Final exam - version 1.1 - 24.01.2025

DatabaseDrugBank is a publicly available and free database of information about drugs (medicinal substances). It was created in 2006 by the team of Craig Knox and David Wishart from the Department of Computer Science and Biological Sciences of the University of Alberta in Canada. It combines data from the fields of chemistry, biochemistry, genetics, pharmacology and pharmacokinetics.

Since access to the full database requires creating an account, filling out a form with a justification for the request for access, and obtaining approval, for the purposes of this credit project you will be provided with the drugbank_partial.xml file with a reduced version of the database. This file contains data for 100 drugs (the full database published on 2024-03-14 has information on over 16,000 drugs).

The project involves analyzing the contents of a shortened version of the database and creating various tables and graphs summarizing the contents of the drug database.

1) Create a data frame that contains the following information for each drug: unique drug identifier in the DrugBank database, drug name, type, description, form in which the drug occurs, indications, mechanism of action, and information on what foods the drug interacts with. (**4 points**)

2) Create a data frame that allows searching for information about all synonyms under which a given drug occurs by DrugBank ID. Write a function that will create and draw a graph of synonyms for a given DrugBank ID using the libraryNetworkX . The generated drawing should be legible. (**4 points**)

3) Create a data frame for pharmaceutical products containing a given drug (drug substance). The frame should contain information on the drug ID, product name, manufacturer, US National Drug Code, form in which the product occurs, method of application, information on the dose, country and agency registering the product. (**4 points**)

4) Create a data frame containing information on all pathways of all types, i.e. signaling, metabolic, etc., with which any drug interacts. Provide the total number of these pathways. (**4 points**)

5) For each signaling/metabolic pathway in the database, provide the drugs that interact with it. The results should be presented in the form of a data frame and in a graphical form of your own design. An example of such a graphic could be a bipartite graph, where the two types of vertices are signaling pathways and drugs, and the individual edges represent the interaction of a given drug with a given signaling pathway. The graphical presentation should be legible and attractive. (**4 points**)

6) For each drug in the database, provide the number of pathways with which the drug interacts. Present the results as a histogram with appropriately labeled axes.
(**4 points**)

7) Create a data frame containing information about the proteins with which the individual drugs interact. These proteins are called targets. The data frame should contain at least the DrugBank ID of the target, information about the external database (*source*, e.g.Swiss-Prot), external database identifier, polypeptide name, polypeptide gene name, GenAtlas ID, chromosome number, cellular location. (**4 points**)

8) Create a pie chart showing the percentage of targets in different parts of the cell. (**4 points**)

9) Create a data frame showing how many drugs have been approved, withdrawn, experimental (*experimental* or *investigational*), and approved for animal treatment. Plot this data in a pie chart. Indicate the number of approved drugs that have not been withdrawn. (**4 points**)

10) Create a data frame containing information about potential interactions of a given drug with other drugs. (**4 points**)

11) Develop a graphical presentation according to your own idea, containing information about a specific gene or genes, medicinal substances that interact with this gene/genes, and pharmaceutical products that contain a given medicinal substance. The choice of whether the graphical presentation is made for a specific gene or all genes at once is left to you. When making your choice, you should be guided by the readability and attractiveness of the graphical presentation. (**7 points**)

12) Propose your own analysis and presentation of drug data. You can obtain additional information from other biomedical and bioinformatic databases available online. However, you should make sure that the database allows for automated data collection by the program. For example, the databaseGene Cards explicitly prohibits this, which is highlighted in red on this pagepage . Example databases are: UniProt (https://www.uniprot.org/), Small Molecule Pathway Database (https://smpdb.ca/), The Human Protein Atlas (https://www.proteinatlas.org/). (**7 points**)

13) Create a simulator that generates a test database of 20,000 drugs. The values of the generated 19,900 drugs in the "DrugBank Id" column should have consecutive numbers, and in the remaining columns the values should be randomly selected from the values of the existing 100 drugs. Save the results in the file drugbank_partial_and_generated.xml. Conduct the analysis according to points 1-12 of the test database. (**7 points**)

14) Prepare unit tests using the pytest library. (**7 points**)

15) Complete point 6 so that you can send the drug ID to your server, which will return the result in response (use fastapi and uvicorn; it is enough to demonstrate sending data using the POST method, via Execute in the documentation)**(4 points)**