IITM-CS6024 : Algorithmic Approaches to Computational Biology — Release Date: Oct 8

Assignment 2 — Due Date : Oct 18, 23:55

Roll No: IONdNTPH+ — Name: Gatacca

- 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. Same rules apply for codes uploaded to HackerRank (i.e., write your own code; we will run plagiarism checks on codes).
- 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.

1. (6 points) [Biological significance] If a pattern of length k appears exactly in every sequence, a simple enumeration of all k-letter patterns that appear in the sequences gives the solution to finding the pattern that occurs frequently.
   (a) (2 points) Why is finding DNA motifs more complex than just matching all k-length patterns in the input DNA sequences?
   (b) (1 point) State two biological reasons, apart from transcription factor (TF) protein binding to DNA, for why motifs occur in DNA sequences.
   (c) (1 point) Do mutations in TF binding motifs always lead to changes in the expression of the downstream gene?
   (d) (2 points) Why are some DNA motifs the same across many species?

2. (6 points) [Solve or search!] Answer YES/NO to these questions on the Euclidean-distance-based clustering problems seen in class. Justify each answer very briefly (one/two sentences either with your own argument or by citing a properly-formatted reference paper/book). Each of the n datapoints is of dimension m, and we seek k clusters.
   (a) (1 point) For all $m > 1$, is k-Means problem NP-hard for general k and polynomial-time solvable for $k = 1$?
   (b) (1 point) For all $m > 1$, is k-Centers problem NP-hard for general k and **not** polynomial-time solvable for $k = 1$?
   (c) (1 point) For $m = 1$ (i.e., all data points lie on a line), is k-Means problem polynomial-time solvable for general k?
   (d) (1 point) For $m = 1$, is k-Centers problem polynomial-time solvable for general k?

(e) (1 point) For $m = 1$ and $k = 1$, does the optimal solution of the k-Centers problem always cost half the distance between the two farthest datapoints?

(f) (1 point) The cost of a solution (SOLN) returned by the FarthestFirstTraversal heuristic seen in class is guaranteed to be at most twice the cost of an optimal solution (OPT) i.e., $SOLN \leq 2\,OPT$. Are there input instances (sets of datapoints) where the solution cost can be the worst value of **exactly** twice the optimal cost i.e., $SOLN = 2\,OPT$?

Note that polynomial-time solvable means the running time is polynomial in n, k and m, and "general k" refers to a k that is not fixed to any value (such as 1) but instead provided as part of the input.

3. (11 points) [Random coding] You are about to implement a variant (Gibbs sampling) of one of the top 10 algorithms of the 20th century (Metropolis algorithm for Monte Carlo).

   (a) (8 points) Solve the HackerRank Challenge: `https://www.hackerrank.com/iitm-cs6024aacb-assignment-2` that asks you to apply Gibbs sampling technique to find motifs in a set of sequences. Please follow random seed, pseudocount, and other instructions carefully to produce the exact same output as the test cases.

   (b) (3 points) Can you find a better-scoring motif for "Sample Input 1" than the one that your current algorithm outputs? What parameters of your algorithm did you tune to achieve this better-scoring motif? Feel free to run your code outside of HackerRank to answer this subquestion.

4. (11 points) p53 is a transcription factor that suppresses tumor growth through regulation of dozens of target genes with diverse biological functions. This master regulator is inactivated in nearly all tumors. Let's try to identify the DNA-binding motif onto which p53 transcription factor binds using a motif-finding tool called MEME (available via Galaxy or stand-alone).

   (a) (7 points) Run MEME on this input fasta file to get 3 candidate motifs. Of these three, which one do you think actually binds to p53 and why? The input fasta file contains a sample of ∼200-500bp sequences bound by p53 in a ChIP-seq experiment[1]. MEME may also have output two other motifs - argue by giving specific reasons whether those motifs could be true binding elements of p53?

   (b) (4 points) Run your Gibbs Sampler code from last question on the sequences provided. Report the motif you found, and did it match the one output by MEME? Report the parameter values you tried and chose for motif width (k), number of iterations (N), and the number of random restarts.

[1] Our initial search for motifs in the 2000-bp promoter sequences of known target genes of the p53 didn't yield expected results, so we resorted to ∼200-500bp p53-bound sequences obtained from the peaks of a genome-wide p53 ChIP-seq experiment described at the p53 BAER resource

5. (6 points) [Research exploratorium]: Provide properly-formatted references for papers in this solution.

   (a) (3 points) Read this paper on how to read a paper https://web.stanford.edu/class/ee384m/Handouts/HowtoReadPaper.pdf. What new tip/trick did you learn from this paper that you didn't already know before?

   (b) (3 points) What are some latest research publications you could find on *de novo* motif finding? Report one such paper based on Gibbs-sampling-like heuristics and another based on machine learning approaches such as clustering, SVM or deep learning. Try to answer this question by looking only at the papers' title/abstract.