Metabolic Diseases: Pathway Analysis

Computational Health Laboratory Project

Simone Marzeddu - Jacopo Raffi

Academic Year 2022-2023

1. Introduction

Metabolic diseases are a class of typically hereditary disorders which affects the natural biochemical processes operated by enzymes, consisting of the conversion of food to energy on a cellular level. The consequences of those metabolic imbalances can be manifold and severe, with symptoms that include (depending on the dysfunctional enzyme): blindness, deafness, convulsions, decreased muscle tone, and even intellectual disability [1].

1.0.1 Project's Goals

The final project discussed in this report, related to the "Computational Health Laboratory" examination, consists of the selection and combination of pathways, found on "Reactome" [2], belonging to a subclass of metabolic diseases together with drug interactions obtained from "DrugBank" [3]. Through the generation of a graph modelling the links between drugs and diseases, the ultimate aim of the project is to analyse the distances between the nodes of the graph, obtaining a ranking capable of highlighting interesting relationships between drugs and each of the diseases of interest.

1.0.2 Document's Structure

The rest of the document is organised as follows:

- **Chapter 2**, where are discussed the main implementations and development steps of the project;
- **Chapter 3**, which discuss the technical aspects about code structure, execution and dependencies;
- **Chapter 4**, where results are discussed and visualised;
- **Chapter 5**, which consist of a conclusive summary and a discussion about future updates for the project.

2. Project Development

Considering the classification presented in "Metabolic Disease", published by "Encyclopedia Britannica" [1], the class of metabolic diseases around which the implementation will be oriented is "disorders of amino acid metabolism". In particular, the diseases selected for the study are twelve: "Phenylketonuria" (PKU), "Tyrosinemia Type I", "Tyrosinemia Type II", "Tyrosinemia Type III", "Homocystinuria", "Hyperglycinemia, "OTC Deficiency", "Cystinuria", "Lisinuric Protein Intolerance" (LPI), "Propionic Acidemia", "Methylmalonic Acidemia" and "Maple Syrup Disease" (MSP).

2.1 Pathways Selection

The first essential step of the project, after an initial theoretical research phase, consisted in selecting, downloading and combining metabolic disease pathways extracted from "Reactome" [2]. In particular, two types of pathways were selected: those relating to the specific twelve diseases highlighted in the previous section, and those relating to "healthy" metabolic processes, so as to integrate information on nodes and edges that were absent in diseases' pathways due to the fact that these may inhibit or alter certain processes. All data downloaded from "Reactome" [2] and generally used in this first phase are collected in the project's "./data" folder.

2.2 Integration of drug interactions

The reference source for known drugs for each of the diseases investigated, as well as for each of the drugs evaluated in the rankings, is "DrugBank" [3]. A design choice applied at this stage was to consider all and only the drugs belonging to the file "pharmacologically_active.csv" found in "Drugbank" [3]. This is because, since the implementation approach is based on distances in a graph, this would penalise in the ranking phase all the drugs that by their very nature do not interact directly with the diseases under examination,

and would therefore make the system heavier and slower, while not providing important contributions to the results.

This step resulted in the addition to the graph of nodes representing the 2220 drugs under consideration for each disease ranking, including the 14 drugs known for the treatment of the 12 metabolic diseases under consideration. Each drug node has edges pointing to each of the targets of the specific drug it models, also represented by nodes in the graph. Finally, at the end of this step, as well as at the end of the production of the final complete graph, each node will represent a specific biological entity, unique in the whole graph and source of edges modelling every useful known interaction of the component, pointing to each of the nodes it interacts with.

2.3 Graph enrichment and completion

Exploiting "BioGRID" [4], the aim of this step was to integrate novel information (node and edges - components and interactions) to enrich the already built graph. In particular, the components selected as a starting point for integration where all the genes present in the graph, both from diseases and drugs. The new data where acquired by performing queries on the server, specifying the genes of interest, and using a file ("BIOGRID-ORGANISM-Homo_sapiens-4.4.221.tab3.txt"), containing all interactions in the human body, subsequently filtered.

2.4 Ranking algorithm

As the focus during development was more directed toward the study of data structures and results interpretation, the implementation of the algorithm for the calculation of rankings is based on a simple and basic approach. Nevertheless, however, the project produced interesting results that will be evaluated and validated in the next sections.

The implemented algorithm calculates, for each drug, the average distance that the node that models it in the graph has with respect to the various components of the disease under examination. This exploiting also the hereditary nature of the metabolic diseases, that permits the approximation on the components (represented by nodes) of those diseases as the genes which are in any sense involved in those diseases' pathways.

