Final Prep

Alignment -either global or local

Given the following strings S1,S2

cctag

catt

1.)

a.)Create the solution matrix for optimal alignment of S1,S2 such that matches=3

Mismatches =-2 and spaces=-1

b.) provide a single optimal global aligment

2.) a.) Create a local alignments solution matrix

b.) provide a single optimal local aligment

Recurrence and solution matrix questions

a.)Create a recurrence relationship for calculating the similarity score between sequences given that the only alignments you all those that have

matches, mismatches, and spaces but only spaces are allowed cross from A, and T characters.

Bayes Method Question

Given covid test has been FDA approved with the following testing:
For people with covid 90% of them showed two lines on the test.
For people not currently infected by covid 5% showed two lines on the test.

The population is currently experiencing a high covid rate of 10%.
You test positive for covid what is the prob that you have covid?

K-means Question:

Suppose you are analyzing gene expression data of two genes across different tissues of an organism. The expression values are measured in two dimensions (log10-transformed counts per million) and are as follows:

Gene	Tissue 1	Tissue 2	Tissue 3	Tissue 4
A	1	1	2	2
B	1	1.5	2	2.5

You want to cluster the tissues based on the similarity of their gene expression profiles using the k-means algorithm with k=2.
Let the initial centroids be C1=(1,1.25) C2=(2,2.25)
a) Assign each tissue to the closest centroid.
b) Recalculate the centroids based on the mean of the tissues assigned to each cluster.
c.) what conclusion can you make, do we need to recluster?

Sequence Assembly Question.

Given the following reads can you create a de Bruin graph to find the Hamiltonian path. (You do not need to find the path).

R1: ACTA

R2:TATAT

R3: ATAT

BLAST Question

(reminder how blast works https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/650/Seeding.html)

Given known sequence database of

S1: ATTACG

S2: TATGTC

S3:ATTTAT

Create the hashtable that will be used for search given kmer length is 3.

Let q=ATCG and let the threshold T=2, assume we use a global alignment where Match=1, Mis=-4 for all miss matches accept for score(A,C)=0.

a.)What are the seed kmers?

b.)name the sequences that have a match your query sequence?

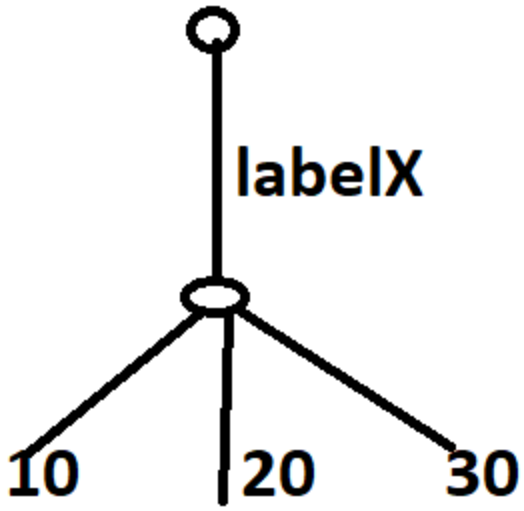
c.) in what cases would you use blast vs global/local alignment.

Suffix-Tree Question

a.)Create a suffix tree for the string

S=ATATATATAT$

b.)



Given |labelX|=7

a.) does S[10]=S[20]? Yes

b.) does S[16]=S[26]? Yes

c.) What is the alphabetical relationship between S[17],S[27],S[37]?

S[17]<S[27]<S[37]

Phylogeny

Initialization:

Assign each x_i into its own cluster C_i

Define one leaf per sequence, height 0

Iteration:

Find two clusters C_i, C_j s.t. d_{ij} is min

Let $C_k = C_i \cup C_j$

Define node connecting C_i, C_j,
& place it at height $d_{ij}/2$

Delete C_i, C_j

Termination:

When two clusters i, j remain,
place root at height $d_{ij}/2$

	a	b	c	d
a	0	2	4	6
b	2	0	4	6
c	4	4	0	6
d	6	6	6	0

Is the above matrix ultrametric? Explain your answer.
Please create a phylogenetic tree using UPGMA algorithm based on the distrance matrix.

I will provide the code ot UPGMA on the test.

Hidden Markov Chain:

1.) Given a HMM, and a path and a set observations can you find P(pi,x).
2.) Can you give the most likely path given an HMM.
3.) Explain the optimality equation of the virterbi algorithm, what is its run-time

Coding:

Write perl code to print all suffixes.
Write perl code to print all kmers.
Write perl code to search for query sequence given some database sequences (program 3)

Write code to calculate D(i,j)

$$D_{ij} = d_{ij} - (r_i + r_j)$$

Where

$$r_i = \frac{1}{|L| - 2} \Sigma_k d_{ik}$$

Where D is an L by L array

Neighbor-Joining- Not on final but fun to try

	a	b	c	d
a	0	3	6	8
b	3	0	5	7
c	6	5	0	8
d	8	7	8	0

- Guaranteed to produce the correct tree if distance is additive
- May produce a good tree even when distance is not additive

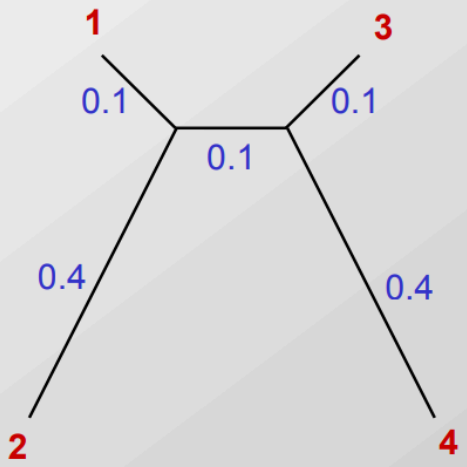
Step 1: Finding neighboring leaves

Define

$D_{ij} = d_{ij} - (r_i + r_j)$

Where

$$r_i = \frac{1}{|L| - 2} \Sigma_k d_{ik}$$



Claim: The above "magic trick" ensures that D_{ij} is minimal **iff** i, j are neighbors
Proof: Beyond the scope of this lecture (Durbin book, p. 189)

Algorithm: Neighbor-joining

Initialization:

Define T to be the set of leaf nodes, one per sequence
Let L = T

Iteration:

Pick i, j s.t. D_{ij} is minimal
Define a new node k, and set $d_{km} = ½ (d_{im} + d_{jm} - d_{ij})$ for all $m \in L$

Add k to T, with edges of lengths $d_{ik} = ½ (d_{ij} + r_i - r_j)$
Remove i, j from L;
Add k to L

Termination:

When L consists of two nodes, i, j, and the edge between them of length d_{ij}