Published online 19 November 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D255–D266***
doi:10.1093/nar/gkp965

An overview of the PubChem BioAssay resource

Yanli Wang, Evan Bolton, Svetlana Dracheva, Karen Karapetyan, Benjamin A. Shoemaker, Tugba O. Suzek, Jiyao Wang, Jewen Xiao, Jian Zhang and Stephen H. Bryant*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA

Received September 11, 2009; Revised October 8, 2009; Accepted October 13, 2009

ABSTRACT

The PubChem BioAssay database (http://pubchem.ncbi.nlm.nih.gov) is a public repository for biological activities of small molecules and small interfering RNAs (siRNAs) hosted by the US National Institutes of Health (NIH). It archives experimental descriptions of assays and biological test results and makes the information freely accessible to the public. A PubChem BioAssay data entry includes an assay description, a summary and detailed test results. Each assay record is linked to the molecular target, whenever possible, and is cross-referenced to other National Center for Biotechnology Information (NCBI) database records. 'Related BioAssays' are identified by examining the assay target relationship and activity profile of commonly tested compounds. A key goal of PubChem BioAssay is to make the biological activity information easily accessible through the NCBI information retrieval system-Entrez, and various web-based PubChem services. An integrated suite of data analysis tools are available to optimize the utility of the chemical structure and biological activity information within PubChem, enabling researchers to aggregate, compare and analyze biological test results contributed by multiple organizations. In this work, we describe the PubChem BioAssay database, including data model, bioassay deposition and utilities that PubChem provides for searching, downloading and analyzing the biological activity information contained therein.

INTRODUCTION

PubChem (http://pubchem.ncbi.nlm.nih.gov) (1,2) (Figure 1) is an open public repository containing chemical structures and biological properties of molecules including small molecules and siRNA reagents. This resource, first available in September 2004, is part of the US National Institutes of Health (NIH) Molecular Libraries Roadmap Initiative. This research program aims to identify and develop chemical probes through high-throughput screening of small molecules that modulate the activity of gene products (4,5) and to accelerate chemical biology research and facilitate drug development by offering biomedical researchers access to the large-scale screening capacity and the biological test results generated via PubChem.

PubChem consists of three interconnected databases: Substance, BioAssay and Compound. The Substance database (primary accession—SID) contains contributed sample descriptions (primarily small molecules) provided by depositors. The BioAssay database (primary accession—AID) contains contributed assay descriptions and associated biological screening results of substances provided by depositors. The Compound database (primary accession—CID) contains the unique chemical structures derived from the Substance database records, thus allowing substance information (e.g. bioassay data) from different depositors to be viewed for unique chemical structures. The PubChem BioAssay system is the repository of the small molecule screening data generated by the Molecular Library Screening Center Network (MLSCN) and the Molecular Library Probe Production Center Network (MLPCN) under the NIH Molecular Libraries Program (MLP) (6,7). PubChem also receives biological property contributions from many other organizations.

Biological test results contained in the PubChem BioAssay database consist of information generated through high-throughput screening experiments, biological and medicinal chemistry research, as well as those extracted from the literature. PubChem BioAssay currently contains over 1700 biochemical and cell-based bioassay screens, containing nearly 60 million biological activity outcomes for several thousand different protein and gene targets. These test results provide biological annotations for more than 750000 unique small molecule chemical structures and tens of thousands of siRNA probes. While the majority of the archived test results were deposited by MLP screening centers, PubChem BioAssay contains biological test results contributed by

*To whom correspondence should be addressed. Tel: +1 301 435 7792; Fax: +1 301 480 9241; Email: bryant@ncbi.nlm.nih.gov

Figure 1. PubChem home page. One can search PubChem Substance, Compound and BioAssay by entering the search term into the input box or access the summary, chemical structure search and bioactivity analysis service using the respective links.

a number of US government organizations, research programs at various academic institutions and individual research laboratories. This includes, for example, human tumor cell line screening data from the Developmental Therapeutic Program (DTP) (8,9) at the US National Cancer Institute (NCI), toxicology data from the DSSTox (10) program at the US Environmental Protection Agency (EPA), biological test results from the US National Institute of Neurological Disorders and Stroke (NINDS) Approved Drug Screening Program and the US National Institute of Mental Health (NIMH) Psychoactive Drug Screening Program (PDSP), as well as anticonvulsant data from NINDS. It also includes ligand–protein binding activity data generated by the targeted high-throughput structural biology experiments at the European Structural Genomics Consortium (11), literature-extracted bioactivity data from the BindingDB (12) project, the IUPHAR (13) project and the PDBBind (14) project, high-throughput screening results from ChemBank (15), and target profiling and phenotypic assays from commercial vendors (16). A new addition to PubChem BioAssay database is biological activity data for siRNA probe reagents. The siRNA screening results currently contained in PubChem BioAssay include high-throughput siRNA screening data contributed by the RNAi Global Initiative (http://www.rnaiglobal.org/), data extracted from the literature by the NCBI Probe resource (http://www.ncbi.nlm.nih.gov/probe) as well as a recent contribution from authors of the journal 'Cell' reporting research results on the identification of clock genes and modifiers (17).

With substantial growth in data volume and diversity, and increasing demand from public users, PubChem faces great challenges comparing to many other chemical biology resources. Collecting, archiving and organizing each individual biological test, as well as bioassay information across multiple screening experiments and across multiple types of reagents (i.e. small molecules versus siRNAs), complement data quantity with diversity and breadth of coverage of available information. Providing effective means to store, retrieve and analyze tens of millions of bioassay outcomes across thousands of bioassays is a considerable task. Developing a robust information-tracking system, providing a user-friendly data deposition interface as well as data analysis tools for public usage all add to the complexity of the effort. Such systems and services, by necessity, must be intuitive to use while still providing advanced search features to help researchers find relevant information and make the scientific connections that transform data into knowledge. PubChem BioAssay accomplishes this by encouraging chemical biology researchers to share their scientific

discoveries through the publically available PubChem resource, by integrating with the rich database resources and information retrieval system at NCBI. Such database integration provides scientists a series of integrated tools and services to analyze biological test results and to extract those biologically interesting molecules and chemical probes contained within PubChem.

PUBCHEM BIOASSAY DATA MODEL

The PubChem BioAssay model is designed to allow unambiguous representation of data produced by various experimental procedures, to support the retrieval of individual information components and to track the biological target and the respective bioactivity outcome. An assay record, represented by a unique PubChem BioAssay accession AID, is organized in two parts, the assay description and the assay results. The assay description includes a name, data source, purpose, experimental protocol, tested reagent category (e.g. small molecule versus siRNA), comment and result/readout descriptions. The PubChem BioAssay archive format provides numerous ways for contributing organizations to annotate a given assay. These annotations include: textual descriptions; target information, including cross-references to GenBank (18) records, name, description, molecule type (e.g. protein versus nucleotide) and taxonomy; qualified cross-references to PubMed citations, three-dimensional protein structures, biosystems and diseases; and URLs back to the depositor's website. Each BioAssay record can contain as many comments and as much descriptive text as needed to provide the overall background of the assay, such as the biological system tested in the assay or the relationship between a disease and the selected therapeutic target. The assay protocol, similar to the method section of a journal publication, helps to explain the actual methodology of the assay.

