Homework 4

CS 181, Fall 2022

Out: Nov. 16
Due: Nov. 22, 11:59 PM

Please upload your solutions on Gradescope. You can use LaTeX or a word document to write up your answers, but we prefer you use LaTeX. You may scan hand-written work or images for parts of solutions **only if** they are extremely clean and legible. Please ensure that your name does not appear anywhere in your handin.

Problem 0: Readings

The following sections of the Barton textbook chapter, which can be found on the resources page of the course website under Chapter 3 Readings. They may be helpful for Problems 3 and 4.

(a) *For Distance Methods, Corrections Are Essential to Convert Measures of Similarity to Evolutionary Distances*, pgs. 27 - 29

(b) *UPGMA and Neighbor-Joining Methods*, pgs. 22 - 26

Problem 1: Building and using suffix trees

The CS 181 TAs take a special trip to the Jurassic Park because entry is free for college students today! When they arrive, it turns out they have to solve a special problem in order to gain entry. They must find all the suffixes of the word, "SAUSAURUS", and they're enlisting your help to do so!

(a) Construct the **expanded** suffix tree T_x for the word $x = \texttt{SAUSAURUS}$. Build the tree one suffix at a time, starting with suf_1. Show the first four partial trees in your construction as well as the complete suffix tree T_x.

(b) Show how to use your suffix tree to return the starting indices of all occurrences of the string `SAU` in the text x. List all of these indices. In general, what is the big-O time complexity of returning the starting indices of all occurrences of a pattern p in a text t? Explain.

(c) Show how to use your suffix tree to determine whether the string `SAUSAGE` occurs in x. Does it occur in x? In general, what is the big-O time complexity of determining whether a pattern p occurs in a text t? Explain.

(d) **Bonus:** Construct the **compacted** suffix tree and the **position** suffix tree for the word $x = \texttt{SAUSAURUS}$. Explain why **position** suffix trees can be stored in $O(n)$ space, whereas **compact** suffix trees require $O(n^2)$ space.

Hints for this section: (1) You can assume that each node stores their children in a dictionary, so checking for a given character in the direct children of a node is constant time. (2) When calculating Big-O time complexity, consider what strings would create the worst-case scenario tree and how many nodes this tree would have.

Problem 2: Longest shared substring

Consider the following problem:

> LONGEST SHARED SUBSTRING PROBLEM
>
> **Input:** Strings u and v.
>
> **Output:** The longest substring that occurs in both u and v.

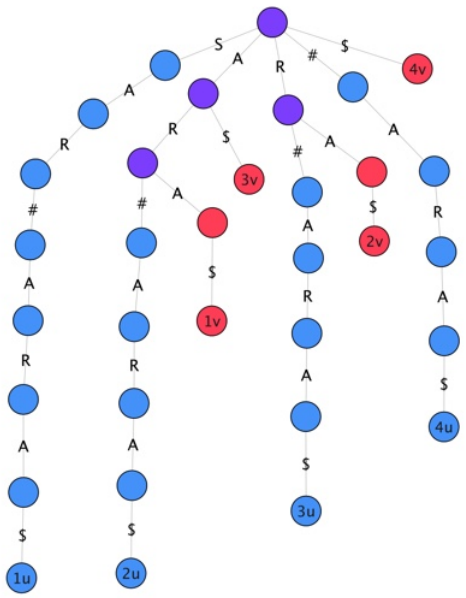
In theory, we could solve this problem by concatenating the two strings and constructing a joint suffix tree for the resulting string. In this problem, we expand upon this idea.

We begin by appending two different end-of-string characters to u and v, giving us $u\#$ and $v\$$. We then build the suffix tree for the concatenation of these two strings, $u\#v\$$. We shall color the leaves of this tree according to the following rule: if a leaf is labeled by the starting position of a suffix starting in u, color it blue; if a leaf is labeled by the starting position of a suffix starting in v, color it red. Note that all the leaves in the tree will be colored either red or blue according to this procedure.

Next we color the internal nodes of the tree according to the following rules:

- A node is colored blue (resp. red) if all the leaves in the subtree rooted at that node are blue (resp. red).
- A node is colored purple if the subtree rooted at that node contains both blue and red leaves.

The following is an example of joint suffix tree for $u\#v\$$ where u = SAR and v = ARA:



For all tasks in this problem, assume that we are using **expanded** suffix trees.

(a) Explain how every path ending in a purple node in the suffix tree of $u\#v\$$ spells out a substring shared by u and v.

(b) Explain how every path ending in a blue (resp. red) node in the suffix tree of $u\#v\$$ spells out a substring that appears in u but not in v (resp. v but not in u). **Note:** If the path contains #, then the substring spelled out by this path ends at #.

Given these two facts, it should be clear how we find the longest shared substring of u and v: We need only examine the strings spelled by paths that lead to purple nodes; the longest such string is precisely the solution to the Longest Shared Substring Problem.

