization (screening center, equipment/reagent manufacturer, etc.), people (who did the research), role (context and actions performed by an entity), and quality (characteristics about an entity). It also includes a *Properties* construct that enables relationships between different concepts. The *Bioassay Component* describes assays and their context. The *Format* describes the biological model system (biological and chemical features of the experimental system). The *Method* describes how the assay is performed. The *Biological Component* describes the biology of the assay. The *Screened Entity Component* describes the chemical or biological substance being tested. The *Endpoint Component* describes results of the biological perturbations. By distilling the assay screening constructs in use and organizing them, BAO can be used to assign relevant categorizations to assays and to identify closely related assays rapidly. BAO is designed specifically to handle biological assay screens, as opposed to more generic modeling of biomedical investigations, and it provides the domain-specific constructs needed such as assay design, detection technologies and standardized endpoints. Future versions of BAO are intended to be harmonized with OBI, so terms are consistently used within BAO, and with BAO enhancing OBI by providing compatible domain-specific assay screening extensions.

Adding structure to legacy biological assay screening results

Almost all publically available bioassay screening data are legacy data and lack the benefit of a standardized vocabulary and structured description (e.g., of assay protocols). This complicates the ability to integrate, search, compare, and analyze biological activity data. Large scale efforts to retrofit the legacy data to include the benefits of standardized vocabulary and other structural improvements are known. Two examples are showcased here.

BioAssay Research Database (BARD)

The BioAssay Research Database (BARD) (21–22), an MLP funded effort, focused on tasks to (re)annotate, (re)organize, and (re)standardize the MLP assay information found in PubChem. The reason for this was simple. The ability to integrate and reuse the MLP data in PubChem was made difficult by the lack of consistent terminology between screening centers in the assay descriptions and reported results. For example, while key assay details were provided, they were found only in human readable textual descriptions, as opposed to

machine friendly annotations, thus hampering computational cross assay analysis. In addition, some MLP assay screening campaign details were missing or difficult to discern, such as differentiation between a confirmatory screen and a counter screen. Furthermore, while the vocabularies necessary to consistently annotate the data did not exist at the start of the project, they exist now (15). Through BARD, the MLP screening centers set about to consider how to best improve the representation of their data for improved reuse.

Given the large corpus of thousands of assays, adding structure to the MLP assay screening data was no small undertaking (21). To get started, a controlled vocabulary and hierarchy of terms was developed. The minimum vocabulary to compare compounds, assays, and results was generated. Leveraging BAO, the BARD vocabulary uses a project-based scheme that describes biological assay descriptions, experimental conditions, and results. The top levels of the terminology include Assay (protocols), Biology (biological system studied by a protocol), Project (grouping instances of protocols), and Result (measurements). A particular emphasis was placed on assay protocols to ensure methodological connections between disparate projects could be readily identified. They developed a Catalog of Assay Protocols (CAP) that handles relationships between results, assay parameters, and experimental choices. Another purpose of CAP was to enable hypothesis generation by both novice and expert users. It is important to note that while the BARD data dictionary uses BAO terms (especially those dealing with assay protocols) it is not an ontology. Rather, it provides a hierarchical set of terms and concepts. This prevents the BARD dictionary from being used directly for inference purposes and other formal machine modeling of concepts; however, a mapping of BARD terms to BAO exists, when possible, allowing an indirect use of BAO in some cases.

Beyond working towards improving the utility of MLP data, BARD produced an open source technology platform that included a desktop application, web-based client, and downloadable database. Given that the MLP program is completed, it is not clear to what extent and in what form BARD as a platform will continue. Given this uncertainty, it would be an encouraging to see these annotations picked up and harnessed by projects like PubChem and ChEMBL.

Open Pharmacological Concepts Triple Store (Open PHACTS)

A more general but related project to improve annotation of legacy biological assay screening results is found within the Open PHACTS platform (23–25). One way annotation of assay data is improved is by focusing on semantic interoperability. It is unique in that it provides an example of trying to prevent duplicitous integration of drug discovery focused public data sources by private and public organizations. As an open drug discovery informatics platform with significant quantities of bioactivity information, Open PHACTS focuses on three primary priorities: [1] providing a sustainable pharmacological information platform to facilitate information sharing; [2] providing accessible tools to explore the pharmacological data space; and [3] providing a quality assessment layer over integrated data.

Using a subset of available public data resources, such as ChEMBL, Open PHACTS strives to address key issues with data access and licensing encountered with public data sources.

Beyond the informatics platform for data integration and access, Open PHACTS attempts to improve upon data quality and annotation. The focus on data quality is handled in different ways. For chemical structures, a chemistry validation and standardization system component (23) provides metrics on chemical structures representation quality, which can be used to provide feedback for erroneous or incompletely defined chemical structures. For names and identifiers, whether they are chemical or biological, Open PHACTS is developing curated data dictionaries by various means and includes a human curation component, supporting crowdsourcing-style approaches. These data dictionaries are used to identify and validate incorporated target, compound, and bioactivity nomenclature data. As a platform, Open PHACTS supports various workflows (24–25) involving chemicals, targets, pathways, and diseases. This data integration of bioactivity information provides links between chemicals and targets.