Multiple test result fields may be specified per assay, each with a unique test identifier (TID), name, description, data type, data unit and annotation for cross-references. The PubChem assigned TID indicates a particular test result or readout when reporting results for a given substance. The number of test readouts are only limited by their potential usability.

Many biological assays employ a dose–response scheme, with a primary endpoint [e.g. IC50 (http://en.wikipedia.org/wiki/IC50)]. PubChem requires this key readout, denoted as an 'active concentration summary', to have micro-molar units and requires the experimental concentrations for the corresponding dose–response readouts (also in micro-molar concentration and referred to as 'tested concentrations') to be designated on the respective test result fields as an attribution. These specialized readouts allow PubChem users to classify and rank hits of a screening test and search bioassay results with specific values or ranges of primary outcomes.

Biological screening data submitted to PubChem are diverse and assay specific. As such, there are no specific requirements on the presence of particular test readouts; however, PubChem requires a summary result for each tested chemical sample. The summary result is 2-fold: bioactivity outcome and bioactivity score. The 'bioactivity outcome' partitions results and includes five categories: chemical probe, active, inactive, inconclusive and unspecified. Criteria and rationale used by the testing organization for summary results, as well as description about possible factors of artifact, are often provided in the assay comment section, aiding the user's interpretation and utilization of the biological data.

The assay result section includes the results for all tested substances. Results reported per substance can include both assay readout and annotations, including target description, comment on the individual biological test result, cross-links to other NCBI resources and URLs to the depositor's website. Assay data are provided in a tabular format, with one tested substance per row and one assay test readout or annotation per column. A substance need not have results reported for all defined test readouts. There is no limit on the count of substance test results in an assay record.

The stage of the biological experiments in PubChem varies. Each assay is classified by the contributing organization according to the stage of the assay project, which is described as the 'activity outcome method' in PubChem. These methods include: 'screening' assay, usually a primary high-throughput assay where the activity outcome is based on percentage inhibition from a single dose; 'confirmatory' assay, typically a low-throughput assay where the activity outcome is based on a dose–response relationship with multiple tested concentrations; 'summary' assay, for validated chemical probes or small molecule leads, summarizing information from multiple assays; and 'other', those assays that do not fit the previous categories. For MLPCN projects, a summary assay is required for each biological screening project to describe the identified chemical probes, report screening steps that lead to the project progress and communicate the bottom line of the screening campaign to the scientific community. A summary assay, therefore, consists of a list of verified chemical probes if identified, a comprehensive text description of the screening campaign and links from this summary assay AID to all associated screening assays deposited in PubChem, to the targeted genes and proteins in GenBank and to the scientific publications describing the screening experiments available in PubMed. To make it easier for the MLP screening centers to create and update summary assays, the PubChem deposition system allows one to create a simple summary template at an early stage and to update the summary assay with any new information and experimental results as the screening project progresses.

It is essential to specify and track the information of the assay target and precisely group and annotate the biological tests based on the respective molecular target. PubChem BioAssay provides several models to do so. The traditional assay model allows for the specification of a single target for the entire assay record, along with associated annotations such as links to the respective gene, taxonomy and biological pathway information.

In this model, the bioactivity outcomes provided in the entire assay dataset are solely for the specific target, for example, to describe the biological effect of the small molecules on the functionality of one enzyme.

PubChem also supports the presentation and annotation of multiple highly related bioactivity outcomes, such as a profiling assay against a panel of molecular targets, in a single assay. Such a panel-type PubChem BioAssay record can contain multiple test readouts and respective bioactivity outcome annotations for each individual target, as well as for an individual cell line or species defined within the 'panel'. Each of such targets, cell lines or species is regarded as a 'panel component' in the data model and a unique panel component identifier (PID) is assigned to each. Description of the experiments, including a component name, general goal, specific experimental protocol and information of assay target, can be provided for each individual panel component. The test results for an individual panel component, which may be multiple, can be designated as a 'bioactivity outcome' or 'active concentration', if necessary, or otherwise treated as regular readouts.

Panel assay results are complex. This expansion of the PubChem BioAssay data model allows for the description of a compound profiling screening test and enables PubChem to record and annotate multiple related bioactivity outcomes under a single AID, which helps facilitate comparison and evaluation of compound bioactivities using the PubChem data analysis profiling tools. To see a panel assay example, one may examine the kinase-profiling assay for AID 1433 (http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=1433).

A third bioassay data model allows one to specify a specific target for each individual tested sample. In this case, a specific test result is defined as containing the assay target. For example, one test result definition may be defined as containing the target identifier, a GenBank Protein GI number, while another test result definition may be defined as containing a short name of the target. This model was introduced originally to support the accommodation of siRNA screening results, where an entire genome may be screened with tens of thousands of siRNA reagents designed for thousands of gene targets of the genome, with one or several siRNA reagents corresponding to each of the targeted genes. Thus, in this situation, the nucleotide or gene target annotation needs to be siRNA specific, and associated with each tested sample. To see an example, one may look at AID 1622, a viability screen of human kinase and cell cycle genes (http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=1622). This data model can also be employed to encode bioactivity information for multiple targets where substance data points across targets are sparsely populated, such as the data contributed by PDBbind in AID 1811 (http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=1811).

In all, the PubChem BioAssay data model supports a comprehensive description of screening experiments and test results by providing flexible schemes to encode bioassay information. Such schemes facilitate public user discovery; for example, the 'bioactivity score' provides the relative activity rank to enable more interesting results to be shown first, while the provision of the bioactivity outcome summary to provide a shortcut from each assay to a list of 'hits' discovered in each screen, empowers the PubChem user to rapidly identify and partition biological assay results of interest. More importantly, the required standard bioassay metrics and annotation of assay targets allow PubChem to provide powerful tools for users to promptly classify and compare results across disparate assays and targets for a given set of chemical samples.

PUBCHEM BIOASSAY DATA SPECIFICATION, DATABASE STRUCTURE AND UPDATE TRACKING

The hierarchical data in the PubChem BioAssay archive are encoded in the data structure ASN.1 notation. All information about a single assay can be contained in a single ASN.1 or equivalent XML data object. It provides separate tagged fields for each aspect of the assay as detailed in the available specification in ASN.1 and XML Schema formats, respectively:

ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem.asn
ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem.xsd

PubChem assay records are stored in a relational database using Microsoft SQL server. The database architecture is designed for efficient storage, tracking and fast retrieval of large-scale biological test results. Test results are keyed on a unique combination of three identifiers: AID, SID and TID. Assay data are partitioned into multiple 'de-normalized' database tables for rapid search and retrieval. Certain summary information, such as tested substances count, unique chemical structures count and TID numerical range are pre-computed to further facilitate efficient retrieval through the data analysis tools.

PubChem allows all aspects of a bioassay to be revised. Updates to a bioassay textual description or annotation are considered to be a 'minor change' and recorded as a description revision. Updates to a substance result are tracked by increasing the 'test result' version, with both duplicate tests and revision to an existing test being considered as test result updates. Major revision to a bioassay is allowed, which includes changes such as addition or removal of test result fields. All bioassay test results must be restated by the data depositor upon such fundamental changes. Whenever a major version is incremented, the description revision and test result version are reset to '1'. Only the current version of description and corresponding test results are shown in the PubChem display system. However, all revisions are archived, tracked and retrievable.

