BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 1 2012, pages 140–141
doi:10.1093/bioinformatics/btr622

Data and text mining

Advance Access publication November 13, 2011

PubChem promiscuity: a web resource for gathering compound promiscuity data from PubChem

Stephanie A. Canny, Yasel Cruz, 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: Promiscuity counts allow for a better understanding of a compound's assay activity profile and drug potential. Although PubChem contains a vast amount of compound and assay data, it currently does not have a convenient or efficient method to obtain in-depth promiscuity counts for compounds. PubChem promiscuity fills this gap. It is a Java servlet that uses NCBI Entrez (eUtils) web services to interact with PubChem and provide promiscuity counts in a variety of categories along with compound descriptors, including PAINS-based functional group detection.

Availability: http://chemutils.florida.scripps.edu/pcpromiscuity

Contact: southern@scripps.edu

Received on July 18, 2011; revised on October 18, 2011; accepted on November 7, 2011

1 INTRODUCTION

A better understanding of a compound's drug potential can be obtained by determining a compound's selectivity for protein targets and how many and what type of assays a compound is active or tested in. Compound promiscuity data are essential to fully exploit the therapeutic potential and minimize the toxic effects of drug candidates (Frye, 2010; Li *et al*., 2010; Rix, 2009; Xie, 2009). Promiscuous compounds are also traditionally problematic in high-throughput screening (HTS) and it is helpful to identify them early in the HTS process (Schürer *et al*., 2011). Gathering compound promiscuity data can help determine if the compound's activity may be affected by a particular assay technology, detection method or interaction with biological targets (Schürer *et al*., 2011). There are additional positive uses for the identification of promiscuous compounds. For instance, finding chemical promiscuity across a family of proteins could provide chemical starting points for targets within that family that lack chemical probes.

PubChem is a public repository of substance information, compound structures and BioActivity data for HTS campaigns. The majority of the data points come from the Molecular Libraries Screening Center Network (MLSCN) under the NIH Molecular Libraries Program (MLP) (Austin, 2004; Li *et al*., 2010). MLP screening campaigns use the Molecular Libraries Small Molecule Repository (MLSMR), which is a collection of over 300 000 compounds (http://mli.nih.gov/mli/compound-repository/). This results in data that is highly amenable for assessing compound promiscuity. PubChem is a valuable resource for studying the promiscuity of compounds (Li *et al*., 2010). PubChem is integrated with Entrez (NCBI's primary search engine) and also has BioActivity Services (http://pubchem.ncbi.nlm.nih.gov/assay) that provide simple promiscuity counts for compounds including active and tested protein and BioAssay counts.

Here, we present a tool that gathers promiscuity counts and compound descriptors for multiple compounds. Results are displayed in a formatted web table for easy interpretation. Although PubChem has extensible search and data management features, to manually obtain the results that the tool provides a user would need to visit five BioActivity Services web pages and complete nine Entrez queries (see Section 2). Each Entrez query requires determining the appropriate search terms and extracting the data from those searches. In comparison, this tool runs queries in batches and provides results in only a few seconds per compound.

2 METHODS

The promiscuity analysis gathers promiscuity counts and chemical descriptors for PubChem compounds or substances. Assay and protein counts are, respectively, hyperlinked to PubChem BioActivity DataTable and Summary results so that the associated assays, proteins and data can be viewed. Sorting and filtering are possible and useful to find and then disregard sparsely tested compounds.

The input to the tool is a list of PubChem compound or substance identifiers (CIDs or SIDs). Various PubChem BioAssay, Compound or Substance queries are then performed using NCBI Entrez Utilities (eUtils) eLink, eSearch and eSummary web services (http://www.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html). Cross-links between databases are specified (http://eutils.ncbi.nlm.nih.gov/corehtml/query/static/entrezlinks.html) as are controlled vocabulary fields and operators (http://www.ncbi.nlm.nih.gov/books/NBK3837/#EntrezHelp.Entrez_Searching_Options) and (http://pubchem.ncbi.nlm.nih.gov/help.html#PubChem_index).

There are two output formats: overall promiscuity counts for each compound (per compound) or individual counts per protein target (per compound and protein). The tool provides all the following data:

Assay counts: are active and total counts for PubChem BioAssays that a compound is tested in. A compound is active in an assay if the compound's PubChem activity outcome (specified by the depositor) is assigned active. Assay counts are found using the link names pccompound_pcassay and pccompound_pcassay_active. MLP and ChEMBL (https://www.ebi.ac.uk/chembldb/index.php) assay counts are provided separately and found using the search term and field combinations 'NIH Molecular Libraries Program'[SourceCategory] and 'ChEMBL'[SourceName] (see Section 4).

Project counts: a 'project' is a set of assays that are PubChem XRef assays to a summary assay. A compound is active in a 'project' if it is active in one or more XRef assays that contain only protein targets that are also targets in the summary assay. A compound is not active in a 'project' if it is active in one or

*To whom correspondence should be addressed.

more XRef assays that have protein targets differing from the targets in the summary assay. The search terms and fields used to find 'project' counts are 'summary'[ActivityOutcomeMethod]. All projects and MLP project counts are provided.

