Bioinformatics, 2017, 1–2
doi: 10.1093/bioinformatics/btx788
Advance Access Publication Date: 18 December 2017
Applications Note

Data and text mining

Rcupcake: an R package for querying and analyzing biomedical data through the BD2K PIC-SURE RESTful API

Alba Gutiérrez-Sacristán[1,2,3], Romain Guedj[1], Gabor Korodi[1], Jason Stedman[1], Laura I. Furlong[2,3], Chirag J. Patel[1], Isaac S. Kohane[1] and Paul Avillach[1,*]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA, [2]Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM), Barcelona 08003, Spain and [3]Department of Experimental and Health Sciences (DCEXS), Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on August 16, 2017; revised on November 20, 2017; editorial decision on November 30, 2017; accepted on December 16, 2017

Abstract

Motivation: In the era of big data and precision medicine, the number of databases containing clinical, environmental, self-reported and biochemical variables is increasing exponentially. Enabling the experts to focus on their research questions rather than on computational data management, access and analysis is one of the most significant challenges nowadays.
Results: We present Rcupcake, an R package that contains a variety of functions for leveraging different databases through the BD2K PIC-SURE RESTful API and facilitating its query, analysis and interpretation. The package offers a variety of analysis and visualization tools, including the study of the phenotype co-occurrence and prevalence, according to multiple layers of data, such as phenome, exposome or genome.
Availability and implementation: The package is implemented in R and is available under Mozilla v2 license from GitHub (https://github.com/hms-dbmi/Rcupcake). Two reproducible case studies are also available (https://github.com/hms-dbmi/Rcupcake-case-studies/blob/master/SSCcaseStudy_v01.ipynb, https://github.com/hms-dbmi/Rcupcake-case-studies/blob/master/NHANEScaseStudy_v01.ipynb).
Contact: paul_avillach@hms.harvard.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The national precision medicine initiative represents a large bet on the hypothesis that multimodal or multidimensional perspectives will be able to identify subpopulations that are more predictable and share common disorders and potentially effective treatments. Testing this hypothesis requires: building an infrastructure capable of such multidimensional representation of patients in all their clinical complexity and genomic variability (National Research Council *et al.*, 2011) and making the access and analysis of the data included in this infrastructure easy and reproducible. That is the reason why the BD2K PIC-SURE RESTful API (http://bd2k-picsure.hms.harvard.edu/) was created (Margolis, 2014). Actually, it provides a unique programmatic interface to access multiple large and complex data sets, containing data from different levels, like phenome, exposome or genome, in a reproducible way to follow the FAIR principles: findability, accessibility, interoperability and reusability (Wilkinson *et al.*, 2016). By providing some tools that make the access and analysis of these multiple layers of data through this API easier, the use of the API will be leveraged.

We created an R package named 'Rcupcake' that provides to the user an easy way to query data through the BD2K PIC-SURE RESTful API and a set of functions to perform descriptive analyses, univariate analyses and comorbidities studies.

2 Materials and methods

The R package 'Rcupcake' includes a set of 18 functions (Supplementary Section 9) for retrieving and analyzing any patient-level data—linked with a patient identifier (genomics, clinical, environmental, etc.). Bringing together the data from different studies on an individual takes considerable effort and a significant investment of time to understand how all the data relates to individual study participants. Rcupcake allows data scientists to generate data analysis for their own database easily.

2.1 Retrieving the data from the BD2K PIC-SURE RESTful API

Data retrieval using the BD2K PIC-SURE RESTful API requires several steps (http://bd2k-picsure.hms.harvard.edu/example-01.html): (i) to get the API key, (ii) to start a secure session, (iii) to create and run a JSON query and (iv) to download the data in one of the available formats.

The Rcupcake package only requires the URL, the access keys to the databases of interest and the location paths of the variables of interest from the end-user to retrieve the data. With this input, the Rcupcake package handles the connection to the servers, the JSON query building and the data downloading. Once done, the data can be exported for further analyses.

