IITM-CS6024 : Algorithmic Approaches to Computational Biology Date: Oct 23

Assignment 5 Due Date : Oct 27, 23:59

Roll No: GATAGAAA Name: ATGCATGCATGC

- Use LaTeX to write-up your solutions, and submit the resulting single pdf file at GradeScope by the due date (Note: As before, **no late submissions** will be allowed, other than one-day late submission with 10% penalty! Within GradeScope, indicate the page number where your solution to each question starts, else we won't be able to grade it!).
- Collaboration is encouraged, but all write-ups must be done individually and independently, and mention your collaborator(s) if any. If you have referred a book or any other online material for obtaining a solution, please cite the source. Again don't copy the source *as is* - you may use the source to understand the solution, but write-up the solution in your own words.
- This question is less of a standard theory/programming assignment, and more of an investigative practical assignment where you're allowed to use web resources and online tools. The story in this question is fictitious, and any resemblance to actual persons, places or things is purely coincidental.

1. **Prologue:** Having completed a major part of the CS6024 course, you have become a popular sought-after bioinformatician by local health care officials. One morning, you receive a call from an IITM hospital staff that she has collected a viral sample from a patient with grave symptoms and sequence of a gene from that viral sample (which is attached as a sequence in FASTA format in the course moodle). She would like you to use any of the bioinformatics analyses you have learnt in the course to help answer:

   The Questions:

   (a) (6 points) Which exact viral species the sample belongs to, whether outbreaks of the same viral species (possibly a different strain) has happened in other countries, and which other viral species is closest to this viral species?
   (*Hint:* NCBI BLAST webtool is your friend! Glossary that may help you interpret BLAST results: cds - coding sequence of a gene that codes for a protein; you may also want to know that UMMC refers to a University in Malaysia)

   (b) (6 points) You may find that at least two other countries had the same viral species outbreak in the past - which of these two countries is closer to the IITM outbreak? Support your answer by providing the length of the longest common subsequence between the IITM viral gene and the same gene from the other country.
   (*Hint:* Use BLAST output from above that reports a local alignment between the query gene's cds and whole genome or cds from another country; EMBOSS suite of sequence extraction/alignment tools could also help), and

(c) (8 points) What is the phylogenetic relationship between the IITM viral gene and the other countries' genes? Please provide your answer as a phylogenetic tree diagram.
(*Hint:* Take top 20 sequences that BLAST outputs, perform multiple alignment of them using MUSCLE or CLUSTALW to get a distance matrix, and then use Neighbor Joining or other algorithms to get the tree; MEGA software conveniently contains both multiple alignment and Neighbor Joining algorithms; other tools like Phylip are also possible).

(d) (bonus 2 points) Which specific research group(s) in the campus or elsewhere in Chennai or Tamil Nadu would you collaborate with to obtain samples from fruit bats and sequence them using Next Generation Sequencing technology to test if fruit bats on the campus could've been the likely carriers of the virus?

Epilogue: Because patients are waiting for treatment and you have various responsibilities ranging from placements to end-sems to co-curricular activities, you do not have the time to write your own program, and hence are encouraged to use freely available bioinformatics software or online web-resources as indicated in the hints above.

2. (8 points) Provide an one-page description of the progress your team has made with the course research project. Elaborate on the implementation methods, problems faced, figures generated (e.g., figures you may be trying to reproduce from the paper you are extending), experiments run and results obtained. Also include details on the direction the project is currently taking, benchmarks for experiments you wish to achieve and figures you wish to generate before the final project presentation/report in the week starting Nov 4th.

3. (2 points) Please find below the link to the post-midterm feedback form for the course CS6024. Use your smail ids to access: https://forms.gle/ReSWabMNeVkBGDcH6. Kindly provide feedback by Assignment deadline. Constructive criticism is encouraged and requested.