Protein target counts: protein target counts are a compound's activity against all the protein targets it has been tested against. A compound is considered active against a protein if the compound is active in at least one assay where the protein is one of the targets of the assay. Protein targets are retrieved from the assay's Entrez eSummary. Assays with no protein target (e.g. cytotoxicity assays) are also provided.

Luciferase, β-lactamase and fluorescence assay counts: luciferase assays have designs based on the 'luciferase-induced conversion of luciferin substrates that result in the emission of light' (Fan and Wood, 2007; Schürer *et al*., 2011). Most β-lactamase assays 'use fluorescence resonance energy transfer (FRET) substrates, resulting in a fluorescence shift upon hydrolysis of the beta-lactam' (Schürer *et al*., 2011; Zlokarnik *et al*., 1998). These assays are found by looking for keywords in the BioAssay protocol and description (e.g. 'luciferase'[AssayProtocol] OR 'luciferase'[AssayDescription]).

Compound descriptors: 'Rule of 5' (Lipinski *et al*., 2001) violation count, smiles and a full set of chemical descriptors available from eUtils eSummary are provided. Chemical descriptor columns are hidden in the web table output.

Functional group detection: SMARTS patterns for pan assay interference compounds (PAINS) (Baell and Holloway, 2010) that are frequent hitters in high-throughput screens were obtained from http://blog.rguha.net/?p=850. These are compared against each compound and detected groups are flagged (see Section 3).

3 IMPLEMENTATION

This tool was developed as a Java (http://www.oracle.com/technetwork/java/index.html) servlet utilizing the previously published PubChemDB Java API (Southern and Griffin, 2010) for eUtils functions. JOELib (http://sourceforge.net/projects/joelib/) was used for SMARTS querying of functional groups. SmartClient (http://www.smartclient.com/product/smartclient.jsp) was used to display the output web table.

4 DISCUSSION

PubChem promiscuity extracts promiscuity counts and compound descriptors from PubChem, and example results of known promiscuous compounds are available on the tool's website. This data can give a picture of a compound's selectivity and drug potential. The promiscuity counts are not easily accessible in PubChem itself, especially if desired for several compounds at once. The resultant data are accessible in a convenient and relatively quick fashion that is based on the restrictions of using NCBI eUtils web services (http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html). Design considerations and limitations are addressed below:

Project counts eliminate artificial inflation of assay counts where a compound is tested multiple times within one screening campaign. With the definition of 'active project' used for this tool, it is possible that probes of a project may not be counted as active if that compound is active in an assay that is part of the project and has a target that is not in the 'PubChem Summary AID'.

Protein counts can show if a compound is protein selective, and assays with no-protein targets can typically point out cytotoxic compounds. If a compound is specified as active in a 'PubChem Panel Assay', it is not possible with NCBI Entrez to determine which individual panels and therefore targets the compound is active against. PubChem promiscuity will return the result that the compound is active against all the targets in the panel assay, but this may not be true.

Assay reporter technologies that are commonly used include luciferase, β-lactamase and fluorescence. If a compound is active in many or all assays that use the same assay reporter technology and across protein targets, this could mean that the compound is interfering with the assay technology. PubChem does not have an annotation for these types of assays and therefore the counts are dependent on the keywords and fields used in the search.

MLP and ChEMBL are two common sources of assays in PubChem. MLP assays are large-scale screening assays, and ChEMBL assays are mostly single points curated from the scientific literature. Also, ChEMBL data may not specify compounds as active or inactive. MLP and ChEMBL assay counts are given separately because of the contrast in the data they provide.

Funding: The Comprehensive Center for Chemical Probe Discovery and Optimization at TSRI (grant 5 U54 MH084512-02) (Hugh Rosen, Principal Investigator).

Conflict of Interest: none declared.

REFERENCES

Austin,C.P. (2004) NIH Molecular Libraries Initiative. *Science*, **306**, 1138–1139.

Baell,J.B. and Holloway,G.A. (2010) New substructure filters for removal of Pan Assay Interference Compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.*, **53**, 2719–2740.

Fan,F. and Wood,K.V. (2007) Bioluminescent assays for high-throughput screening. *Assay Drug Dev. Technol.*, **5**, 127–136.

Frye,S.V. (2010) The art of the chemical probe. *Nat. Chem. Biol.*, **6**, 159–161.

Li,Q. *et al*. (2010) PubChem as a public resource for drug discovery. *Drug Discov. Today*, **15**, 1052–1057.

Lipinski,C.A. *et al*. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, **46**, 3–26.

Rix,U. and Superti-Furga,G. (2009) Target profiling of small molecules by chemical proteomics. *Nat. Chem. Biol.*, **5**, 616–624.

Schürer,S. *et al*. (2011) BioAssay ontology annotations facilitate cross-analysis of diverse high-throughput screening data sets. *J. Biomol. Screen.*, **16**, 415–426.

Southern,M.R. and Griffin,P.R. (2011) A Java API for working with PubChem datasets. *Bioinformatics*, **27**, 741–742.

Xie,L. *et al*. (2009) Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. *PLoS Comput. Biol.*, **5**, e1000387.

Zlokarnik,G. *et al*. (1998) Quantitation of transcription and clonal selection of single living cells with beta-lactamase as reporter. *Science*, **279**, 84–88.