Bioinformatics, 33(3), 2017, 464–466
doi: 10.1093/bioinformatics/btw640
Advance Access Publication Date: 14 October 2016
Applications Note

Data and text mining

CIIPro: a new read-across portal to fill data gaps using public large-scale chemical and biological data

Daniel P. Russo[1], Marlene T. Kim[1,2], Wenyi Wang[1], Daniel Pinolini[1], Sunil Shende[1,3], Judy Strickland[4], Thomas Hartung[5,6] and Hao Zhu[1,2,*]

[1]The Rutgers Center for Computational and Integrative Biology, Camden, NJ, 08102, USA, [2]Department of Chemistry, [3]Department of Computer Science, Rutgers University, Camden, NJ 08102, USA, [4]ILS, Research Triangle Park, NC 27709, USA, [5]Johns Hopkins Bloomberg School of Public Health, Center for Alternatives to Animal Testing (CAAT), Baltimore, MD, 21205, USA and [6]University of Konstanz, CAAT-Europe, Konstanz, Germany

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on July 8, 2016; revised on September 29, 2016; accepted on October 5, 2016

Abstract

Summary: We have developed a public Chemical *In vitro–In vivo* Profiling (CIIPro) portal, which can automatically extract *in vitro* biological data from public resources (i.e. PubChem) for user-supplied compounds. For compounds with *in vivo* target activity data (e.g. animal toxicity testing results), the integrated cheminformatics algorithm will optimize the extracted biological data using *in vitro–in vivo* correlations. The resulting *in vitro* biological data for target compounds can be used for read-across risk assessment of target compounds. Additionally, the CIIPro portal can identify the most similar compounds based on their optimized bioprofiles. The CIIPro portal provides new powerful assessment capabilities to the scientific community and can be easily integrated with other cheminformatics tools.

Availability and Implementation: ciipro.rutgers.edu.

Contact: danrusso@scarletmail.rutgers.edu or hao.zhu99@rutgers.edu

1 Introduction

There currently exists an enormous amount of biological data available to researchers through public repositories (e.g. PubChem, ChEMBL; Zhu *et al.*, 2014). Computational methods to utilize public bioassay data for toxicity prediction and assessment are being developed (Kim *et al.*, 2014, 2016; Low *et al.*, 2013; Wang *et al.*, 2015). However, identification of relevant assay data to incorporate into an assessment requires extensive manual reading and searching.

Read-across is a technique for filling data gaps for a target chemical by interpolating from data for other substances within the same group or 'category' (Patlewicz, 2014). The use of biological data in read-across has attracted attention (Ball et al., 2016; Low *et al.*, 2013; Zhu, 2016). In earlier studies, traditional pairwise similarity calculations (e.g. Tanimoto similarity) were used to compare chemicals using their biological data (Low *et al.*, 2013). However, these metrics require biological data to be available for all chemicals, limiting their use. In addition, the missing data creates certain reliability issues. For example, the similarity of two compounds can be considered to be more reliable when more biological data are available.

Here, we introduce the Chemical *In vitro–In vivo* Profiling (CIIPro) portal. CIIPro is a versatile workspace for users to profile compounds of interest with biological data from public resources (i.e. PubChem) and use these data for read-across assessment. The profiling and read-across approaches integrated into this portal have been used to develop multiple predictive models for complex bioactivities (Kim *et al.*, 2016; Ribay *et al.*, 2016; Wang *et al.*, 2015; Zhang *et al.*, 2014; Zhu *et al.*, 2009, 2014).

