Project 3

Akshay Kumar Varanasi (av32826)

Instructions

After completing this Jupyter notebook, please convert it to pdf and submit both the pdf and the original notebook on Canvas **no later than 4:00 pm on May 9, 2019**. The two documents will be graded jointly, so they must be consistent (as in, don't change the Jupyter notebook without also updating the converted pdf!).

All results presented **must** have corresponding code. Any answers/results given without the corresponding python code that generated the result will be considered absent. All code reported in your final project document should work properly.

Before submitting the Jupyter notebook part, please re-run all cells by clicking "Kernel" and selecting "Restart & Run All."

The project consists of two problems. For both problems, please follow these guidelines:

- Final output needs to be nicely formatted and human readable. For example, if your result is a count, don't just print the value of the count, print "The count is: ...".
- For each problem, limit your total code to less than 100 lines.
- Write comments and explanatory text, so we understand what you are doing.
- Do not print out large datasets, such as an entire genome, or a list of all genes in a genome, etc.
- Verify that nothing of importance (code, comments, other text) is cut off in your final pdf.

Problem 1

The bacteria called *Salmonella enterica* Typhimurium are pathogenic bacteria closely related to *E. coli*. They cause typhoid fever in humans. There are many different *S. enterica* Typhimurium strains, and here we will compare two such strains, LT2 and CT18. LT2 is the canonical strain that is most commonly used as a reference. CT18 is another widely used reference.

Before we can work with these two genomes, we need to download them. Note: Running the next cell may take a few minutes.

In [1]:

```
from Bio import Entrez
Entrez.email = "akshayvaranasi@utexas.edu" # put your email here

# Download S. enterica strain LT2 and write into file "S_enterica_LT2.gb":
download_handle = Entrez.efetch(db="nucleotide", id="NC_003197", rettype="gbwit
hparts", retmode="text")
out_handle = open("S_enterica_LT2.gb", "w")
out_handle.write(download_handle.read())
download_handle.close()
out_handle.close()
print("Downloaded S. enterica LT2")

# Download S. enterica strain CT18 and write into file "S_enterica_CT18.gb":
download_handle = Entrez.efetch(db="nucleotide", id="NC_003198", rettype="gbwit
hparts", retmode="text")
out_handle = open("S_enterica_CT18.gb", "w")
out_handle.write(download_handle.read())
download_handle.close()
out_handle.close()
print("Downloaded S. enterica CT18")
```

```
Downloaded S. enterica LT2
Downloaded S. enterica CT18
```

Problem 1a (30 pts): How many named protein-coding genes are in *S. enterica* LT2? And how many of these genes have synonyms in *S. enterica* CT18?

Hint: Gene names have been defined for the LT2 strain. You can find these names in the "gene" qualifier of CDS features. When equivalent genes exist in CT18, they are listed under the "gene_synonym" qualifer of the CDS features. As an example, manually open the two genome files and look for the "thrL" gene in each genome.

Answer: After reading the two files using SeqIO, we loop over features in each record. In the loop, we look at feature type whether it is CDS or not. If it is CDS, we look for gene in qualifiers and store the names along with the count of them in a dictionary. Since total number of such named genes was asked in the first part, we need to sum the counts of all the different named genes stored in the dictionary. For later part, we look at both the dictionaries and see if there are names (keys) which are in both the dictionaries i.e we are checking if the gene has a synonym or not. After finding such names (keys), we add the counts of those names (keys) in the dictionary of named genes in LT2 as we need the count of how many of them have synonyms.

In [2]:

```
from Bio import SeqIO

# read in the LT2 genome
in_handle = open("S_enterica_LT2.gb", "r")
record_LT2 = SeqIO.read(in_handle, "genbank")
in_handle.close()

# read in the CT18 genome
in_handle = open("S_enterica_CT18.gb", "r")
record_CT18 = SeqIO.read(in_handle, "genbank")
in_handle.close()

# Dictionary to keep the count of gene names in LT2
prot_names={}


for feature in record_LT2.features:                    # Loop over the features
    if feature.type =='CDS':                           # Check the feature type
is CDS or not
        if 'gene' in feature.qualifiers:               # Check if gene is in qu
alifiers
            name = feature.qualifiers['gene'][0]       # name of the genes
            if name in prot_names:                     # check if the name is a
lready in the dictionary
                prot_names[name]=prot_names[name]+1    # if it is then increase
the count
            else:
                prot_names[name]=1                     # else initialize it wit
h 1


print("There are",sum(prot_names.values()),"named protein-coding genes in S. en
terica LT2")


# Dictionary to keep the count of gene names in CT18
prot_syn_names={}


for feature in record_CT18.features:                           # Loop over the
features
    if feature.type =='CDS':                                   # Check the feat
ure type is CDS or not
        if 'gene_synonym' in feature.qualifiers:               # Check if gene
is in qualifiers
            name = feature.qualifiers['gene_synonym'][0]       # name of the ge
nes
            if name in prot_syn_names:                         # check if the n
ame is already in the dictionary
                prot_syn_names[name]=prot_syn_names[name]+1    # if it is then
increase the count
            else:
                prot_syn_names[name]=1                         # else initializ
e it with 1

count1 = 0
for i in prot_names.keys():
    if i in prot_syn_names.keys(): # checking ig the same name is in other dict
ionary
        count1+=prot_names[i]      # if it is then add the count of that name i
n the original dictionary

print("No of named genes which have synonyms in S. enterica CT18 are",count1)
```

```
There are 3242 named protein-coding genes in S. enterica LT2
No.of named genes which have synonyms in S. enterica CT18 are 1515
```

No.of named protein-coding genes in S. enterica LT2 are 3242 and number of genes having synonyms in S. enterica CT18 is 1515.

