bio1_coursework2_instructions

November 8, 2023

Introduction

An Exploration of GenCC the Gene Curation Consortium Database of Gene-Disease Mappings

In this coursework we are going to put into practice what you've been learning about biological databases, ontologies, and network analysis. In the process you will find and retrieve flat-files and perform integrative analyses with the data.

In **Part One** we focus on some summary analysis of data from the Gene Curation Coalition (GenCC). In **Part Two** we look at "Modes of Inheritance" and use the Human Phenotype Ontology. In **Part Three** we use the MONDO disease ontology to look at subsets of the GenCC data and gene-disease connections. In **Part Four** we build a simple disease network based on genes associated in common between pairs of diseases, cluster the network and plot the resulting graph.

All that you need to successfully complete this project has been covered in the course using web-tools and/or programming approaches. This coursework relies more on python coding than the first one. Please do ask questions in the forum, but please do not reveal answers to the coursework when you do so. It is fine for you to use either approach or a hybrid of both, but however you aim to undertake the work everyone needs to write up a report of the research. Throughout the instructions, below, we guide you through the project and provide tips, especially where there is an opportunity to do something a little more complex if you choose. We hope you enjoy the project and exploring in a real-world example how you can bring together different sources of information to anayse and gain insight into a research problem.

Research Instructions

Part One - *Summary Analysis of GenCC Data - parsing flat files (**Total 20 marks**)*

For the first part of the coursework we will be using a flat-file downloaded from the Genomics Curation Consortium (GenCC) who curate relationships between genes and diseases. Please use this link to download the file.

- **Task 1** - Using the GenCC flat-file calculate the number of genes asscoiated with each disease. Present a table of the top 10 diseases with the most genes associated with them. The table should include two columns, the first the disease name and the second the number of genes associated to the disease (**3 marks**)
- **Task 2** - Plot the table data from Task 1 as a barchart with appropriate labels (**2 marks**)
- **Task 3** - In the GenCC data each entry is tagged with a confidence categories of `Definitive`, `Strong`, `Supportive`, `Moderate`, `Limited`,`Disputed`, `Refuted`, or `No Known Disease Relationship`. Calculate the number of GenCC entries that are in each of these

categories and present the results in a table with the rows in the same order as the preceding sentence (**3 marks**)

- **Task 4** - Plot the table data from Task 3 as a barchart with appropriate labels (**2 marks**)
- **Task 5** - The GenCC data has a column `submitted_as_assertion_criteria_url` that can contain provenance information to support the entry. Some of the entries are supported here by PubMed identifiers (PMID:1234), or by unique Digital Object Identifiers (DOIs) that are URLs, others have no such provenance. Calculate the number of entries that fall into each of these categories and present the results in a table. Check that the total adds up to the total number of rows in the data. (**5 marks**)
- up to **5 marks** for including one additional piece of analysis.

Possible extensions here might be to: - The GenCC data file also includes submission dates, you could analyse this to show the timeline over which entries were added. - There are other sources of gene-disease relationships available online, you could perform a similar summary analysis on one of these. - For the PubMed entries you find in Task 5 you could retrieve the paper meta-data and look at when the papers were published. How do these dates relate to when the entries were made in GenCC?

Part Two - *Modes of Inheritance - working with an ontology (**Total 15 marks**)*

The GenCC schema uses CURIEs (compact URIs) which have the form prefix:identifier, for example `HGNC:10896` which is a Human Genome Nomenclature Committee (HGNC) gene identifier. We are going to extract information about the mode of inheritance (MOI) for the diseases in GenCC and use ontologies to add information to our analyses.

- **Task 1** - How many entries are there in the GenCC data and what percentage of them have an MOI CURIE? (**2 marks**)
- **Task 2** - Finding meta-data for the `moi_curie` entries (**7 marks**) In the `moi_curie` column different modes of inheritance are represented by accession ids from the human phenotype ontology e.g. `HP:0000006`
  - Step 1 - Create a table of all of the unique categories in this field of the GenCC data in the first column and the mode of inheritance name in the second column.
  - *(HINT) - In order to do this you should download the human phenotype ontology file from the NCBO BioPortal and use it to find matches to the `moi_currie` terms in the GenCC data. We did something similar in the notebooks in weeks 7 and 8 using the python library `pronto`*
  - Step 2 - Present the results in a table
- **Task 3** - Merge the table data from Task 2 with the GenCC data and then plot a barchart with `mode of inheritance name` agains number of genes for that MOI (**4 marks**)
  - Step 1 - Create a `Pandas` dataframe containing the `moi_curie`,`mode of inheritance name` and merge this with a dataframe containing the GenCC data to add the `mode of inheritance name` to all of the rows that have a valid `moi_curie`.
- **Task 4** - Why do you think the distribution of MOIs looks as it does in the barchart? (**2 marks**)

Part Three - *Disease Groupings and their Associated Genes - working with an ontology (**Total 20 marks**)*

In this part we will use the MONDO disease ontology to select a specific subset of GenCC diseases

and find their associated genes.

