Midterm Exam

CS 4390/5390

17 October 2019

Please put your name on each page of this test. You can write on the front and back of pages, but please keep answers to a single question on a single sheet of paper. If extra room is needed please get extra paper from the instructor and ensure that the problem number, and your name are on it, as well as a page number.

This assessment is open book and open computer, but you are not allowed to communicate with other people either inside or outside of this class.

By signing below you are certifying that you did not have assistance from anyone else on this exam, that you did not help anyone else, and that if you know of or come to learn of someone who did gain assistance from another person you will inform the instructor immediately.

Key

Printed Name

Signature Date

Problem	Score
1	/2
2	/1
3	/2
4	/3
5	/3
6	/2
7	/2
8	/1
9	/1
10	/2
11	/1
Total	/20

1. (2 points) Given the table below which was created using the Smith-Waterman algorithm for local alignment, (a) identify the local alignment score, and (b) perform trace-back to find the optimal alignment.

		G	A	A	C	G	G
	0	0	0	0	0	0	0
A	0	0	↖6	↖6	←2	0	0
T	0	0	↑2	↖4	↖4	0	0
A	0	0	↖6	↖8	←4	↖2	0
C	0	0	↑2	↖↑4	↖14	↖←10	↖←6
G	0	↖6	←2	0	↑10	↖20	↖←16
T	0	↑2	↖4	0	↑6	↑16	↖18
T	0	0	0	↖2	↑2	↑12	↖↑14

Optimal Local Alignment Score:

20.

Optimal Local Alignment (note not all of the spaced will be used)

														A	T	A	C	G
														A	-	A	C	G

2. (1 point) Given the Needleman-Wunsch table below, find the optimal global alignment for the two sequences.

		A	A	C	A	T	G	A
	0	←-2	←-4	←-6	←-8	←-10	←-12	←-14
A	↑-2	↖5	↖←3	←1	↖←-1	←-3	←-5	↖←-7
A	↑-4	↖↑3	↖10	←8	↖←6	←4	←2	↖←0
T	↑-6	↑1	↑8	↖8	↖←6	↖11	←9	←7
C	↑-8	↑-1	↑6	↖13	←11	←↑9	↖9	↖←7
G	↑-10	↑-3	↑4	↑11	↖11	↖←9	↖14	←12
T	↑-12	↑-5	↑2	↑9	↖↑9	↖16	←14	↖←12
A	↑-14	↖↑-7	↖↑0	←↑7	↖14	↑14	↖14	↖19

Optimal Global Alignment (note not all of the spaced will be used)

											A	A	T	C	G	T	-	A
											A	A	-	C	A	T	G	A

3. (2 point) Given the following partially completed computation of the Z-value algorithm, compute the rest of the values using the $O(n)$ time algorithm we discussed in class. Describe how you arrived at each value.

	C	G	T	C	G	T	A	C	G	T	C	G	A	C
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Z_i	-	0	0	3	0	0	0	5	0	0	2	0	0	1

coppied Z_2

coppied Z_3

coppied Z_5

~~already checked~~ compare S[13] = S[1]