The above approach makes use of a single suffix tree to solve the Longest Shared Substring Problem. It is also possible to solve this problem using two suffix trees, one for each of u and v.

(c) Describe an algorithm that solves the Longest Shared Substring Problem using the two suffix trees T_u and T_v.

(d) Compare (in as much detail as possible) the worst case space- and time-efficiency of the algorithm that uses a single suffix tree for $u\#v\$$ with the worst case space- and time-efficiency of the algorithm you designed in (c) above.

Problem 3: Jukes-Cantor Distance

(a) To get the feel of how to calculate the Jukes-Cantor distance, consider two biological sequences that have accumulated mutations over time. Below are snapshots of the two sequences at 4 different points in evolutionary time:

	$t = 1$	$t = 2$	$t = 3$	$t = 4$
Sequence 1	AGGTCA	AGCTCA	AGCTCA	AACTCA
Sequence 2	AGGTCA	AGGTCA	TGGTCA	TGGTCC

For each timepoint t, calculate the fraction of sites that are different between the two sequences, λ_t. Then use this value to calculate the Jukes-Cantor distance D_t between each sequence.

(b) Now consider a fifth timepoint:

	$t = 5$
Sequence 1	AACTGA
Sequence 2	TGGTCC

Compute λ as above. What happens when you try to use this value to calculate the Jukes-Cantor distance?

(c) Propose a correction to the Jukes-Cantor distance that avoids the issue above. What result do you get when you calculate your new distance at $t = 5$? Explain in two or three sentences the potential benefits and drawbacks of your new distance. It may be helpful to read up a little on the *Infinite Sites Model*.

Problem 4: Neighbor Joining

During a special trip to Maine, the CS181 TAs take a short boat ride to the New England Revived Dinosaur Sanctuary. One of the world's most non-extinct dinosaur sanctuaries, NERDS is abundant with T-Rexs, Brontosauruses, Velociraptors, and other dino life. They are lucky enough to see a Hannahsaurus hatch! Wanting to better understand their evolution over the decades, one of the TAs decides to build an evolutionary tree for various dinosaurs and their relatives.

	Justinosaurus	Corinnosaurus	Sorinosaurus	Chicken	Serenasaurus
Justinosaurus					
Corinnosaurus	135				
Sorinosaurus	159	24			
Chicken	213	78	54		
Serenasaurus	177	42	18	36	

Use the distance matrix above and the neighbor joining algorithm to construct an evolutionary tree for the dinos above. In a sentence or two, comment on Serenasaurus's place in the evolution of dinosaurs.

Problem 5: Ethics of HIV Phylogenetics

Molecular HIV Surveillance (MHS) refers to the use of phylogenetic methods to detect and understand patterns of HIV transmission, and particularly to detect clusters of individuals with genetically-linked strains of HIV.

One application of MHS data is to HIV cluster detection and response (CDR), which is defined by the CDC as "the actions taken to prevent further transmission when a growing cluster of HIV transmission and its associated risk network is detected." CDR has been identified as a crucial pillar of a federal initiative to End the HIV Epidemic by 2030.

Review Molldrem and Smith (2020), an article which criticizes the use of Molecular HIV Surveillance (MHS) data in HIV cluster detection and response (CDR).

a. Read the section titled "History and Context for MHS and CDR". Then answer the following questions.

   i. What characteristic of HIV makes phylogenetics a possible tool for studying the virus? (1-2 sentences)

   ii. Briefly explain how HIV surveillance data is used in CDR programs in your own words. (1-2 sentences)

b. Read the section "Theoretical Framing: From 'Critical Bioethics' to 'Bioethics of the Oppressed"'. Then answer the following questions.

   i. Explain your understanding of "Data Justice" as a critical framework (2-3 sentences). (Optional: If you would like further reading, you can take a look at page 874 of this article)

   ii. How does "Data Justice" as a framework for analyzing ethical issues differ from ethical theories we have discussed previously in the class (3-4 sentences)?

iii. The article explains that blood samples from HIV-positive patients are routinely collected in order to identify whether a patient's particular strain of HIV is resistant to certain antiviral drugs. However, the "[r]e-uses of clinical HIV data for public health surveillance and prevention do not require consent." Do you believe this practice is ethical? Please use at least one of the three ethical frameworks (utilitarianism, Kantian deontology, Data Justice) in justifying your response (4-5 sentences).

c. Read the section "Case 2: Determining Directionality of Transmission".

i. Explain how the way phylogenetic relationships are represented in tools such as HIV-TRACE may imply the directionality of HIV transmission (1-2 sentences).

ii. Brainstorm one or two ways you could redesign tools like HIV-TRACE to make them less vulnerable to misinterpretation such as directionality of transmission (3-4 sentences). (Your answer may consider any aspect of the tool, including the visual representation of the phylogenetic relationships. We're not looking for a specific, or fully-fleshed out solution here, just brief thoughts which reflect thoughtful engagement with the question).