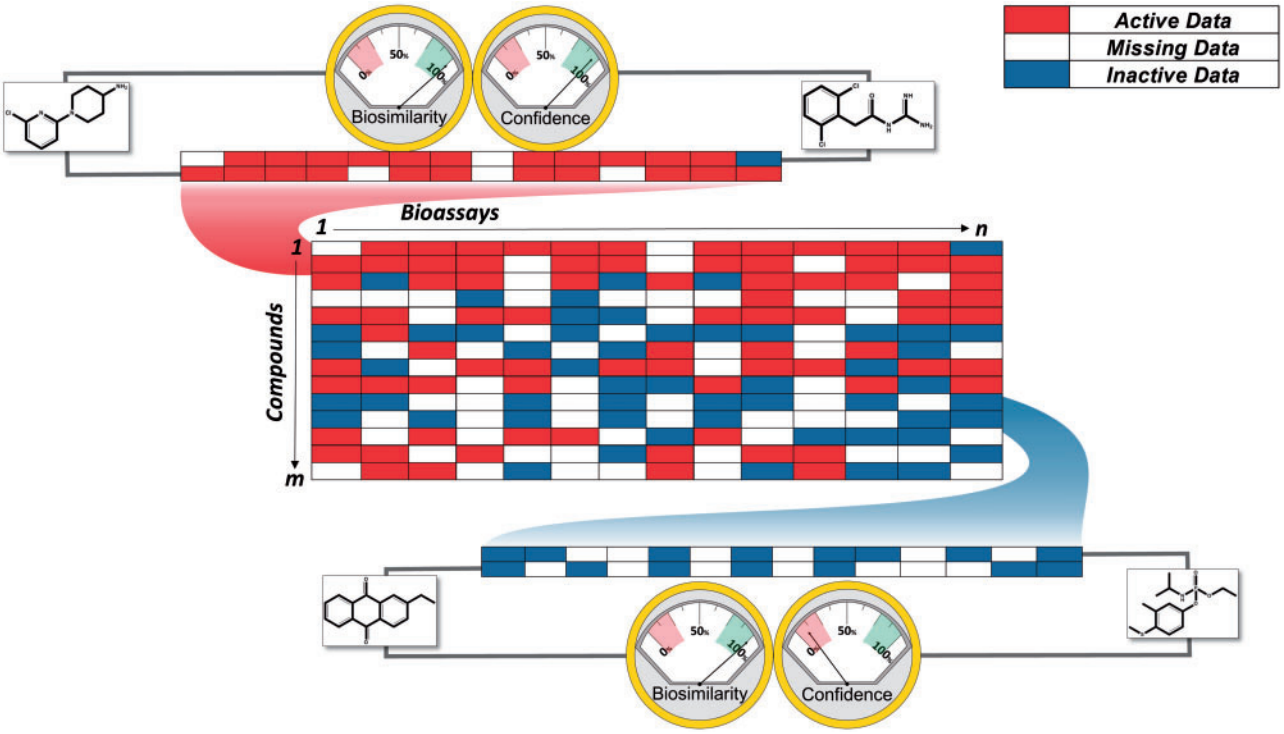

Fig. 1. Biosimilarity reliability of bioprofiles in CIIPro: active responses (red) are given higher weights than inactives (blue); the existence of missing data (white) also needs to be considered

2 Methods

2.1 CIIPro overview

CIIPro is a Python-based web portal built using a variety of open-source libraries and is freely accessible via ciipro.rutgers.edu. Bioassay data is supplied by PubChem, the largest public chemical data source and was downloaded using a File Transfer Protocol (FTP) offered through PubChem available at: ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay. Compounds' responses in bioassays were classified using PubChem's default activity classifications (active, inactive and inconclusive) and stored in a non-relational database, MongoDB. Since the PubChem database is continually updated, the updates are being incorporated to the CIIPro database on a monthly basis. All code is available at www.github.com/russodanielp/ciipro.

2.2 Biosimilarity

Biosimilarity between two molecules is calculated by a weighted similarity metric (Equation 1).

$$\mathit{Biosimilarity}\ (A,\ B) = \frac{|A_a \cap B_a| + |A_i \cap B_i| \cdot w}{|A_a \cap B_a| + |A_i \cap B_i| \cdot w + \ |A_a \cap B_i| + |A_i \cap B_a|} \quad (1)$$

Here, A_a and B_a represent the sets of active responses for compounds A and B, respectively. Similarly, A_i and $B_{i,}$ represent the sets of inactive responses. Our previous work showed the biosimilarity to rely on active data more than inactive data, due to the biased public data (i.e. much more inactive responses than active; Zhang *et al.*, 2014; Zhu *et al.*, 2009, 2014). The variable w, defined as the ratio of active data to inactive data in the target compound bioprofiles, ranges from 0 to 1, giving inactive data a fraction of the weight of active data. In our previous studies the variable w was much lower than 1 (Ribay *et al.*, 2016).