- **Task 1** - Download the MONDO ontology file and find the MONDO accession id for the term `nervous system disorder` (**1 mark**)
- **Task 2** - Create a list of all the MONDO terms below this node of the MONDO ontology, how many are there? (**1 mark**)
  - *(HINT) - use the `pronto` function `subclasses()` - details*
- **Task 3** - Present a table of the first 10 terms with the first column containing the `MONDO_ID` and the second column the `MONDO_NAME` (**2 marks**)
- **Task 4** - How many genes are linked to the diseases in the list of MONDO IDs you created in Task 2 (**8 marks**)
  - Step 1 - Use the MONDO list (which contains all nervous system disorder `MONDO IDs`) you created in Task 2 to select rows from the original GenCC data with matching MONDO IDs.
  - Step 2 - Create a new restricted GenCC dataset `NSD_GenCC` containing only those rows and count the number of genes per disease
  - Step 3 - Present the results in a table showing the top 10 diseases by gene count with the columns `MONDO ID`,`Disease Name`,`Gene Count`
- **Task 5** - Using the `NSD_GenCC` dataset created in Task 4 present a table showing the top10 genes by number of entries with the columns `Gene`,`Number of restricted GenCC entries` (**2 marks**)
- **Task 6** - How many unique Genes are there in the `NSD_GenCC` dataset? (**1 mark**)
- up to **5 marks** for including one additional piece of analysis.

Possible extensions here might be to: - perform the same analyses but for different disease groupings e.g. `digestive system disorder` or `disorder of the visual system` etc. - look at the disribution of diseases across the MONDO ontology, are particular disease groupings more densely represented? - look at the distribution of gene evidence in GenCC by disease groupings, are some disease groupings more highly represented than others, why might that be?

Part Four - * Building a Simple GenCC Disease Network - working with networks (**Total 20 marks**)*

In this part we will evaluate how many genes our `NSD_GenCC` diseases share in common and use this to build a simple disease network connecting diseases by the number associated genes they share.

- **Task 1** - For every pair of diseases in the `NSD_GenCC` dataset count the number of genes they have in common. Present a table of the top 10 disease pairs and their common gene count with the columns `Disease Pair`, `Gene_Count` with the format of the first column `MONDO:ID1,MONDO:ID2` e.g. `MONDO:0100038,MONDO:0000508` (**6 marks**)
- *(HINT) - use the `itertools` function `combinations` to create a list of pairs
- *(HINT) - you can use a `for` loop to go through the combinations or a `string literal` which is faster
- *(HINT) - in Python you can use `set` comparison `set_a & set_b` to find the intersection between two sets

- **Task 2** - Build a network from the disease pair data created in Task 1 with diseases as nodes and only creating edges between nodes if they share three or more genes in common. Do not include unconnected (orphan) nodes. How many edges are there in the graph? (**6 marks**)
- **Task 3** - Cluster the network using the `networkx` function `algorithms.community.greedy_modularity_communities` with default settings - details. Present a table of the communities sorted by community size with the columns `Community Number`,`Number of Diseases` (**3 marks**)
- *(HINT) - the clustering algorithm returns a `list` of communities with each element containing a `set` of the diseases in that community
- **Task 4** - Plot the network with nodes labelled with the disease name and coloured by the cluster they belong to, layout the graph so that clusters are reasonably well separated (**5 marks**)
- *(HINT) - look at the course notebook for networks where we plot clusters in this way for a different example

Structure of the Report

Write a report of your project including figures, tables and results from your research in Parts One-Four above in the relevant sections. The report should follow the structure described below.

- Introduction
- Data & Methods
- Results
  - Part One - Summary Analysis of GenCC Data
  - Part Two – Modes of Inheritance
  - Part Three – Disease Groupings & their Associated Genes
  - Part Four - Building a Simple GenCC Disease Network
- Discussion
- References
- Appendices (optional)

Your report should use an 11-point font for the main text and be no more than 8 pages in total (including everything except the references and optional appendices). You should not include code in the main text of the report, but can include them in the appendices if you choose. Please do not paste screenshots of results into the report, but instead generate appropriate tables and figures that are properly labelled and formatted with sufficiently detailed legends to understand their content and key interpretation. Make sure that you document your work correctly using official accession ids, refrence the databases and resources you use, detail the methods and approaches you use specifying parameters, algorithms, or tools that you use as appropriate. When referring to information you have found either online or in publications site them appropriately in the report text and include them in a consistently formatted bibliography at the end of the report. You may use whichever referencing style you prefer.

Marking Scheme

Your mark for this coursework assignment will contribute 50% to your mark for the Bioinformatics 1 course.

The work will be marked out of 100 with the following scheme:

Overall Report [25 marks] - Data, Methods, Presentation of Results and their interpretation clearly described paying close attention to keeping track of versions and dates. Think of how able

another researcher would be to look at your report and repeat what you have done - reproducibility [10 marks] - Correct overall structure of the report [3 marks] - References and any Appendices in good order [2 marks] - Clear, well-labelled figures and tables throughout [10 marks]

Part One [20 marks] - 15 marks for successfully completing the tasks. - up to 5 marks for including one additional piece of analysis.

Part Two [15 marks] - 15 marks for successfully completing the tasks.

Part Three [20 marks] - 15 marks for successfully completing the tasks. - up to 5 marks for including one additional piece of analysis.

Part Three [20 marks] - 20 marks for successfully completing the tasks.