drugverse-RConsortium-Proposal

Mohammed Ali

2019-10-13

Signatories

Project team

- Mohammed Ali, R&D Developer @ Upland Software
- Ali Ezzat, Chief Data Scientist @ Synapse Analytics
- Nada Mamdouh, Teaching Assistant @ Modern Academy
- João Torres, Linguistic Administrator @ Council of the European Union

Contributors

Contributors are all drugverse users who reported bugs in or requested features to be added to drugverse.

Consulted

The Problem

At present, there is no single R package (or series of packages) that acts as a one-stop shop when it comes to pharmaceutical research activities. Practitioners in the pharmaceutical industry are currently forced to work with different tools for handling the different (data science) tasks. Oftentimes, these tasks require extra effort from the practitioners to develop their own code that does what they need depending on the situation.

As far as we know, there is no cohesive, well-integrated series of packages in R that tackles tasks that are typical of a data science workflow in the field of drug discovery. Ideally, this series of packages would:

- gather data from disparate data sources and be able to wrangle them despite the different formats they may come in
- explore the data from various aspects to come up with findings that are relevant to pharmaceutical researchers
- building/implementing prediction models tailored for the challenges in the field of drug discovery

Having such a series of packages would significantly reduce the time and effort required by pharmaceutical researchers in the drug discovery pipeline. Furthermore, since many researchers use R in their daily work, we imagine that these packages would be useful for them and the drug discovery community.

The proposal

drugverse is an R ecosystem for drug discovery that provides pharmaceutical researchers with different tools that mimic the data science life cycle and helps them to achieve their missions in solving challenges pertaining to drug development.

Overview

Following is what drugverse would be capable of doing upon its completion:

- wrangle data obtained from different drugs-related data sources and formats into forms that are more analysis-ready under R,

- explore and mine drugs-related data in Shiny applications containing user-friendly interfaces and interactive visualizations,
- run prediction algorithms (e.g. similarity based algorithms, feature based algorithms, semantic based algorithms, etc) for the prediction of different associations (e.g. drug-target interactions, drug-disease associations, etc.)
- validate prediction results both quantitatively (via latest known statistical validation methods) and qualitatively (e.g. viewing highest-ranked predictions and highlighting those that are already known in the literature) in intuitive and easy-to-understand reports.

Following the development of drugverse:

- The R community would benefit from packages that perform data retrieval, data wrangling and building prediction models for drug development
- Practitioners in drug discovery would benefit from visualization facilities provided in drugverse that would enable them to explore data in an interactive manner
- Practitioners would also be provided with facilities to implement cross validation strategies for use in drug discovery efforts

Detail

drugverse proposed tools are designed to handle the different stages of drug discovery data science projects lifecycle as follow:

- **dbparser (currently at version 1.0.4, updated regularly)**: It is an R package. Its main purpose is to parse the DrugBank database which is downloadable in XML format into more than 75 R dataframes. These dataframes can then be explored and analyzed as desired by the user. This package further provides the facility of saving the parsed data into a given database (e.g. MySQL).
- **DrugMiner (under development)**: **DrugMiner** is an R Shiny application that can be used to do the following:
  - DrugMiner is working on parsed data from **dbparser**
  - Displaying an interactive drug-target network.
  - Augmenting the visualized network with drug-enzyme, drug-transporter and drug-carrier relations.
  - Selecting certain drugs to perform specific tasks on them using one of the other tools that are available in drugverse such as the Drug-Target Interaction Predictor (see below).
- **Drug Target Interaction Predictor (under development)**: Drug-Target Interaction Predictor is an R package that provides implementations of different algorithms for predicting drug-target interactions. These algorithms can be categorized as follows:
  - Similarity-based methods
  - Bipartite local models
  - Matrix factorization methods
  - Feature-based methods
  - Network-based methods
- **Byakugan (in planning phase, development to commence shortly)**: Byakugan is an R Shiny application that provides the following features:
  - Extracting pharmacological information of drugs and biomedical documents using NLP and text mining techniques from different online resources to find implicit co-occurrent compound–protein relations.
  - Building its own knowledge base from the extracted information
  - Answering queries about drugs and targets based on the knowledge parsed from the different data sources (e.g. article abstracts, online biological databases, RDF databases, etc.)

