CSE4065 – Spring 2020
Computational Genomics
DNA Replication & Assignment I

Betül Boz

betul.demiroz@marmara.edu.tr

Announcement

Dear students;

The whole world is going through a battle with COVID-19 virus.

Please stay at home and keep safe.

I have previously told you about your first assignment, which will still be used as your next assignment. Your first assignment is updated and will be about COVID-19.

In this assignment, you are asked to find the most frequent k-mers in two COVID-19 genome sequences taken from two patients in Nepal and Wuhan.

You already have all the slides needed for this assignment on Canvas - Lecture 3: DNA Replication, but I also added them to this presentation.

Assignment I

What to do?

- First get the whole genome sequences from:
  - Nepal: https://www.ncbi.nlm.nih.gov/nuccore/MT072688
  - WUHAN: https://www.ncbi.nlm.nih.gov/nuccore/MN988668
- Write an algorithm that finds the most frequent k-mers in two sequences.
  - k lenght is unknown so start with 5 and continue incrementing k until you no longer can find frequent k-mers (strings of lenght k appearing more than once)
  - Write a report and give the frequent k-mers for varying values of k.
- Do not forget to also check for reverse complements

- Also compare the results you obtained from two patients, are the frequent k-mers the same or different?

What to submit?

- Your source code written in any programming language.
- A report showing your results and a comparison between two patients to see whether there are mutations in the virus or not.

- It might be hard to work as a group but if you can, you can work as groups of 2-3 people.

- If you have any questions, please send an e-mail to me.

- Due to: 13.4.2020