Team 1 Gene Prediction Homework Assignment [updated 3/9/18]

All tools necessary for this assignment are already installed on the server.

Please email this assignment as LastName_FirstName_GP.pdf or .docx to clsmith@gatech.edu by midnight 3/15/18 along with any additional files related to this assignment.

Problem 1. Gene Prediction (15 points)

a) Briefly describe what gene prediction is and why it is important. (5 points)
b) How is gene prediction different in prokaryotes vs eukaryotes and why are open reading frames (ORFs) important? (5 points)
c) In the scope of our project, is it necessary to look at regions of non-coding RNA? (5 points)

Problem 2: Using homology-based tools (20 points)
Use the genome assembly SRR3982229 (located at */projects/data/team1_genomeAssembly/reference_based_assembly/assembly_50/SRR3982229* on the server). Use the reference genome NZ_CP007727.1 (located at */projects/data/projects/data/team1_GenePrediction/validation/reference_genes/NZ_CP007727.1.fa* in the server).

a) What is the genus and species of the assembly? (5 points)
b) Run BLAST via the command line. What command(s) did you use and why? Also explain any flags used. (10 points)
   Note: you will first have to make a blast database.
c) Briefly interpret the results, focus on parameters you considered useful. (5 points)

Bonus (5 points): According to the report, do you consider this particular blast result good or bad? Explain your reasoning.

Problem 3: Using ab initio tools. (20 points)

a) Run Prodigal on the same genome as in Problem 2. Make sure you get three output files: a gene coordinates file (main output), a nucleotide sequence file, and a protein translations file. Save the **nucleotide sequence** file as 'nucleotide_sequence.fa' and attach it to the email. (5 points)
b) List the command used. (5 points)

c) Open the nucleotide sequences file and look at the description line for a gene. Paste the line for the first gene below. Can you explain the information on that line? (5 points)

d) Compare the nucleotides from the 1st gene with the nucleotides in the original assembly sequence. Why **don't** they match? Now compare the ones from the 3rd gene and the assembly. Why **do** those match?
Hint: one of the descriptors in part c) explains why. Prodigal did something additional to the first gene sequence it didn't do to the third. (5 points)

Problem 4. Non-coding RNA. (20 points)

a) From the main approaches used for gene prediction, which is used for discriminating between coding and noncoding regions? Why would we be interested in noncoding regions? (10 points)
b) What are the main problems one faces when choosing a tool for searching for noncoding RNA? How would you propose to level out these inconveniences? (10 points)

Problem 5. Validate your results. (25 points)

a) Sensitivity and specificity are critical when considering gene prediction tools. Define these terms. (5 points)
b) How would you obtain the factors that needed to calculate sensitivity and specificity for the ab initio tool you used (Hint: maybe with the help of blast?) (5 points)
c) Perform blast and write down the command you used, calculate and report the sensitivity and PPV of the ab initio tool you used. (15 points)
Commands that may be useful here:
    1. makeblastdb -in reference.fa -dbtype nucl
    2. blastn -db reference.fa -query predicted.fa -out out_file_name -outfmt 6 -max_target_seqs 1