BioE 131/231
Fall 2018

Lab Assignment #4

Overview and references

You will search several remote biological databases for gene sequences, annotations, and relationships between genes. You will then import the data you've found into a database that you will design.

For this lab, you are only required to import data of a particular type (e.g., a gene sequence, annotation, or pathway) into an appropriated-designed table. However, the true power of a relational database and SQL comes from the use of associative tables that link different data types together. Extra credit will be given for the design and use of associative tables that link together the data you've retrieved.

You are expected to keep a thorough record of everything you did in your notebook. Create a folder in your home directory for each lab, and keep all your files there. Try to create a directory hierarchy that makes sense, like the one we went over in Lab 1. Copy and paste any terminal commands you used into a Markdown section and explain what the input was, what the tool did, and what the output was. Plot any results in-line and explain them.

An iPython notebook containing your analysis and your SQLite database is due at the beginning of lab section next week. These can be uploaded to bCourses.

SQLite

https://www.tutorialspoint.com/sqlite/index.htm
https://docs.python.org/3/library/sqlite3.html

Relational Databases

https://en.wikipedia.org/wiki/Relational_database
https://mariadb.com/kb/en/library/relational-databases-basic-terms/
https://en.wikipedia.org/wiki/Associative_entity

BioPython

https://biopython.org/DIST/docs/api/Bio.Entrez-module.html

Biological Databases

https://genome.ucsc.edu/
https://www.ncbi.nlm.nih.gov/search/
https://reactome.org/
https://pfam.xfam.org/
https://www.genome.jp/kegg/

Background

The lab you're working in studies metabolism. Over the years, many students and post-docs have studied enzymes from organisms across the tree of life, from flies, worms, mice, and *E. coli*. Your PI wants you to build a database of these enzymes that future researchers in the lab can use going forward.

Database Design

For this assignment, we'll focus on three pathways: glycolysis, the citric acid cycle, and the pentose phosphate pathway. We will consider genes in these pathways from *Drosophila*, *E. coli*, and humans. At a minimum, your database should contain a table with rows for genes and a table with rows for pathways.

Gene Table. What fields belong in this table? Start with name, description, organism, and nucleotide sequence. Additional fields might include chromosome, start and end position, strand, and translated sequence. For eukaryotes, the nucleotide sequence should be the spliced mRNA and the coordinates should span the entire locus.

Pathway Table. What fields belong in this table? Perhaps just name and description.

Enzyme Table. How about this one? I can think of name, function, and enzyme commission (EC) number. Multiple genes encode enzymes that perform the same function, so there ought to be fewer enzymes than genes.

Associative Tables. I can think of at least three associative tables that make sense here.

1. There are relationships between enzymes and pathways—some enzymes belong to certain pathways. Do any belong to multiple pathways? Is this a one-to-many or many-to-many relationship?
2. There is an order to enzymes within pathways. How can the order be represented in a table?
3. Genes in the gene table encode enzymes in the enzyme table. How can this be represented? Is this a one-to-one, one-to-many, or many-to-many relationship, and in which direction?

Creating your database and tables

Using the example from the PowerPoint as a guide, sketch out your database schema in a Markdown section of your iPython notebook. Copy this example command once per each table you're designing, and modify the fields to suit your needs:

```
CREATE TABLE genes (id INT PRIMARY KEY ASC, name TEXT, description TEXT,
chromosome TEXT, start INT, end INT, strand VARCHAR(1));
```

Once your tables are designed, try creating your database by following the example in the PowerPoint. This is a good opportunity to locate any bugs in your schema. You might have to delete and re-create your database file each time you run the section of your code that creates the database.

Searching databases for data

A good place to start will be this map of glycolysis from KEGG: https://www.genome.jp/kegg-bin/show_pathway?map00010

There are a lot of genes in these pathways and how thoroughly you curate your database is up to you, but for the sake of time, pick 4 enzymes from glycolysis, TCA cycle, and pentose phosphate. Gather information and sequences from *Drosophila*, *E. coli*, and human for those enzymes. Try UCSC, Entrez, KEGG, and Reactome. Notice how each database presents the information in a different way. Do you have any preference? Which seems to be the most complete?

If you're clever, you can automate this entire process using Bio.Entrez, which can be used to pull entire annotations for each enzyme from each organism—and more. Try asking it to return data as GenBank and using Bio.SeqIO to read the returned GenBank file.

Any data that you've pulled from a web browser should be noted in your iPython notebook.

INSERTing your data into your database

At a minimum, you'll be creating 4 enzymes * 3 pathways * 3 organisms = 36 rows in a your gene table. If you've managed to organize the data you've gathered into a table, or pulled it directly into Python using Bio.Entrez, a `for` loop over a list can be used to insert your data into your database. Otherwise, you can create one line of Python for each entry. Your INSERT commands will look something like this:

```
INSERT INTO genes (name, description, chromosome, start, end, strand) VALUES
('BRCA1', 'Breast Cancer 1', 'chr17', 43044295, 43170245, '-');
```

Happy hunting!