BIOASSAY DEPOSITION AND UPDATE

PubChem is an open repository and any organization may contribute data. The only caveat is that contributed information must be real (e.g. the substance data must exist and the primary assay data experimentally derived) and not theoretical or virtual in nature. To deposit data, one

can visit the PubChem Deposition Gateway website (http://pubchem.ncbi.nlm.nih.gov/deposit/) and fill out the web form for a PubChem deposition account. When applying for a deposition account, one needs to provide a valid contact address, bonafide organization information and specify a name for the data source, which is the data attribution describing the resource to which all future depositions will be referred. Furthermore, one must agree to and is bound by the 'PubChem Data Transfer Agreement' (http://pubchem.ncbi.nlm.nih.gov/deposit/dta.html), where the depositor retains rights to their data but gives PubChem the ability to display and redistribute provided information. A hold-until date may also be specified for an assay submission. This may be desirable for journal-article authors, for example, to make auxiliary data in PubChem available as of the publication date of the article. If a hold-until date is specified, the deposited assay record will be uploaded immediately but the data content is made available for search and public access only upon that date.

It is straightforward to deposit data. To submit a biological assay, an appropriate description must be provided. These may be conveniently entered using the webforms in the PubChem Deposition Gateway. Assay test results for substances may be uploaded using a comma-separated value (CSV) formatted file. The uploaded data resemble a data table, with each column in the CSV file corresponding to a particular test result and each row representing all of the test results for a single substance record. Prior to submission of assay results, a description of the substances tested in an assay must be provided and publicly available in PubChem. Substance deposition can be made using the industry standard SD format (http://www.symyx.com/downloads/public/ctfile/ctfile.jsp). The SD file should contain predefined PubChem SD fields to be accepted by the PubChem deposition system. The only requirement is that each substance in the SD file must have a unique registry ID. In addition to web-based submission, XML-based submissions and automated FTP data transfers are also possible. Trained staff is available to assist and detailed instruction for depositing data into PubChem is available (http://pubchem.ncbi.nlm.nih.gov/deposit/deposit_help.html).

Upon data submission, automated data validation is performed to ensure that the submitted assay data follows the data type description of the readout, for example. Contributing organizations may preview assay description and data presented by PubChem assay summary service prior to commit their data to be loaded into PubChem, whereby PubChem staff perform a final check of the data submission and approve the deposition for release into PubChem. If not an update, an AID is assigned at load time and a final notification is sent to the depositor after contributed data are publicly available.

All updates on PubChem BioAssay records, including minor changes, major revisions and test result additions, may be done through the deposition interface or using the FTP data transfer scheme. Individual test results as well as an entire assay record can be revoked. Similarly, a PubChem substance record can be revoked, though all associated assay results must be revoked first. Assay version, description version or test result version may change depending on the update type, but the assay AID remains the same.

It should be noted that biological test results refer to substances deposited by the same contributor to associate test results with the physical samples tested. Assay data may also refer to substances deposited by another contributor, but only when the latter is designated as a collaborating sample provider. In such collaborations, the substance depositor must agree to associate unique substance identifiers (SIDs) with the physical samples provided to any assay depositor. The NIH Molecular Libraries Program (MLP) organizes one such collaboration, under which the NIH Molecular Libraries Small Molecule Repository (MLSMR) associates new PubChem SIDs with any newly purchased, newly synthesized or newly purified physical sample. Each MLP screening center can test substance samples in this collection and refer to MLSMR substance descriptions when depositing screening results in PubChem.

INTEGRATION WITH PUBCHEM SUBSTANCE, COMPOUND, TARGET AND OTHER NCBI DATABASES

Each assay record is linked to all tested substances. A substance is in turn linked, whenever possible, to a corresponding compound. Separate links are created from each assay to subsets of chemical probes, active and inactive substances and compounds. Similarly, substances and compounds are linked to the bioassays where they are tested, as well as to the subsets of bioassays where they are chemical probes, active or inactive. These shortcuts allow one to rapidly retrieve a list of active substances or compounds directly using the Entrez search system (http://www.ncbi.nlm.nih.gov/sites/gquery), the text-based global interface for searching all NCBI databases (3). For example, to find out how many substances are considered active in a pair of assays with related protein targets, one could retrieve the list of active substances for each assay and then find the intersection of the lists by performing a Boolean 'AND' operation in the Entrez history facility. Alternatively, one can find out the subset of substance records that are active in a primary screening assay but inactive in a corresponding confirmatory assay. Navigating bioassay results in the same manner using compound records to aggregate on depositor substance descriptions creates a broader range of bioactivity profile available for the same chemical structure.

Assay target information allows PubChem to provide specialized data analysis tools and interfaces to navigate bioactivity results across protein targets, to examine compound selectivity, to identify off-target effects and to filter out promiscuous compounds. To date, more than half of all PubChem BioAssay records contain a discrete molecular target with a link to the corresponding GenBank protein and nucleotide records. Assay targets are further annotated with the NCBI BioSystems (http://www.ncbi.nlm.nih.gov/biosystems) biological pathway information, helping to facilitate understanding

of the biological role of the target and providing a context for the effect of bioactive compounds. Additionally, protein targets are linked to the curated NCBI Conserved Domain Database (CDD) (19) (http://www.ncbi.nlm.nih.gov/cdd) and to three-dimensional structures of closely related proteins contained in the NCBI Molecular Modeling Database (MMDB) (20) (http://www.ncbi.nlm.nih.gov/structure). Such links help to identify conserved active sites in proteins and protein–ligand interactions.

Searching PubChem BioAssay targets via the NCBI BLAST service (21,22) (http://blast.ncbi.nlm.nih.gov) provides an additional way for a molecular biologist to discover chemical probes or effective siRNA reagents contained in PubChem. Sequences of assay targets are used to construct a BLAST database for a protein similarity search. The BLAST service now highlights assay targets among the hits when reporting search results. Each of such BLAST hits is linked back to the associated assay records in PubChem, thus facilitating, for example, the discovery of bioactive compounds, including the highly potent and selective chemical probes identified for a specific target.

RELATED BIOASSAYS

For a public repository with biological test results contributed by many organizations, assays can be related in a number of ways. For example, different laboratories may submit assays that screen for inhibitors of the same protein target. The same laboratory may also provide different assay records for the same target, e.g. one for primary screening data and others reporting follow-up experiments. Given the breadth of tested targets per compound and compounds tested per target, PubChem provides the means to readily identify such assay relationships and suggests a focus for subsequent data mining.

Three computationally derived methodologies are provided for deriving inter-assay relationships in PubChem BioAssay: 'Related BioAssays by Activity Overlap', 'Related BioAssays by Protein Target Similarity' and 'Related BioAssays by BioSystems via Protein (or Gene) Target'. In the method 'Related BioAssays by Activity Overlap', one assay is considered related to another if there is at least one chemical structure active in both assays. Assays neighbored using this method are ordered according to the similarities of their activity profiles. Such methodology does not allow direct conclusions to be made on the relationship between assays; however, it may allow one to rapidly identify, and thereby avoid, promiscuous inhibitors, or to help discover more complex target-based relationships.