2.5 Validation

After the ranking phase, a series of four validations were conducted: first a set of two qualitative validations and than another set of two statistical validations:

- **first qualitative test**: the purpose of the test is to verify that known drugs are at the top of the rankings for their respective (target) diseases. 79.17% of drugs qualify in high rankings for their target diseases;
- **second qualitative test**: the objective of the test is to verify that for each known drug, the mean rank achieved in its target disease ranking is higher than the mean rank achieved by the same drug in non target diseases rankings. A total of 3 out of 14 drugs fail this test;
- **first statistical test**: the aim of the test is the same as the previous one, but in this case a statistical validation is performed through a t-test. The test is performed on two vector: the first containing the rankings of known drugs towards the respective target diseases, while the second contains the rankings of the same drugs but for diseases not yet related to these drugs. The p-value obtained for this test is lesser than 0.05, which tests that, for diseases under consideration, known treatment drugs are better classified than unknown ones;
- **second statistical test**: via t-test, this test shows that, considering all the rankings, the 14 known drugs have a better rank than randomly taken drugs. The test is performed on two vector: the first is identical to the first of the previous test, while the second one contains the 1680 ranks (14 different randomly extracted drugs in the 12 diseases rankings, repeating the process in 10 times iterations). As a result, the p-value computed is lesser than 0.05, which demonstrate that known drugs achieve a better rank respect to randomly drawn drugs.

3. Technicity and Dependencies

This chapter provides a brief overview of the key support functions, implemented during the development of the project, named as follows:

- **biopax2igraph** reads a "bioPAX level 3" file and transform it into a "igraph" object;
- **csv2igraph** reads a "DrugBank"'s csv file, and computes the associated graph;
- **splitNodes** replaces the graph's nodes that represent a "Complex" element with its components;
- **interaction2igraph** transforms "BioGRID" [4] interactions in nodes and edges of the graph;
- **interaction2igraph2** reads a "tab3.txt" file and returns the related graph;
- **ranking** computes the drug ranking for a specific disease by computing for each drug the mean distance to disease's components, the rank and the percentage of total outclassed drugs;
- **test1Ranking** performs a t-test to verify that the already known drugs achieve better ranking on their target diseases respect that on non target diseases;
- **test2Ranking** executes a t-test to evaluate that the already known drugs achieve a better ranking than random drugs.

The external libraries used for the project are "rBioPaxparser" [5] for reading biopax files, "igraph" [6] for manipulating and analysing graphs and "biogridr" [7] for extracting the interactions of biological components. The main script to start the execution is "pathway_analysis.R" which exploits the support functions located in "graph_lib.R". The code is made public at the following link `https://github.com/JacopoRaffi/Pathways_Analysis`.

4. Results

Once the obtained rankings have been validated, it is possible to observe the resulting information. In this section, the best classified drugs for the diseases under investigation will be discussed in detail. Before highlighting the results, however, it is necessary to state that all the results of an analysis such as the one performed by the project remain only an intuition that is meant to represent a guideline for potential biological studies based on the evidences obtained from this first approach.

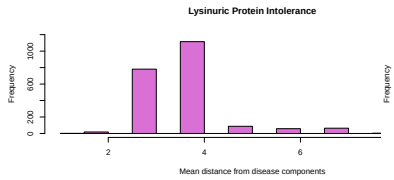
The vast majority of drugs with excellent ranks turn out to be **anti-cancer treatments**, examples being "Entinostat", "Vorinostat" and "Belinostat", detected in the ranking for "Propionic Acidemia" [3].

Also of interest are the findings of treatments for "**Alzheimer**" and "**Parkinson**", found in the rankings of "Methylmalonic Acidemia" and "Propionic Acidemia", respectively, represented by the drugs "Gantenerumab" and "Amantadine" [3].

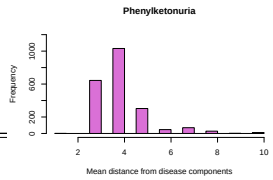
Another case is that of "Rabies Immune Globuline", a solution of antibodies used to prevent "**Rabies**" after an exposure, found to be both a good candidate in the rankings of "Homocystinuria" and "Propionic Acidemia" [3].

Other findings are also confirmed or at least supported by certain correlations, interactions and symptoms already found in the scientific literature when referring to the diseases in analysis. For example, "Vitamin B12" and "Hydroxocobalamin" were found in the highest positions for "Methylmalonic Acidemia", a disease that appears to be related [8] to the deficiency of "Vitamin B12", of which "Hydroxocobalamin" is a synthetic substitute [3]. Similarly, correlations have been found between drugs that seems related to "Hypercalcemia", such as "Zoledronic Acid" and "Pamidronic Acid", in the ranking of "Hyperglycinemia", a metabolic disease that seems to have correlations [9] with the former [3].

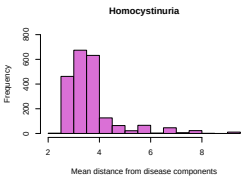
Finally, an honourable mention goes to "Foreskin Keratinocyte (Neonatal)", a drug that ranks highly for the vast majority of analysed diseases [3].



