Bioinformatics, 34(1), 2018, 131–133
doi: 10.1093/bioinformatics/btx566
Advance Access Publication Date: 7 September 2017
Applications Note

OXFORD

Databases and ontologies

An automated tool for obtaining QSAR-ready series of compounds using semantic web technologies

Oriol López-Massaguer, Ferran Sanz and Manuel Pastor*

Research Programme on Biomedical Informatics (GRIB), Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Dept. of Experimental and Health Sciences, Universitat Pompeu Fabra, 08003 Barcelona, Spain

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on July 25, 2017; revised on September 1, 2017; editorial decision on September 4, 2017; accepted on September 6, 2017

Abstract

Summary: We describe an application (Collector) for obtaining series of compounds annotated with bioactivity data, ready to be used for the development of quantitative structure-activity relationships (QSAR) models. The tool extracts data from the 'Open Pharmacological Space' (OPS) developed by the Open PHACTS project, using as input a valid name of the biological target. Collector uses the OPS ontologies for expanding the query using all known target synonyms and extracts compounds with bioactivity data against the target from multiple sources. The extracted data can be filtered to retain only drug-like compounds and the bioactivities can be automatically summarised to assign a single value per compound, yielding data ready to be used for QSAR modeling. The data obtained is locally stored facilitating the traceability and auditability of the process. Collector was used successfully for the development of models for toxicity endpoints within the eTOX project.

Availability and implementation: The software is available at http://phi.upf.edu/collector. The source code is located at https://github.com/phi-grib/Collector and is free for use under the GPL3 license. The web version is hosted at http://collector.upf.edu/.

Contact: manuel.pastor@upf.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In biomedical research and, in particular, in drug development, there is an increasing need of re-using existing experimental data (Machina *et al.*, 2013; Mons *et al.*, 2011). Often, this data is stored in isolated data silos, either public [ChEMBL (Gaulton *et al.*, 2012), UniProt (The UniProt Consortium, 2013) WikiPathways (Kelder *et al.*, 2012), etc.] or private (e.g. pharma companies data only accessible within research consortia). Accessing public data and processing it is a complex process, involving sources with heterogeneous data formats, diverse licensing models and different levels of quality. Initiatives like FAIR (Findable, Accessible, Interoperable, Reusable) aim to overcome these issues by establishing clear principles of data management and stewardship (Wilkinson *et al.*, 2016).

In an attempt to mitigate this problem IMI project Open PHACTS (OPS) (Gray *et al.*, 2014; Groth *et al.*, 2014; Ratnam *et al.*, 2014) built a software platform specifically designed to facilitate drug development by integrating relevant data sources using semantic web technologies (Allemang and Hendler, 2011; Berners-Lee *et al.*, 2001; Parsons, 2009). The OPS platforms contains a comprehensive collection of relevant information, including (August 2017) 11 538 targets, 14 675 320 activities and 1 735 442 compounds.

One of the applications of the data obtained through OPS is the development of QSAR models, able to predict the biological properties of new compounds. The starting point for building such models is the compilation of a series of compounds (the training series) for which the bioactivity value is known. Probably the main

Fig. 1. Collector web GUI showing results of a job (detail)

determinant of the quality of any QSAR model is the adequacy of the training series, particularly its size and the structural diversity. The software described here, so called Collector, aims to facilitate this task by extracting series of compounds from OPS, together with the result of pharmacological/toxicological experiments (IC_{50}, pK_d, etc.) and to apply customizable curation filters, producing series of compounds in a format well suited for the development of QSAR models. Thanks to the use of the OPS platform, the licensing conditions of the extracted data is very well defined (Ratnam *et al.*, 2014).

This tool was developed as an 'exemplar application' of OPS, to be applied for developing models for predicting toxicology-related outcomes in the framework of the eTOX IMI project (Sanz *et al.*, 2015). Some of the models developed will be briefly described here and in full detail as Supplementary Information.

2 Methods and implementation

Collector works as a data extraction engine, using OPS API https://dev.openphacts.org/. The system was developed using Scala programming language (Odersky *et al.*, 2004) and the Play web framework (Hilton *et al.*, 2014). The data extracted is stored in a PostgreSQL (Stonebraker *et al.*, 1986) https://www.postgresql.org/ local database. RDKit http://www.rdkit.org was used as a chemistry toolkit to manipulate molecular structures.

