Plan Overview

A Data Management Plan created using DMPTuuli

Title: Development of methods and tools supporting drug repurposing

Creator:zia ur rehman

Principal Investigator: Zia ur Rehman

Data Manager: Zia ur Rehman

Project Administrator: Zia ur Rehman

Affiliation: University of Helsinki

Funder: The Research Council of Finland (former The Academy of Finland)

Template: Academy of Finland data management plan guidelines (2021-2023)

ORCID iD: 0000-0003-2435-9862

Project abstract:

Drug development process consumes 8-12 years and approximately one billion US dollars in terms of costs. Due to high finances and time costs required by traditional drug discovery paradigm, repurposing of old drugs to treat cancer as well as rare diseases is becoming popular. Computational approaches are largely data-driven and involve systematic analysis of different data types leading to the formulation of repurposing hypotheses. These data types include chemical structures, adverse event profiles, drug-target interactions (DTI), pathways, gene disease associations, genomics, proteomics, and transcriptomics. Extracting and integrating these heterogeneous data types is challenging, and there is insufficient data, especially on DTIs, which may result in misleading conclusions.

This project will develop open-sourced packages (like RDKit and OpenBabel) to automatically extract the comprehensive chemical, genomics, and pharmacological data types from public resources (WP1). Data extraction packages will be cost effective, easy to use, and scalable with other tools. Based on these packages, we will develop computational drug repurposing tools to explore new therapeutic potential for approved and investigational compounds (WP2). Currently, 11% of the human proteome is targeted by small molecules. We will develop **BERT** based text mining models to improve protein target coverage across approved and investigational compounds (WP3). Furthermore, we will device scoring methods to rank the mined protein targets, link those with crowd-sourced platform (**DrugTargetCommons**) and assist manual curators in extracting quantitative bioactivity measurements (WP4). Finally, we will experimentally validate (WP5) the top predicted indications to assess the translational potential of the computational methods.

This project will generate a great scientific benefit through the development of useful tools to explore alternative drug indications from different perspectives, and aid in the drug discovery process.

ID: 19477

Start date: 01-09-2022

End date: 31-08-2026

Last modified: 05-07-2022

Grant number / URL: 351507

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Development of methods and tools supporting drug repurposing

1. General description of data

1.1 What kinds of data is your research based on? What data will be collected, produced or reused? What file formats will the data be in? Additionally, give a rough estimate of the size of the data produced/collected.

Most of the datasets in this project are publicly available in different databases. Some of these can datasets are available as stand-alone tabular files, a few as SQL dumps, while others as available through Application Programming Interface (API). Types of datasets used in this project are chemical structures, adverse event profiles, drug-target interactions (DTI), gene-disease associations, genomics, proteomics, transcriptomics, and PubMed literature (for text mining).
AML patient samples are obtained from the HUCH hematology clinic, under ethical approval on Oct 12, 2010 (No. 239/13/03/00/2010). This project does not need the patient's identifiers, so we store the patient-derived cancer sample data with anonymous identifiers.

Data type	Source	File format	Sensitivity (controller)	Size estimate
Chemical structures	Data reused from https://pubchem.ncbi.nlm.nih.gov/ and https://www.rdkit.org/	.txt, .csv	No	20GB
PubMed annotations and articles	Some data will be reused from https://www.ncbi.nlm.nih.gov/research/pubtator/, while new drug-target annotations will be produced	.txt, .csv	No	40GB
Adverse event profiles	Data reused from: http://sideeffects.embl.de/	.txt, .csv	No	400MB
Drug-target interactions	Data reused from: https://www.ebi.ac.uk/chembl/ and https://go.drugbank.com/	.txt, .csv, SQL dumps	No	20GB
Disease gene associations data	Data reused from: https://www.opentargets.org/ and https://www.disgenet.org/	.txt, .csv	No	20GB
Genomics, proteomics, transcriptomics	Data reused from: https://depmap.org/portal/ , https://sites.broadinstitute.org/ccle/ and https://www.cancerrxgene.org/	.txt, .csv	No	1GB
Drug-target interactions	Data produced	.txt, .csv	No	1GB
Computional drug repurposing	Data produced	Xlsx, csv, jpgs	No	1GB
Drug annotations	Data produced	Xlsx, csv, jpgs	No	1 GB

1.2 How will the consistency and quality of data be controlled?

In this project, I do not need to access personal data based on the minimization principle, and the whole analysis procedure is personal-data-independent. All the patient information will be handled anonymously. Data scientists involved in the project will be responsible for handling the data, and they know validation methods. They will be familiar with the quality control pipeline, such as noise detection, data normalization, etc.
Newly mined drug-target interactions or indications will be cross-validated using computational approaches, and top predictions will be validated experimentally to have quality control for the produced data.

2. Ethical and legal compliance

2.1 What legal issues are related to your data management? (For example, GDPR and other legislation affecting data processing.)

AML patient samples are obtained from the HUCH hematology clinic, under ethical approval on Oct 12, 2010 (No. 239/13/03/00/2010). This project does not need the patient's identifiers, so we store the patient-derived cancer sample data with anonymous identifiers. This project does not need to save or work with any sensitive data.

2.2 How will you manage the rights of the data you use, produce and share?

In this project, I will follow an open data policy. The produced data by this project will be released to the public under license: Creative Commons Attribution-Share Alike 4.0 International: https://creativecommons.org/licenses/by-sa/4.0/.
Also, I will only redistribute existing open data if its license allows redistribution. I will personally contact the authors for permission if the license is not explicitly represented. We have ethical approval for using patient-derived cancer samples.

3. Documentation and metadata

3. How will you document your data in order to make the data findable, accessible, interoperable and re-usable for you and others? What kind of metadata standards, README files or other documentation will you use to help others to understand and use your data?

All the produced datasets will be made publicly available using web applications (findable and accessible). We will provide a detailed user guide on data types and accessing methods. Moreover, we will provide programmatic access over produced datasets using an application programming interface (API) to make datasets reusable for translational researchers and developers of new tools (Interoperable and reusable).

4. Storage and backup during the research project

4.1 Where will your data be stored, and how will the data be backed up?

I will take advantage of the IT services University of Helsinki provides and maintains. I will use the UH options for storing primary documents, e.g., dedicated physical servers. The IT team at the University of Helsinki (at Meilahti campus) provides a data backup option. The data will be stored behind the firewall in protected databases with access control.

4.2 Who will be responsible for controlling access to your data, and how will secured access be controlled?

I will be responsible for controlling data access with the IT team's help.

5. Opening, publishing and archiving the data after the research project

5.1 What part of the data can be made openly available or published? Where and when will the data, or its metadata, be made available?

All data produced by my lab experiments, described in the research plan, will be made openly available and published in a data journal such as Data in Brief or Scientific Data. Data will be released together with the final publication of each working plan in this project.

5.2 Where will data with long-term value be archived, and for how long?

The data with long-term value will be stored at the servers of the University of Helsinki. My group will develop web applications (hosted at university servers) to provide access to data to the public.

6. Data management responsibilities and resources

6.1 Who (for example role, position, and institution) will be responsible for data management?

The principal investigator and the hired Postdoc will be responsible for many data management tasks, including producing metadata, anonymizing, arranging, transferring data, etc., during the research project life cycle.

6.2 What resources will be required for your data management procedures to ensure that the data can be opened and preserved according to FAIR principles (Findable, Accessible, Interoperable, Re-usable)?

Our estimated resources for data management are as follows; financial: One-tenth of Grant, time: four hours per experiment each week.