1. Use the *Needleman-Wunch* dynamic programming table for $S =$ `CTACTGTGT` and $T =$ `CACCCCTGTG` below to the next questions.

		C	T	A	C	T	G	T	G	T
	0	←-0.5	←-1	←-1.5	←-2	←-2.5	←-3	←-3.5	←-4	←-4.5
C	↑-0.5	↖5	←4.5	↖4	↖3.5	←3	←2.5	←2	←1.5	←1
A	↑-1	↑4.5	↑↖4	↖9.5	←9	←8.5	←8	←7.5	←7	←6.5
C	↑-1.5	↑4	↑↖3.5	↑9	↖14.5	←14	←13.5	←13	←12.5	←12
C	↑-2	↖3.5	↑↖3	↑8.5	↑↖14	↑↖13.5	↑↖13	↑↖12.5	↑↖12	↑↖11.5
C	↑-2.5	↑3	↑↖2.5	↑8	↑↖13.5	↑↖13	↑↖12.5	↑↖12	↑↖11.5	↑↖11
C	↑-3	↑2.5	↑↖2	↑7.5	↑↖13	↑↖12.5	↑↖12	↑↖11.5	↑↖11	↑↖10.5
T	↑-3.5	↑2	↖7.5	↑7	↑12.5	↖18	←17.5	↖17	←16.5	↖16
G	↑-4	↑1.5	↑7	↑↖6.5	↑12	↑17.5	↖23	←22.5	↖22	←21.5
T	↑-4.5	↑1	↑6.5	↑↖6	↑11.5	↑17	↑22.5	↖28	←27.5	↖27
G	↑-5	↑0.5	↑6	↑↖5.5	↑11	↑16.5	↑22	↑27.5	↖33	←32.5

10pt (a) How many co-optimal alignments of the two strings are there?

10pt (b) What is the optimal alignment of $S[1....3]$ and $T[1...5]$? (note these are prefixes `CTA` and `CACCC`)

10pt (c) What is the mismatch penalty used to construct the table? match score? indel penalty?

15pt/10pt (d) † Using *only the scores in the table above* is it possible to determine the score of the optimal alignment of $S[4...9]$ and $T[6...10]$? Why or why not?

a) 4

CTAC---TGTG
C-ACCCCTGTG

CTA-C--TGTG
C-ACCCCTGTG

CTA--C-TGTG
C-ACCCCTGTG

CTA---CTGTG
C-ACCCCTGTG

b) CTA----
C-ACCCC

c) $\forall x \neq y \in \Sigma$

$\delta(x,x) = 5$

$\delta(x,y) = -1$

$\delta(-,x) = \delta(x,-) = -0.5$

d) No, Global alignment only calculates the optimal alignments of prefixes of the strings.

2. Use the alignments and plots below to answer the following questions:

10 pt (a) Calculate the accuracies for the following groups of alignments, the reference alignment is provided at the top of each column.

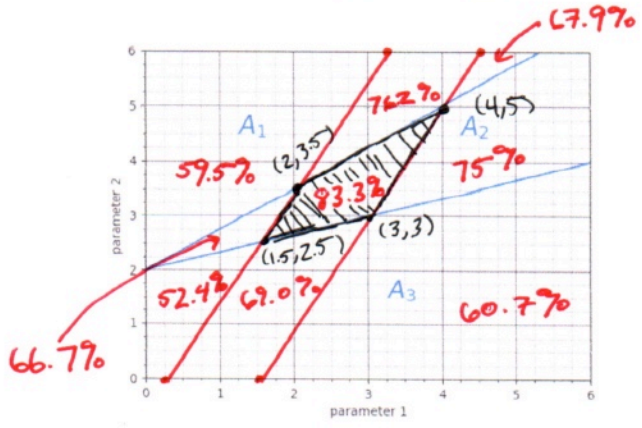
20 pt (b) Given the accuracies and the parameter decompositions shown in the figures, what is the region of the parameter space (identify the corners of the polygon) that provides the best alignments on average across these two pairs of sequences.

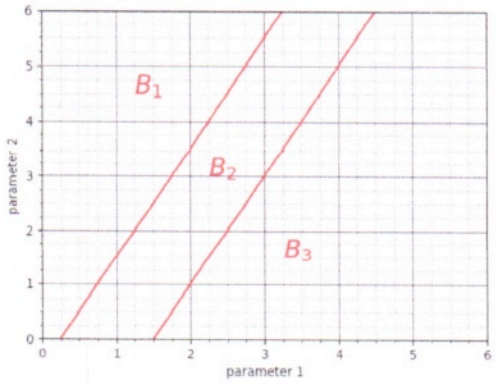
Remember accuracy, with respect to a reference, is the fraction of columns from the reference that are recovered in a computed alignment; and that each region of the plots corresponds to a set of parameters that produce the labeled alignment.

