A peer-reviewed version of this preprint was published in PeerJ on 9 May 2016.

View the peer-reviewed version (peerj.com/articles/cs-61), which is the preferred citable publication unless you specifically need to cite this preprint.

Clark AM, Litterman NK, Kranz JE, Gund P, Gregory K, Bunin BA. 2016. BioAssay templates for the semantic web. PeerJ Computer Science 2:e61 https://doi.org/10.7717/peerj-cs.61

BioAssay templates for the semantic web

Alex M Clark, Nadia K Litterman, Janice E Kranz, Peter Gund, Kellan Gregory, Barry A Bunin

Annotation of bioassay protocols using semantic web vocabulary is a way to make experiment descriptions machine-readable. Protocols are communicated using concise scientific English, which precludes most kinds of analysis by software algorithms. Given the availability of a sufficiently expressive ontology, some or all of the pertinent information can be captured by asserting a series of facts, expressed as semantic web triples (subject, predicate, object). With appropriate annotation, assays can be searched, clustered, tagged and evaluated in a multitude of ways, analogous to other segments of drug discovery informatics. The BioAssay Ontology (BAO) has been previously designed for this express purpose, and provides a layered hierarchy of meaningful terms which can be linked to. Currently the biggest challenge is the issue of content creation: scientists cannot be expected to use the BAO effectively without having access to software tools that make it straightforward to use the vocabulary in a canonical way. We have sought to remove this barrier by: (1) defining a bioassay template data model; (2) creating a software tool for experts to create or modify templates to suit their needs; and (3) designing a common assay template (CAT) to leverage the most value from the BAO terms. The CAT was carefully assembled by biologists in order to find a balance between the maximum amount of information captured vs. low degrees of freedom in order to keep the user experience as simple as possible. The data format that we use for describing templates and corresponding annotations is the native format of the semantic web (RDF triples), and we demonstrate some of the ways that generated content can be meaningfully queried using the SPARQL language. We have made all of these materials available as open source (http://github.com/cdd/bioassay-template), in order to encourage community input and use within diverse projects, including but not limited to our own commercial electronic lab notebook products.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1932v1 | CC-BY 4.0 Open Access | rec: 7 Apr 2016, publ: 7 Apr 2016

BioAssay Templates for the Semantic Web

Alex M. Clark (corresponding author)
Nadia K. Litterman
Janice E. Kranz
Peter Gund
Kellan Gregory
Barry A. Bunin

Abstract

Annotation of bioassay protocols using semantic web vocabulary is a way to make experiment descriptions machine-readable. Protocols are communicated using concise scientific English, which precludes most kinds of analysis by software algorithms. Given the availability of a sufficiently expressive ontology, some or all of the pertinent information can be captured by asserting a series of facts, expressed as semantic web triples (subject, predicate, object). With appropriate annotation, assays can be searched, clustered, tagged and evaluated in a multitude of ways, analogous to other segments of drug discovery informatics. The BioAssay Ontology (BAO) has been previously designed for this express purpose, and provides a layered hierarchy of meaningful terms which can be linked to. Currently the biggest challenge is the issue of content creation: scientists cannot be expected to use the BAO effectively without having access to software tools that make it straightforward to use the vocabulary in a canonical way. We have sought to remove this barrier by: (1) defining a bioassay template data model; (2) creating a software tool for experts to create or modify templates to suit their needs; and (3) designing a common assay template (CAT) to leverage the most value from the BAO terms. The CAT was carefully assembled by biologists in order to find a balance between the maximum amount of information captured vs. low degrees of freedom in order to keep the user experience as simple as possible. The data format that we use for describing templates and corresponding annotations is the native format of the semantic web (RDF triples), and we demonstrate some of the ways that generated content can be meaningfully queried using the SPARQL language. We have made all of these materials available as open source (http://github.com/cdd/bioassay-template), in order to encourage community input and use within diverse projects, including but not limited to our own commercial electronic lab notebook products.

Introduction

One of the major problems currently being faced by biologists charged with the task of performing experimental assays on pharmaceutically interesting molecules is the information burden involved with handling collections of assay descriptions. Individual laboratories may carry out hundreds or even thousands of screening experiments each year. Each of these experiments involves a protocol, and any two experiments may be identical, similar, or completely different. The typical practice for describing bioassay protocols, for both external communication and internal record keeping, is to use concise scientific English, which is the

most universally human readable method of communication, assuming the recipient is familiar with the relevant jargon.

Unfortunately this method is not scalable. Even given the availability of an expert, it is often quite difficult and time-consuming to read two assay description paragraphs and provide a metric for the degree to which two protocols differ. There are many workflow scenarios where comparison of protocols is necessary, e.g. searching through a collection of previous experiments, or making a judgment call as to whether two batches of small molecule measurements are comparable. Attempting to use software to assist with such tasks, when the substrate is unconstrained text, results in solutions that are crude at best.

While these issues with scalability could be described as a relatively minor nuisance in a small laboratory, the field of drug discovery has lately been undergoing a renaissance of open data. [1,2,3,4] Services such as PubChem provide a truly massive resource;[5] PubChem alone provides more than a million unique bioassay descriptions, and is growing rapidly.[6,7] Such data are supplemented by carefully curated resources like ChEMBL,[8] which are much smaller but have strict quality control mechanisms in place. What these services have in common is that their bioassay protocols have very little machine-readable content. In many cases, information about the target, and the kind and units of the measurements, have been abstracted out and represented in a marked up format, but all of the remaining particulars of the protocol are ensconced within English grammar, if at all.

In order to address this problem, the BioAssay Ontology (BAO) was devised.[9,10,11] The BAO, which includes relevant components from other ontologies, is a semantic web vocabulary that contains thousands of terms for biological assay screening concepts, arranged in a series of layered class hierarchies. The BAO is extensive and detailed, and easily extensible. The vocabulary is sufficiently expressive to be used for describing biological assays in a systematic way, yet it has seen limited use. Influential projects such as PubChem,[12] ChEMBL,[13] BARD[14] and OpenPHACTS[15] make use of the ontology, but the level of description in each is shallow, using only a small fraction of the terms.

There are a number of factors holding back scientists from using the BAO and related ontologies to describe their assays in detail, with perhaps the most substantial being the lack of software that makes the annotation process fast and convenient. Because it is based on the semantic web, BAO concepts are expressed as triples, of the form [*subject*, *predicate*, *object*]. There are no hard rules about how this is applied, which is a characteristic of the semantic web, and is both an asset and a liability. The simplest way to consider annotating a particular feature of an assay, e.g. the biological process, is to compose a triple of a form such as [*assay ID*, *biological process*, *viral genome replication*]. Each of these 3 fields is a uniform resource indicator (URI), which points to a globally unique object with established meaning. In this case, *assay ID* would correspond to an identifier that the user has created for the assay description; *biological process* corresponds to a specific property in the BAO that is used to link assays and the biological process that is being affected; and *viral genome replication* refers to a class in the BAO, which identifies a specific instance of a biological process, which is in turn inherited from a sequence of increasingly general classes, and may also be linked to any other node within the greater semantic web, such as the extensive Gene Ontology (GO)[16].

In principle, screening biologists can use the properties and classes from the BAO to annotate their assays intelligently in a machine readable format that is compatible with the universe of the semantic web. If large numbers of assays were sufficiently annotated, biologists and other drug discovery scientists could perform advanced searches and filtering that would enable better interpretation of results, enhanced building of machine-learning models, and uncovering of experimental artifacts. Despite the clear benefits of semantic annotation, the BAO remains

largely unused, the primary reason being its lack of accessibility. The BAO and its linked dependencies are large, and can be expected to keep growing as they are extended to capture more biological concepts. For an interactive view onto these terms, the site http://bioportal.bioontology.org/ontologies/BAO should be used to peruse the hierarchy.[17] Figure 1 shows two snapshots of part of the BAO hierarchy, using the BioPortal resource. The *classes* (Figure 1a) that make up the ontology contain the bulk of the terms and provide most of the expressive value, while the *properties* (Figure 1b) are used to provide context. The class hierarchy is in places many levels deep, and although it is arranged in a logical pattern, it is nonetheless necessary to be familiar with the entire layout in order to meaningfully annotate an assay protocol. Even an expert biologist familiar with the entire ontology would be presented with multiple degrees of freedom for deciding how to annotate a protocol; this is a fundamental problem for machine readability, which requires uniform consistency.

In our previous work we addressed the end-user problem, and invented technology that applies to the scenario when a user is presented with plain English text, and is charged with the task of selecting the appropriate semantic annotations. Our solution involved a hybrid approach that combined natural language processing with machine learning based on training data, with an intuitive interface that helps the user select the correct annotations, leaving the final choice in the hands of the scientist.[18] During this process we found that the challenge that we were unable to fully overcome was the burden of creating new training data. The BAO vocabulary defines more than 2500 classes, in addition to properties and terms from other ontologies, all of which can be expected to grow as the BAO is increasingly used for more biological content.

Considering each term as it applies to a given assay requires a high level of expertise of the BAO itself. For example, the NIH's Molecular Libraries Program's bioassay database, known as the BARD, employed dedicated research staff to annotate more than two thousand assays.[19] The absence of clear and straightforward guidance as to which terms to use under what circumstances is preventing adoption of the BAO by drug discovery scientists. For our model building efforts, we made use of a training data set made up of 1066 PubChem bioassays that each had more than a hundred terms associated with them,[20,21] although not all of the annotations were able to be matched to ontology terms. For purposes of creating additional training data, we experienced considerable difficulty finding what we considered to be canonical annotations for any given assay.

The BAO is essentially a vocabulary that is capable of describing many assay properties, but it lacks instructions on its use. This is an issue that we have undertaken to solve, and in this article we describe our approach to providing this critical missing component.

We describe a data model called the BioAssay Template (BAT), which consists of a small number of terms which are organized to describe *how* the BAO and linked ontologies should be used to describe a particular kind of bioassay. A template is essentially a gateway to the overall ontology, which divides the assay annotation process into a fixed hierarchy of *assignments*, each of which has a prescribed list of *values*, which are cherry-picked from the overall ontology.

The BAT vocabulary can be used to create any number of templates, which can be customized to suit the task at hand. As a starting point, we have created what we refer to as the *common assay template* (CAT). CAT is an annotation recipe that is intended to capture the major properties that most biologists need to describe their assays and that enables most drug discovery scientists to have a basic understanding of an assay and its results.

A condensed summary of this template is shown in Figure 2. Unlike the class hierarchy of the BAO, the tree structure of the CAT is flat. While the data model allows groups and subgroups, our current template errs on the side of simplicity, and includes just 16 different assignments,

each of which is associated directly with the top-level assay, and each of which has a list of associated values (examples shown in Figure 2).

A template can be customized as necessary, and once it is ready, it can be used to define the way in which assays are annotated. The data model is designed to enable software to compose a user interface: presenting each of the categories, and making use of the selected values as the options that are made available to the user. It is essentially a way to restrict and simplify the large scope of the BAO, reduce the degrees of freedom, and remove ambiguity. Having curated the assignments and values so that the lists consist of the minimum number of relevant possibilities, each of them decorated by a meaningful label and a more detailed description, it becomes possible to design a user experience that is suitable for a scientist who is an expert in the field, but does not necessarily know anything about semantic web concepts.

