# Homework 3

You have been tasked with identifying examples where there genes in query organism which differ the most between two strain.

We will use as an example the Bacteria **Serratia marcescens**. There are 3000+ genomes in genbank for this species https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=615 You will download the coding sequence file (DNA) of the genes - The reference genome we will compare to is ELP1.10. The CDS file is here.

- The query genome is one strain I chose randomly: CH3. You will download its CDS file for comparison.

I've already started a version of a script for you to download called `00_download_and_blast.sh` in the template.

1. Update this bash script to index the database and run BLAST. It should produce a file named `serratia_search.blastn.tsv`. Make sure your `git add serratia_search.blastn.tsv` and git commit the `serratia_search.blastn.tsv` file

## Update the scripts created for you

### script1.py

Count the number of sequences in the file - Function: `calculate_seqcount(filename)` - Returns: Number of sequences in the query fasta file

### script2.py

Implements a function to parse BLAST table and return count of query sequences which have matches. - Function: `calculate_num_matched_sequences(filename,min_identity)` - Returns: number of sequences that have a match with at least min_identity threshold

The program will also take an argument `--ratio` which will also use the total number of sequences in the query file to calculate the fraction of query sequences which have a match.

### script3.py

Implements the function to read the BLAST table result and find the number of query sequences where the BEST match has a percent identity lower than the input threshold (default is 80%) - Function: `find_lowID_best_hits(blastfile, threshold)` - Returns: Names of sequences and their percent identity to best match, but only those where threshold is met

## How to Run

Run each script individually:

```
python3 script1.py
python3 script2.py
python3 script3.py
```

## Expected Output

- `script1.py`: Should print the number of sequences in the query file
- `script2.py`: Should print number of sequences with matches in blast result file
- `script3.py`: Should print number of sequences which are BEST match but have less than 80% identity to the query sequence

## For Students

Complete the TODO sections in each script to implement the required functionality. The scripts are pre-filled with working implementations as examples, but you should understand how they work.

Following using these scripts with their current names will allow the results to be tested with google classroom automatic grader.