In the method 'Related BioAssays by Protein Target Similarity', related bioassays are identified by checking sequence similarity significance between protein targets. To this end, all unique assay target protein sequences in PubChem are compared to each other by performing a BLAST search, using an *E*-value threshold of 0.01 as the cutoff for pertinent similarity. This neighboring basis is useful to identify bioassays screening similar targets and allows one: to group compounds tested against the same or related targets; to isolate chemical agents with distinct biological effects, such as agonists and antagonists; to compile a list of inhibitors identified in different bioassays; and more.

In the method 'Related BioAssays by BioSystems via Protein (or Gene) Target', common biological pathways of the respective targets are examined. Related bioassays are identified when the protein or gene targets are involved in the same biological pathway found within the NCBI BioSystems database. Such a relationship allows users to utilize this neighboring relationship to aggregate assay results and identify compounds affecting a common pathway. This type of neighboring relationship can act as a bridge between small molecule screens against a discrete protein target or between interfering siRNA gene expression knockout experiments involved in the same pathway for disease-oriented therapeutic research.

Independent of computed bioassay neighboring, 'Related BioAssays' may be specified by the assay depositor. Normally, these relationships are provided when further confirmatory or counter-screenings are performed, thus providing the means to rapidly identify all screens involved in the same screening campaign. Typically, a 'Summary' assay is defined within such a grouping that provides an overview of how each assay is involved in the overall effort and recaps the findings.

The 'Related BioAssays Summary' tool, accessible from the BioAssay Summary service for an assay record, allows one to examine or to select one of the pre-computed or depositor-specified assay–assay relationships. Interestingly, there is often clear overlap between the respective related bioassays identified by activity overlap and by sharing protein target similarity. One bioassay example to better understand the neighboring relationships in PubChem was contributed by the NIH Chemical Genomics Center (NCGC), AID 1751 (http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=1751), which looks for inhibitors and activators of human muscle isoform 2 pyruvate kinase. Examination of the neighboring relationships for protein target by selecting the 'Target Similarity' tab (Figure 2a) indicates that there are several assay records containing the same targets (AIDs 1540, 1631 and 1634). In addition, two sets (AIDs 954, 958, 1542 and 1780; AIDs 1541, 1543, 1781, 1782) have high sequence identities: 96% and 72%, respectively. Several of these assays (AIDs 1504, 1631, 1541, 1543 and 1782) are also found among the neighbor lists when examining the related assays based on the 'Related BioAssays by Activity Overlap' method due to two commonly tested and commonly active compounds (with CIDs 650361 and 654375, respectively). These homologous pyruvate kinases with high sequence similarity level appear to bind some of the same compounds, potentially due to interaction with conserved binding pocket amino acid residues among the protein targets. Looking across the bioactivity results of such related assays may help one to identify and separate

non-selective versus selective inhibitors. Furthermore, checking assays with targets from the same biological pathways (Figure 2b), shows that the protein targets of AID 1751 and several related assays discussed above are all involved in the Type II diabetes mellitus pathway as annotated by the KEGG resource (23). Altogether, assay neighboring relationships and the web-based summary service provide the means to facilitate scientific discovery by combining related information available within PubChem and Entrez.

SEARCH, PUBLIC ACCESS AND DOWNLOAD

PubChem BioAssay can be searched using Entrez at (http://www.ncbi.nlm.nih.gov/pcassay) or through the PubChem web page at (http://pubchem.ncbi.nlm.nih.gov) (Figure 1). Queries are made using particular terms or keywords to locate and display matching assays. A default search will match the term to all indexed fields. For example, a search for 'Parkinson's disease' finds 19 records that were contributed by different laboratories (Figure 3). Entrez provides a document summary (DocSum) for each matching BioAssay record. This includes the assay name, data source, target description and links to all integrated resources. One can access the full assay record using the 'Summary' link or retrieve actual test results using the 'Data' link. In addition to index searching, Entrez provides cross-links for each assay record in the search results to related records in other Entrez databases, found in link menus called 'Related BioAssays', 'Chemicals', 'Target', 'Literature' and 'Other Links' to the right of each search result.