In order to explore this approach, we have created a software package called the BioAssay Schema Editor, which is open source and available via GitHub. It is written using Java 8, and runs on the major desktop platforms (Windows, Mac & Linux). The software implements the data model that we describe in this article.

Our priorities for this work are to: (1) establish a data model for bioassay templates; (2) create an intuitive software package for editing these templates and using them to annotate real data; and (3) collaboratively establish a CAT for general purpose use. We have put a considerable amount of effort into the user interface for editing templates, even though we expect only a small fraction of biologists will ever be directly involved in editing them. We have also invested significant effort towards developing a one-size-fits-most template, the CAT. Our goal with the CAT was to enable capture of ~80% of the most commonly used terms, and present them in a logical and concise way, so that a large proportion of users will be able to use it as-is to add a significant amount of value to their protocol data. In addition, the CAT can act as a starting point for modification if scientists would like to tailor the template.

Scientists working in research groups that routinely make use of terms that are not included in the CAT can elect to start with an existing template and add the missing assignments and values, and also delete whole groups of content that do not apply to their research. A research group may accumulate a collection of task-specific templates, allowing their scientists to pick the most appropriate one. By ensuring that the editor software is easy to use, runs on all platforms, and is open source, we hope to ensure that this option is quite practical for any research group with access to basic information technology expertise. We intend to encourage the community to make use of these resources, both as standalone tools and interoperating with the electronic lab notebook software that we are presently designing.

One of the implicit advantages of using semantic web technology as the underlying data format (triples), and a well established set of reference terms (the BAO and various linked ontologies), is that even if two scientists are annotating assays with different templates, it is highly likely that many or most of the terms will overlap, even if the templates were created from scratch. Since the final deliverable for an annotated assay is the semantic web, it means that the output can be subjected to the entire universe of software designed to work with RDF triple stores.[22] As more assays are annotated, the scope and power of queries and informatics approaches for enhancing drug discovery projects are similarly increased. With a large corpus of annotated assays available, scientists will be able to make better use of prior work for understanding structure activity relationships, uncovering experimental artifacts, building machine-learning models, and reducing duplicated efforts.

Methods

Data Model

The semantic description of templates and annotations uses a small number of additional URIs, each of which has the root stem http://bioassayontology.org/bat, and is denoted using the Turtle-style[23] abbreviated prefix "bat".

The hierarchical model for describing a template is shown in Figure 3. Parent:child relationships denoted by an arrow indicate one-to-many relationships, while the properties listed in the boxes underneath the nodes are one-to-one relationships. A template definition begins with the *root*, which is distinguished by being of type bat:BioAssayTemplate. The root is also of type bat:Group, and has some number of child nodes, which are themselves either assignments or subgroups.

An assignment node has several scalar properties, including label and description, and it also refers to a *property* resource. These are typically mapped to URI resources found within the BAO (e.g. http://www.bioassayontology.org/bao#BAO_0000205, label: "has assay format"). Each assignment has some number of values associated with it, and these make up the list of available options. Each value is primarily identified by the resource that it maps to, which is typically found in the BAO (e.g. http://www.bioassayontology.org/bao#BAO_0000219, label: "cell based format"). Besides the label and description, which are customizable within the template data model, the reference URI has its own implied class hierarchy (e.g. "cell based format" is a subclass of "assay format"), which is not encoded in the template data model, but is inferred once it is paired with the BAO and its linked ontologies.

The schema for annotation of assays is shown in Figure 4. The assay is given a distinct URI, and is associated with several properties such as label and description. The template is recorded, as is an optional reference to the origin of the assay (which may be a semantic web resource, or a DOI link to a journal article). The free-text description of the assay can also be recorded using the *hasParagraph* predicate.

The assay is associated with some number of annotations, which are primarily linked to assignments within the corresponding template. For annotations that assert a URI link, the *hasValue* predicate typically corresponds to one of the available values that was prescribed for the assignment in the template definition, and generally refers to a term defined in the BAO, though custom references can be used - or the annotation may be specified using the *hasLiteral* predicate instead, which means that the user has entered data in a different form, typically text or a numeric value. The *hasProperty* predicate is generally copied from the corresponding assignment.

When annotating an assay, each assignment may be used any number of times, i.e. zero instances means that it has been left blank, while asserting two or more triples means that all of the values apply. The relationship between assays and annotations has no nesting: the intrinsic group/sub-group structure of any particular annotation can be inferred from the template, since the *usesTemplate* and *isAssignment* predicates refer to the origins in the template.

Software

The BioAssay Schema Editor is available from GitHub (https://github.com/cdd/bioassay-template) and may be used under the terms of the Gnu Public License 2.0.[24] The code is written using Java 8, and the user interface is based on JavaFX. Semantic web functionality is implemented by incorporating the Apache Jena library.[25] The project includes a snapshot of the BioAssay Ontology[26] and some of the linked ontologies, as well as the latest version of the

common assay template schema. It should be assumed that the project will continue to evolve until well after the publication date of this article.

The application operates on a datafile referred to as a *schema*, which is represented as a collection of triples (in Turtle format, with the extension .ttl). A schema is expected to include a single template, for which the root node is of type bat:BioAssayTemplate, and may optionally contain any number of assays that have been (or will be) annotated using that same template. Triples are used as the serialization format in order that the editable files can be used as-is by a Triple store, and become a part of the semantic web with no further modification.

Figure 5 shows the main window for the application, which has loaded a contemporary version of the *common assay template (CAT)*, and has several accompanying assays awaiting annotation. The components that make up the template are shown as a hierarchy on the left hand side of the panel. Selecting any of the groups or assignments causes the detail view on the right to be filled in with the corresponding content.

Adding, deleting, renaming etc. of groups, assignments and values is fairly mundane, and follows standard desktop user interface design patterns. Selecting URI values for properties and values requires a more specific interface, and is composed by summarizing the BAO vocabulary, which is loaded into the application at the beginning. Resources can be selected using a dialog box that can present the list of options in a flat list, with an optional search box for restricting the list (Figure 6a) or by using the hierarchy view that shows the position in the BAO ontology (Figure 6b). The dialog box can also be used to add multiple values at once, which is particularly convenient when a branch of the BAO encompasses multiple terms that are all valid options. When a resource is selected, its label and description are imported from the BAO into the template: these values can be edited after the fact, but by default they are the same as in the underlying vocabulary.

The primary role of the schema editor is to provide a convenient way to edit templates, but in support of this goal, it also provides an interface to use the template to annotate assays. The interface can be used for generating training data (e.g. for model generation), but it is mainly intended as a way to ‘test drive’ the current template. Because the annotation process is directly derived from the template, having the two editing processes side by side is advantageous when the template is being designed. For example, the operator can begin annotating an assay, and if a value is missing from one of the assignments, or a new kind of assignment turns out to be necessary, this can be added to the template within the same editing session.

Figure 7a shows an example of an assay that has been annotated. The detail view has a placeholder for description text, which is particularly useful when the content has been imported from some external source, and the annotations are being made by converting the protocol text into semantic annotations. Clicking on any of the annotation buttons brings up a panel of options (Figure 7b) that represent the prescribed values for the assignment. Each of the assignments can be left blank, annotated once, or given multiple values. The ideal use case is when the value (or values) occurs within the list of prescribed values, but since the data model allows any URI, the user interface also allows the user to insert a custom URI. In cases where no URI is listed in the template (e.g. a concept that does not have an established URI), it is possible to add plain text for any of the assignment annotations. While this has no meaning from a machine-learning point of view, it can serve as a convenient placeholder for terms that will be invented in the future.

Results

Templates

We set out to create a common assay template (CAT) that includes the basic details essential to defining any bioassay: assay type, format, target and biology, results and pharmacology, and other details. The CAT was developed with the opposing goals of identifying assignments that (1) would be limited in number in order to be not overly burdensome vs. (2) comprehensively cover the majority of the information contained in written descriptions of bioassays. We also considered the type of information that would be utilized by an end user attempting to search, filter, and aggregate assays by their bioassay annotations. For example, details such as the assay footprint (plate type), assay kit, and detection instrument were included because they may be useful terms for identifying experimental artifacts. Biological process and other target-related information were included to enable aggregating results across similar drug discovery projects for model-building and other applications. Finally, we limited assignments to those where the BAO offered sufficient options for possible values. Since the goal of the project is to generate machine-readable assay annotations, we avoided assignments where BAO terms were not available, such as those characterizing *in vivo* assays, and especially assignments whose values would be very specific for each assay, such as negative and positive controls. These areas will be addressed in the future once the underlying vocabulary (BAO or otherwise) is available sufficient to expand the domain. Similarly, the CAT falls short of capturing detailed protocol steps. In its present incarnation, it cannot be considered as a complete replacement for the text that is typically used to describe an assay, though we do intend to pursue this level of detail in future work. For the present, we are primarily concerned with utilizing the rich vocabulary within the BAO to achieve maximum impact with minimum additional burden on the end user workflow.

To develop the CAT, we used the following process: first, biologists independently considered each of the terms available in the BAO and prioritized assignments for the CAT. Each assignment was associated with a number of possible values based on the BAO hierarchy. Then, quantitative and qualitative approaches were used to determine if the prioritized assignments included in the CAT were sufficient to fully describe most assays. For the quantitative approach, we assessed the set of 1066 PubChem bioassays[27] that were previously annotated by hand by BAO experts.[28] In that exercise, the BAO experts aimed to fully annotate each assay, capturing all applicable information for more than a hundred different categories or terms. If there was not an applicable value, the assignment or category was left blank. We analyzed the use of the BAO terms to assess the utility and comprehensiveness of the assignments included in the CAT compared to the remaining terms. We found that the 16 CAT assignments were annotated in 81% of the 1066 PubChem assays compared to 33% for the remaining terms. We also found that 95% of the values for CAT assignments were BAO terms rather than literal or non-URI based terms, compared to 63% in the remaining categories. These results suggested that the CAT includes assignments that are both relevant to the majority of assays as represented in PubChem and well covered by the BAO.

For an in-depth qualitative assessment of the CAT, biologists annotated a wide variety of assays, encompassing different assay types (e.g., cell viability, enzyme activity, binding, and ADMET), assay formats (e.g., cell-based, biochemical, microsome, organism, tissue, etc.), and assay design methods (e.g., ATP quantitation, cell number, immunoassays, gene expression, radioligand binding, etc), as summarized in Table 1. We found that in many cases, both from assay descriptions available from PubChem and from in-house screening assay descriptions, the CAT captured much of the relevant information. For example, annotating an assay for cell viability (PubChem ID 427) shows that all but two of the 16 CAT assignments are readily

annotated from the short descriptive information provided (Figure 8). 'Target' is left blank, as it is not applicable (this assay aims solely to identify cytotoxic compounds); 'Detection Instrument' was not noted. Similarly, as shown in Figure 9, all applicable CAT assignments (15 of the 16) are annotated from the description of a competitive binding assay (PubChem ID 440). Figure 9 also illustrates that multiple values can be annotated for a single assignment, enabling content from complex assays to be captured. Together, these two examples highlight that both cell-based and biochemical assays can be extremely well-suited to be annotated using the CAT.

