Gene Prediction
Team-II
Homework Assignment [50 pts]

INSTRUCTIONS

- Please do not run anything related to this homework assignment on the server. None of the questions require you to run any commands expect I.2.c (Ab Initio section).
- This homework is due Tuesday, March 27th at 1:25PM.

I GENERAL GENE PREDICTION

1. Please describe the different file formats in which you can output the results of gene prediction (example: GFF2, GFF3, etc). What information can you extract from each file type? **[5 pts]**

2. Explain what Sensitivity and Precision are. How do they help us understand the difference between the tools we tested? **[5 pts]**

3. Why is gene prediction easier in prokaryotes as compared to eukaryotes ? **[2 pts]**

II COMPARATIVE METHODS

1. What are the advantages and disadvantages of the comparative methods for gene prediction ? **[5 pts]**
2. What are the steps to find genes using BLAST+ command line tool ? Explain the following :
    a. BLAST+ commands to do so and what they mean **[4 pts]**
    b. BLAST+ output on querying a genome against the database. Mention the type of output format with the description of what each column header means. **[4 pts]**

III AB-INITIO

1. Richard Dawkins holds the belief that if there's life elsewhere in the universe, it will possess a self-replicating entity, at its fundamental level - equivalent to our

gene but not necessarily as we know it. This entity will be capable of influencing its own probability of replicating, with slight errors.
Assume that many, many years from now, Elon Musk's car is returning back to earth. Scientists screen the car and find that some unknown alien species (microscopic) had hitch-hiked to earth and were found to be alive. They sequence all the organisms and find the following:

1. Genomic content was composed of 3 types of bases - X,Y and Z (not necessarily nucleotides)
2. X-Y and X^*-Z formed complementary pairs, where X^* is a tautomeric form of X
3. "Codons" are 3-bases long
4. A functional region which, putatively, is involved in metabolic processes during extreme conditions.
5. The alignment of the above region in 10 organisms looks like this -

Sequence 1:	XYZYZYZXZ
Sequence 2:	XZYYZYZXZ
Sequence 3:	YZZYZYZXZ
Sequence 4:	XYZYZXZXY
Sequence 5:	XZZYZYYXZ
Sequence 6:	XYXYZYZZZ
Sequence 7:	XYXYZYYXZ
Sequence 8:	XYZYZYZXZ
Sequence 9:	XXZYZYZXX
Sequence 10:	XYZYZXZXZ

Corresponding PWM (relative frequencies)-

Pos	1	2	3	4	5	6	7	8	9
X	0.9	0.1		0	0		0	0.9	0.1
Y	0.1		0.1	1.0		0.8	0.2	0	0.1
Z	0	0.3	0.7	0	1.0	0	0.8	0.1	0.8

Using the above information, answer the following -

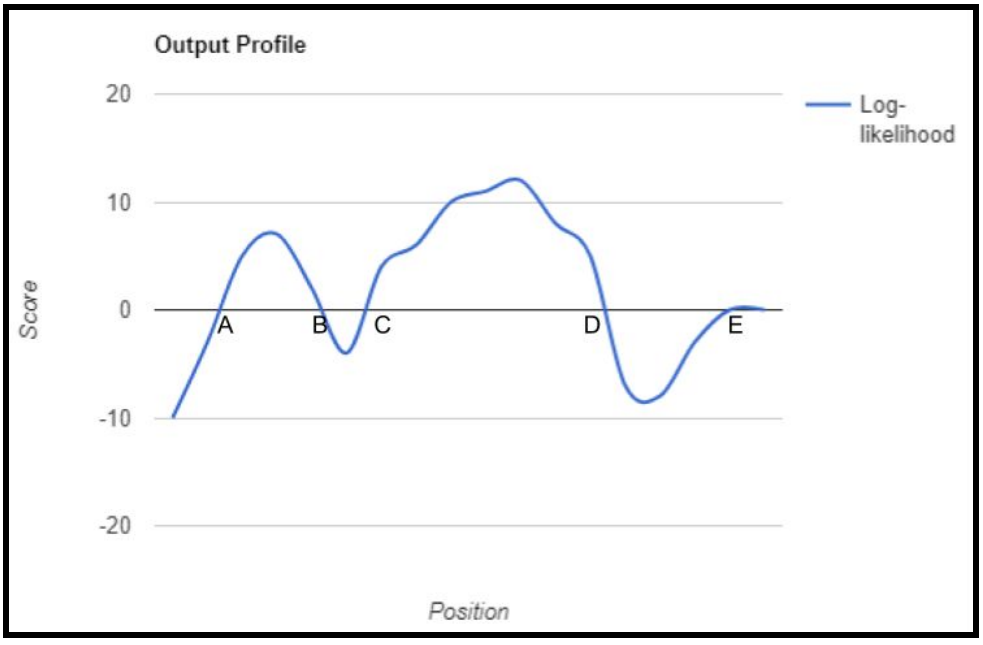
i) What is the probability of a sequence XXZYZYZXZ being in a random site? **[2 pts]**

