BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 5 2011, pages 741–742
doi:10.1093/bioinformatics/btq715

Databases and ontologies

Advance Access publication January 6, 2011

A Java API for working with PubChem datasets

Mark R. Southern* and Patrick R. Griffin

Translational Research Institute and Molecular Therapeutics, The Scripps Research Institute, , Scripps Florida, 130 Scripps Way, Jupiter, FL 33458, USA

Associate Editor: Martin Bishop

ABSTRACT

Summary: PubChem is a public repository of chemical structures and associated biological activities. The PubChem BioAssay database contains assay descriptions, conditions and readouts and biological screening results that have been submitted by the biomedical research community. The PubChem web site and Power User Gateway (PUG) web service allow users to interact with the data and raw files are available via FTP.

These resources are helpful to many but there can also be great benefit by using a software API to manipulate the data. Here, we describe a Java API with entity objects mapped to the PubChem Schema and with wrapper functions for calling the NCBI eUtilities and PubChem PUG web services. PubChem BioAssays and associated chemical compounds can then be queried and manipulated in a local relational database. Features include chemical structure searching and generation and display of curve fits from stored dose–response experiments, something that is not yet available within PubChem itself. The aim is to provide researchers with a fast, consistent, queryable local resource from which to manipulate PubChem BioAssays in a database agnostic manner. It is not intended as an end user tool but to provide a platform for further automation and tools development.

Availability: http://code.google.com/p/pubchemdb

Contact: southern@scripps.edu

Received on October 22, 2010; revised on December 2, 2010; accepted on December 19, 2010

1 INTRODUCTION

PubChem (Bolton *et al*., 2008; Wang *et al*., 2009) (http://pubchem.ncbi.nlm.nih.gov) is a public repository of chemical structures and associated biological activities. It was launched as part of the Molecular Libraries Roadmap (Zerhouni, 2003) from the National Institutes of Health (NIH), which aims to increase the discovery and use of chemical probes through high-throughput screening of small molecules (Zerhouni, 2003, 2006). It is composed of three interconnected databases. The Compound database contains unique chemical structures. The Substance database contains batch/sample level descriptions of these chemical structures. The BioAssay database (Wang *et al*., 2010) is the repository of biological screening results. As of September 2010, over 72 million chemical structures have been deposited by 128 distinct organizations and 462 752 BioAssays with over 90 million data points have been deposited by 43 organizations. These include governmental, academic and private entities (http://pubchem.ncbi.nlm.nih.gov/sources/sources.cgi).

*To whom correspondence should be addressed.

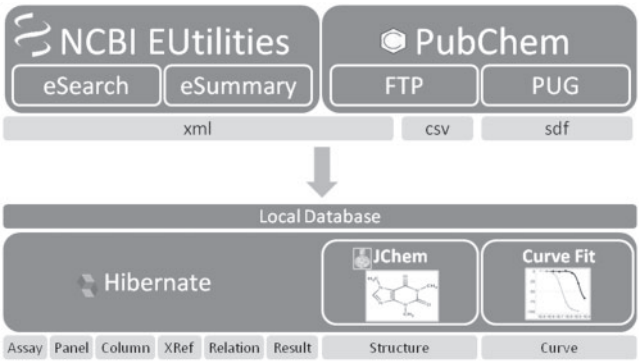

Fig. 1. Depiction of the NCBI files and web services used in building and manipulating PubChem datasets locally via a Java/Hibernate-based API. The PubChem Schema is mapped to database-aware Java objects.

2 DATA MODEL

The PubChem BioAssay data model has been described previously (Wang *et al*., 2010). Here we implement a data model with the Hibernate (http://www.hibernate.org) persistence framework that maps the PubChem Schema (ftp://ftp.ncbi.nlm.nih.gov/pubchem/data_spec/pubchem.xsd) as a Java object-orientated domain model.

The Java entity objects are depicted in Figure 1 and are as follows:

Assay: details such as the assay Id (AID), assay name, description, project category, protocol, whether it is small molecule or RNAi based and the number of total/active/inactive/inconclusive substances and compounds (SIDs and CIDs).

Panel: allows more than one 'assay' or 'target'to be specified within one BioAssay. Provides panel name, order number, description and result column type.

Column: stores definitions of the tested readouts within the assay. For example, an IC50, % inhibition or any other observed measurement or statistic. Details are name, description, data type and unit of measure. Each can be assessed at a particular tested concentration. In dose–response assays, a column containing the 'active' concentration, which produces 50% of the maximum activity can be flagged.

XRef: stores links to other NCBI databases as specified in the BioAssay data model, including the protein or nucleotide target of the assay, publications in PubMed, taxonomy links, etc.

Result: stores the data values for all Columns of a single PubChem Substance (SID), including the PubChem Activity Outcome and Score. With some simple logic, an attempt is made to identify the

primary, most important (e.g. IC50, % inhibition) result definition and any qualifier ('=', '⩾', etc.) that it might have.