However, there were some cases where the CAT was less effective in capturing important information. For example, 14 of the 16 CAT assignments could be annotated for PubChem ID 488847, some with multiple values; however, the 'big picture' view of this rather complex primary assay is not as readily apparent from its 'CAT profile' as from a single sentence in the description (Figure 10). In addition, this PubChem record had extensive technical details such as reagent components, liquid handling volumes and instruments, times of incubation and plate processing steps, which could be important for identifying matching assays or interpreting the results. Another example of a poor fit for the CAT, as noted earlier, are *in vivo* assays. These are largely beyond the scope of this effort, which is currently constrained to terms defined by the BAO: key parameters such as route of administration, dose, dose units, type of model (e.g. xenograft, disease) are not well represented. These and other limitations will be addressed in the future by adding or extending the underlying ontologies.

Finally, as noted earlier, we designed the CAT to be a 'one-size-fits-most' template. A summary of assignments for the complete set of assays annotated in the course of developing the CAT shows we have achieved this (Table 1). One consequence of this 'one-size-fits-most' strategy is that certain attributes (such as those highlighted in green or red in Figures 8 and 9) have been omitted. Depending on one's perspective, these types of data (such as positive and negative controls, data processing/normalization steps, relevant disease indication, and specific protocol details such as pre-incubation of compounds with the target, time or temperature of an assay) could be viewed as essential. We decided to exclude this type of information from the CAT because of irregularity of appearance in bioassay descriptions, the lack of coverage by the BAO, or incompatibility with the current data model. Expanding into this area is an opportunity for future development, and it should be noted that the CAT may be used as a starting point for templates that provide a set of assignment options that are customized for subcategories of assays, or even specific projects. We believe the next immediate step should be to apply our CAT to a large (>10,000) set of assays, both to facilitate new meta-analyses and to identify potential gaps in annotation revealed by such studies.

PubChem

Possibly the most voluminous source of openly accessible bioassay data can be found on PubChem, which hosts more than 1.1 million assay records at the time of publication, and is growing rapidly. These are individually associated with the chemical structures of the compounds for which the measurements were made. Each of the assays is decorated with several descriptive fields that are essentially plain text, and which are populated by contributors during the upload process, or in some cases by an import script transferring data from other sources. While many of the entries contain a significant amount of detail, the phrasing style and level of detail varies considerably, often erring on the side of too little or too much information about the assay protocol.

Nonetheless, the PubChem assay collection represents one of the best and most convenient sources of data for annotation purposes, and for this reason we have added a feature to the BioAssay Template editor that explicitly searches for PubChem records, as shown in Figure 11.

The dialog box allows the user to type in a PubChem Assay ID number, or to hit the button labelled *Random*, which picks an arbitrary assay from the entire collection, and fills in the corresponding text and URI of origin. While a large proportion of assays loaded into PubChem contain only sparse tags about the data source, or the abstract of the corresponding publication, there are a significant number of records that contain lengthy descriptions of the assay. The dialog box provides an opportunity for the user to tidy up the text (e.g. removing irrelevant content) prior to importing it into the schema. The content is then added to the list of assays being annotated within the schema model, whereby the origin is recorded as a link to the assay, and the text is associated using the *hasParagraph* predicate. Once the text is augmented with annotations using the current template, it becomes a useful entry for training data. This is one of our main strategies for generating a corpus of data for machine-learning purposes, which will ultimately find its way into a user friendly ELN for bioassay annotation.

Analysis

Because the data model we describe is based on semantic web triples, and the file format that is used by the BioAssay Schema Editor is made up of triples (in Turtle format), it means that any templates and assay annotations can be loaded directly into a triple store database, and queried using SPARQL queries. Content can be hosted on private servers for local use, or it can be exposed to the greater web of connected data. The supplementary information (Section 1) describes a configuration script for the open source Apache Fuseki Jena server which can be used to load the BioAssay Ontology, its related ontologies, and some number of files saved with the BioAssay Schema Editor, which can then be served up as read-only content.

Once the content is available via a SPARQL endpoint, there are a number of boilerplate queries that can be used to extract summary and specific information. Fetching a list of all bioassay templates can be accomplished using the following query:

```
PREFIX bat: <http://www.bioassayontology.org/bat#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?template ?label ?descr WHERE
{
    ?template a bat:BioAssayTemplate ; rdfs:label ?label .
    OPTIONAL {?template bat:hasDescription ?descr .}
}
```

The above query identifies any resource that is tagged as having the *BioAssayTemplate* type. Obtaining information about the assignments that are associated with a template can be done by looking for resources of type *Group* that are associated with it. Obtaining a summary list of assignments that are attached to the top level (i.e. not within a subgroup) can be accomplished with a query similar to the following (using the same prefixes as above) which explicitly references the common assay template:

```
SELECT ?assn ?label ?descr ?property ?numValues
{
    <http://www.bioassayontology.org/bas#CommonAssayTemplate>
                        bat:hasAssignment ?assn .
    ?assn a bat:Assignment ;
        rdfs:label ?label ;
        bat:hasProperty ?property .
    OPTIONAL {?assn bat:hasDescription ?descr .}
    {
        SELECT ?assn (COUNT(?value) as ?numValues) WHERE
        {
            ?assn bat:hasValue ?value .
        }
        GROUP BY ?assn
    }
}
ORDER BY ?label
```

Similarly, assignments with one level of nesting can be obtained with a slightly longer query, which explicitly inserts a subgroup in between the template and assignment:

```
SELECT ?group ?glabel ?assn ?label ?descr ?property ?numValues
{
    <http://www.bioassayontology.org/bas#CommonAssayTemplate>
                        bat:hasGroup ?group .
    ?group a bat:Group ;
        rdfs:label ?glabel ;
        bat:hasAssignment ?assn .
    ?assn a bat:Assignment ;
        rdfs:label ?label ;
        bat:hasProperty ?property .
    {
        SELECT ?assn (COUNT(?value) as ?numValues) WHERE
        {
            ?assn bat:hasValue ?value .
        }
        GROUP BY ?assn
    }
}
ORDER BY ?glabel ?label
```

To query for information about the prescribed values for assignment (in this case the bioassay assignment from the common assay template), the following query can be used:

```
SELECT ?property ?value ?label
{
    <http://www.bioassayontology.org/bas#Bioassay>
        bat:hasProperty ?property ;
        bat:hasValue
        [
            bat:mapsTo ?value ;
            rdfs:label ?label
        ] .
}
```

The query specifically pulls out the *property* field, which is typically a link into the BAO property terms, and the *value* field, which is typically a link into the BAO classes. Pursuing either of these resources provides a wealth of implicit information, partly from the hierarchical nature of the BAO terms, and the unlimited opportunities for these terms to be linked to other semantic resources.

To obtain a list of assays that have been annotated using one of the templates, the following query can be used:

```
SELECT ?assay ?label ?descr ?template WHERE
{
    ?assay a bat:BioAssayDescription ;
        rdfs:label ?label ;
        bat:usesTemplate ?template .
    OPTIONAL {?assay bat:hasDescription ?descr .}
}
```

Obtaining all of the annotations for such an assay can be done with:

```
SELECT ?assn ?label ?property ?value ?literal ?group WHERE
{
    <http://www.bioassayontology.org/bas#ExampleAssay>
                        bat:hasAnnotation ?annot .

    ?annot bat:isAssignment ?assn ;
        rdfs:label ?label ;
        bat:hasProperty ?property .
    OPTIONAL {?annot bat:hasValue ?value}
    OPTIONAL {?annot bat:hasLiteral ?literal}
    ?group a bat:Group ; bat:hasAssignment ?assn .
}
```

Because annotations are directly attached to an assay description, hierarchical information about the nature of the assignment can be obtained by further investigating the template definition of the assignment (*?assn*) or either of the linked BAO terms (*?property* and *?value*).

Conclusion

We have developed a data model and interactive tool that can be used to narrow the degrees of freedom from the BioAssay Ontology (BAO) and its linked dependencies. This has been done in

order to facilitate content creation activities, so that semantic annotation of assay protocols can be carried out by a domain expert with no corresponding expertise with the underlying ontology. We have provided a proof of concept tool that creates a user interface based on the template data model, and made this available to the community as open source.

The data model that we have created follows a simplistic pattern, where elementary facts can be asserted. By leveraging the implied value of the underlying ontology, a small collection of a dozen or so such annotations provides a significant amount of machine-readable context about the assay. While insufficient to completely define an assay protocol experiment, this stands in contrast to the standard practice of providing essentially zero machine-readable information (i.e. plain English text with quasi-standardized jargon).

We have made available the *common assay template* (CAT) which was designed by biologists with the objective of leveraging the BAO to provide the largest amount of useful, relevant, machine-readable information with the fewest number of additional data points needing to be captured by the originating scientist. The CAT is expected to be useful for a wide variety of sorting, filtering, and data aggregating tasks that drug discovery scientists need to be able to carry out on a large scale, but currently cannot due to the absence of machine-readable annotations.

The CAT prioritizes 16 assignments that biologists consider most central to describing their assays and reporting assay results. Annotations for these assignments will enable biologists to ask complex queries. For example, one could ask if there are systematic differences in cell-based versus biochemical-based assays for a certain target class, such as kinases. One could determine if a certain assay set-up, such as 96-well plates using a spectrophotometer were likely to have a higher hit rate. Similarly, one could identify if a certain compound or class of compounds is active in multiple assays, and if those assays assess similar biological processes or if the activity is likely to be an artifact.

By focusing on 16 assignments out of more than a hundred options available in the BAO, the CAT is meant to impose a minimal burden for annotating scientists. Our goal is to make annotating assays simple and easy so that the practice may be generally adopted. Templates are malleable and scientists can easily include other assignments.

One critical type of information that is not included in the current framework is protocol steps, which would be essential for directly comparing two assays. In the future, it would be useful if this information were machine-readable. However, semantic technology using a simplistic data model like the BAT cannot capture sequences of information. Capturing procedural or protocol steps would require the development of a more complex data model. Under the current system, we imagine that queries using annotations from the CAT will allow scientists to hone in on similar assays, but for the moment, experts will still need to read the full assay descriptions to make decisions about combining different assays' data sets.

We have carried out this work in the context of a much larger scope, which is to provide scientists with tools to easily annotate bioassays and other related experiments in a way that is complete and machine-readable. Given that the standard industry practice does not involve adding any machine readable data to assay protocols, and that there are currently no widely available tools to do so with a user experience that is sufficiently painless for mass adoption, we have taken an incremental approach. This additional work has been done in order that we can continue with our previous work that was focused on using machine learning techniques to accelerate manual assignment of assays.[15] Our immediate follow-up goals are to make use of the CAT to gather a large corpus of training data, both from active users of CDD Vault, and from existing repositories such as PubChem. This training data will be used to ensure that our

enterprise ELN tools will be supported by machine learning technology as soon as they are unveiled.