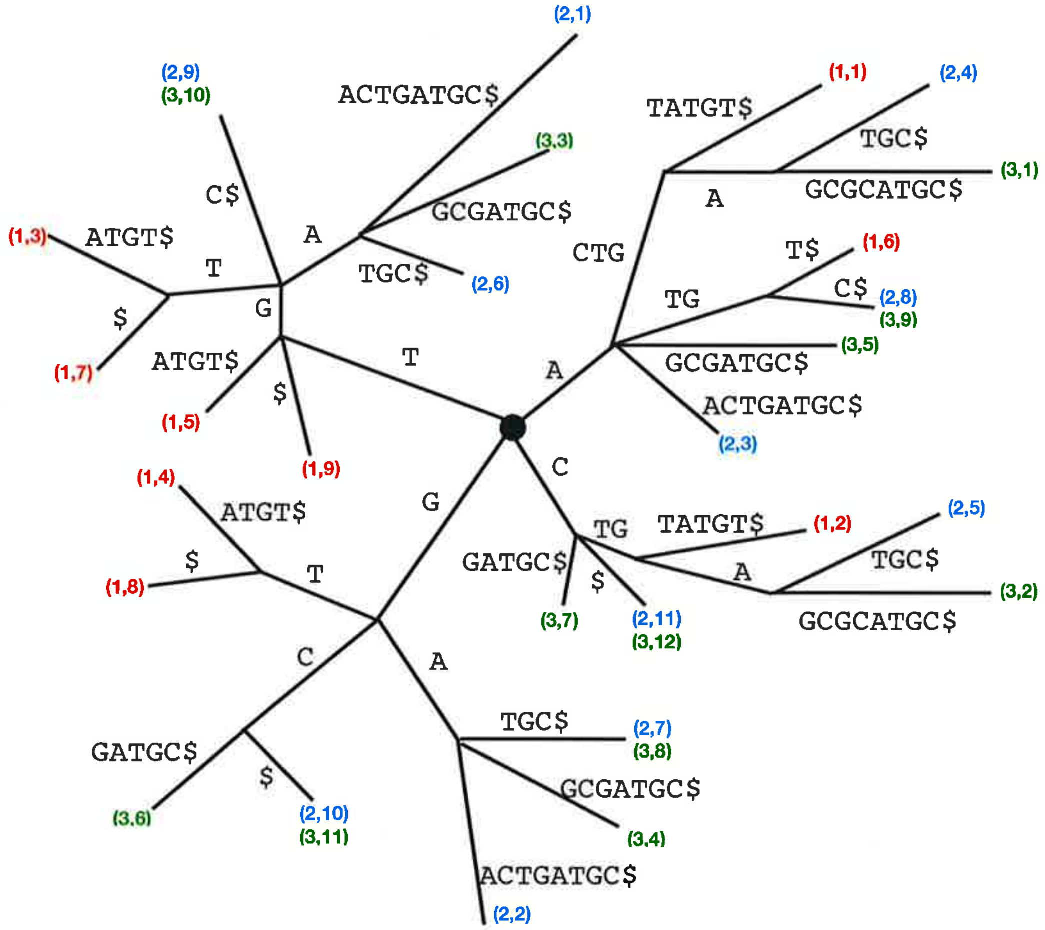
compare S[14] = S[1]

$Z_4 > 5-3$, So check

S[13] = S[3].

it does not so the value is 2

4. (3 points) From the suffix tree below: (a) determine if the string ACTG is in the input set of sequences, and explain your reasoning; (b) find the longest substring that occurs in all of the sequences *twice*, and explain your reasoning; (c) list the missing suffix links.



(a) yes, a path from the root labeled "A", "CTG" exists

(b) "TG", deepest of "G", "TG", "T", "A"

(c) ACTG → CTG

ACTGA → CTGA

ATG → TG

CTG → TG

CTGA → TGA

GA → A

GC → C

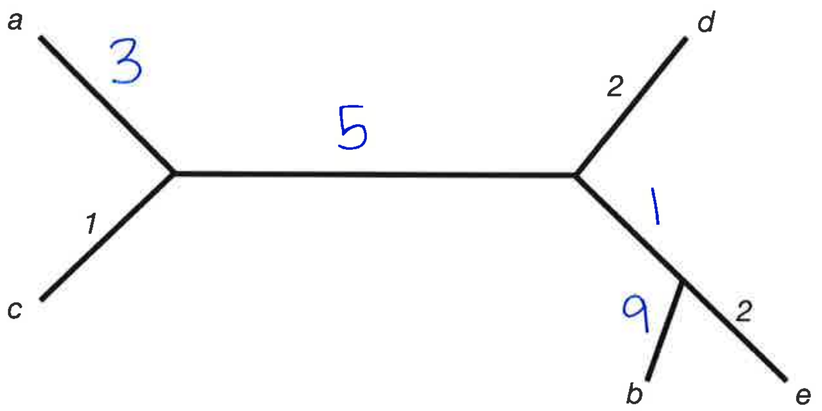
GT → T

TG → G

TGA → GA

5. (3 points) (a) Place all of the **additive** labels (leaves and edge values) on the tree below, using the given table D. (note the edges are purposefully not to scale.) (b) How do you find the value of $D(i, j)$ for leaves i and j from the tree? (c) Is D also ultrametric or min-ultrametric? explain your answer.

D	a	b	c	d	e
a	0	~~17~~ 18	4	10	11
b		0	16	12	11
c			0	8	9
d				0	5
e					0



(b) Sum the edge weights along the path from i to j.

