TEAM 2 FUNCTIONAL ANNOTATION HOMEWORK
DUE: April 10th, 2018

General Functional Annotation (16 pts.)

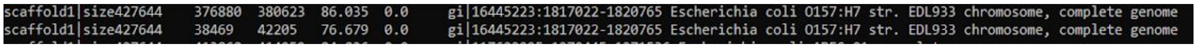
1. What is the difference between functional annotation and gene prediction? (2pts)
2. What should you consider when choosing functional annotation tools? (3pts)
3. As you already know, we are dealing with a large number of genomes. What steps can we take to make our approach scalable? Name one algorithm that can help us achieve this with protein sequences. What is the underlying principle of the algorithm? (3 pts)
4. Give a simple explanation of Hidden Markov Model(HMM).(5pts)
5. Dr. Boswell's Lab is working on the organism *Borrelia americana*. The organism is usually susceptible to most drugs and is known to undergo transformation readily. The lab has isolated a novel strain of the organism, *Borrelia americana* XYZ123 which has been categorized to be a "superbug". The lab sequenced the organism's genomic DNA. You were assigned the job to functionally annotate the genes, however, upon analysis, you did not come across any relevant gene that could cause multiple drug resistance. How could you explain this phenomenon? (3 pts)

Prokka (8 pts.)

1. What is the mandatory parameter involving the file type as input while running PROKKA?(1 pt)
2. What are the 2 stages of annotation while running PROKKA? Give a brief explanation. (3 pts)
3. What is the difference between CDS and ORF? How can you get noncoding RNAs in PROKKA? (Just give the parameters you will use while running the code)(2pts)
4. What is the default parameter for the command line options for the following flags in PROKKA:- (2pts)
   --kingdom:
   --gcode:
   --fast:
   --rfam:
   --cpus:
   --norrna:
   --notrna:

Specific Tools (26 pts.)

1. Questions on PHASTER.
    a. What does PHASTER annotate in bacterial genome and why may those genes be important? (2pt)
    b. Does it use homology or ab-initio method to annotate? (1pt)
    c. How to use PHASTER on command line (Please provide one example of uploading sequence and retrieving result from PHASTER) (2pt)?
2. Describe the criteria for a “perfect match” in the RGI toolkit from CARD. Does RGI use homology based or ab-initio methods for these matches? (2 pts)
3. Describe the criteria for a “strict match” in the RGI toolkit from CARD. Does RGI use homology based or ab-initio methods for these matches? (2 pts)
4. Describe the criteria for a “loose match” in the RGI toolkit from CARD. Does RGI use homology based or ab-initio methods for these matches? (2 pts)
5. Using the command blastn -query $query -db $database -outfmt '6 qseqid qstart qend pident evalue stitle' -out out.txt we get the following output:



```
scaffold1|size427644    376880  380623  86.035  0.0     gi|16445223:1817022-1820765 Escherichia coli O157:H7 str. EDL933 chromosome, complete genome
scaffold1|size427644    38469   42205   76.679  0.0     gi|16445223:1817022-1820765 Escherichia coli O157:H7 str. EDL933 chromosome, complete genome
```

    Help us to troubleshoot this problem to get rid of duplicate hits by modifying the above code. (1 pt)
6. CRISPR is known for its key role in the bacterial defense system. What are they exactly? Briefly explain why they need to be annotated in this study? (4 pts)
7. How to increase the specificity of PILER-CR by setting its main parameters? (1 pt)
8. eggNOG delivers annotations via two primary pipelines.
    a. Give the two and what are the primary differences between them? (3 pts).
    b. Why is DIAMOND (a BLAST alternative) able to perform alignments in such an efficient manner? (2 pts)
9. There are **3** operons in the following table, remember that operon is a cluster of genes that get transcribed together. Assume operon ID is assigned consecutively. Please help us to fill in the missing Operon IDs (4pts).

OPERON ID	GI #	Product
1143278	402778467	mannitol operon repressor
	402778468	mannitol-1-phosphate 5-dehydrogenase
	402778469	PTS system mannitol-specific transporter subunit IIC
	402778471	Signal recognition particle receptor protein FtsY

	402778472	cell division transporter ATP-binding protein FtsE
	402778473	cell division protein FtsX
	402778478	ribose ABC transporter permease RbsC
	402778479	D-xylose transport ATP-binding protein XylG
	402778480	Xylose ABC transporter, periplasmic xylose-binding protein XylF

Extra Credit

What name did Jacob use to introduce himself during the first presentation?