We are also pursuing options for extending the BioAssay Template (BAT) data model so that it is capable of capturing more sophisticated information about assays, e.g. linking to other ontologies to cover more types of assays; adding terminology for capturing quantities; addition of indefinite numbers of preparation steps; dependent assignment types, etc. One critical step when we enable connecting with other ontologies will be the ability to link the 'Target' to a unique identifier such as geneid or UniProtID. Each unique target identifier can be associated with a rich array of corresponding GO terms, of which a subset are mapped into the default selection of BAO classes. This will enable comparison of assays based on specific targets and related biological processes or molecular functions. While our first objective is *horizontal* scaling, i.e. ensuring that all assay protocols have semantic annotations that make a large portion of the content machine-readable, pursuing *vertical* scaling is also of great interest, i.e. making it possible for the semantic annotations to replace the need for use of English text.[29] This brings about some exciting possibilities beyond just improvement of searching and matching, such as uploading protocols to robotic assay machinery, or making the publication process multi-lingual, thus alleviating a considerable burden to non-native English speakers. Pursuing this goal will require significant additions to the BAO itself, as well as making increased use of borrowed terms from other ontologies.

The technology that we have described in this article has been created for the purpose of improving the electronic lab notebook (ELN) technology that is offered by Collaborative Drug Discovery, Inc. (CDD), and we have begun work on a web-based interface for using templates such as the CAT for annotating assay protocols.[30] We have disclosed all of the underlying methods, data and open source code because we welcome participation by anyone and everyone. While CDD is a privately held for-profit company, it is our firm belief that improvement to this particular aspect of scientific research is a positive sum game, and we have more to gain by sharing than by keeping our technology entirely proprietary.

Supporting Materials

The BioAssay Schema Editor is publicly available from GitHub (https://github.com/cdd/bioassay-template). The source code for the application is available under the terms of the Gnu Public License (GPL) v2, which requires that derived works must also be similarly open. The underlying semantic data model for the template and assay annotation, as well as the common assay template (CAT), are public domain: they are not copyrighted, and no restrictions are placed on their use. The BioAssay Ontology (BAO) is available from the corresponding site (http://bioassayontology.org/bioassayontology) under the Creative Commons Attribution License v3.

Tables

Table 1. Representation of Common Assay Template in Sample Assay Set

CAT Assignment	Test Assays (of 43) With at Least 1 Value	# of Unique Values Annotated
bioassay type	43 (100%)	24 of 88
assay format	43 (100%)	6 of 19
assay design method	43 (100%)	20 of 76
assay cell line	24 (55.8%)	15 of 95
organism	41 (95.3%)	11 of 65
biological process	40 (93.0%)	28 of 54
target	32 (74.4%)	13 of 38
assay mode of action	43 (100%)	8 of 13
result	41 (100%)	16 of 94
result unit of measurement	32 (74.4%)	6 of 56
assay screening campaign stage	40 (93.0%)	8 of 23
assay footprint	36 (83.7%)	5 of 20
assay kit	9 (20.9%)	5 of 93
physical detection method	42 (97.7%)	11 of 51
detection instrument	26 (60.5%)	9 of 97
perturbagen type	20 (46.5%)	3 of 9

Figure Captions

Figure 1: A selection of the BioAssay Ontology hierarchy, visualized using BioPortal (http://bioportal.bioontology.org): (a) classes and (b) properties.

Figure 2: An overview of the *common assay template* (CAT) at the time of publication.

Figure 3: BioAssay Template data model, which is used to describe a template.

Figure 4: Data model for annotated assays, which is used to apply a template to a specific assay.

Figure 5: A snapshot of the BioAssay Schema Editor. On the left hand side the current template is shown at the top (with its hierarchy of groups and assignments), and any assays currently in progress shown underneath. The panel on the right shows the details for an assignment - *assay format* - and the prescribed values that are associated with it.

Figure 6: A snapshot of the two main tabs used for locating a value in the BioAssay Ontology. The left hand side (a) shows the list view, which is flat, while the right hand side (b) shows the values in context of the actual hierarchy of the underlying ontology.

Figure 7: A snapshot of the annotation interface that is available within the template editor (a). The current template can be applied to specific assays within the same overall user interface, which is a convenient way to evaluate its suitability. Selecting any of the assignments brings up a dialog box presenting all of the prescribed values (b).

Figure 8. Example of PubChem Assay text ideally suited for annotation with the CAT. Left: Text from description in PubChem Assay ID 427: yellow = information captured in CAT, green = information not captured but possible for a future version (e.g., controls, data processing), red= information beyond the scope of BAO (technical details) **Right:** CAT assignments in BioAssay Schema Editor.

Figure 9. Example of PubChem Assay text ideally suited for annotation with the CAT. Left: Text from description in PubChem Assay ID 440: yellow = information captured in CAT, pink = information added as 'literal' values (i.e., too specific to exist as a BAO entry, but deemed valuable), green = information not captured but possible for a future version (e.g., controls, data processing), red= information beyond the scope of BAO (technical details). **Right:** CAT assignments in BioAssay Schema Editor. Annotations added as 'literal' values are highlighted yellow and contained in single quotes. Note that multiple values for a single CAT assignment can be annotated (*target biological process, assay mode of action, assay screening campaign stage, perturbagen type*).

Figure 10. Example of an assay partially suited for annotation with the CAT. Left: Text from description in PubChem Assay ID 488847: yellow = information captured in CAT, pink= information added as 'literal' values (i.e., too specific to exist as a BAO entry, but deemed valuable), green = information not captured but possible for a future version (e.g., controls, labels of target and ligand, assay quality data (Z')), red= information beyond the scope of BAO (technical details). **Right:** CATvalues assigned in the BioAssay Schema Editor capture key parameters of the assay yet do not capture the complexity of the assay articulated in the single sentence (arrow): "a flow cytometry protein interaction assay to screen for compounds that compete with RNA binding to GRK2".

Figure 11: Dialog box for random lookup of assays from PubChem.

References

1. A.M. Clark, A.J. Williams, S. Ekins: Machines first, humans second: on the importance of algorithmic interpretation of open chemistry data, Journal of Cheminformatics, 7:9 (2015) DOI 10.1186/s13321-015-0057-7

2. A. Hersey, S. Senger, J.P. Overington: Open data for drug discovery: learning from the biological community. *Future Medicinal Chemistry*, **4**, 1865-1867 (2012) DOI 10.4155/fmc.12.159

3. G.F. Ecker, B. Williams-Jones: Editorial: Open Innovation in Drug Discovery. *Molecular Informatics*, **31**, 519-520 (2012) DOI 10.1002/minf.201280004

4. A.J. Williams, J. Wilbanks, S. Ekins S: Why Open Drug Discovery Needs Four Simple Rules for Licensing Data and Models. *PLoS Computational Biology* , **8**:e1002706 (2012) DOI 10.1371/journal.pcbi.1002706

5. K.Y. Helal, M. Maciejewski, E. Gregori-Puigjane, M. Glick, A.M. Wassermann: Public Domain HTS Fingerprints: Design and Evaluation of Compound Bioactivity Profiles from

PubChem's Bioassay Repository. *Journal of Chemical Information and Modeling*, just accepted (2016) DOI 10.1021/acs.jcim.5b00498
6. E. Bolton: Reporting biological assay screening results for maximum impact, *Drug Discovery Today: Technologies*, **14**, 31-36 (2015) DOI 10.1016/j.ddtec.2015.03.004
7. It should be noted that the majority of the first million PubChem assays do not contain detailed experimental assay descriptions. Contributors such as the Broad Institute and organizations affiliated with the Molecular Libraries Screening Center can be selected by browsing the sources: http://pubchem.ncbi.nlm.nih.gov/sources/sources.cgi
8. A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J.P. Overington: ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, D1100-D1107 (2012) DOI 10.1093/nar/gkr777
9. The materials for the BioAssay Ontology can be found at http://bioassayontology.org
10. S. Abeyruwan, U.D. Vempati, H. Küçük, U. Visser, A. Koleti, A. Mir, K. Sakurai, C. Chung, J. Bittker, P. Clemons, S. Brudz, A. Siripala, A. Morales, M. Romacker, D. Twomey, S. Bureeva, V. Lemmon, S.C. Schürer: Evolving BioAssay Ontology (BAO): Modularization, Integration and Applications, *Journal of Biomedical Semantics*, **5**, 1:S5 (2014) DOI 10.1186/2041-1480-5-S1-S5
11. U.D. Vempati, M.J. Przydzial, C. Chung, S. Abeyruwan, A. Mir, K. Sakurai, U. Visser, V.P. Lemmon, S.C. Schürer: Formalization, annotation and analysis of diverse drug and probe screening assay datasets using the BioAssay Ontology (BAO). *PLoS One*, e49198 (2012) DOI: 10:1371/journal.pone.0049198
12. S. Kim, P.A. Thiessen, E.E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B.A. Shoemaker, J. Wang, B. Yu, J. Zhang, S.H. Bryant: PubChem Substance and Compound databases. *Nucleic Acids Research*, D1202-13 (2016) DOI: 10.1093/nar/gkv951
13. E.L. Willighagen, A. Waagmeester, O. Spjuth, P. Ansell, A.J. Williams, V. Tkachenko, J. Hastings, B. Chen, D.J. Wild: The ChEMBL database as linked open data. *Journal of Cheminformatics*, **5**:23 (2013) DOI: 10.1186/1758-2946-5-23
14. A. de Souza, J.A. Bittker, D.L. Lahr, S. Brudz, S. Chatwin, T.I. Oprea, A. Waller, J.J. Yang, N. Southall, R. Guha, S.C. Schürer, U.D. Vempati, M.R. Southern, E.S. Dawson, P.A. Clemons, T.D.Y. Chung: 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. *Journal of Biomedical Screening*, **19**, 614-627 (2014) DOI: 10.1177/1087057113517139
15. A.J. Williams, L. Harland, P. Groth, S. Pettifer, C. Chichester, E.L. Willighagen, C.T. Evelo, 7, N. Blomberg, G. Ecker, C. Goble, B. Mons: Open PHACTS: semantic interoperability for drug discovery. *Drug Discovery Today*, **17**, 1188-1198 (2012) DOI: 10.1016/j.drudis.2012.05.016
16. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. (2015) *Nucl Acids Res* **43** Database issue D1049–D1056 Online at Nucleic Acids Research
17. It can also be browsed and edited using software such as Protégé, which can be found at http://protege.stanford.edu
18. A.M. Clark, B.A. Bunin, N.K. Litterman, S.C. Schürer, U. Visser: Fast and accurate semantic annotation of bioassays exploiting a hybrid of machine learning and user confirmation. *PeerJ*, 2:e524 (2014) DOI 10.7717/peerj.524
19. A. de Souza, J.A. Bittker, D.L. Lahr, S. Brudz, S. Chatwin, T.I. Oprea, A. Waller, J.J. Yang, N. Southall, R. Guha, S.C. Schürer, U.D. Vempati, M.R. Southern, E.S. Dawson, P.A. Clemons, T.D.Y. Chung: 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. *Journal of Biomedical Screening*, **19**, 614-627 (2014) DOI: 10.1177/1087057113517139
20. Y. Wang Y, T. Suzek, J. Zhang, J. Wang, S. He, T. Cheng, B.A. Shoemaker, A. Gindulyte, S.H. Bryant: PubChem BioAssay: 2014 update. *Nucleic Acids Research*, D1075-82 (2014) DOI 10.1093/nar/gkt978
21. S.C. Schürer, U. Vempati, R. Smith, M. Southern, V. Lemmon: BioAssay ontology annotations facilitate cross-analysis of diverse high-throughput screening data sets. *Journal of Biomolecular Screening*, 16:415–426 (2011) DOI 10.1177/1087057111400191
22. See W3C Resource Description Framework: http://www.w3.org/RDF
23. See W3C RDF Turtle: http://www.w3.org/TR/turtle
24. Gnu Public License 2.0: http://www.gnu.org/licenses/gpl-2.0.en.html: the license allows anyone to use the source code for any purpose, on the condition that products making use of it must be made available under a license that is at least as open. Copyright for the project is held by Collaborative Drug Discovery, Inc.
25. See Apache Jena project: http://jena.apache.org
26. Downloadable OWL files for the BioAssay Ontology: http://bioassayontology.org/bioassayontology
27. Y. Wang Y, T. Suzek, J. Zhang, J. Wang, S. He, T. Cheng, B.A. Shoemaker, A. Gindulyte, S.H. Bryant: PubChem BioAssay: 2014 update. *Nucleic Acids Research*, D1075-82 (2014) DOI 10.1093/nar/gkt978
28. S.C. Schurer, U. Vempati, R. Smith, M. Southern, V. Lemmon: BioAssay ontology annotations facilitate cross-analysis of diverse high-throughput screening data sets. *Journal of Biomolecular Screening*, 16:415–426 (2011) DOI 10.1177/1087057111400191
29. L.N. Soldatova, D. Nadis, R.D. King, P.S. Basu, E. Haddi, V. Baumié, N.J. Saunders, W. Marwan, B.B. Rudkin: EXACT2: the semantics of biomedical protocols. *BMC Bioinformatics*, **15**, S5 (2014) DOI: 10.1186/1471-2105-15-S14-S5
30. A preliminary version of the web interface can be found at http://bioassayexpress.com. At the time of writing this service is in an early pre-alpha phase, but will be updated as the project progresses.