2.2 Data analyses and visualization

Rcupcake provides support for the study and visualization of some attributes contained in the datasets, enabling researchers to explore and characterize the retrieved data in three different ways: (i) descriptive analysis, (ii) univariate analysis and (iii) co-occurrence analysis.

The descriptive analysis can be performed using three different functions:

- ***demographic.summary*** performs the description of two main demographic characteristics— age and gender—of the population under study (examples and detailed information in Supplementary Section 5).
- ***phenotype.summary*** performs a descriptive analysis of all the variables of interest for the whole population and/or according to one variable (exposure, mutation, etc.) (examples and detailed information in Supplementary Section 6).
- ***phenotype.prevalence*** provides the proportion of all the variables of interest with their confidence intervals (examples and detailed information in the Supplementary Section 7.1).

The univariate analysis can be performed using the function:

- ***comparison2b2*** performs the univariate analysis between two variables. The function determines whether the variables are binary or continuous and computes a Fisher's test, a *t*-test or a Pearson correlation test.

The co-occurrence analyses and visualization can be performed using three functions:

- ***co.occurrence*** assesses the co-occurrence between two variables (usually diseases). Different co-occurrence tests and cut-off values can be chosen (examples and detailed information in the Supplementary Section 7). The co-occurrences analyses can also be performed according to a variable (Exposure) such as an exposome variable or a gene mutation in order to compare the results.
- ***cooc.network*** *and* ***cooc.heatmap*** create, respectively, a network plot or a heatmaps plot to visualize the results of the co-occurrence analyses.

3 Example applications of the Rcupcake package

To show the functionality and features of Rcupcake, two different datasets accessible with the BD2K PIC-SURE RESTful API are used as demonstration case studies, NHANES (Patel *et al.*, 2016) and Simon Simplex Collection (SSC) (Fischbach and Lord, 2010). NHANES is a public available database that contains phenome and exposome data for more than 41 474 patients while SSC encloses phenome and genome information from 2600 simplex families where there is one proband suffering from autism and needs a signed data-use agreement. The Rcupcake package enables the user to exploit the data from both of them, following the same easy structure (Supplementary Jupyter Notebooks). We have specifically illustrated this using the open source Jupyter notebook, which has been widely adopted across multiple data science disciplines both for its usefulness in keeping a record of data analyses and also for allowing reproducibility of the studies by any party, given access to these notebooks.

The two case studies show a step by step retrieval and analysis process for each of these databases. The steps described are (i) how to query the database through the BD2K PIC-SURE API RESTful API using R, (ii) how to perform a descriptive analysis of the population and the variables retrieved, (iii) how to perform a univariate analysis with the retrieved data and (iv) how to perform a co-occurrence analyses and visualize the results.

4 Conclusion

The Rcupcakge package combined with the BD2K PIC-SURE RESTful API facilitates the data access and analysis to any patient-level data linked with a patient identifier available through the API with R and provides a solid base for scalable and reproducible analysis. Rcupcake package is especially suited for researchers to analyze multiple layers of data, enabling the combination of thousands of different variables—clinical, environmental, self-reported, biochemical—all in one set of data structures that are queryable. Moreover, it can be integrated with other R packages available for follow-up analyses. It represents a step further to improve data-driven discovery, understanding and interpretation.

Funding

This work has been supported by the NIH BD2K grant U54HG007963; NIH Data Commons OT3 OD025466-01; NHLBI OT3 HL142480-01.

Conflict of Interest: none declared.

References

Fischbach,G.D. and Lord,C. (2010) The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron*, **68**, 192–195.

Margolis,R. *et al.* (2014) The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J. Am. Med. Inform. Assoc.*, **21**, 957–958.

National Research Council. *et al.* (2011) *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. National Academies Press, Washington, D.C.

Patel,C.J. *et al.* (2016) A database of human exposomes and phenomes from the US National Health and Nutrition Examination Survey. *Sci. Data*, **3**, 160096.

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