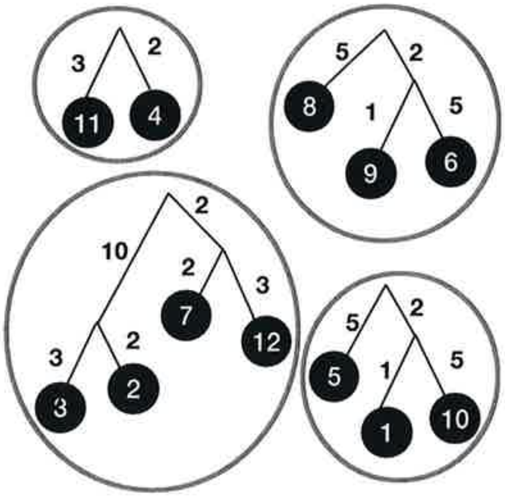
(c) no, The table D contains too many values for it to induce an utrametric or min-ultrametric tree.

6. (2 points) (a) Given the partial neighbor-joining procedure, complete the tree. (b) What does the *argmax* in the procedure do?

$n=12 \quad n-2=10$

U		D	{1, 5, 10}	{2, 3, 7, 12}	{4, 11}	{6, 8, 9}
1.9	A	{1, 5, 10}	0	6	4	9
3.6	B	{2, 3, 7, 12}	0.5	0	18	12
2.9	C	{4, 11}	-0.8	11.5	0	7
2.8	D	{6, 8, 9}	4.3	5.6	1.3	0

$D - U_A - U_B$



Branch Lengths

$$\frac{4+1.9-2.9}{2} = 1.5 \quad A$$

$$\frac{4+2.9-1.9}{2} = 2.5 \quad C$$

$$\begin{array}{c|ccc|c} & A,C & B & D & U \\ \hline A,C & & 10 & 6 & 1.6 \\ B & 6.2 & & 12 & 2.2 \\ D & \boxed{2.6} & 8 & & 1.8 \end{array}$$

New Dist

$$\frac{6+1.8-4}{2} = 10$$

$$\frac{9+7-4}{2} = 6$$

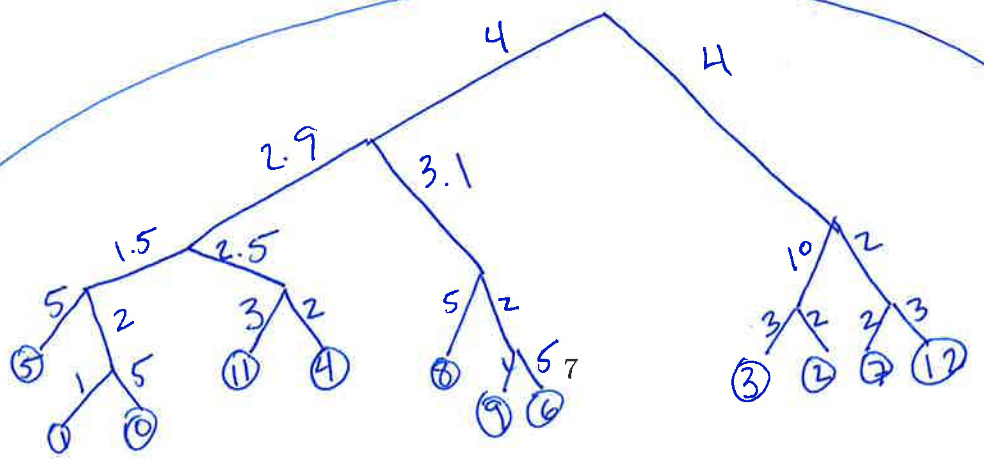
Branch Lengths

$$\frac{6+1.6-1.8}{2} = 2.9 \quad A,C$$

$$\frac{6+1.8-1.6}{2} = 3.1 \quad D$$

New Dist

$$\frac{10+12-6}{2} = 8$$



7. (2 points) (a) using the replacement matrix below, find the neighbors of the k-mers in the sequence with scores greater than 20 for $k = 3$. (b) what does it mean when BLAST finds hits to this set of k-mers in the database.