ii) What is the probability that the above sequence is in a functional site? [HINT: You'll need to complete the matrix and use relative frequencies.] **[4 pts]**

iii) What is the log-likelihood of the gene being in a functional region? **[1 pt]**

iv) Scientists say that these organisms are likely to have a higher error rate during replication. Do you agree? Give reason. **[1 pt]**

v) Scientists have developed a software to predict functional sites for this species, based on log-likelihood. Higher the value of log-likelihood, higher the chances of the region being a functional site. The following is an output profile from the software. Find the functional site(s).
(Give your answer as "Start_Positon -- End_Position". Eg A -- D) **[2 pts]**



[Note: All of the above is a hypothetical situation. Scientists believe that an odd number of bases is impossible for complementary pairing.]

2. Answer the following -
    a. What are the input and output files for GeneMarkS, Glimmer and Prodigal? **[1 pt]**
    b. How would you find consensus among the results from each of the above mentioned tools' output files? **[1 pt]**
    c. Help us to convert the below .predict file (output from GLIMMER) to standard GFF3 format (ideally write a script). You can post a screenshot of the converted GFF3. [Note - For this question you will need to install GLIMMER on your local system.] **[3 pts]**

ORF	START	STOP	FRAME	SCORE
orf00001	822	382	-1	12.07
orf00002	1380	922	-1	12.88
orf00005	1490	2524	+2	13.53
orf00006	3976	2528	-2	15.65
orf00007	5472	3973	-1	14.81
orf00008	5690	7558	+2	16.19
orf00010	7744	8163	+1	10.36
orf00011	8174	9679	+2	16.00
orf00013	9685	10650	+1	14.90

IV NON-CODING RNA

1. Please describe the different types of ncRNA. How are they different? Why do we need to predict ncRNA genes? Why does each type of ncRNA require a different tool? What is the potential role of ncRNA in terms of antibiotic resistance

    [5 pts]

V OUR GROUP

1. Please describe our proposed pipeline. What tools did we decide to use and why? What is the input? What is the expected output? **[5 pts]**

Have Issues? Need Help? Stuck somewhere? Dozed off during presentations?
Contact us -
Pro Tip: Use the following as subject -
"Gene Prediction Team 2 HW Related"

Points of Contact -

General Gene Prediction -	Any of the group members
Comparative Methods -	Parisa (pyz@gatech.edu) Sini (sini.nagpal@gatech.edu) Qi (zhanqi1010@outlook.com)
Ab Initio -	Sarthak (sarthaksharma@gatech.edu) Ayush (asemwal3@gatech.edu) Rong (rjin38@gatech.edu)
Non-Coding RNA -	Beatriz (beatrizsaldana@gatech.edu) Michael (sng32@gatech.edu) Jiani (jlong96@gatech.edu)
Our Group -	Any of the group members