|  |  |  |
| --- | --- | --- |
| *Name* | *Surname* | *ID* |

## Midterm test No. 3

### 22 / 12 / 2020

Please answer all questions below and submit this document in **PDF format** by **12:30 of the 12 January 2021** (three weeks after) to **damiano.piovesan@unipd.it**.

Each student is assigned a different **list of UniProt accessions** from the human organism. The entire exercise is based on the analysis of that set of proteins. For each question **concisely explain all passages** **(max 5 rows)** necessary to reproduce the results (e.g. parameters, database queries, algorithms, etc.). Optionally, if relevant, you can provide source code (not necessary).

Student **assignments** are available[**here**](https://docs.google.com/spreadsheets/d/1S2W0qaA0OLrcFzfaiCiLkOMx3aPqEolOKHQqRaV8sVk/edit?usp=sharing).

Lists of **UniProt accessions** are available [**here**](https://drive.google.com/file/d/1WVYhxUWXTx0hb1d84_2iR5WkTfGZvI02/view?usp=sharing). (Columns: assignment ID, comma separated UniProt accessions).

## Questions

1. Paste below your assignment ID.
2. DownloadhumanGO annotations (GAF format) from GOA at EBI ([ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/](http://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/)).
   1. How many annotations are associated with your list of proteins?
   2. How many annotations are associated with the rest of human proteins?
   3. How many different (unique) terms are associated with your proteins?
   4. How many different (unique) terms are associated with the rest of human proteins?
   5. How many leaf terms are associated with your proteins?
3. Which are the most abundant GO terms considering your list of proteins? Provide no more than 10 terms separately for each namespace (sub-ontology).
4. Which are the most abundant GO terms considering your list of proteins after integrating ancestors terms? Provide no more than 10 terms separately for each namespace (sub-ontology).
5. For each term *GOi* build a confusion matrix as defined below and calculate the “fold increase” within your set of proteins compared to the rest of human proteins.

**Hint**: The fold increase can be calculated dividing the ratio *having-/not-having the property* of the *selected* with the ratio *having-/not-having* of the *not selected*.

|  |  |  |
| --- | --- | --- |
|  | **Having the property** | **Not having the property** |
| **Selected** | No. proteins with *GOi* in your set | No. proteins without *GOi* in your set |
| **Not selected** | No. proteins with *GOi* in the rest of human proteins | No. proteins without *GOi* in the rest of human proteins |

* 1. How many terms have a fold increase larger than 1.
  2. Report terms with the largest fold increase, 10 terms for each namespace (sub-ontology).

1. For each confusion matrix generated above, calculate the enrichment with Fisher’s exact test.

**Hint:** You can use <https://pypi.org/project/fisher/>

* 1. Which P-value between the left- and right-tail tells which are the enriched terms in your set of proteins?

**Hint**: Make a comparison with terms with high “fold increase” calculated above if you are not sure.

* 1. Report enriched terms in the selected set and the corresponding P-values (left-/right-/two-tails).