Hiring task 2023

Malay (mbasu@kumc.edu)

Table of contents

1 Task

You need to write a computer program that takes a `GMT` file as an input and replace the gene names (or symbol) in the file with Entrez Ids that effectively creates another `GMT` file with Entrez ID.

Entrez ID (or gene ID) is a unique number that is given to any gene in the NCBI Entrez database. A `GMT` file is a tab-delimited text file (columns are separated by tabs). These files are used in pathway analysis. The file format is described in the data section below.

2 Data

There are two files you need to complete this task. The files may have been compressed using `gzip` software. Preferably, you should not unzip these files at all. Most modern computer languages should be able to read these files as it is.

1. `Homo_sapiens.gene_info.gz` . This is a tab-delimited text file that contains information about all the genes in the human genome. The file contains a lot of information, but we are interested in only the following columns: `2, 3, 5`. These are `GeneID (Entrez Id)`, `Symbol` and `Synonyms`. Create a mapping of `Symbol` to `GeneId`. You have to be particularly careful about the `Synonym` column which may have multiple `Symbol` names separated by `|`. Like this `AACT|ACT|GIG24|GIG25`. You need to extract the

individual gene names from this column add them to the `Symbol` to `GeneId` map. If you are interested in more about this file format check here: https://ftp.ncbi.nih.gov/gene/DATA/README.

2. `h.all.v2023.1.Hs.symbols.gmt`. This is a gene matrix transposed file. This is a tab-delimited text file. This type of file is used for pathway analysis. It is a very simple file format, however, each line can be of different lengths. This file, therefore, should be read line by line, not by columns. If you split each line on tab, it will create a number of values (or fields). The first two values are "`pathway name`" and"`pathway description`". All subsequent values are gene names that belong to the particular pathway. Your goal is to replace the gene names with Entrez ID extracted from the first "gene_info" file.

3 Specification

1. Your program should read the `gene_info` file and create a unique mapping of all gene (symbol) names to their corresponding Entrez ID. You need to be comprehensive to include all symbols both including those in the `Synonyms` and `Symbol` columns.
2. Read `GMT` file and proceed to the next step.
3. Write a file (or print in the terminal) a new `GMT` file where symbols have been replaced with Entrez IDs.

4 Rules

1. You have 48 hours to finish the task.
2. You can use any computer language of your choice. However, you get bonus points for finishing this task in `R`.
3. The program must be a stand-alone script, runnable from the command line. No Jupyter, iPython notebook, or Rmarkdown allowed.
4. Strictly follow the submission instruction. The only way your submission is accepted is through a `GitHub` link. No other way of submission is admissible.

5 How to submit

You should create a new repository on `GitHub`. Put your program only (not any other file) into this repository and send me a direct link to the file (not the whole repository) to my email address mbasu@kumc.edu. Use reply to the email that I sent initiating this task. The email must reach me within 48 hours of the initial email. So it is better to be safe and avoid submitting your answer at the very last moment. Make sure you received a reply from me acknowledging your submission.

All the best!
