UPPSALA
UNIVERSITET

Thesis Plan

Metabolomic fetcher and analysis

Rajmund Csombordi Supervisor: Sara Younes
Subject Reader: ?

November 27, 2019

1 Introduction

1.1 Metabolites

Metabolomics is the scientific study of chemical processes involving metabolites, intermediate end products of metabolic pathways. This field of study has great importance due to several diseases being associated with metabolism disorders. Any metabolic disorder or disease is linked to the imbalance of metabolism and consequently changes in the proportions of metabolites.

For instance, the root causes for Type 1 Diabetes - the most common metabolic disease - is still unknown. In this type, Immune T cells attack beta cells of the pancreas, causing insulin production cease to exist.

1.2 Metabolome databases

Metabolome databases store various data regarding to metabolites and metabolic pathways. The problem with metabolomics being a new field of study is that information is scattered across these databases, and they do not provide a standardized and easy to access interface for researchers. Means of data access have to be established before every project and this slows down precious delivery time of research results.

Additional challenges occur when we try to deal with Metabolome databases. Some of the databases use very unusual formats (e.g. mysql dumps) as ways of delivering their data. Moreover, data export is not even consistent within one database's coverage.

To make things worse, foreign reference identifiers may be missing, making it difficult, sometimes impossible to find the link between two records in different databases of the same metabolite. Some reference identifiers are present, however they link to the wrong metabolite.

A unified package that was able to handle all of these scenarios and covered all use-cases of a data fetcher would be highly beneficial for the metabolomics research community. We expect a high number of users and citations for this project.

2 Work

In this thesis project I would create an R package that allows the users to fetch and accumulate metabolome data from several databases and present them in a standardized way.

2.1 Databases to cover

- HMDB - Human Metabolome Database
- ChEBI - Chemical Entities of Biological Interest
- KEGG - Kyoto Encyclopedia of Genes and Genomes

2.2 Preparation

Initial interviews with several researchers would be made to discover their use-cases and needs with such databases. This exploration of intents would be done in order to make a package that suits real user needs and not what I would assume they'd like. Additional features could be implemented based on these needs. Further inspirations could be drawn from other similar tools such as Biomart.

I would make a best-effort scanning of available database formats to see what types of data I'd have to deal with during the project. Quick spike tests would be made to prepare initial prototypes for data handling.

2.3 Data fetching

A prototype would be developed that is capable of accessing all kinds of data from the mentioned databases. Furthermore, the package would enable a local caching where the queried metabolome data would be stored in a standardized way.

In the second part of the development, I would extend the package's features to make it work as seamlessly as possible in the eyes of an end user. Several iterations of verification by the supervisor and tailoring of features is expected in this part.

2.4 Enrichment analysis

In the final part of the project, using the newly produced tool, an arbitrary subset of metabolome data associated with a certain disease would be queried and used in an enrichment analysis, as well as in differentiation analysis of metabolites.