Problem 1b (20 pts): How many of the named genes in LT2 without a synonym in CT18 have their product listed as "hypothetical protein"?

Answer: Way we go about is we loop over features in LT2 and see whether the feature type is CDS. If it is, we look at gene in qualifiers within the feature as they contain the names of the genes and then we check if it has synonym in CT18 by checking the names (keys) in the dictionary which we made earlier. If the gene name does not have synonyms, then we see if the product in the qualifier is "Hypothetical protein" or not. If it is "Hypothetical protein" we increase the count by 1. In the end, we want the value of the count as it is the number of named genes in LT2 without a synonym in CT18 which have their product listed as hypothetical protein.

In [3]:

```
# initialize the count
count=0


for feature in record_LT2.features:                            # loop over the
features
    if feature.type =='CDS':                                   # check if feat
ure type is CDS
        if 'gene' in feature.qualifiers:                       # check if gene
field is there
            name = feature.qualifiers['gene'][0]               #  name of the
gene
            if name not in prot_syn_names:                     # check if it h
as synonym
                if 'product' in feature.qualifiers:            # check for pro
duct field
                    if feature.qualifiers['product'][0]=='hypothetical protein'
: # check if it is hypothetical protein
                        count+=1
# increase the count


print("No.of named genes in LT2 without synonym in CT18 having their product li
sted as hypothetical proteins are",count)
```

```
No.of named genes in LT2 without synonym in CT18 having their product listed as
hypothetical proteins are 244
```

As we can see there are 244 named genes which have product listed as hypothetical proteins

Problem 2

(50 pts)

Ask a question about the genomes from Problem 1 and then write python code that generates an answer. The question does not have to be conceptual, and it can be about only one of the two genomes or about the two genomes jointly.

For full credit, the answer code must meet the following conditions:

- contains at least one `for` loop
- contains at least one `if` statement
- uses at least one list or dictionary
- uses at least one regular expression

Question: . There are different types of proteins listed in "product" field like binding proteins and there are various types of binding proteins. We are interested in DNA,RNA and ATP related binding proteins as they bind them.

How many DNA related binding proteins are in *S. enterica* LT2 and *S. enterica* CT18? Similary find out how many RNA and ATP related binding proteins are in both?

Answer: Since we want binding proteins related to DNA,RNA and ATP, we first search for "binding protein" in "Product" field with feature type "CDS". If we find it, then we search for "DNA" or "RNA" or "ATP" terms in the "Product" field and increase the count of each term stored in the dictionary if we find them. This will give the count of binding proteins related to each DNA, RNA and ATP.

In [4]:

```
# importing the re library
import re

# Dictionary containing counts of each term
protein_count={'DNA':0,'RNA':0,'ATP':0}


for feature in record_LT2.features:
    if feature.type=="CDS":
        if "product" in feature.qualifiers: # see if "product" is there or not.
            product = feature.qualifiers["product"][0]
            match = re.search(r"binding protein", product) # search for binding
protein
            if match:
                match2 = re.search(r"((DNA|RNA)|ATP)", product) # search for te
rms DNA or RNA or ATP
                if match2:
                    protein_count[match2.group()]+=1 #increase the count of the
respective term
for i in protein_count:
    print("No.of binding proteins related to",i,"in LT2 is",protein_count[i])
```

```
No.of binding proteins related to DNA in LT2 is 10
No.of binding proteins related to RNA in LT2 is 3
No.of binding proteins related to ATP in LT2 is 64
```

In [5]:

```
# Dictionary containing counts of each term
protein_count={'DNA':0,'RNA':0,'ATP':0}

for feature in record_CT18.features:
    if feature.type=="CDS":
        if "product" in feature.qualifiers: # see if "product" is there or not.
            product = feature.qualifiers["product"][0]
            match = re.search(r"binding protein", product) # search for binding
protein
            if match:
                match2 = re.search(r"((DNA|RNA)|ATP)", product) # search for te
rms DNA or RNA or ATP

                if match2:
                    protein_count[match2.group()]+=1 #increase the count of the
respective term
for i in protein_count:
    print("No.of binding proteins related to",i,"in CT18 is",protein_count[i])
```

```
No.of binding proteins related to DNA in CT18 is 23
No.of binding proteins related to RNA in CT18 is 3
No.of binding proteins related to ATP in CT18 is 77
```

From the result, we find that there are 10 "DNA", 3 "RNA" and 64 "ATP" related binding proteins in *S. enterica* LT2 and there are 23 "DNA", 3 "RNA" and 77 "ATP" related binding proteins in *S. enterica* CT18. So we conclude that, there are more such binding proteins in CT18 than LT2 genome.