BIOMEDICAL DATA JOURNAL
datajournals.eu

Kissa M & Tsatsaronis G. *Biomed Data J.* 2015; 1(2): 10-12
http://dx.doi.org/10.11610/bmdj.01202

ISSN 2367-5322

***datajournals.eu*: The European Open Data portal**

Data Paper

A Benchmark Dataset for Computational Drug Repositioning

Maria Kissa [a], ***George Tsatsaronis*** (✉) [a]

[a] *Bioinformatics Group, Biotechnology Centre (BIOTEC), Technische Universität Dresden, Tatzberg 47-49,*
01307, Dresden , Germany
http://www.biotec.tu-dresden.de/

ARTICLE INFO:

RECEIVED: 09 Aug 2015
REVISED: 20 Sep 2015
ACCEPTED: 30 Sep 2015
ONLINE: 08 Oct 2015

KEYWORDS:

Drug Repositioning
Benchmark Dataset
Text Mining
Computational Methods

ABSTRACT

As the amount of the publicly available scientific literature increases, efficient text mining techniques are required to aid the biomedical knowledge workers in extracting the most important information from text and, ideally, to train models that may automatically suggest novel hypotheses. In this latter direction, the perspective of the employment of text mining techniques for computational drug repositioning, i.e., predicting new indications for existing drugs, is constantly attracting attention. One of the main obstacles for the development and establishment of such techniques is the systems' evaluation, as there is lack of benchmark datasets for performance evaluation. In this paper we introduce such a dataset for the evaluation of systems that perform computational drug repositioning. The dataset comprises 54 drug repositioning cases, the information for which was manually compiled, curated and integrated. The new indications for the reported cases have been approved by FDA or EU RUS in the period 1955-2013.

Overview

At the research frontier of drug design, several computational techniques have attracted the interest of researchers and pharmaceutical companies in the past two decades, such as drug repositioning, protein-ligand docking and scoring algorithms, and virtual screening. As it has been estimated that the time required to develop a new drug *de novo* ranges between 10 and 17 years, with the chances being only 1:5 000 and respective costs estimated at 4 billion US dollars, drug repositioning methods have attracted great attention, since the cost of a repurposing programme is significantly lower than *de novo* R&D for drug development and the cycle time is significantly shorter.[1]

In addition, it is estimated that drug repositioning accounts for approximately 30 percent of the newly FDA approved drugs and vaccines in recent years.[2] In contrast to the target-based repositioning methods, which cannot be used to identify new mechanisms beyond the known targets, the knowledge-based methods, which apply bioinformatics or chemoinformatics approaches to integrate the available information on drugs, targets and diseases, have the advantage that they incorporate known information for predicting unknown mechanisms beyond the known targets, e.g., un-

✉ Corresponding Author: Tel.: +49 (0) 351 796-5780; Fax: +49 (0) 351 796-5775
E-mail: george.tsatsaronis@biotec.tu-dresden.de

known targets for drugs and unknown drug-drug similarities.

Many of these methods actually apply machine learning for the task, i.e., using FDA approved labels or known side effects to compute drug-drug similarity or drug-target similarity, and produce models which can often lead to valuable predictions.[3,4]

However, a main obstacle for the establishment of such methods is the absence of a benchmark dataset which can be used for evaluating the performance of the methods, as well as their comparison. Given that more of these methods will appear in the literature in the near future, in this paper we introduce a dataset which has already been used for the evaluation of a text mining method for drug repositioning,[3] and which could constitute a basis for benchmarking these methods.

Method

For the creation of the dataset, we manually mined the literature and compiled the set based on *U.S. FDA*, *Wikipedia* and other web resources. Only drugs which have a *DrugBank* identifier were considered. Old and new indications are reported along with the year of approval of each drug's new indication. In addition, the source of the information, which constituted the basis for the reported information, is listed. A total of 54 repositioning cases have been compiled, for which the new indications for the respective drugs have been approved by FDA or EU RUS in the period 1955-2013.

Data Records

The dataset records are presented in a Microsoft excel file (.xlsx). Each record contains seven columns: The first column is the DrugBank ID of the drug; the second is the drug name (label name); the third is name of the old/original indication; the fourth is the name of the new indication; the fifth is the year that the new indication was approved; the sixth is the status (it can take the values FDA APPROVED or EU RUS APPROVED); and the seventh are the sources of the information (URLs), separated by a comma.

Validation

This dataset has been used in the recent past for the validation of a computational drug repositioning method based on text mining.[3]

Use and potential reuse

The dataset was compiled considering the way knowledge-based computational drug repositioning methods work, e.g. based on the analysis of literature and text mining techniques. It is therefore appropriate for re-usage by methods which mainly operate on the concept of analysing the literature, in order to predict/extract the reported cases.

A straightforward way of using the dataset can be by prioritizing predictions of new indications for a specific drug and, based on the list (i.e. the method can operate on the list of the reported drugs), mean average precision, recall and F-Measure can be estimated, as well as ROC curves. Another possibility is to see the dataset as a qrel file (queries and relevant documents), similar to the way researchers in the field of information retrieval (IR) operate. From this perspective, the queries could be the drugs, and the new indications – the expected relevant retrieved documents. In this way, all of the traditional evaluation measures from the IR field can be applied. Finally, the dataset can also be used reversely, e.g. systems that have obtained a diagnosis (i.e. the new indications) and need to find appropriate drugs for the treatment of the patients. However, this dataset was made for, and its primary value is as a drug repositioning dataset used in a way thoroughly described in our previous work.[3]

Conclusions

In this paper we have introduced a dataset for the evaluation of computational drug repositioning methods. We envisage the adoption of this dataset as a basis for the comparative evaluation of knowledge-based methods that are able to produce novel predictions for the repositioning of existing drugs.

Acknowledgements

The authors would like to acknowledge funding by the EU OpenScienceLink project and the BMWi GeneCloud project.

References

[1] Arrowsmith J, Harrison R. Drug Repositioning: The Business Case and Current Strategies to Repurpose Shelved Candidates and Marketed Drugs. In: Barratt MJ and Frail DE, editors. Drug Repositioning: Bringing New Life to Shelved Assets and Existing Drugs. Hoboken, NJ: Wiley, 2012. p. 9-32.

[2] Jin G, Wong S. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. Drug Discovery Today. 2014 May;19(5): 637-44. DOI: 10.1016/j.drudis.2013.11.005.

[3] Kissa M, Tsatsaronis G, Schroeder M. Prediction of drug gene associations via ontological profile similarity with application to drug repositioning. Methods. 2015 Mar;74:71-82. DOI: 10.1016/j.ymeth.2014.11.017.

[4] Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug Target Identification Using Side-Effect Similarity. Science. 321(5886);2008: 263-66. DOI: 10.1126/science.1158140.