1

A selection of the BioAssay Ontology hierarchy, visualized using BioPortal (http://bioportal.bioontology.org): (a) classes and (b) properties.

Figure 1: A selection of the BioAssay Ontology hierarchy, visualized using BioPortal (http://bioportal.bioontology.org): (a) classes and (b) properties.

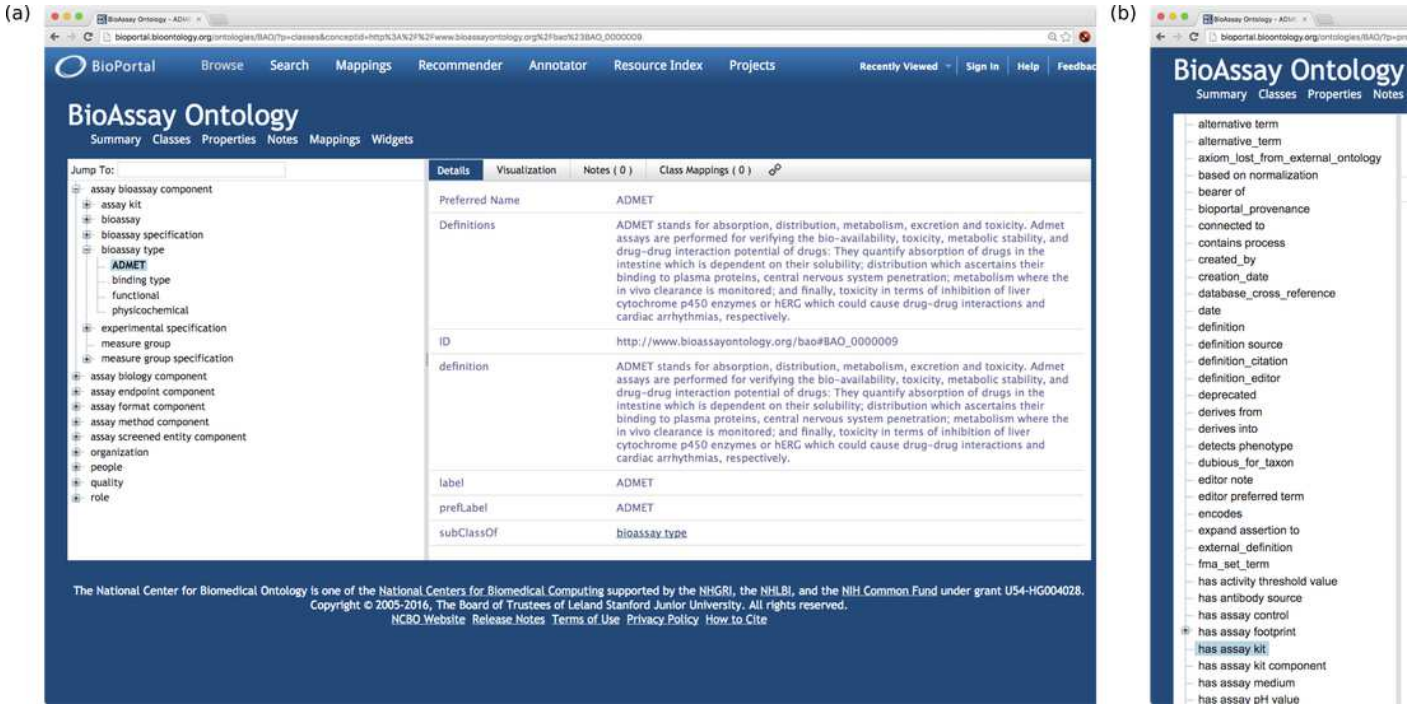


Figure 2(on next page)

An overview of the *common assay template* (CAT)

Figure 2: An overview of the *common assay template* (CAT) at the time of publication.

common assay template

URI: http://www.bioassayontology.org/bas#

bioassay type → (has bioassay)
- ADMET
- apoptosis assay
- beta galactosidase enzyme activity assay
- beta galactosidase reporter gene assay
- beta lactamase reporter gene assay
- binding assay
- bioavailability assay
- calcium redistribution assay
- cAMP redistribution assay
- (+ 79 more)

target → (has biological macromolecule)
- adhesion
- carbohydrate
- chaperone
- cytosolic protein
- enzyme
- enzyme regulator
- G protein
- G protein coupled receptor
- generic hydrolase
- (+ 29 more)

assay kit → (uses assay kit)
- Adapta Universal Kinase Assay Kit
- ADP Glo Kinase Assay
- ADP Hunter Plus
- AlphaScreen cAMP assay kit
- AlphaScreen cGMP Detection
- AlphaScreen GST detection kit
- AlphaScreen IgG detection kit
- AlphaScreen Phosphotyrosine Assay Kit
- Alphascreen second messenger IP1 detection kit
- (+ 84 more)

assay format → (has assay format)
- biochemical format
- cell based format
- cell membrane format
- cell-free format
- cytosol format
- microsome format
- mitochondrion format
- nuclear extract format
- nucleic acid format
- (+ 10 more)

assay mode of action → (has mode of action)
- activation
- agonism
- antagonism
- competitive binding
- inhibition
- irreversible binding
- ligand binding mode of action
- ligand function mode of action
- modulation
- (+ 4 more)

physical detection method → (has detection method)
- absorbance
- alphascreen
- atomic absorption spectrophotometry
- bio layer interferometry
- bioluminescence
- brightfield microscopy
- carbon nanotube based sensor
- chemiluminescence
- circular dichroism
- (+ 42 more)

assay design method → (has assay design method)
- antigen down assay
- ATP quantitation
- ATP quantitation using luciferase
- beta galactosidase induction
- beta lactamase induction
- binding assessment method
- caspase activity determination
- cell cycle progression assessment method
- cell movement measurement method
- (+ 67 more)

result → (has result)
- 50 percent activation
- 50 percent inhibition
- 80 percent inhibition
- 90 percent inhibition
- AC10 absolute
- AC1000 absolute
- AC26 absolute
- AC35 absolute
- AC40 absolute
- (+ 85 more)

detection instrument → (uses detection instrument)
- 3i Marianas
- 8453 UV-Visible Spectrophotometer
- Acumen
- AlphaQuest reader
- AMINCO-Bowman Series 2 Luminescence Spectrometer
- Analyst HT
- API 4000 LC/MS/MS System
- Applied biosystems 8200
- ArrayScan 3.1 HCS Reader
- (+ 88 more)

assay cell line → (is cell line of)
- 293 cell
- 293T/17 cell
- A2780
- A549 cell
- ACHN cell
- AML12 cell
- BA/F3 cell
- BJ
- BSC-1
- (+ 86 more)

result unit of measurement → (has unit of measurement)
- angstrom
- catalytic (activity) concentration unit
- cell concentration unit
- cells per milliliter
- centimeter
- century
- concentration unit
- concentration unit
- counts per second
- (+ 47 more)

perturbagen type → (has perturbagen)
- compound library
- DIVERSet
- LOPAC 1280
- miRNA library
- MLSMR library
- NINDS library
- shRNA library
- siRNA library
- The NatProd Collection

organism → (has organism)
- Arabidopsis thaliana
- bacterium
- Bluetongue virus 10
- Bos taurus
- Caenorhabditis elegans
- Candida albicans
- Canis lupus familiaris
- cellular organisms
- Chlorocebus aethiops
- (+ 56 more)

assay screening campaign stage → (has assay stage)
- alternate assay conditions
- alternate assay format
- alternate assay type
- alternate cell line assay
- alternate confirmatory assay
- alternate organism assay
- alternate target assay
- compound aggregation assay
- compound fluorescence assay
- (+ 14 more)

biological process → (has biological process)
- absence
- alternative mRNA splicing, via spliceosome
- ambiguous
- apoptotic process
- autophagy
- biofilm formation
- calcium-mediated signaling using intracellular calcium source_bao
- cAMP-mediated signaling_BAO
- cell cycle
- (+ 45 more)

assay footprint → (has assay footprint)
- 1536 well plate
- 24 well plate
- 384 well plate
- 96 well plate
- array
- cuvette
- gene array
- HYPER flask
- microplate
- (+ 11 more)

Figure 3(on next page)

BioAssay Template data model

Figure 3: BioAssay Template data model, which is used to describe a template.