2.3 Biosimilarity reliability

As shown in Figure 1, the biosimilarity between two compounds that share a large number of active responses is more meaningful than that generated from two compounds that share a relatively small number of inactive responses, although biosimilarity scores for both cases will be close to 1.0. To address this issue, a confidence value of the biosimilarity calculations is calculated as shown in Equation 2.

$$\mathit{Confidence}\ (A,\ B) = |A_a \cap B_a| + |A_i \cap B_i| \cdot w + |A_a \cap B_i| + |A_i \cap B_a| \quad (2)$$

The confidence value represents the number of assays that have results for compounds A and B but gives less weight to the assays that only have inactive results for both compounds (Equation 2). Thus, the confidence value increases when there are more active data used to calculate the biosimilarity.

3 Features

3.1 CIIProfiler: biological response profiling

Under the *CIIProfiler* tab, *in vitro* biological data can be extracted to create a bioprofile for the target compounds. Bioprofiles can be optimized by removing assays with too few active responses within target compounds. The remaining assays can be ranked by their correlations with the target *in vivo* activity provided by users. Additionally, the optimized bioprofiles can be visualized by a bioprofile heat-map, as seen in Figure 1.

3.2 CIIP predictor

Under the *CIIP Predictor* tab, read-across can be performed by using chemical similarity and/or biosimilarity. The prediction for a new compound is made based on its chemical and biological nearest neighbors. The prediction result can be visualized by a similarity chart, along with the associated similarity and confidence values.

4 Conclusions

Although currently there are still many features that can be added into this portal (e.g. more data sources than PubChem), CIIPro is the first available public tool to take advantage of the dynamic biological big data landscape for the purpose of read-across predictions. Compared to the existing hybrid approaches (Low *et al.*, 2013), the CIIPro portal provides a new read-across strategy to deal with missing data and biased data issues when using public data sources.

Acknowledgements

We would like to acknowledge and thank Ms. Catherine Sprankle at Integrated Laboratory Systems, Inc., for her help in developing the portal.

Funding

This research was supported in part by National Institutes of Health (NIH) grants P30ES005022 and R15ES023148, and the Johns Hopkins Center for Alternatives to Animal Testing (CAAT) grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH and CAAT. JS provides technical support to the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods under NIEHS contract HHSN273201500010C, but does not represent NIEHS, the National Toxicology Program, or the official positions of any US Federal agency.

Conflict of Interest: none declared.

References

Ball,N. *et al.* (2016) Toward Good Read-Across Practice (GRAP) guidance. *Altex*, **33**, 149–166.

Kim,M.T. *et al.* (2014) Critical evaluation of human oral bioavailability for pharmaceutical drugs by using various cheminformatics approaches. *Pharm. Res.*, **31**, 1002–1014.

Kim,M.T. *et al.* (2016) Mechanism profiling of hepatotoxicity caused by oxidative stress using the antioxidant response element reporter gene assay models and big data. *Environ. Health Perspect.*, **124**, 634–641.

Low,Y. *et al.* (2013) Integrative chemical–biological read-across approach for chemical hazard classification. *Chem. Res. Toxicol.*, **26**, 1199–1208.

Ribay,K. *et al.* (2016) Predictive modeling of estrogen receptor binding agents using advanced cheminformatics tools and massive public data. *Front. Environ. Sci.*, **4**, 12.

Patlewicz,G. *et al.* (2014) Read-across approaches – misconceptions, promises and challenges ahead. *Altex*, **31**, 387–396.

Wang,W. *et al.* (2015) Developing enhanced blood–brain barrier permeability models: integrating external bio-assay data in QSAR modeling. *Pharm. Res.*, **32**, 3055–3065.

Zhang,J. *et al.* (2014) Profiling animal toxicants by automatically mining public bioassay data: a big data approach for computational toxicology. *PLoS ONE*, **9**, e99863.

Zhu,H. *et al.* (2009) A novel two-step hierarchical quantitative structure-activity relationship modeling work flow for predicting acute toxicity of chemicals in rodents. *Environ. Health Perspect.*, **117**, 1257–1264.

Zhu,H. *et al.* (2014) Big data in chemical toxicity research: the use of high-throughput screening assays to identify potential toxicants. *Chem. Res. Toxicol.*, **27**, 1643–1651.

Zhu,H. (2016) Supporting read-across using biological data. *Altex*, **33**, 167–182.