Resource Description Framework (RDF) for biological screening data

RDF is a general framework for data interchange on the Web, and part of World Wide Web Consortium (W3C) specifications. It breaks down knowledge into machine readable discrete pieces, called “triples”. A “triple” is a trio of “subject-predicate-object”. For example, in the phrase “atorvastatin may treat hypercholesterolemia,” the subject is “atorvastatin”, the predicate is “may treat”, and the object is “hypercholesterolemia.” The benefit of RDF is that information becomes uniquely addressable using a Uniform Resource Identifier (URI) to name each part of the “subject-predicate-object” triple. These URIs are often web URLs. The RDF statements provide information about things much like a chemical bond describes the nature of the connection between two atoms. In the case of biological screening data, RDF formatted data using ontologies provides the means to analyze information and their interrelationships. It also helps to make it easier to find, share, and combine information, improving the utility and ability of researchers to combine public and private data. This is helped by the availability of open source RDF “triplestores” using the SPARQL query language. Both ChEMBL (26) and PubChem (27) now offer RDF formatted data.

ChEMBL RDF

In late 2013 EBI released a Resource Description Framework (RDF) platform for linked open data (28). ChEMBL is part of the EBI RDF platform and helps to support (in part) the Open PHACTS project. The data model is described here (26) and uses an internally developed ChEMBL Core Ontology (CCO) to describe entities (such as substances, assays, targets, and documents) and includes use of multiple ontologies, including BAO. The EBI RDF platform includes downloadable content, a Linked Data browser, and a SPARQL endpoint.

PubChem RDF

In early 2014, PubChem introduced the PubChemRDF project (27). It currently handles a broad set of fifteen interlinked primary subdomains and their interrelationships as depicted in Figure 1. Similar to ChEMBL RDF, PubChem RDF harnesses ontological frameworks to help facilitate PubChem data sharing, analysis and integration across scientific domains. Unlike ChEMBL RDF, PubChem RDF did not develop a local ontology, rather it

extensively leverages existing ontologies, except when an ontological description is not available, and a PubChem vocabulary term is used. PubChemRDF includes summary biological activity screening data and uses BAO concepts of *BioAssay*, *MeasureGroup*, and *Endpoint* and other OBI terms to organize and describe biological screening data. Despite organizational differences, PubChemRDF resembles ChEMBL RDF in the degree of information provided. In terms of capabilities PubChemRDF does not provide a SPARQL endpoint and must be downloaded and loaded/queried on local computing infrastructure. In many ways, PubChemRDF is a high level overview of PubChem biological screening data, lacking most of the details about the experiment as can be found within the BARD system. In addition, PubChemRDF lacks the extensive integration of this data as found within the Open PHACTS system. However, PubChemRDF provides substantially more biological screening data than that found in resources like ChEMBL, BARD, and Open PHACTS.

Conclusions

A great wealth of open biological screening data is available in resources like PubChem and ChEMBL. Ontologies (such as BAO and OBI) can help to standardize the description of these experiments and their readouts and can be found in the semantic RDF descriptions in ChEMBL and PubChem. In addition, BARD and Open PHACTS are helping to improve the quality and improve the annotation of PubChem and ChEMBL, respectively. To help ensure adequate reporting of biological screening data, minimum reporting guidance and standards, such as MIABE, are available. If followed, the quality of public (and private) biological screening data will be greatly enhanced for community reuse. If LIMS and ELN providers work with the ontology community describing experiments (such as OBI), one can imagine the ability to export experiment details in a standardized and machine readable format for inclusion with publications will become a trivial exercise, dramatically improving the ability to integrate, compare, combine, and analyze experimental data between scientists, institutes, and data archives.

Acknowledgement

This research was supported [in part] by the Intramural Research Program of the NIH, National Library of Medicine.

References

1. Austin C, et al. NIH Molecular Libraries Initiative. *Science*. 2004; 306:1138–1139. [PubMed: 15542455]
2. Yang Y, et al. PubChem BioAssay: 2014 update. *Nucleic Acids Res*. 2014; 42:D1075–D1082. Database issue, s.l. [PubMed: 24198245]
3. Bolton, E., et al. *Annual Reports in Computational Chemistry*. Vol. 4. Oxford: Elsevier; 2008. PubChem: Integrated Platform of Small Molecules and Biological Activities; p. 217-240.12
4. Wang Y, et al. PubChem's BioAssay Database. *Nucleic Acids Res*. 2012 Jan;40:400–412. 1, s.l.
5. Bento A, et al. The ChEMBL bioactivity database: an update. *Nucleic Acids Res*. 2014; 42:1083–1090. s.l.
6. Ertl P, Selzer P, Muhlbacker J. Web-Based Cheminformatics Tools Deployed via Corporate Intranets. *Drug Discov. Today: BIOSILICO*. 2004; 2:201–207. 5, s.l.
7. Rojnuckarin A, et al. ArQioLogist: An Integrated Decision Support Tool for Lead Optimization. *J. Chem. Inf. Model*. 2005; 45:2–9. 1, s.l. [PubMed: 15667123]