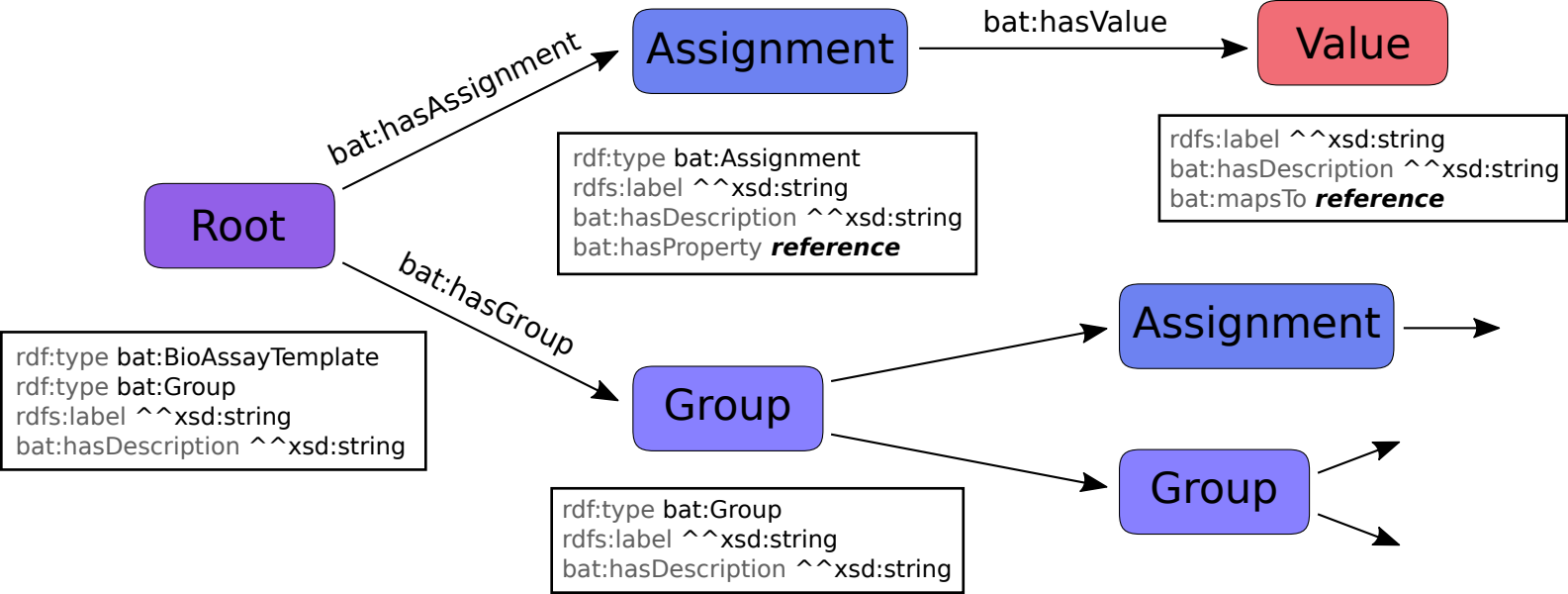
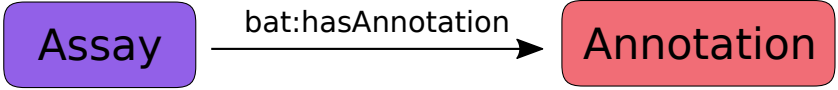

Root
bat:hasAssignment
Assignment
bat:hasValue
Value
rdf:type bat:Assignment
rdfs:label ^^xsd:string
bat:hasDescription ^^xsd:string
bat:hasProperty reference
rdfs:label ^^xsd:string
bat:hasDescription ^^xsd:string
bat:mapsTo reference
bat:hasGroup
rdf:type bat:BioAssayTemplate
rdf:type bat:Group
rdfs:label ^^xsd:string
bat:hasDescription ^^xsd:string
Group
Assignment
Group
rdf:type bat:Group
rdfs:label ^^xsd:string
bat:hasDescription ^^xsd:string

Figure 4(on next page)

Data model for annotated assays

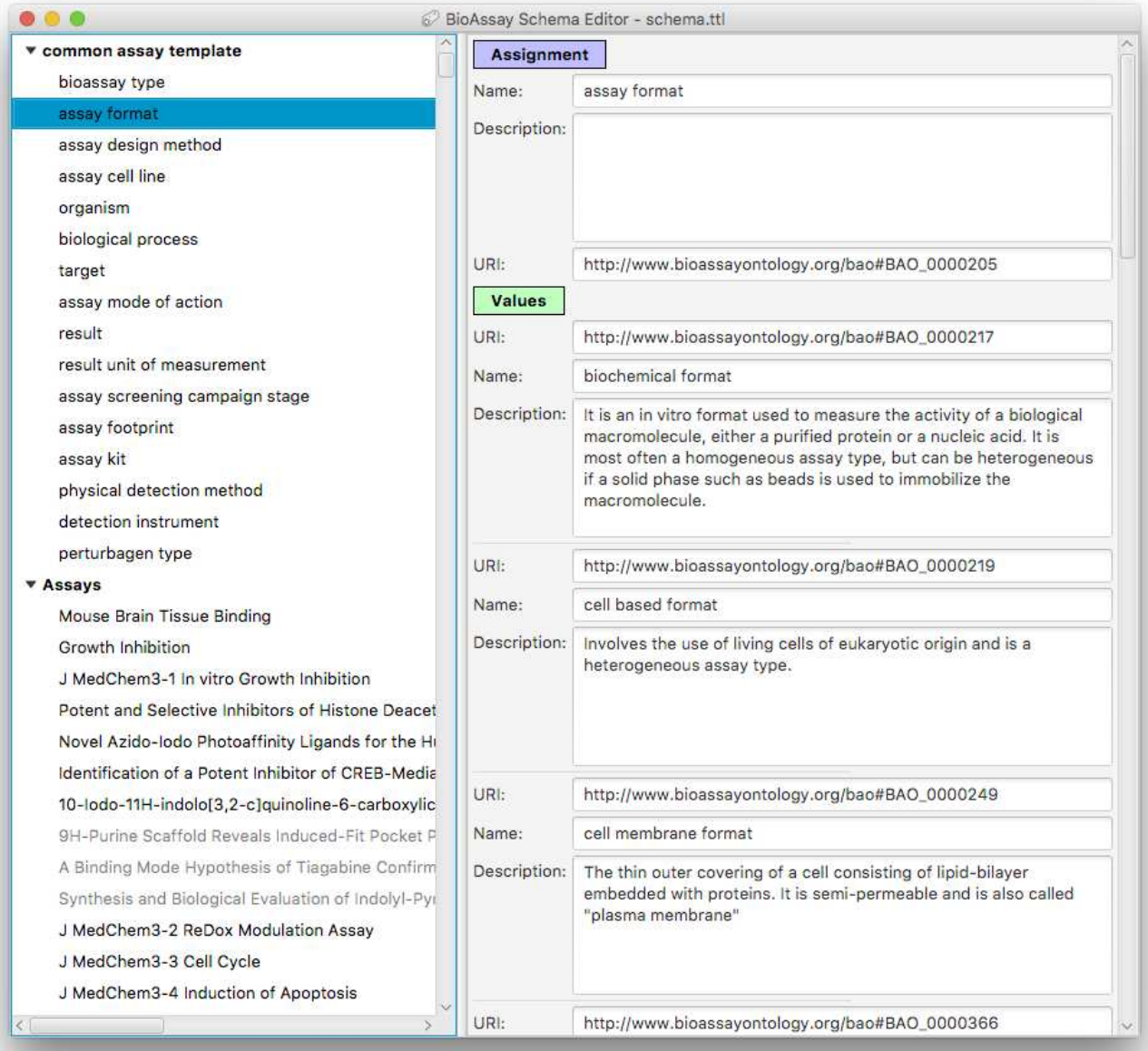
Figure 4: Data model for annotated assays, which is used to apply a template to a specific assay.


Assay
bat:hasAnnotation
Annotation
rdf:type bat:BioAssayDescription
rdfs:label ^^xsd:string
bat:hasDescription ^^xsd:string
bat:usesTemplate reference
bat:hasParagraph ^^xsd:string
bat:hasOrigin reference
isAssignment reference
rdfs:label ^^xsd:string
bat:hasDescription ^^xsd:string
bat:hasProperty reference
bat:hasValue reference
bat:hasLiteral literal

5

A snapshot of the BioAssay Schema Editor

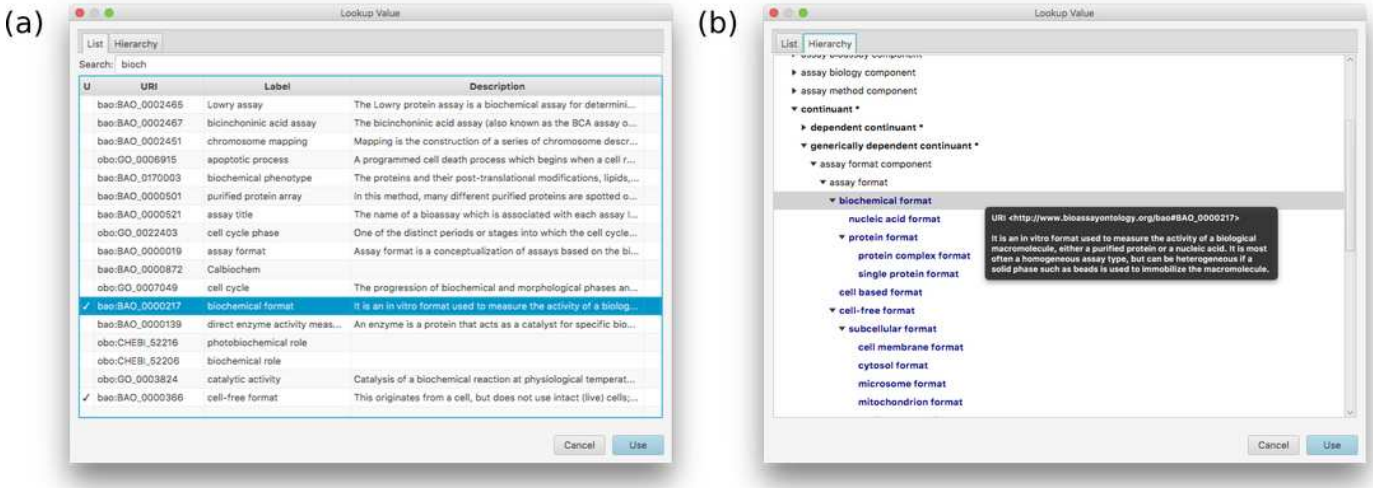
Figure 5: A snapshot of the BioAssay Schema Editor. On the left hand side the current template is shown at the top (with its hierarchy of groups and assignments), and any assays currently in progress shown underneath. The panel on the right shows the details for an assignment - *assay format* - and the prescribed values that are associated with it.



