Project 2: Annotation of Klebsiella pneumoniae

Xiaojun Gao

Note: For all programs to run normally, please add "./" before executable name.

Part 1: Genome Annotation

Part a:

First, copy the original fasta file from data base and rename to Kleb.fa:

```
cp /opt/ccb/data/Klebsiella_genome_data/Kleb_largest_contig.fa Kleb.fa
```

Next, run long-orfs with the designated restriction commands.

Output to file "long-orf-result":

```
long-orfs -n -t 1.15 Kleb.fa long-orf-result
```

Then, run extract and omit stop codons and redirect output to file "Kleb.train":

```
extract -t Kleb.fa long-orf-result > Kleb.train
```

Part b:

First, run build-icm and output to file "Kleb.icm":

```
build-icm -r Kleb.icm < Kleb.train
```

Next, run glimmer 2 with designated restriction commands and output to file beginning in "tag" (i.e. tag.predict, tag.detail)

```
glimmer3 -o 50 -g 110 -t 30 largest_config.fasta Kleb.icm tag
```

Part c:

Use the program 1c.cpp to convert tag.predict into intermediate.gtf.

Run the program as the following:

```
g++ 1c.cpp -o myProgram
./myProgram tag.predict intermediate.gtf
```

Next, run gffread to create protein sequences and output to protein.fasta:

```
gffread -g Kleb.fa -y protein.fasta intermediate.gtf
```

Part d:

First, copy the reference file from data base and name it accordingly:

```
cp /opt/ccb/data/Klebsiella_genome_data/GenBank_reference_genome/
GCF_002752995.1_ASM275299v1_protein.faa
GCF_002752995.1_ASM275299v1_protein.faa
```

Next, run mummer on the file with designated restriction command.

Output to file "mummer_output":

```
mummer -l 10 GCF_002752995.1_ASM275299v1_protein.faa protein.fasta >
mummer_output
```

Finally, to generate the final gtf file, we would use the program 1d.cpp. Note that this program requires GCF_002752995.1_ASM275299v1_protein.faa as the exact name and within the folder of the script.

Run the script as following, output to file final_CA.gtf.

```
g++ -o myProgram 1d.cpp
./myProgram mummer_output final_CA.gtf
```

Disclaimer: for this part of the project, I am doing as what the CA asked for in Piazza instead of the answer from the TA. The script result will produce as the requirement of the CA. However, I have also attached the gtf file according to the TA's requirement as final.gtf.

The associated code submitted are: 1c.cpp, 1d.cpp.

The 2 gtf files to view is: intermediate.gtf, final_CA.gtf (or final.gtf, whichever satisfies the requirement)

Part2: Composition of Genes

Run the program with the following commands, where orfs.in can be replaced by any inputfile name and out.txt can be replaced by any output file name.

```
g++ -o myProgram 2.cpp
./myProgram < orfs.in > out.txt
```