BIOI 3500, *Advanced Bioinformatics Programming*
Assignment 9

Assignment 9 :: 50 points :: Due: 11:59 p.m. Friday, April 9th 2021

Objectives

This assignment will give you the opportunity to practice using **psycopg2**, a Python PostgreSQL API.

Description

In order to do this assignment, you will need to create in your VM another table called ***gene2refseq*** in the ***pubmed*** database, along the lines of what we did to create the other two tables. The files needed (data file and SQL schema, respectively) are in Canvas under **FILES/Assignment09 (gene2ref.small, createGene2refseqTable.sql).**

Using the database tables ***geneinfo***, ***gene2pubmed*** and ***gene2refeq,*** write a procedural Python (version 3) program, **netID_Assign09.py** where **netID** is your UNO NetID, that queries the PostgreSQL database and displays:

(1) gene symbol (***geneinfo*** table); (2) taxonomy ID (***geneinfo*** table); (3) gene ID (***geneinfo*** table); (4) RNA accession number(s) (***gene2refseq*** table); (5) protein accession number(s) (***gene2refseq*** table); and (6) pubmed ID(s) for a gene with a given symbol (***gene2pubmed*** table).

Use the following command-line options:

Option	Function
i	Specifies the name of the input file.
o	Specifies the name of the output file. If this option is not used, output should be sent to standard output (i.e., the screen).

Input will consist of a list of gene symbols, one per line. Use file **sample3Gene.txt** (that contains 3 random gene symbols) in Canvas to test your program and compare with the given ".out" file.

The output should contain one line per gene ID. Output should be in tab-separated fields as follows : gene symbol, taxonomy ID, gene ID, RNA accession number(s), protein accession number(s), PubMed ID(s). Separate multiple entries in a given field with pipes (**|**). Indicate empty fields with a dash (**-**). All entries in a field must be unique. Genes in multiple organisms should be displayed in order of ascending taxonomy id. If a field has multiple entries, they should be given in ascending order.

A sample output for three gene symbols is given below. Although lines wrap in the example there are no internal newline characters. For additional gene symbols you can use the **sampleGenes.txt** file on Canvas.

```
ATP6V1B2	9606	526	NM_001693.4|XR_002956632.1|XR_002956633.1	NP_001684.2
	33144569
MIR506	9606	574511	NR_030233.1	-	33291316
YTA7	559292	853186	NM_001181399.1	NP_011786.1	33301732
```