6

A snapshot of the two main tabs used for locating a value in the BioAssay Ontology

Figure 6: A snapshot of the two main tabs used for locating a value in the BioAssay Ontology. The left hand side (a) shows the list view, which is flat, while the right hand side (b) shows the values in context of the actual hierarchy of the underlying ontology.



7

A snapshot of the annotation interface that is available within the template editor

Figure 7: A snapshot of the annotation interface that is available within the template editor (a). The current template can be applied to specific assays within the same overall user interface, which is a convenient way to evaluate its suitability. Selecting any of the assignments brings up a dialog box presenting all of the prescribed values (b).

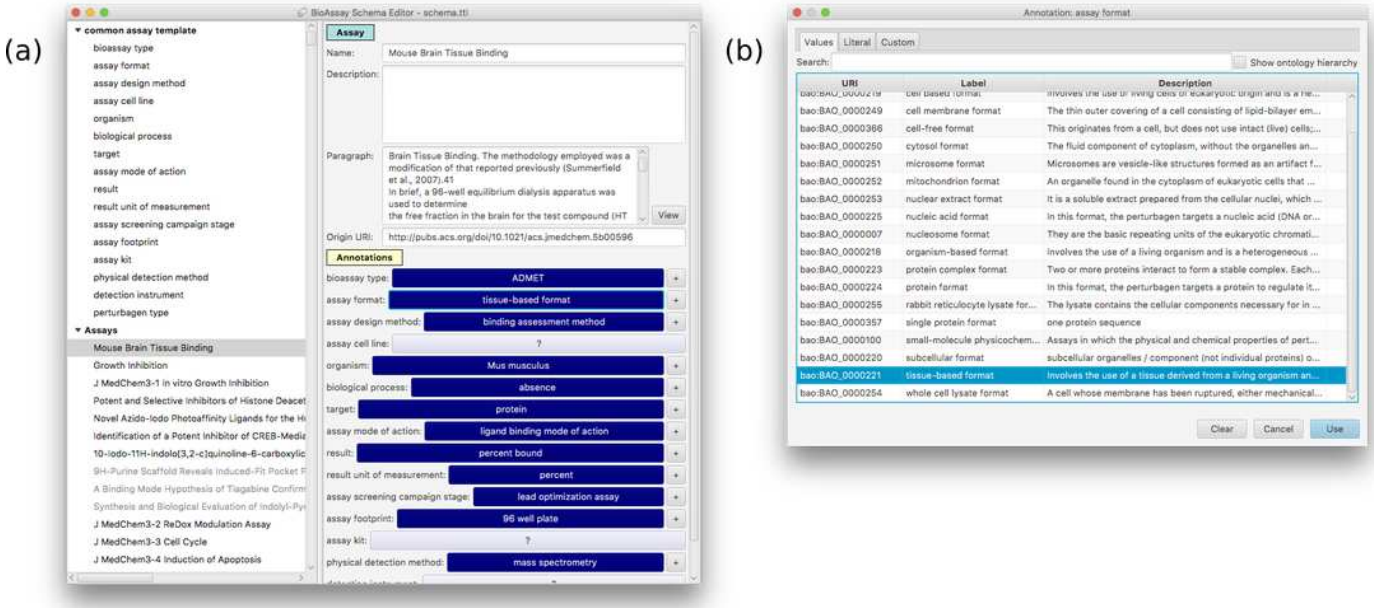


Figure 8(on next page)

First example of PubChem Assay text ideally suited for annotation with the CAT

Figure 8. Example of PubChem Assay text ideally suited for annotation with the CAT. Left: Text from description in PubChem Assay ID 427: yellow = information captured in CAT, green = information not captured but possible for a future version (e.g., controls, data processing), red= information beyond the scope of BAO (technical details) **Right:** CAT assignments in BioAssay Schema Editor.

We have developed a 1536-well cell-based assay for quantitative high throughput screening (qHTS) against a number of cell lines to determine in vitro cytotoxicity of small molecules. This particular assay uses the Hek 293 cell line which is derived from human embryonic kidney cells (transformed with adenovirus). The CellTiter-Glo luminescent cell viability assay (Promega) is a homogeneous method to measure the number of viable cells in culture. The end point readout of this assay is based on quantitation of intracellular ATP, an indicator of metabolic activity, using the luciferase reaction. Luciferase catalyzes the oxidation of beetle Luciferin to oxyluciferin and light in the presence of ATP. The luminescent signal is proportional to amount of ATP present. Using the CellTiter-Glo luminescent cell viability assay, the amount of cellular ATP was measured in the Hek293 cell line with complete culture medium following compound treatment for 40 hours. The assay was performed in opaque white Kalypsys 1536-well plates. In the screen, tamoxifen and doxorubicin were used as positive controls. Library compounds were measured for their ability to cause acute toxicity in the cell line, as reflected by a decrease in intracellular ATP levels, in a concentration-dependent manner. Data were normalized to the controls for basal activity (DMSO only) and 100% inhibition (100 uM tamoxifen). AC50 values were determined from concentration-response data modeled with the standard Hill equation.

Key

Annotated with URI

Added as literal

Not annotated: missed opportunity

Requires more advanced template model

PubChem Assay (ID 427)

Origin: http://pubchem.ncbi.nlm.nih.gov/bioassay/427

bioassay type —(has bioassay)→ cell viability assay

assay format —(has assay format)→ cell based format

assay design method —(has assay design method)→ ATP quantitation using luciferase

assay cell line —(is cell line of)→ HEK293

organism —(has organism)→ Homo sapiens

biological process —(has biological process)→ cell death

target —(has biological macromolecule)→ (not assigned)

assay mode of action —(has mode of action)→ modulation

result —(has result)→ AC50

result unit of measurement —(has unit of measurement)→ (not assigned)

assay screening campaign stage —(has assay stage)→ primary assay

assay footprint —(has assay footprint)→ 1536 well plate

assay kit —(uses assay kit)→ CellTiter-Glo Luminescent Cell Viability Assay

physical detection method —(has detection method)→ luminescence method

detection instrument —(uses detection instrument)→ (not assigned)

perturbagen type —(has perturbagen)→ compound library

Figure 9(on next page)

Second example of PubChem Assay text ideally suited for annotation with the CAT

Figure 9. Example of PubChem Assay text ideally suited for annotation with the CAT. Left: Text from description in PubChem Assay ID 440: yellow = information captured in CAT, pink = information added as 'literal' values (i.e., too specific to exist as a BAO entry, but deemed valuable), green = information not captured but possible for a future version (e.g., controls, data processing), red= information beyond the scope of BAO (technical details). **Right:** CAT assignments in BioAssay Schema Editor. Annotations added as 'literal' values are highlighted yellow and contained in single quotes. Note that multiple values for a single CAT assignment can be annotated (*target biological process, assay mode of action, assay screening campaign stage, perturbagen type*).

The assay reported here uses flow cytometry to measure test compound competition with a high-affinity fluorescent ligand for binding to human FPR. The assay was performed in a "duplex" format in which U937 cells expressing FPR were tested together with a Rat Basophilic Leukemia (RBL) cell line that expressed the related receptor, FPRL1. The FPR-expressing cells were stained with a red-fluorescent dye, FURA-red, to allow them to be distinguished from the FPRL1-expressing cells during flow cytometric analysis. A fluorescein label was conjugated to the lysine residue of the peptide, WKYMVm (WPep), to produce a fluorescent ligand (WPep-FITC) that bound FPR and FPRL-1 with high affinity. Dissociation constants (Kd) for binding of WPep-FITC to FPR and FPRL1 were determined to be 10 nM and 8 nM, respectively. WPep-FITC was used as the fluorescent ligand in the duplex FPR-FPRL1 assay to determine compound activity for both receptors. A set of 9,993 compounds, designated the 10K Set Type 1 (10KST1), and a separate set of 16,322 compounds, designated the 17K Set Type 1 (17KST1), was obtained from the Molecular Libraries Small Molecule Repository (MLSMR) maintained by Discovery Partners International in conjunction with the NIH Molecular Libraries Screening Center Network. There was an overlap of 2,595 compounds common to the two sets so that the total number of unique compounds evaluated in these two sets was 23,720. An additional 586 compounds were cherry picked from the remainder of the MLSMR compound collection on the basis of a previously described virtual screening approach for predicting FPR activity.

The primary high throughput screening (HTS) assay was performed in 384 well format. Test compounds were assessed at a single concentration of 6.7 microM for the ability to inhibit fluorescent ligand binding, detected as a decrease in cell fluorescence due to displacement of fluorescent ligand from FPR. The FPRL1 primary HTS assay results obtained in parallel in the same wells have been reported separately (AID 441) and represent counter-screen data with which to determine selectivity and specificity of compounds with FPR binding activity identified in this report. Likewise, FPR binding results reported here represent counter-screen data with which to determine the selectivity and specificity of compounds identified to have FPRL1 binding activity in the FPRL1 primary HTS assay report (AID 441) For assay performance, additions to wells were in sequence as follows: 1) test compounds and control reagents (5 microL/well); 2) a combination of FPR- and FPRL1-expressing cell lines (10^7/mL each, 5 microL/well); 3) (after 30 min, 4 degrees C incubation) fluorescent peptide (5 microL/well). After an additional 45 min, 4 degrees C incubation, plates were immediately analyzed by flow cytometry. The assay response range was defined by replicate control wells containing unlabeled receptor-blocking peptide (positive control) or buffer (negative control). fMLFF (4Pep) was used as the FPR-blocking peptide, unlabeled WPep as the FPRL1-blocking peptide. The assay was homogeneous in that cells, compounds and fluorescent peptide were added in sequence and the wells subsequently analyzed without intervening wash steps. The HyperCyt high throughput flow cytometry platform was used to sequentially sample cells from wells of 384-well microplates (2 microL/sample) for flow cytometer presentation at a rate of 40 samples/min. The resulting time-resolved data files were analyzed with IDLeQuery software to determine compound activity in each well.

PubChem Assay (ID 440)

Origin: http://pubchem.ncbi.nlm.nih.gov/bioassay/440

bioassay type —(has bioassay)→ protein-small molecule interaction assay

assay format —(has assay format)→ cell based format

assay design method —(has assay design method)→ fluorescent ligand binding method

assay cell line —(is cell line of)→ U-937 cell

organism —(has organism)→ Homo sapiens

biological process —(has biological process)→ neutrophil activation
G-protein coupled receptor signaling pathway

target —(has biological macromolecule)→ G protein coupled receptor
"FPR"

assay mode of action —(has mode of action)→ inhibition
ligand binding mode of action
competitive binding

result —(has result)→ percent inhibition

result unit of measurement —(has unit of measurement)→ percent

assay screening campaign stage —(has assay stage)→ primary assay
counter screening assay

assay footprint —(has assay footprint)→ 384 well plate

assay kit —(uses assay kit)→ (not assigned)

physical detection method —(has detection method)→ flow cytometry

detection instrument —(uses detection instrument)→ HyperCyt High Throughput Flow Cytometry System

perturbagen type —(has perturbagen)→ MLSMR library
"17K Set Type 1 (17KST1)"
"10K Set Type 1 (10KST1)"

Figure 10 (on next page)

Example of an assay partially suited for annotation with the CAT

Figure 10. Example of an assay partially suited for annotation with the CAT. Left: Text from description in PubChem Assay ID 488847: yellow = information captured in CAT, pink= information added as 'literal' values (i.e., too specific to exist as a BAO entry, but deemed valuable), green = information not captured but possible for a future version (e.g., controls, labels of target and ligand, assay quality data (Z')), red= information beyond the scope of BAO (technical details). **Right:** CATvalues assigned in the BioAssay Schema Editor capture key parameters of the assay yet do not capture the complexity of the assay articulated in the single sentence (arrow): "a flow cytometry protein interaction assay to screen for compounds that compete with RNA binding to GRK2".