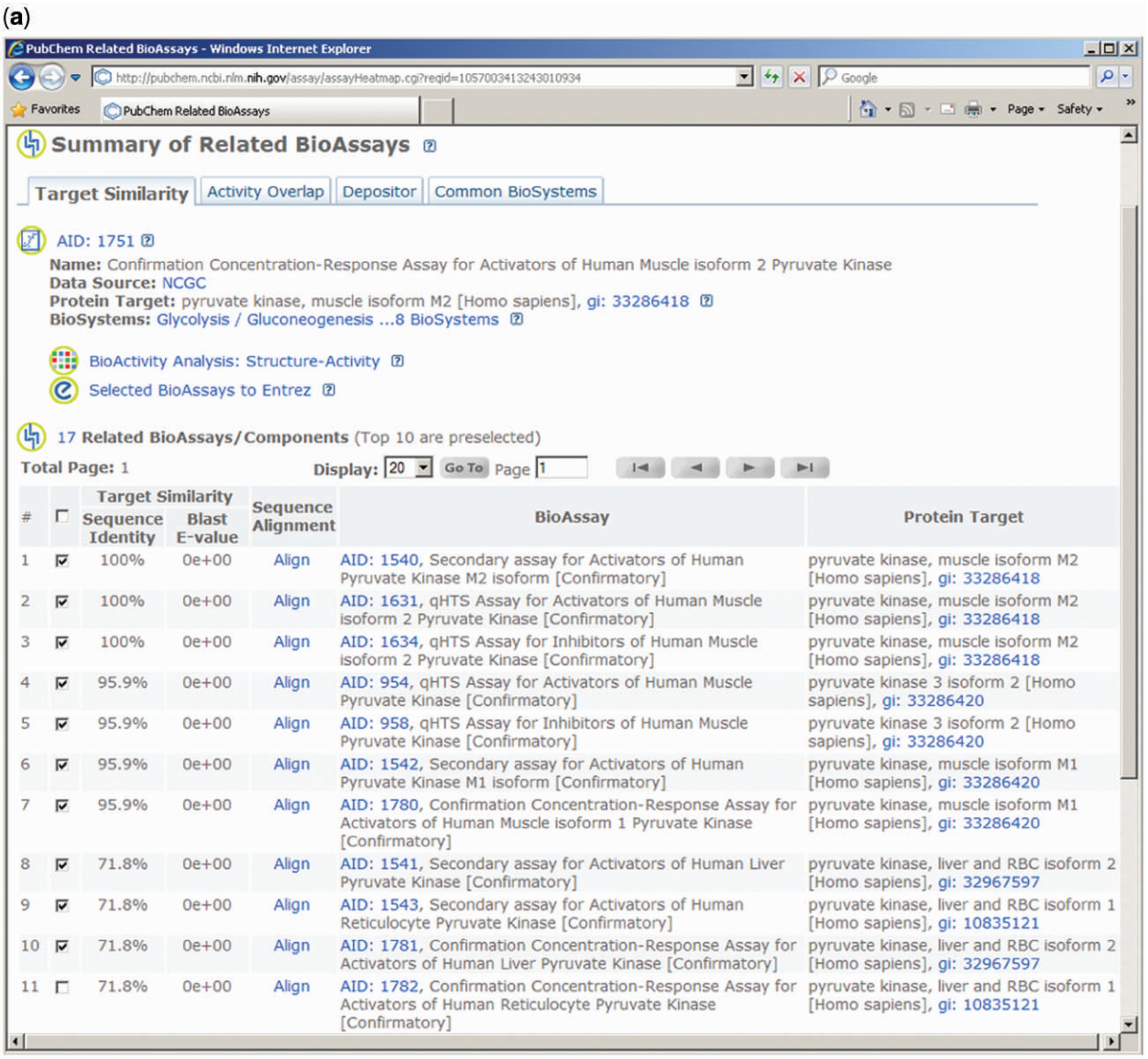
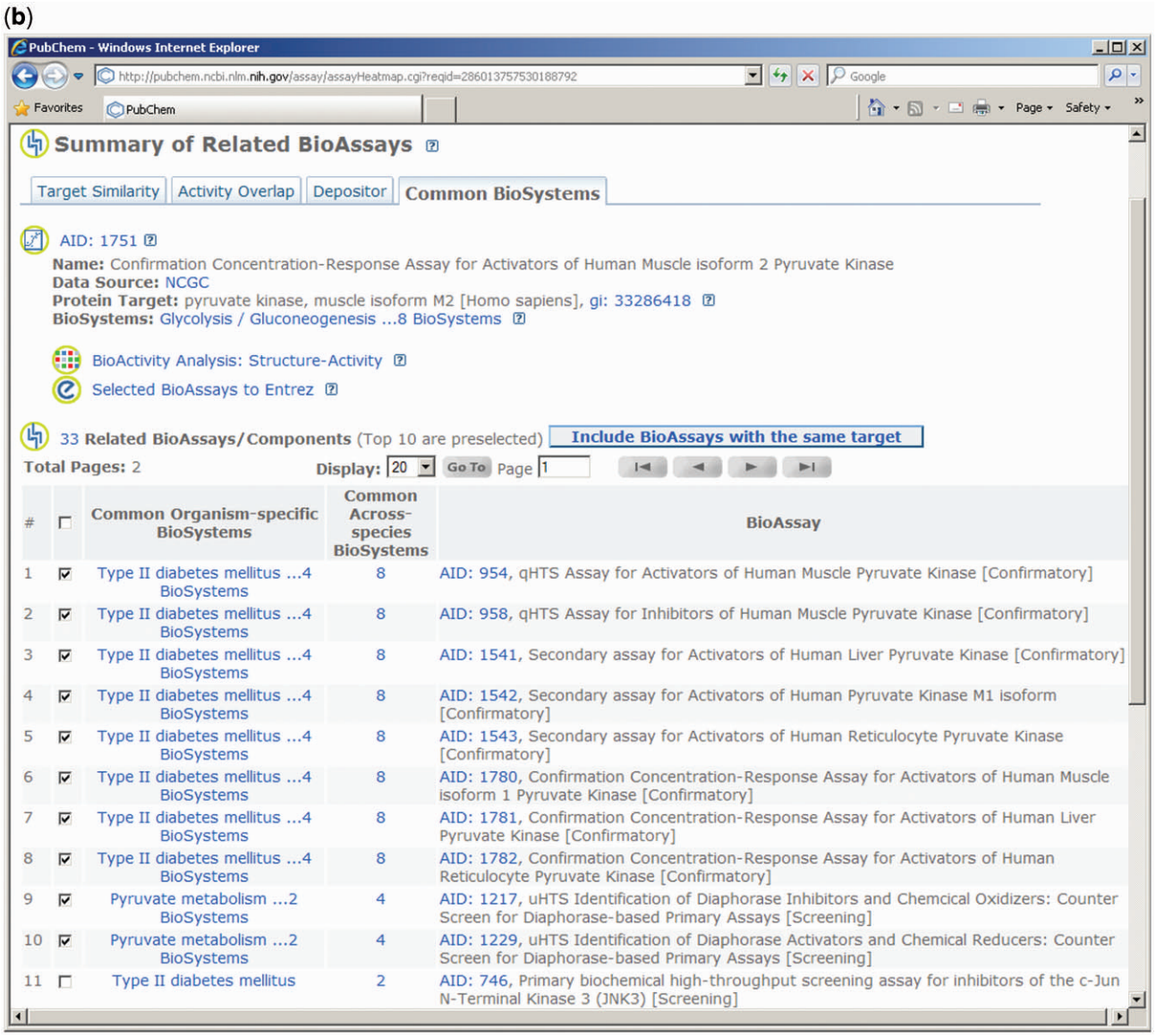


Figure 2. (**a** and **b**) Tool for visualizing related bioassay relationships. One can examine the four types of derived bioassay neighboring relationships for AID 1751 targeting human muscle isoform 2 pyruvate kinase using the provided 'Target Similarity', 'Activity Overlap', 'Depositor' and 'Common BioSystems' tabs.

(b)



#	Common Organism-specific BioSystems	Common Across-species BioSystems	BioAssay
1	Type II diabetes mellitus ...4 BioSystems	8	AID: 954, qHTS Assay for Activators of Human Muscle Pyruvate Kinase [Confirmatory]
2	Type II diabetes mellitus ...4 BioSystems	8	AID: 958, qHTS Assay for Inhibitors of Human Muscle Pyruvate Kinase [Confirmatory]
3	Type II diabetes mellitus ...4 BioSystems	8	AID: 1541, Secondary assay for Activators of Human Liver Pyruvate Kinase [Confirmatory]
4	Type II diabetes mellitus ...4 BioSystems	8	AID: 1542, Secondary assay for Activators of Human Pyruvate Kinase M1 isoform [Confirmatory]
5	Type II diabetes mellitus ...4 BioSystems	8	AID: 1543, Secondary assay for Activators of Human Reticulocyte Pyruvate Kinase [Confirmatory]
6	Type II diabetes mellitus ...4 BioSystems	8	AID: 1780, Confirmation Concentration-Response Assay for Activators of Human Muscle isoform 1 Pyruvate Kinase [Confirmatory]
7	Type II diabetes mellitus ...4 BioSystems	8	AID: 1781, Confirmation Concentration-Response Assay for Activators of Human Liver Pyruvate Kinase [Confirmatory]
8	Type II diabetes mellitus ...4 BioSystems	8	AID: 1782, Confirmation Concentration-Response Assay for Activators of Human Reticulocyte Pyruvate Kinase [Confirmatory]
9	Pyruvate metabolism ...2 BioSystems	4	AID: 1217, uHTS Identification of Diaphorase Inhibitors and Chemcical Oxidizers: Counter Screen for Diaphorase-based Primary Assays [Screening]
10	Pyruvate metabolism ...2 BioSystems	4	AID: 1229, uHTS Identification of Diaphorase Activators and Chemical Reducers: Counter Screen for Diaphorase-based Primary Assays [Screening]
11	Type II diabetes mellitus	2	AID: 746, Primary biochemical high-throughput screening assay for inhibitors of the c-Jun N-Terminal Kinase 3 (JNK3) [Screening]

Figure 2. Continued.

Links to the set of all search results are found in the 'Links' line above the results list. These links include related assays, active/inactive chemicals, molecular targets, biosystems, PubMed articles, genes, taxonomy, OMIM, etc. For example, one may find out how many targets have been tested among the 19 bioassay experiments related to 'Parkinson's disease' by clicking the 'Target|Protein Target' link on the top of the page, or one may retrieve the active compounds discovered by the 'JNK3 AlphaScreen Assay' (AID 530) via the 'Chemicals|PubChem Compound, Active' link associated with AID 530.