Relation: stores the relationships between Bioassays. These are depositor specified or have been calculated from target or activity similarities.

Structure: stores chemical structures. JChem technology from ChemAxon (http://chemaxon.com) is utilized to handle chemical structure processing and search.

Curve: stores curve fits calculated from dose—response experiments. Once the curve fit parameters and equation have been determined and stored, the curves can be plotted without further processing. The stand-alone curve fitting library used in this process (http://code.google.com/p/curve-fit/) was derived from the software source code of NCGC CurveFit (http://www.ncgc.nih.gov/pub/openhts/curvefit/), a 'United States Government Work'.

3 ASSEMBLING THE DATA

Several different mechanisms are employed to assemble the data and instructions for running these tasks as shell scripts for the purpose of creating a database mirror are included in the software distribution (see 'Availability' section).

Firstly, the assay descriptions in XML format and assay data in comma separated values (CSV) format are mirrored locally from the PubChem FTP site (ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/CSV/). This is optional as the files can be downloaded on the fly but has advantages in terms of perceived speed and network reliability.

To start the build, a list of AIDs to process is obtained. This can either be user provided, a complete set in the case of a full build, or just the AIDs that have changed since last time, in the case of an incremental build. The NCBI eSearch utility is used to query for the list of AIDs and PubChem maintains several date filters on which to search. For example, to obtain a list of all the AIDs deposited in 2010, you could use the query; '*2010/01/01[DepositDate] : 2010/12/31[DepositDate]*'.

The XML description files are then used to populate the Assays, Columns, Panels and XRefs, determining if the version of an assay has changed and thus flagging the assay data to be reloaded. The NCBI Entrez Utilities (eUtils) provide additional information:

eSearch utility: provides a list of embargoed assays and determines which are small molecule based and which are nucleotide based.

eSummary utility: obtains total/active/inactive/inconclusive CID/SID counts.

The BioAssay CSV data files are then used to populate the assay Results. These are faster to process and consume fewer resources than using the XML data files. Because of the degree of variation between the Columns of different Bioassays, the PubChem Activity Score and Activity Outcome, SID and CID are modeled as separate database columns but the other result definitions are stored in a single column as a one line CSV file. In the Java/Hibernate data model, these are elegantly hidden behind a Hibernate UserType API that makes them appear as discrete values. Outside of the Hibernate world, these are available to any downstream software that can process CSV.

After the data files have been processed, a SQL query is run to determine which new chemical structures need to be fetched and these are then downloaded via the PubChem PUG soap service and inserted into the database via JChem from ChemAxon.

The relationships between Bioassays are processed from pairwise relationships found in text files from the AssayNeighbors folder on the PubChem FTP site (ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/AssayNeighbors/).

Finally, for each downloaded Result from dose–response Bioassays, a curve fit can be calculated and stored in the database. Curve plots can be generated from the stored curves with no further analytical processing.

4 DISCUSSION

We present a Java API for working with PubChem datasets. It assembles data from several sources (NCBI web services and flat files) and is backed by the Hibernate persistence framework. It utilizes a relational database for storage and has built in querying/search capabilities. It integrates chemical functionality such as structure processing and search via JChem from ChemAxon and can also handle dose–response curves via code derived from NCGC CurveFit.

We enable programmatic access to PubChem features such as BioAssay Panels, Columns and Results, providing data types, units of measure and associating dose responses to specific curve plots. The PubChem power user gateway (PUG) web services do not describe the PubChem Schema as completely and such detail is also not available in PubChem's CSV downloads, but only on the BioAssay web pages.

Additionally, a local relational mirror of the PubChem BioAssay database provides the data in a way that was previously not available. This can be of great benefit to researchers and software developers as relational databases are a widespread, familiar technology and can open up the dataset to many other tools and technologies than can interact via web browser or web-based protocols with inherent latency and statelessness. It can also serve as an integration platform as annotations and other information can be added to the database schema by adding additional tables and fields as necessary.

In summary, this is a toolkit for software developers to work with PubChem Bioassay datasets in a consistent manner.

Funding: The Comprehensive Center for Chemical Probe Discovery and Optimization at TSRI (grant5 U54 MH084512-02) (Hugh Rosen, Principal Investigator).

Conflict of Interest: none declared.

REFERENCES

Bolton,E.E. *et al*. (2008) PubChem: integrated platform of small molecules and biological activities. *Annu. Rep. Comput. Chem.*, **4**, 217–241.

Wang,Y. *et al*. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.

Wang,Y. *et al*. (2010) An overview of the PubChem BioAssay resource. *Nucleic Acids Res.*, **38**, D255–D266.

Zerhouni,E. (2003) Medicine: the NIH Roadmap. *Science*, **302**, 63–72.

Zerhouni,E.A. (2006) Clinical research at a crossroads: the NIH roadmap. *J. Investig. Med.*, **54**, 171–173.