Project plan

Start-up phase

All the code is hosted in GitHub Dainanahan Account in which each tool is hosted in its own reprository (i.e. dbparser). Moreover, each tool has the following:

- MIT License, i.e. dbparser license
- Code of Conduct, i.e. dbparser code of conduct
- Reporting issues, i.e. dbparser issues reporting
- Tool Site and Documentation, i.e. dbparser site

Technical delivery

drugverse tools are R packages and Shiny R applications detailed as follow:

- **dbparser** (already released) –> an R package with different functionalities to parse different elements of DrugBank xml database.
- **DrugMiner** (Expected to be released on December 2020) –> an R package that builds plots that are designed specifically for drug dicovery process in addition to an R shiny application that presents this plots and reports in an interactive and atrractive way.
- **Drug Target Interaction Predictor** (Expected to be released on February 2020)–> an R package that implement different machine learning algorithms in which are designed specifically for Drug Discovery process along with different cross validations methods.
- **Byakugan** (Expected to be released on February 2021)–> an R package that:
  - retrieve and parse related drugs documents (i.e. articles, books, papers).
  - apply different NLP and semantic algorithms for these documents to retrieve meaningful information.
  - build semantic database with that information.
  - build webservices to connect to other online drug semantic databases to enrich later phases of mining and prediction.
  - apply mining and prediction algorithms on semantic data.
  - In addition to the R package, an R shiny application will present the above features to the user with interactive and attractive plots and reports.

Other aspects

The following activities are intended to publish drugverse tools:

- present drugverse in rPharma conference 2020,
- write different papers for drugverse diferent tools and we intend to present them related Drug Discover conferences,
- publish a site for each tool (i.e. dbparser)
- publish posts about the tools in different channels (i.e. ResearchGate, Twitter, LinkedIn),
- announce packages on different R communites channels.

Requirements

People

It is required to have the dedication of a minimum of:

- three full-time data scientists
- two part-time pharamceutical researchers

- one part-time web developer
- one part-time DevOps engineer

Processes

At present, drugverse is being developed under the MIT license, and the link below has our code of conduct: https://github.com/Dainanahan/dbparser/blob/master/CODE_OF_CONDUCT.md

We are meeting regularly for the sake of:

- discussing drugverse's proposed features and how to implement them
- responding to the latest issues raised by users of drugverse

Tools & Tech

We intend to host the project on the cloud (RStudio hosting service) to host different web applications and data stores pertaining to drugverse.

Funding

We kindly request a funding total of **55,000 USD** to fund this project to its completion. The breakdown is as follows:

- **Maintenance and adding features for *dbparser*:** 5,000 USD
- **Implementation, Testing and Release of *Drug Miner*:** 10,000 USD
- **Implementation, Testing and Release of *Drug target interaction predictor*:** 10,000 USD
- **Implementation, Testing and Release of *Byakugan*:** 15,000 USD
- **Publishing articles at conferences and journals (plus travel expenses):** 10,000 USD
- **Cloud hosting:** 5,000 USD

Summary

Costs of this project consist of those for:

- paying the salaries of the personnel working on this project
- hosting the project on the cloud
- covering article publication and conference attending fees

Success

Definition of done

The drugverse series of packages would all be successfully uploaded to CRAN, and a Shiny R server that is using drugverse would be hosted on Shiny's hosting service (R Studio).

Measuring success

- The packages would easily accessible and downloadable from CRAN
- The users of drugverse and/or its Shiny R server would be able to provide us with feedback via drugverse's GitHub pages and via email
- Many users would fork/star drugverse's repository on GitHub as well as submitting pull requests to be approved
- We will have published a few papers on the different features provided in drugverse
- Users of drugverse would cite our papers in their work

Future work

We have the following short-term objectives in mind:

- Conduct further research into taking advantage of heterogeneous information sources for enhancing the accuracy of the different prediction algorithms.
- Further our self-branding on social media (LinkedIn, ResearchGate, etc.) to improve our reach as well as attend conferences to show our work (see next section for more details).
- Make use of ensemble learning and deep learning to further improve accuracy of the prediction models.
- Utilizing Big Data technologies and tools in order to support the huge amounts of data that we will eventually be dealing with.
- Conduct a search for individuals who are willing to contribute to drugverse and who preferably possess a background in pharmaceuticals.

Key risks

- It would hurt the project, if people who are working on it (especially the full-timers) leave for whatever reason.
- Some of the employees reside in other countries and communicating with them may not always be efficient online.
- Untimely delivery of drugverse's tools
- Costs of salaries and web-hosting services might not be adequately covered by the provided funds,