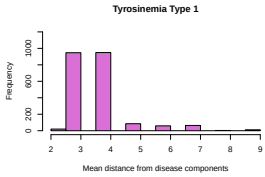
(a) LPI



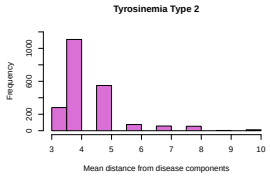
(b) Phenylketonuria



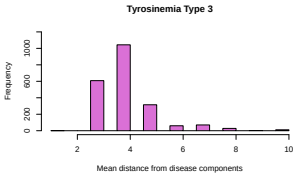
(c) Homocystinuria



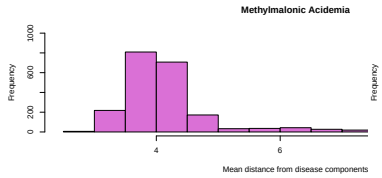
(d) Tyrosinemia Type I



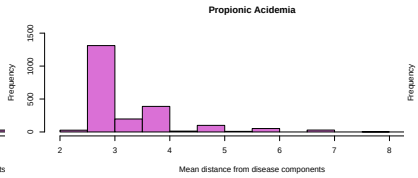
(e) Tyrosinemia Type II



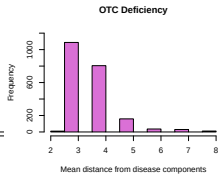
(f) Tyrosinemia Type III



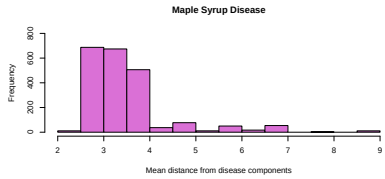
(g) Methylmalonic A.



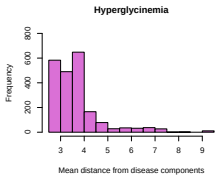
(h) Propionic Acidemia



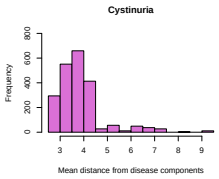
(i) OTC Deficiency



(j) Maple Syrup Disease



(k) Hyperglycinemia



(l) Cystinuria

Figure 4.1: Drugs mean distance frequency distribution for each disease's ranking.

5. Conclusions

By selecting pathways related to disorders of amino acid metabolism from "Reactome" and enriching the obtained graphs with information on drugs, healthy metabolic processes and general interactions of the human organism via "Biogrid", it was possible to produce a ranking of the 2220 drugs obtained from "DrugBank" for each of the 12 identified metabolic diseases. A valuable feature of the resulting rankings is the distribution of the percentages of drugs per rank. In fact, thanks to the bell curve of distribution, the drugs which achieved the best ranks are generally in very low percentage considered the total number of drugs. a key feature to make future studies in the field of Drug Repurposing specific and targeted.

5.1 Future Plans

During the development, as well as following the analysis of the results, several insights emerged that could be taken into consideration for future approaches and directions for the project.

Thanks to its modular and extensible nature, in fact, the implementation would lend itself to the application of analysis on different sets of drugs, taken for instance from other databases, representing other drug categories or considering in general not only the drugs listed in the "pharmacologically_active.csv" file from "DrugBank". It would also be relevant to perform ranking even with less basic algorithms, potentially even disengaging from the use of distances as a ranking parameter, exploring instead factors such as flow or approaches similar to the "PageRank" algorithm.

Finally, of even greater importance would be the submission of the results obtained from rankings to experts in Biology and metabolic diseases, in order to obtain a feedback on the system's performance, as well as fulfilling the purpose of research in the field of Drug Repurposing.

Bibliography

[1] Enns and Gregory. *Metabolic Disease.* "Encyclopedia Britannica", 2019.

[2] Griss J, Viteri G, Sidiropoulos K, Nguyen V, Fabregat A, and Hermjakob H. *Efficient Multi-Omics Comparative Pathway Analysis.* "ReactomeGSA", 2020.

[3] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. "DrugBank 5.0: a major update to the DrugBank database for 2018". In: *Nucleic Acids Research* 46.D1 (Nov. 2017), pp. D1074–D1082. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1037. eprint: https://academic.oup.com/nar/article-pdf/46/D1/D1074/23162116/gkx1037.pdf. URL: https://doi.org/10.1093/nar/gkx1037.

[4] "BioGRID: a general repository for interaction datasets". In: *Nucleic Acids Research* 34.suppl$_1$ (Jan. 2006), pp. D535–D539. ISSN: 0305-1048. DOI: 10.1093/nar/gkj109. eprint: https://academic.oup.com/nar/article-pdf/34/suppl_1/D535/3925435/gkj109.pdf. URL: https://doi.org/10.1093/nar/gkj109.

[5] *rBiopaxParser.* https://github.com/frankkramer-lab/rBiopaxParser.

[6] *igraph.* https://github.com/igraph/igraph.

[7] *biogridr.* https://github.com/npjc/biogridr.

[8] Xiaoyan Zhou, Yazhou Cui, and Jinxiang Han. "Methylmalonic acidemia: Current status and research priorities". In: *Intractable Rare Dis Res* (2018).

[9] S Nagasaka, T Murakami, T Uchikawa, S E Ishikawa, and T Saito. "Effect of glycemic control on calcium and phosphorus handling and parathyroid hormone level in patients with non-insulin-dependent diabetes mellitus". In: *Endocr J* (1995).