Assay Background and Significance:
A small family of G protein-coupled receptor (GPCR) kinases (GRKs) negatively regulates heterotrimeric G protein signaling by phosphorylating multiple sites in the cytoplasmic loops and tails of activated GPCRs [Krupnick, et al. 1998]. Through this process, cells adapt to persistent stimuli that act at GPCRs and protect themselves from damage incurred by sustained signaling. GRKs can also play maladaptive roles in human disease. GRK2 is overexpressed during heart failure, which not only uncouples cardiac receptors from the central nervous system, but also promotes the release of excessive amounts of catecholamines from the adrenal gland [Vatner, et al 1996]. Inhibition of GRK2 by transgenic peptides prevents cardiac failure in mouse models [Rockman, et al. 1998], suggesting that GRK2 is an excellent target for the treatment of heart disease. However, selective small molecule inhibitors of GRKs have not been reported, perhaps due to high homology among the active sites of GRKs and other AGC kinases. Over the last six years, our lab has made significant progress in understanding the structure and function of GRKs, and we are currently investigating the molecular basis for the selective inhibition of GRK2 by a high affinity RNA aptamer [Tse and Boger, 2005]. Preliminary crystallographic studies of this complex demonstrate that the aptamer binds primarily to the large lobe of the kinase domain, where it blocks the entrance to the nucleotide binding site of the kinase domain. In the HTS assay reported here, an RNA aptamer is used in a displacement assay to identify small molecules that bind to regions on GRK2 outside of its active site that are also critical for activity. This is a robust flow cytometry protein interaction assay to screen for compounds that compete with RNA binding to GRK2. Using activity-based secondary screens, we will confirm which hits derived from HTS campaigns exhibit direct binding to GRK2 and inhibit kinase activity. These compounds will be further characterized to establish membrane permeability, their mode of inhibition, and their selectivity for GRK2. Although all active molecules are of interest, small molecules that do not exhibit competitive inhibition with ATP are of particular importance because they would likely represent novel and selective therapeutic leads for the treatment of heart disease.
GRK2 protein is biotinylated using biotinamidohexanoic acid N-hydroxysuccinimide ester(Sigma). The RNA aptamer is fluorescently labeled on the 3'end with carboxyfluorescein (synthesized and labeled byIDT). Streptavidin-coated beads (Spherotech) are incubated with biotinylated GRK2 (bGRK2) at a final concentration of 2 nM for 30 minutes. The BioTek Microflow liquid dispenser is used to dispense 4 microL of assay buffer to all but column 1 of a 384-well assay plate. The positive (blocked) control containing 50X unlabeled RNA aptamer in assay buffer is dispensed to column 1 by a Microflow liquid dispenser (BiotTek, USA). Compounds (10 microM in-well concentration) are transferred to assay wells via 100 nanoL pintool transfer on the Biomek FX liquid dispenser (Beckman Coulter, USA. A total of 3 microL of bead suspension is dispensed into assay wells using the Nanoquot liquid dispenser (BioTek, USA). Plates are incubated at RT for 30 min. 3 microL FAM-C13.28 aptamer (final concentration 2 nanoM, supplied by the assay provider) is added to assay wells using the Microflow liquid dispenser. The reaction is incubated for one hour at RT. In this flow cytometry-based HTS [Kuckuck, et al. 2001] a CyAn flow cytometer (Dako / Beckman Coulter) interfaced with a HyperCyt (IntelliCyt, USA) auto-sampler is used to measure the median fluorescence intensity associated with bead-bound bGRK2.

Calculation:
For plates that passed the Z' test (Z'>.30) a compound was considered active if the PERCENT_RESPONSE > .40. The Z' mean for all the plates was 0.8 with a standard deviation of 0.2.

The 40% cutoff corresponds to about three times the standard deviation of PERCENT_RESPONSE from 'non-fluorescent' test compounds. Negative PERCENT_RESPONSE is primarily due to test compounds with innate fluorescence.

PUBCHEM_ACTIVITY_SCORE = PERCENT_RESPONSE
PUBCHEM_ACTIVITY_OUTCOME = 2 (or ACTIVE) if PUBCHEM_ACTIVITY_SCORE > 40, otherwise the PUBCHEM_ACTIVITY_OUTCOME = 1 (or INACTIVE).

PubChem Assay (ID 488847)

Origin: https://pubchem.ncbi.nlm.nih.gov/bioassay/488847

bioassay type —(has bioassay)→ protein-RNA interaction assay
protein-small molecule interaction assay

assay format —(has assay format)→ biochemical format

assay design method —(has assay design method)→ binding assessment method

assay cell line —(is cell line of)→ (not assigned)

organism —(has organism)→ Homo sapiens

biological process —(has biological process)→ G-protein coupled receptor signaling pathway

target —(has biological macromolecule)→ kinase
"GRK2"

assay mode of action —(has mode of action)→ competitive binding

result —(has result)→ percent response

result unit of measurement —(has unit of measurement)→ percent

assay screening campaign stage —(has assay stage)→ primary assay

assay footprint —(has assay footprint)→ 384 well plate

assay kit —(uses assay kit)→ (not assigned)

physical detection method —(has detection method)→ flow cytometry

detection instrument —(uses detection instrument)→ CyAn Flow Cytometer

perturbagen type —(has perturbagen)→ compound library

11

Dialog box for random lookup of assays from PubChem

Figure 11: Dialog box for random lookup of assays from PubChem.

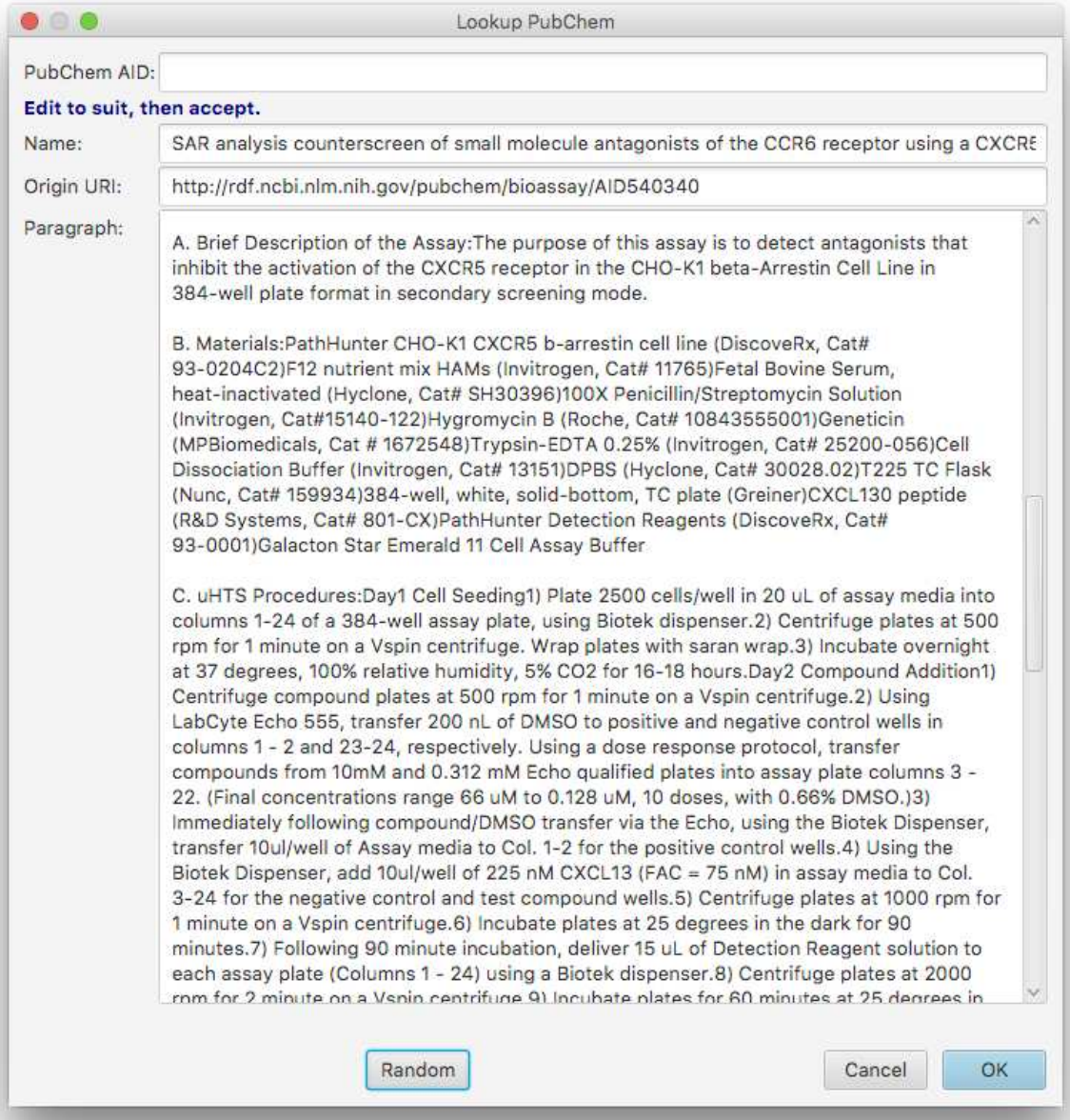


Table 1 (on next page)

Representation of Common Assay Template in Sample Assay Set

Table 1. Representation of Common **Assay T**emplate in Sample Assay Set[b]

Table 1. Representation of Common Assay Template in Sample Assay Set

CAT Assignment	Test Assays (of 43) With at Least 1 Value	# of Unique Values Annotated
bioassay type	43 (100%)	24 of 88
assay format	43 (100%)	6 of 19
assay design method	43 (100%)	20 of 76
assay cell line	24 (55.8%)	15 of 95
organism	41 (95.3%)	11 of 65
biological process	40 (93.0%)	28 of 54
target	32 (74.4%)	13 of 38
assay mode of action	43 (100%)	8 of 13
result	41 (100%)	16 of 94
result unit of measurement	32 (74.4%)	6 of 56
assay screening campaign stage	40 (93.0%)	8 of 23
assay footprint	36 (83.7%)	5 of 20
assay kit	9 (20.9%)	5 of 93
physical detection method	42 (97.7%)	11 of 51
detection instrument	26 (60.5%)	9 of 97
perturbagen type	20 (46.5%)	3 of 9