3 Results

In a typical Collector session, the user specifies the target of interest entering a valid identifier, like the UniprotID or the target name. During the query execution, the target identifier is expanded to search OPS for compounds annotated against this target, using any synonym term stored in the platform. The query can optionally incorporate filters for discarding compounds not meeting certain criteria, for example, Lipinski Rule of 5 (Lipinski *et al.*, 2001) or containing unusual elements (different from H, C, N, O, S, P, Cl, I, Br, F). Other possible filters relate to the kind of biological annotation (activity, inhibition, K_i or IC_{50}). The filtering protocol is customizable and the user can define additional filters for special requirements.

At execution time, the filters are applied in the sequence defined by the filtering protocol and their effect on the series size is recorded. The series obtained after each job is stored locally in a PostgreSQL database and tagged with the execution data for easy retrieval and traceability.

Often this list contains multiple annotations for the same compound. For this reason, Collector offers the possibility to aggregate all the annotations at compound level by computing the median of all the values found, thus producing a list with a single line per compound.

The series can be exported as a CSV file with the structures as SMILES, or as a SDFile with the annotations as separate fields. These files can be used directly as input in model development frameworks like eTOXlab (Carrio *et al.*, 2015).

Collector provides both web GUI (Fig. 1) and command line interfaces. The GUI is interactive and shows graphically the effect of the different filters on the size of the series obtained and a histogram representing the distribution of the biological property values. The structures of the compounds extracted and filtered out are also represented in an interactive table, with links to the original sources. The command line interface allows an easy integration of Collector in scripts and workflows.

Collector has been used in the eTOX project for the development of QSAR models against targets considered of toxicological interest (anti-targets). A full description of the series collected and the models developed is included as Supplementary Information.

Funding

This project was developed under the *Innovative Medicines Initiative Joint Undertaking Open PHACTS Project*, grant agreement number 115191 and eTOX project, grant agreement n° 115002, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013) and EFPIA companies' in kind contributions.

Conflict of Interest: none declared.

References

Allemang,D. and Hendler,J. (2011) *Semantic web for the working ontologist: effective modeling in RDFS and OWL*. 2nd ed. Elsevier, Waltham, MA.

Berners-Lee,T. *et al.* (2001) The Semantic Web. *Sci. Am.*, **284**, 34–43.

Carrio,P. *et al.* (2015) eTOXlab, an open source modeling framework for implementing predictive models in production environments. *J. Cheminf.*, **7**, 8.

Gaulton,A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.

Gray,A.J.G. *et al.* (2014) Applying linked data approaches to pharmacology: architectural decisions and implementation. *Semant. Web*, **5**, 101–113.

Groth,P. *et al.* (2014) API-centric Linked Data integration: The Open PHACTS Discovery Platform case study. *Web Semant. Sci. Serv. Agents World Wide Web*, **29**, 12–18.

Hilton,P. *et al.* (2014) *Play for Scala: Covers Play 2 Manning*.

Kelder,T. *et al.* (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.*, **40**, D1301–D1307.

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.

Machina,H.K. *et al.* (2013) Effective integration of informatics tools to enhance the drug discovery process. *Ind. Eng. Chem. Res.*, **52**, 16547–16554.

Mons,B. *et al.* (2011) The value of data. *Nat. Genet.*, **43**, 281–283.

Odersky,M. *et al.* (2004) An overview of the Scala programming language. Technical Report IC/2004/640, EPFL Lausanne.

Parsons,S. (2009) A Semantic Web Primer, Second Edition by Antoniou Grigoris and Harmelen Frank van, MIT Press. *Knowl. Eng. Rev.*, **24**, 415.

Ratnam,J. *et al*. (2014) The application of the Open Pharmacological Concepts Triple Store (Open PHACTS) to support drug discovery research. *PLoS One*, **9**, e115460.

Sanz,F. *et al*. (2015) Integrative modeling strategies for predicting drug toxicities at the eTOX project. *Mol. Inf*., **34**, 477–484.

Stonebraker,M. *et al*. (1986) The design of POSTGRES. In: *Proceedings of the 1986 ACM SIGMOD International Conference on Management of Data – SIGMOD '86*. ACM Press, New York, New York, USA, pp. 340–355.

The UniProt Consortium (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res*., **41**, D43–D47.

Wilkinson,M.D. *et al*. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.