In Entrez, the scope of a search can be restricted to a particular search field (e.g. target name). To do this, specify the name of the field in brackets after the search term. For example, to search for assays for the AKT1 target, one could enter 'AKT1[ProteinTargetName]'. Alternatively, search fields can be chosen from a pulldown menu on the advanced search page found by clicking on the 'Limits' tab in the DocSum view. The 'Limits' facility helps to build a query from a subset of available search fields and filters (Figure 4). For example, one can easily identify 'Summary' assays containing a chemical probe report by selecting the 'Summary, Probe' option under the 'BioAssay Type' section and clicking 'Go' at the top of the page. Queries can also be built up to refine a search. For example, a query in PubChem BioAssay of 'aspirin AND "NIH Molecular Libraries Program"[SourceCategory]' will retrieve all assays deposited by the NIH-funded screening centers where aspirin was tested. To keep things simpler for complex searches, the 'History' tab can be used to combine previous searches with Boolean logic. There is also a 'Preview/Index' tab to find a detailed listing of all search terms being indexed. A list of all search fields and link filters are in the PubChem Help page

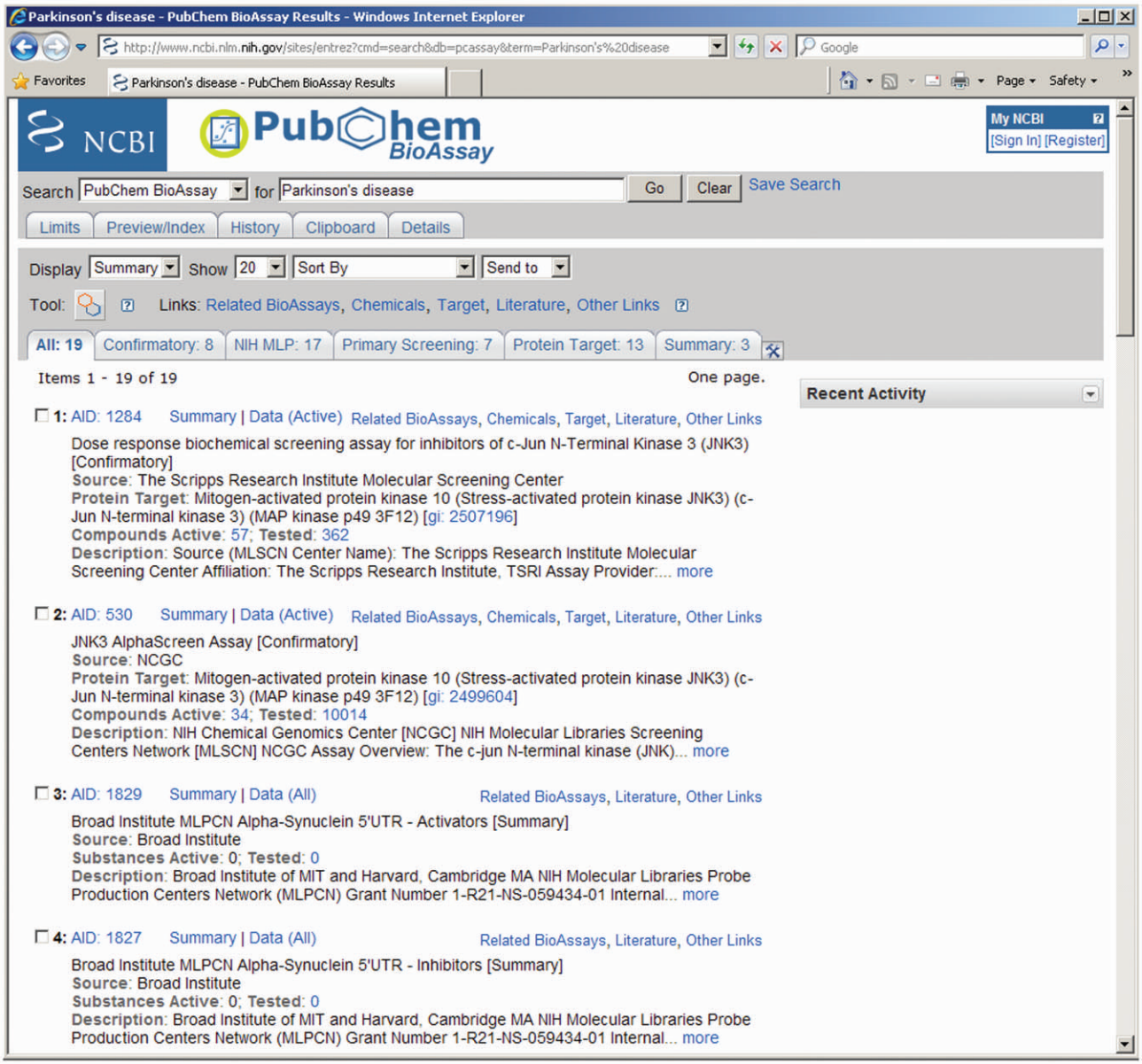


Figure 3. Search of the PubChem BioAssay database using the keywords 'Parkinson's disease' returns, among others, a series of related assays screening for the various targets associated with the disease process. Cross-links with other database records such as active compounds and protein targets are available via the links pop-up menus at the top of the DocSum page for the entire search result list and are available via link menus on the right side of each entry in the DocSum for the individual bioassay record.

(http://pubchem.ncbi.nlm.nih.gov/help.html#PubChem_index). Documentation for general use of the NCBI Entrez system is also available (http://www.ncbi.nlm.nih.gov/Database/index.html).

One can directly access individual assay records through the BioAssay Summary service by providing a valid PubChem BioAssay accession (AID). A direct URL to a given bioassay can easily be constructed, such as for assay AID 123 (http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=123). Similarly, direct URLs of summaries for substance SID 123 (http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?sid=123) and compound CID 123 (http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=123) are easy to construct as well.

A bioassay summary for each assay record provides a summary of all deposited assay information, including descriptions, annotations and assay results (Figure 5). It allows one to retrieve and download test results as well as chemical structures of tested small molecules through the 'BioAssay Results|Data Table' links. It also provides a central entry point to begin an analysis of active compounds using the bioactivity analysis tool through the 'BioActive Compounds|BioActivity Summary' or 'BioActive Compounds|Structure-Activity Analysis' link. It allows one to learn more about the target by following the various annotations, to explore assay neighboring relationships through the 'Related BioAssays|Summary' links or to review depositor-provided cross-references.

