| Master Degree: Data Science a.y. 2024/2025 | 27 November 2024 |
| --- | --- |
| BIOINFORMATICS AND NETWORK MEDICINE | |

Putative disease gene identification and drug repurposing for <DISEASE>

A. Student, B. Student, C. Student

GROUP <nn>

Delivered: day Month year

**ABSTRACT**

Brief and clear paragraph: it should be a miniature of the manuscript including a brief description of scientific issue and aim, methods, main result, and conclusion.

**INTRODUCTION**

This section includes: the definition of the scientific issue; the summary of main results in the scientific literature (state of the art) and of the principles needed for the comprehension of the new hypothesis; the identification of the hypothesis, the disclosure of the relevance of the results and the overview of contents of following sections. In the introduction the aims of the study must be clearly stated. It does not include conclusions and recommendations.

**MATERIALS AND METHODS**

Describe in this section the experimental procedures and resources, data analysis procedures and statistical methods. Give enough detail to replicate the experiment but do not overwhelm the reader with too many details. For what concerns the specific work, please follow these main steps:

1. **PPI and GDA data gathering and interactome reconstruction**
   1. **Download PPIs** from BioGRID latest release to build the human interactome
      * use “all organisms” tab3 file, unzip and get “Homo sapiens” only
      * filter out all non-human interactions, i.e., both “organism A” and “B” fields must be = 9606 (Homo sapiens)
      * keep only “physical” interactions” (“Experimental System Type” = physical)
      * purge out redundant and self loops
      * isolate the largest connected component (interactome LCC)
   2. **Gather gene-disease associations** (GDAs) of the disease assigned is provided (DisGeNet starting this year is behind a paywall) along with these instructions in the file “DISEASES\_Summary\_GDA\_CURATED\_<id>.tsv
      * For each disease, GDAs are reported in the column “Gene”
      * verify on [HGNC](https://www.genenames.org/tools/multi-symbol-checker/) the correctness of gene names, report deviations and inconsistencies, resolve conflicting information, if any, also by checking the columns UnitProt, geneEnsemblIDs, geneNcbiID, geneNcbiType
   3. **Compute and characterize the disease LCC and its basic network measures**
      * Check for the presence of disease genes in the interactome LCC (as from point 1.1) and identify the disease interactome by getting the interactions among disease genes only, and the disease LCC (i.e. the LCC of the disease interactome)
      * Summarize the GDA-related data as in table 1
      * Compute the following network metrics on the disease LCC:
        + Node degree
        + Betweenness centrality
        + Eigenvector centrality
        + Closeness centrality
        + ratio Betweenness/Node degree
      * Report in a table the above network measures of the first 50 disease genes in the disease LCC ordered for node degree from higher to lower, as in table 2
      * Represent node degree and node betweenness in a scatterplot

*Table 1 Summary of GDAs and basic network data*

| disease name | UMLS disease ID | MeSH disease class | number of associated genes | number of genes present in the interactome | LCC size of the disease interactome |
| --- | --- | --- | --- | --- | --- |
| Blood Coagulation Disorders | C0005779 | C02 | 267 | 250 | 180 |

*Table 2 Main network metrics of disease LCC genes*

| Ranking | Gene name | Degree | Betweenness | Eigenvector centrality | Closeness centrality | ratio Betw./Degree |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | EWSR1 | 91 | 0.138182 | 0.001465 | 0.296849 | 0.0015 |
| 2 | BRCA1 | 81 | 0.104515 | 0.0395061 | 0.284398 | 0.0013 |
| 3 | … | … | … | … | … | … |