8. Agrafiotis D, et al. Advanced Biological and Chemical Discovery (ABCD): Centralizing Discovery Knowledge in an Inherently Delocalized World. *J. Chem. Inf. Model.* 2009; 47:1999–2014. 6, s.l. [PubMed: 17973472]
9. Sander T, et al. OSIRIS, and Entirely In-House Developed Drug Discovery Informatics System. *J. Chem. Inf. Model.* 2009; 49:232–246. 2, s.l. [PubMed: 19434825]
10. Muresan S, et al. Making Every SAR Point Count: The Development of Chemistry Connect for the Large-Scale Integration of Structure and Bioactivity Data. *Drug Discov. Today.* 2011; 16:1019–1030. 23–24, s.l. [PubMed: 22024215]
11. Sittampalam, G., et al., editors. *Assay Guidance Manual*. Bethesda: Eli Lilly & Company and the National Center for Advancing Translational Sciences; 2004.
12. Orchard S, et al. Minimum information about a bioactive entity (MIABE). *Nat Rev Drug Discov.* 2011 Aug 9.10:661–669. [PubMed: 21878981]
13. Inglese I, Shamu C, Guy R Kiplin. Reporting data from high-throughput screening of small-molecule libraries. *Nature Chemical Biology.* 2007:438–441. [PubMed: 17637769]
14. Brinkman R, et al. Modeling biomedical experimental processes with OBI. *J Biomed Semantics.* 2010 Jun 22.1(Suppl 1)
15. Abeyruwan S, et al. Evolving BioAssay Ontology (BAO): modularization, integration and applications. *J Biomed Semantics.* 2014 Jun 3.5(Suppl 1):S5. Proceedings of the Bio-Ontologies Spec Interest G. [PubMed: 25093074]
16. Visser U, et al. BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. *BMC Bioinformatics.* 2011 Jun 24.12:257. [PubMed: 21702939]
17. Schürer S, et al. BioAssay ontology annotations facilitate cross-analysis of diverse high-throughput screening data sets. *J Biomol Screen.* 2011 Apr 4.16:415–426. [PubMed: 21471461]
18. Vempati U, et al. Formalization, annotation and analysis of diverse drug and probe screening assay datasets using the BioAssay Ontology (BAO). *PLoS One.* 2012; 11:e49198. s.l. 7. [PubMed: 23155465]
19. Clark A, et al. Fast and accurate semantic annotation of bioassays exploiting a hybrid of machine learning and user confirmation. *PeerJ.* 2014 Aug 14.2:e524. [PubMed: 25165633]
20. Balderud L, et al. Using the BioAssay Ontology for Analyzing High-Throughput Screening Data. *J Biomol Screen.* 2015 Mar 3.20:402–415. [PubMed: 25512330]
21. de Souza A, et al. An Overview of the Challenges in Designing, Integrating, and Delivering BARD: A Public Chemical-Biology Resource and Query Portal for Multiple Organizations, Locations, and Disciplines. *J Biomol Screen.* 2014 Jan 17.19:614–627. 5. [PubMed: 24441647]
22. Howe E, et al. BioAssay Research Database (BARD): chemical biology and probe-development enabled by structured metadata and result types. *Nucleic Acids Res.* 2015 Jan.43:D1163–D1170. Database issue. [PubMed: 25477388]
23. Williams A, et al. Open PHACTS: semantic interoperability for drug discovery. *Drug Discov Today.* 2012 Nov 21–22.17:1188–1198. [PubMed: 22683805]
24. Chichester C, et al. Drug discovery FAQs: workflows for answering multidomain drug discovery questions. *Drug Discov Today.* 2014 Nov 20. Epub ahead of print.
25. Ratnam J, et al. The application of the open pharmacological concepts triple store (open PHACTS) to support drug discovery research. *PLoS One.* 2015; 9:e115460. [PubMed: 25522365]
26. ChEMBL documentation. EBI. [Online] <http://www.ebi.ac.uk/rdf/documentation/chembl>.
27. PubChemRDF Release Notes. PubChem. [Online] <https://pubchem.ncbi.nlm.nih.gov/rdf/>.
28. Jupp S, et al. The EBI RDF platform: linked open data for the life sciences. *Bioinformatics.* 2014 May 1.30:1338–1339. 9. [PubMed: 24413672]

Figure 1.
Fifteen primary subdomains of PubChemRDF and their linking relationships. The context of the relationship is omitted for clarity.