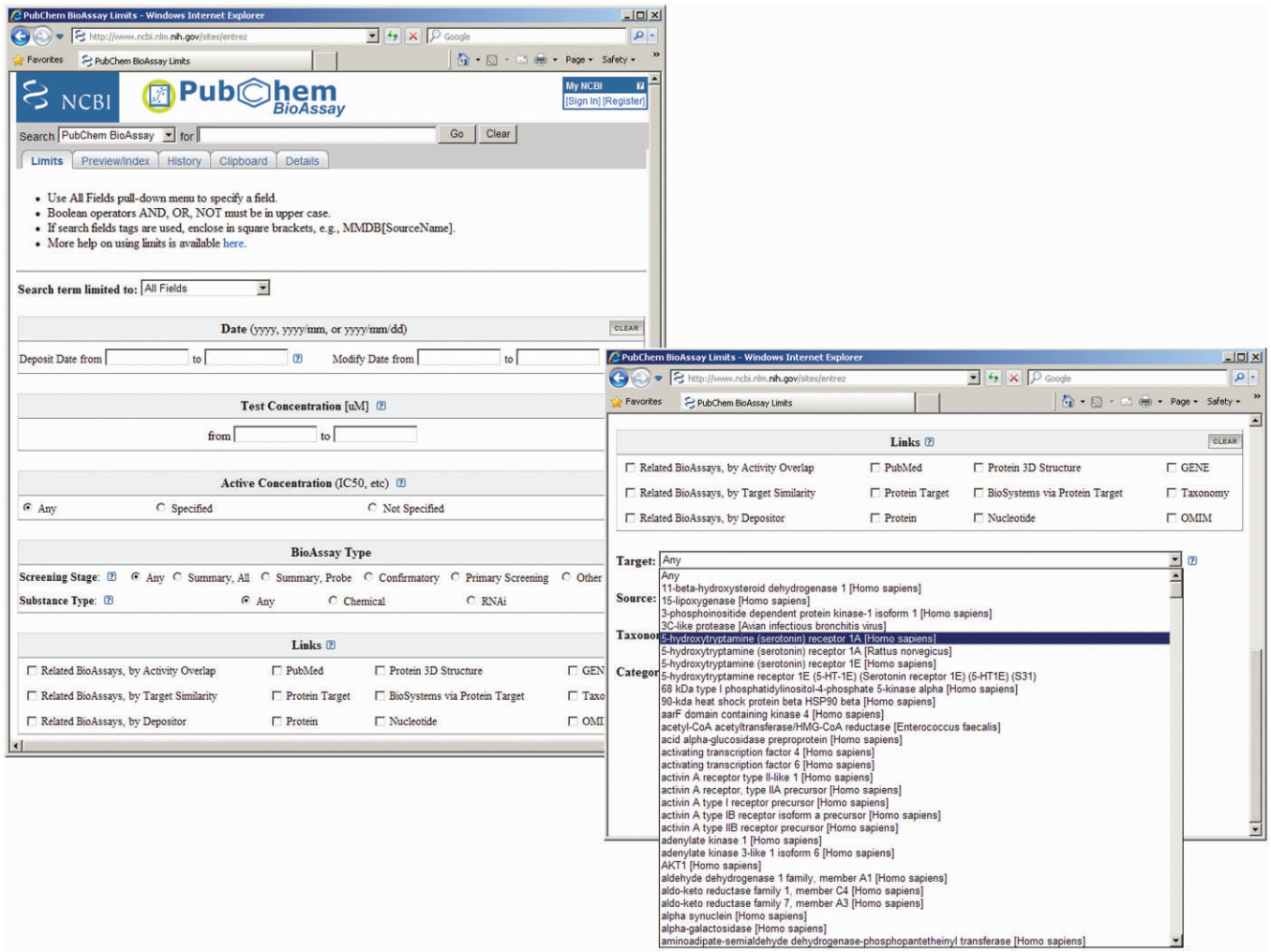


Figure 4. A screenshot for the BioAssay 'Limits' page. One can use this facility to retrieve bioassay records for a particular target using the 'Target' menu, to collect a list of bioassays records that have IC50 reported using the 'Active Concentration|Specified' option, to identify siRNA screening experiments using the 'BioAssay Type|Substance Type|RNAi' feature, or to find out assays targeting a specific protein by selecting one from the 'Target' menu.

PubChem BioAssay data are publicly accessible and can be downloaded via the PubChem FTP site (ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay). PubChem BioAssay descriptions and data are stored in ASN, XML and CSV file formats. A concise view of the data consisting of activity outcome, score and active concentration is available under the 'Concise' directory. Pairwise lists of related assays based on activity overlap, target similarity, common biological pathway or depositor provided similarity are available under the AssayNeighbors directory. Data structure and fields contained within the FTP files are modified when there is a schema change for the BioAssay data specification or a request for additional fields to be dumped from the PubChem BioAssay database. Cron jobs for dumping BioAssay FTP files run weekly in full and daily in incremental modes. Assay descriptions and data table can also be retrieved and downloaded through a programmatic interface using the PubChem PUG/SOAP facilities (http://pubchem.ncbi.nlm.nih.gov/pug/pughelp.html).

BIOACTIVITY ANALYSIS TOOLS

When identifying a potential molecular probe or lead candidate against a given target, researchers need to examine the biological activity results of a screening campaign. Substances that are deemed active in a primary screen are tested in subsequent confirmatory and counter screens, either repudiating or confirming the value of the initial positive test result. In PubChem, a substance is often found to be tested in multiple biological assays and screened against various targets; thereby, it is important for researchers to aggregate and compare the assay data to understand the biological activity profile of a small molecule of interest. Performing such analyses can be laborious and time consuming without appropriate tools to rapidly summarize biological tests.

To aid assay data analysis, PubChem provides a suite of web-based bioactivity analysis tools allowing one to download and search individual test results, compare biological activity data from multiple screenings, examine