Reference A	Reference B
ATG-CTGGAT -TGA-TCGAT	TTGTGTCC-- TT-T-TCCAA

	A_1	A_2	A_3		B_1	B_2	B_3
	ATG-CT-GGAT -TGA-TC-GAT	ATGCTGGAT -TGATCGAT	ATG--CTGGAT -TGATC--GAT		TTGTGTCC TTTTCCAA	TTGTGTCC-- TTTT--CCAA	TTGTGTCC-- --TTTTCCAA

Accuracies:	A_1	A_2	A_3	B_1	B_2	B_3
	6/7 85.7%	7/7 100%	5/7 71.4%	2/6 33.3%	4/6 66.6%	3/6 50%





best polygon (1.5, 2.5)
(2, 3.5)
(4, 5)
(3, 3)

3. Use the suffix tree below to answer the next questions.

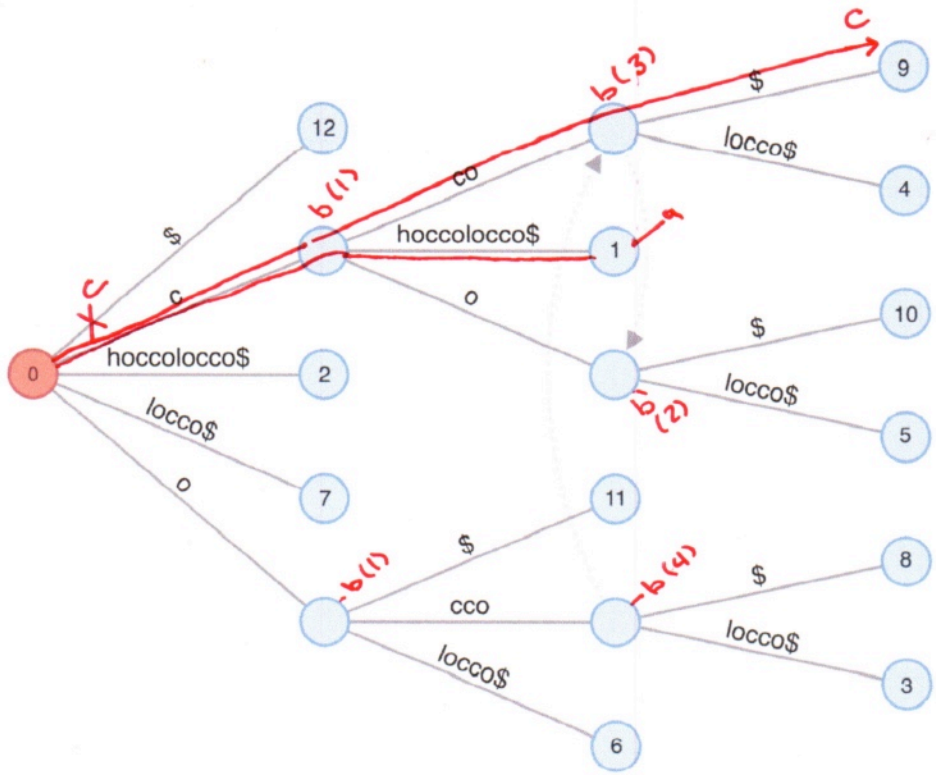


Figure 1: Suffix tree for question 3

5pt (a) What is the full string for which this is the suffix tree?

10pt (b) What is the longest sub-string that occurs 2 times?

10pt (c) What is the lexicographically smallest suffix that is strictly longer than 1 character? (that is, its more than just $)

a) C hoccolocco$

b) occo

c) cco$

4. Below is an ILP for pairwise global sequence alignment with several constraints missing. The scoring uses a match score of α, a mismatch penalty of β and an indel penalty of γ. Define these constraints.

$$\text{maximize} \quad \alpha \sum_{i,j} X_{ij} - \beta \sum_{i,j} Y_{ij} - \gamma \left(\sum_i Z_i^S + \sum_j Z_j^T \right)$$

$$\text{subject to} \quad \sum_j X_{ij} + \sum_j Y_{ij} + Z_i^S = 1 \qquad \forall i$$

10 pt $\sum_i X_{ij} + \sum_i Y_{ij} + Z_j^T = 1$ $\forall j$

10 pt $Y_{ij} = \emptyset$ $\forall i, j : S[i] = T[j]$

10 pt $X_{ij} = \emptyset$ $\forall i, j : S[i] \neq T[j]$

10 pt $(X_{ij} + Y_{ij}) + (X_{i'j'} + Y_{i'j'}) \leq 1$ $\forall i < i', j > j'$

$$X_{ij} \in \{0, 1\}, \qquad \forall i, j$$

$$Y_{ij} \in \{0, 1\}, \qquad \forall i, j$$

$$Z_i^S \in \{0, 1\}, \qquad \forall i$$

$$Z_j^T \in \{0, 1\}, \qquad \forall j$$

Hints:

- The first constraint enforces that each position i in S can only be a match, a mismatch or a deletion, the second constraint will do something similar but for T.
- The 3rd and 4th constraints decide if a position is a match or mismatch, remember when defining an ILP we usually say what a value *can't be* based on the information in the problem.
- The 5th constraint will be similar to the one we defined for LCS, but here we have two variables that can define a match between any two indexes.

20pt 5. † True or False: When computing the sum of pairs score of a multiple sequence alignment

$$id + 2mt + 2ms = L\binom{k}{2},$$

where id is the total number of indels, ms is the total number of mismatches, mt is the total number of matches, k is the number of strings aligned, and L is the length of the alignment itself. Justify your answer.

5pt 6. (bonus for all) What is the item that is the answer to question 3a? (You can use google for this one)

5) False, the total number of pairs of positions is indeed $L\binom{K}{2}$ but, when calculating sum-of-pairs we remove colomns in the induced alignments that are columns only. Therefore it can be less.

(bonus: its *bounded* by this ammount.)

6) Choccolocco is a town in Alabama.