CIS 554 Project 2

Wenyuan Yu

*Answer each of the following questions about the E. coli genome given the information in the Genbank report.*

The following python statements should run first.

***>>> from Bio import SeqIO***

**>>> recs = list(SeqIO.parse("NC\_000913.gbk", "genbank"))**

**>>> ecoli = recs[0]**

*1. What is the total size in bp (base pairs) of the chromosome? How did you determine this number?*

**>>> len(ecoli.seq)**

**4641652**

There are 4641652 bp of the chromosome.

*2. How many CDS features are there?*

**>>> CDS\_index = 0**

**>>> for index, record in enumerate(ecoli.features):**

**if (record.type == "CDS"):**

**CDS\_index += 1**

**>>> print CDS\_index**

**4321**

There are total 4321 CDS features.

*3. What are the coordinates (starting and ending location) of the gene named ileV?*

**for record in ecoli.features:**

**if (record.type == "tRNA"):**

**if (record.qualifiers['gene'][0] == "ileV"):**

**print(record.qualifiers['gene'], record.location)**

**(['ileV'], FeatureLocation(ExactPosition(225380), ExactPosition(225457), strand=1))**

Thus, Gene *ilev* Starting from 225381, ending at 225457.

*4. ileV is a gene for a tRNA sequence -- note that instead of a CDS feature there is a tRNA feature. How many tRNA genes are there in all?*

**>>> tRNA\_index =0**

**>>> for index, record in enumerate(ecoli.features):**

**if (record.type == "tRNA"): tRNA\_index += 1**

**>>> print tRNA\_index**

**89**

There are total 89 tRNA genes.

*5. The sum of the number of CDS and tRNA features is less than the total number of genes. What other kinds of features can you find? Do they all add up to the number of things labeled as “genes”?*

5.1. First we need found out how many “genes” are there in the gene bank.

**>>> gene\_index =0**

**>>> for index, record in enumerate(ecoli.features):**

**if (record.type == "gene"): gene\_index += 1**

**>>> print gene\_index**

**4497**

5.2. gene\_index (4497) > CDS\_index (4321) + tRNA\_index (89). So there must be other features in the genes. We can run the following statement to find distinct features.

**>>> gene\_bank = set ()**

**>>> for index, record in enumerate(ecoli.features):**

**if (record.type not in output):**

**gene\_bank.add(record.type)**

**>>> print gene\_bank**

**set(['misc\_feature', 'mobile\_element', 'repeat\_region', 'rep\_origin', 'tRNA', 'tmRNA', 'source', 'rRNA', 'CDS', 'gene', 'ncRNA'])**

5.3. We can use the same method of calculating the number of CDS features and genes to calculate the total number of rest of genes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Genes types | CDS | *tRNA* | *rRNA* | *ncRNA* | *tmRNA* |
| numbers | 4321 | 89 | 22 | 63 | 2 |

The number of total genes of “CDS”, “rRNA”, “tRNA”, “ncRNA” and “tmRNA” is 4497, which is the same as the number of “genes” (4497).

In project one, I was not be able to find the “tmRNA”. With the help of program, it is much easier to get the information.

*6. Does the E. coli genome contain any pseudogenes?*

If a gene is pseudogene, under the “note” of ecoli.features[i].qualifier, there is “pseudogene” in the note. We can write a python program to find if “pseudogenes” is contained in the string. From the output of the program, we can confirm that the *E. coli* genome contains pseudogenes.

**>>> for record in (ecoli.features):**

**if ("note" in record.qualifiers):**

**if ("pseudogene" in record.qualifiers["note"][0]):**

**print record.qualifiers["note"][0]**

Part of output is shown below.

**pseudogene**

**pseudogene**

**pseudogene, H repeat-associated protein**

**pseudogene, flagellar system protein, promoterless fragment; flagellar biosynthesis**

**pseudogene, lateral flagellar motor protein fragment**

**pseudogene**

**pseudogene, predicted peptide chain release factor homolog; probable peptide chain release factor**

**pseudogene, YdiA family**

**pseudogene, CP4-6 putative prophage remnant;Phage or Prophage Related**

**pseudogene, IS911 transposase A;IS, phage, Tn; Transposon-related functions; extrachromosomal; transposon related**

**pseudogene, CP4-6 prophage; predicted ferric transporter subunit;Phage or Prophage Related**

**pseudogene, predicted IS3 family transposase**

**pseudogene, CP4-6 prophage;Phage or Prophage Related**

**pseudogene, LysR family, fragment;putative regulator; Not classified; putative transcriptional regulator LYSR-type**