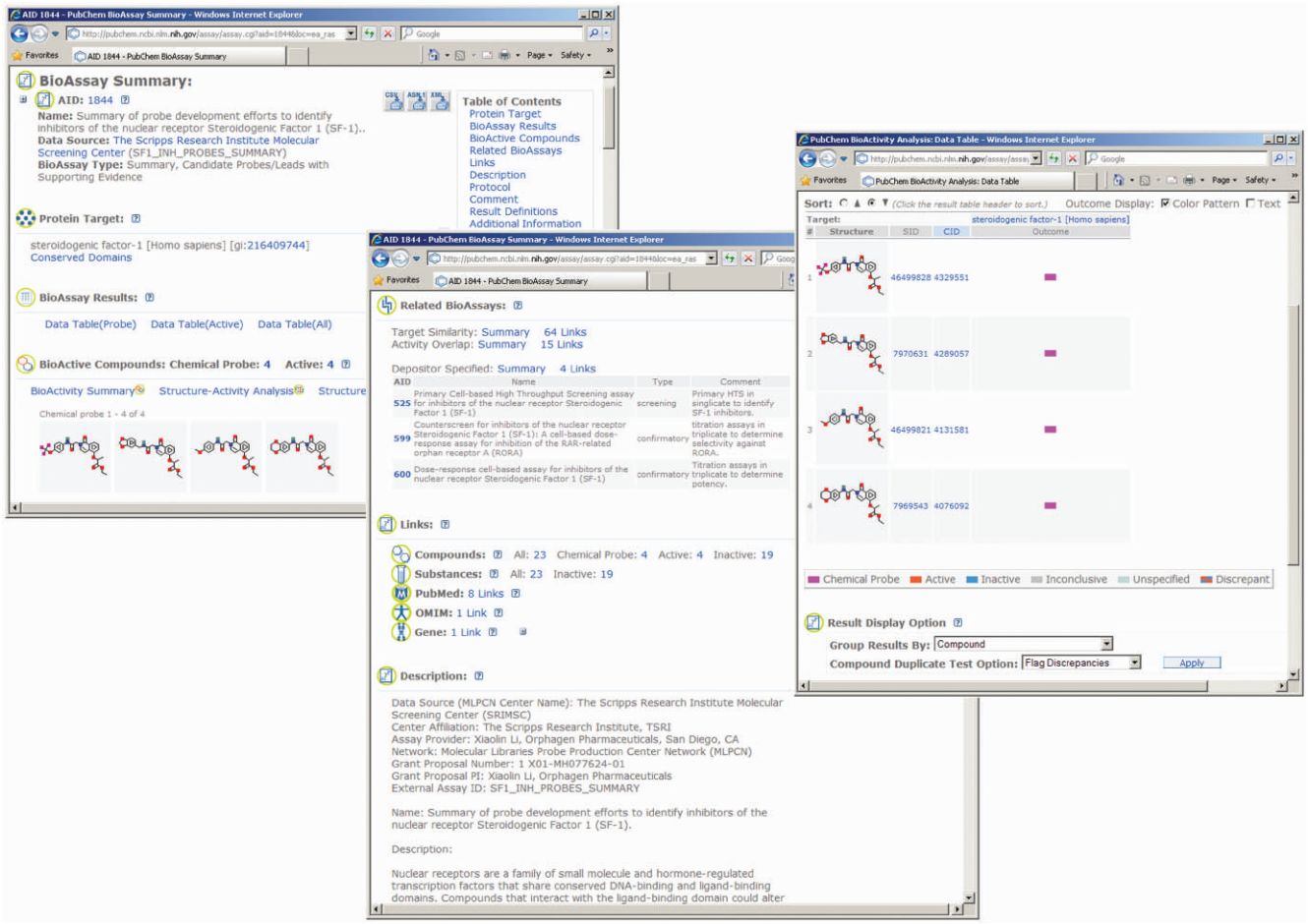


Figure 5. A summary view of PubChem BioAssay AID 1844 contributed by one of the screening centers of the NIH Molecular Libraries Program. This assay summarizes the process for identifying inhibitors of the nuclear receptor Steroidogenic Factor 1 (SF-1) and reports four chemical probes identified by this assay project. Multiple links are provided to allow one to retrieve and analyze testing results using available tools.

target selectivity or explore structure–activity relationships for compounds of interest. Part of the bioactivity analysis tools can be accessed directly through links provided at the PubChem web page at http://pubchem.ncbi.nlm.nih.gov/ (Figure 1). Detailed descriptions about the major components of the data analysis tools and their integration with the rest of the NCBI information system have been provided in a previous work (2).

CONCLUSION

PubChem BioAssay provides a comprehensive platform to select and summarize the bioactivities of tested substances. Additional tools support structure–activity analysis and download of selected results. The PubChem BioAssay system is continually refined with incremental performance and interface improvements and addition of new components, tools and services. PubChem is designed to serve as an archive for the large volume of screening data generated by the NIH Molecular Libraries Program (MLP), as well as for bioactivity information contributed by the broader scientific community. One future direction is to collect additional bioactivity results from the research community by working with other laboratories performing screening and medicinal chemistry research and with journal publishers and authors. Other future directions include better integration with additional information resources available at NCBI, developing facilities to better analyze siRNA screening results, providing better programmatic access to PubChem BioAssay resources and better emphasis of the bioactive chemical probe information generated by the MLP. PubChem welcomes further contributions of bioactivity results for tested substances, including bioactivity results for both small molecules and siRNA probes.

FUNDING

Intramural Research program of the National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Bolton,E.E., Wang,Y., Thiessen,P.A. and Bryant,S.H. (2008) PubChem: integrated platform of small molecules and biological activities. *Annu. Rep. Comput. Chem.*, **4**, Chapter 12, 217–241.

2. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.
3. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
4. Zerhouni,E. (2003) Medicine. The NIH Roadmap. *Science*, **302**, 63–72.
5. Zerhouni,E.A. (2006) Clinical research at a crossroads: the NIH roadmap. *J. Invest. Med.*, **54**, 171–173.
6. Austin,C.P., Brady,L.S., Insel,T.R. and Collins,F.S. (2004) NIH Molecular Libraries Initiative. *Science*, **306**, 1138–1139.
7. Lazo,J.S., Brady,L.S. and Dingledine,R. (2007) Building a pharmacological lexicon: small molecule discovery in academia. *Mol. Pharmacol.*, **72**, 1–7.
8. Driscoll,J.S. (1984) The preclinical new drug research program of the National Cancer Institute. *Cancer Treat. Rep.*, **68**, 63–76.
9. Zaharevitz,D.W., Holbeck,S.L., Bowerman,C. and Svetlik,P.A. (2002) COMPARE: a web accessible tool for investigating mechanisms of cell growth inhibition. *J. Mol. Graph. Model.*, **20**, 297–303.
10. Richard,A.M., Gold,L.S. and Nicklaus,M.C. (2006) Chemical structure indexing of toxicity data on the internet: moving toward a flat world. *Curr. Opin. Drug Discov. Dev.*, **9**, 314–325.
11. Marsden,B.D. and Knapp,S. (2008) Doing more than just the structure-structural genomics in kinase drug discovery. *Curr. Opin. Chem. Biol.*, **12**, 40–45.
12. Liu,T., Lin,Y., Wen,X., Jorissen,R.N. and Gilson,M.K. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.
13. Harmar,A.J., Hills,R.A., Rosser,E.M., Jones,M., Buneman,O.P., Dunbar,D.R., Greenhill,S.D., Hale,V.A., Sharman,J.L., Bonner,T.I. *et al.* (2009) IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels. *Nucleic Acids Res.*, **37**, D680–D685.
14. Wang,R., Fang,X., Lu,Y. and Wang,S. (2004) The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.*, **47**, 2977–2980.
15. Seiler,K.P., George,G.A., Happ,M.P., Bodycombe,N.E., Carrinski,H.A., Norton,S., Brudz,S., Sullivan,J.P., Muhlich,J., Serrano,M. *et al.* (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, **36**, D351–D359.
16. Karaman,M.W., Herrgard,S., Treiber,D.K., Gallant,P., Atteridge,C.E., Campbell,B.T., Chan,K.W., Ciceri,P., Davis,M.I., Edeen,P.T. *et al.* (2008) A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.*, **26**, 127–132.
17. Zhang,E.E., Liu,A.C., Hirota,T., Miraglia,L.J., Welch,G., Pongsawakul,P.Y., Liu,X., Atwood,A., Huss,J.W. 3rd, Janes,J. *et al.* (2009) A genome-wide RNAi screen for modifiers of the circadian clock in human cells. *Cell*, **139**, 199–210.
18. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
19. Marchler-Bauer,A., Anderson,J.B., Chitsaz,F., Derbyshire,M.K., DeWeese-Scott,C., Fong,J.H., Geer,L.Y., Geer,R.C., Gonzales,N.R., Gwadz,M. *et al.* (2009) CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res.*, **37**, D205–D210.
20. Wang,Y., Addess,K.J., Chen,J., Geer,L.Y., He,J., He,S., Lu,S., Madej,T., Marchler-Bauer,A., Thiessen,P.A. *et al.* (2007) MMDB: annotating protein sequences with Entrez's 3D-structure database. *Nucleic Acids Res.*, **35**, D298–D300.
21. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
22. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
23. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.