NSCEVW

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-2	-1	-1	1	-1	-1	0	-2	-1	-2	-1	-1	-3	0	1	1	-4	-3	0
R	-2	7	0	-2	-3	2	-1	-2	1	-3	-3	4	-2	-4	-1	-1	-1	-3	-2	-3
N	-1	0	6	3	-2	1	1	0	1	-4	-4	1	-3	-4	-2	1	0	-5	-2	-3
D	-1	-2	3	6	-4	1	3	-1	0	-5	-5	0	-4	-6	-1	0	-1	-6	-4	-4
C	1	-3	-2	-4	12	-3	-4	-2	-2	0	-3	-4	-1	-3	-3	1	0	-6	0	1
Q	-1	2	1	1	-3	5	2	-2	2	-3	-2	2	-1	-3	-1	0	0	-6	-3	-2
E	-1	-1	1	3	-4	2	5	-1	0	-4	-4	1	-3	-5	-1	0	-1	-6	-3	-3
G	0	-2	0	-1	-2	-2	-1	8	-2	-6	-5	-2	-4	-5	-2	0	-2	-5	-5	-4
H	-2	1	1	0	-2	2	0	-2	8	-3	-2	0	-3	0	-2	0	-1	-1	3	-3
I	-1	-3	-4	-5	0	-3	-4	-6	-3	5	3	-3	2	0	-4	-3	-1	-2	-2	4
L	-2	-3	-4	-5	-3	-2	-4	-5	-2	3	5	-3	3	2	-3	-3	-2	-1	-1	2
K	-1	4	1	0	-4	2	1	-2	0	-3	-3	5	-2	-5	-1	0	0	-4	-3	-3
M	-1	-2	-3	-4	-1	-1	-3	-4	-3	2	3	-2	6	1	-3	-2	-1	-3	-2	2
F	-3	-4	-4	-6	-3	-3	-5	-5	0	0	2	-5	1	8	-4	-3	-3	3	5	-1
P	0	-1	-2	-1	-3	-1	-1	-2	-2	-4	-3	-1	-3	-4	9	0	-1	-4	-5	-3
S	1	-1	1	0	1	0	0	0	0	-3	-3	0	-2	-3	0	4	2	-4	-2	-2
T	1	-1	0	-1	0	0	-1	-2	-1	-1	-2	0	-1	-3	-1	2	4	-5	-3	0
W	-4	-3	-5	-6	-6	-6	-6	-5	-1	-2	-1	-4	-3	3	-4	-4	-5	15	4	-4
Y	-3	-2	-2	-4	0	-3	-3	-5	3	-2	-1	-3	-2	5	-5	-2	-3	4	9	-2
V	0	-3	-3	-4	1	-2	-3	-4	-3	4	2	-3	2	-1	-3	-2	0	-4	-2	4

NSC: NSC 22, NTC 20

SCE: SCE 21

CEV: CEV 21, CEI 21

EVW: EVW 24, NVW 20, DVW 22, QVW 21, KVW 20,

EAW 20, ECW 21, EIW 24, ELW 22, EMW 22, ETW 20,

NIW 20, DIW 22, DLW 20, DMW 20, QIW 21, KIW 20

8. (1 point) What is the sum-of-pairs score of the following multiple sequence alignment using the global scoring with affine scoring model with the following parameters:

match	10
mismatch	-5
indel	-1
gap	-3

M	T	I	G
5	1	0	0
3	0	2	2
3	1	2	2
3	1	2	1
3	0	2	2

```
AGTGCA
AG--TA
AGAGCA
A--GTA
AG-GT-
```

M	T	I	G
3	1	2	1
3	1	2	1
3	1	2	1
3	0	2	2
3	1	2	2

(10 x 32) − (5 x 7) − (1 x 18) − (3x14)

320 − 35 − 18 − 42

320 − 95

225

9. (1 point) Given the pairwise alignments between the 4 sequences, and using sequence A as the star-center, create the multiple alignment using the center-star method.

A :	-AG-GTT-T	A :	-AGG-TTT	A :	AG-GTT-T	A :	-AGG-TT-T
B :	CTGTGTTGT	D :	CTGGATTT	C :	AGTGTTGT	E :	CAGGATTGT

A - A G - G - T T - T

B C T G T G - T T G T

C - A G T G - T T G T

D C T G - G A T T - T

E C A G - G A T T G T

10. (2 points) How would we modify the Needleman-Wunsch algorithm if we wanted to allow for any character in S to be repeated aligned as many times as we want in place.

For example when aligning $S =$ `AGA` with $T =$ `GGGGGA`, an optimal alignment would repeat the `G` in S 5 times to give the alignment:

```
AGGGGGA
-GGGGGA
```

In reality, the middle `G` is is being aligned with all of the `G`s in T.

Modify the recorrance by adding an extra term

$$V(i,j) = \max \begin{cases} V(i-1,j-1) + \delta(S[i],T[j]) \\ V(i-1,j) + \delta(S[i], '-') \\ V(i,j-1) + \delta(\cancel{S}', T[j]) \\ V(i,j-1) + \delta(S[i],T[j]) \end{cases}$$

allows for the best match/mismatch from last char of S.

On backtrack, follow links that may move non-diagonally but still output a column w/ two characters.

11. (1 point) What is the term for computation applied to biological problems? (hint: it is in the name of the course.)

Computational Biology