1. **Comparative analysis of the disease genes identification algorithms**
   1. Use the following algorithms to infer and validate (point 2.2) putative disease genes:
      * DIAMOnD (python code available [here](https://github.com/dinaghiassian/DIAMOnD)), default parameters
      * DiaBLE: starting from the DIAMOnD code, change the universe size used in the hypergeometric function (see slides [4.1](https://classroom.google.com/c/NzEzNDg0MDExNzgy/m/NzMwNzAwMzYzMDcw/details))
      * Diffusion-based algorithm, diffusion times (arbitrary unit): t=0.002, 0.005, 0.01
   2. Computational validation
      * Perform a 5-fold cross validation: split the disease genes set *S0* into 5 subsets. Each time, select one subset as probe set *SP* and the remaining four subsets as training set *ST*. Run the algorithm of your choice using the *ST* sets and check the output for genes in the *SP* set. IMPORTANT: in the Diffusion ranking you will find the seed genes (heat input=1) and the putative disease genes (heat input=0) predicted by the algorithm mixed together: clearly you must eliminate the seed genes (heat input=1) before carrying out the cross validation. This is also valid for point 3.1 if the best performing algorithm is Diffusion
      * Compute the following performance metrics:
        + precision (average SD)
        + recall (average SD)
        + F1-score (average SD)
      * Provide the performance measures selecting the top 50 positions and the top X positions where X = ()n, (¼)n, (½)n, n, with n=number of known GDAs (i.e., number of disease’s seed genes)
2. **Putative disease gene identification**
   1. According to the performance metrics obtained in the validation phase at point 2:
      * select the best performing algorithm and apply the process to predict new putative disease genes using all known GDAs as seed genes
      * obtain a list of 100 putative disease genes
   2. Enrichment analysis
      * perform the enrichment analysis (via [EnrichR](https://maayanlab.cloud/Enrichr/): GO-BP, GO-MF, GO-CC, Reactome and KEGG pathways) over the putative disease genes
      * perform the enrichment analysis (via [EnrichR](https://maayanlab.cloud/Enrichr/): GO-BP, GO-MF, GO-CC, Reactome and KEGG pathways) for the original disease genes gathered at point 1.2
      * evaluate the overlap (if any) between enriched functions (i.e. terms associated with *adjusted p-value<0.05*) of original disease genes and putative disease genes
3. **Drug repurposing**
   1. Drug identification
      * Selectthe first 20 putative disease genes in the ranking obtained at point 3.1
      * Use DGIdb ([dgidb.org](https://dgidb.org)) latest “interactions.tsv” file to associate such 20 genes (column “gene\_name”) to approved drugs (column “drug\_name”, column “approved”=True) (access the db file manually or via API at your own convenience).
      * Compile a ranking of identified drugs, starting with the drug associated with the most of the above 20 genes, moving downward.
   2. Clinical Trials validation
      * Take the first three drugs and for each of them taken individually check on <https://clinicaltrials.gov> if there are clinical trials testing the drug (search field: “other terms”) for the disease of interest (search field: “condition or disease”) and note the number of clinical trials, optionally reporting the title and ID (NTC…) of one of them that cites drug name and disease name.
4. ***OPTIONAL TASK***

Run PROCONSUL (temp=1) and compare top 20 genes with top 20 from the best performer algorithm chosen at point 3.1.

**RESULTS AND DISCUSSION**

Describe the obtained results with the help of figures and tables, when needed.

For each algorithm, report the performance metrics and justify your choice.

Provide a discussion of the results (e.g. interpretation, considerations, relationship of your findings with other published findings) and implications that can be drawn from your findings.

Example of possible figures: putative disease module (disease genes + putative disease genes).

**AUTHOR CONTRIBUTIONS**

Provide a short description of the contribution to the project of each author, for example: A.S.: data gathering; A.S., B.S., C.S.: algorithm implementation; A.S., B.S., C.S.: task x; B.S., C.S.: cross-validation; A.S., B.S., C.S.: writing—original draft preparation, A.S., B.S.: writing—review & editing,

**REFERENCES**

Use the following format:

Author A, et al.. Title. Journal name, year; volume number: from page-to page. DOI. PMID.

example:

Einstein A. Time, space, and gravitation. Science, 1920; 51(1305): 8-10. doi: 10.1126/science.51.1305.8. PMID: 17820331.

**FURTHER INSTRUCTIONS (do not include this section in the report)**

Produce a short report (up to 6 pages, please use Arial text size from 9 to 11) following the above guidelines and this template.

Please specify the used programming language, the available functions and, if necessary, those implemented. Do acknowledge any source you used, such as software code, third party figures, cited text and articles, contribution by non-authors, etc.

**Project evaluation criteria**

The mark will be based on the clarity of writing, accuracy of the methods’ description, completeness, and appropriate presentation of results (including quality of figures and tables, that must have self-explanatory captions), appropriate discussion of the outcome of comparisons, overall structure and format of the report.

**Submission procedure**

Each group will submit to [manuela.petti@uniroma1.it](mailto:manuela.petti@uniroma1.it) and [paolo.tieri@uniroma1.it](mailto:paolo.tieri@uniroma1.it) an email stating in the Subject field: *BNM 2024 project group <nn>*, where <nn> is the two-digit group number and with all the co-authors in cc, and attaching two files:

1. a PDF file containing the report;
2. a compressed archive (zip, rar, 7z) containing, coherently divided by typology in subfolders, the software code used to perform the analyses, the full results, and, if present/appropriate/of interest, intermediate results or other relevant material (unedited figures, raw output files, etc). Do not include downloaded data in the archive. Please manage and check the final dimensions of the email, if larger than 10MB use a file sender (FileSender, WeTranfer, etc.).

The attached files must be named according to the following scheme:

BNM\_proj\_group\_<nn>.<ext>,

where <ext> is either ‘pdf’ or {‘zip’|‘rar’|‘7z’}, e.g. BNM\_proj\_group\_01.pdf

| **Delivery date deadline: Mon. 15 January 2025 h 23:59** |
| --- |