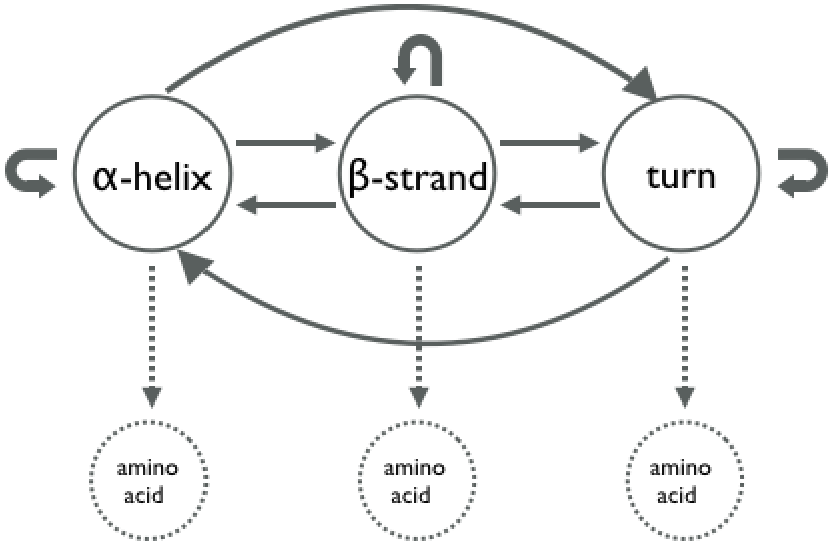
Homework 4

Dieu My Nguyen | MCDB 5520 | Mar 23, 2018

1. You consider using an HMM approach to model protein secondary structure prediction. The straight-forward approach uses three secondary structure confirmations: “α-helix”, “β-strand”, and “turn” as the hidden states emitting observable amino acids. It is assumed that the frequencies/probabilities of each of the twenty amino acids can be determined from experimental data for each of those confirmations.

a) (4pt) Draw the state diagram (circles and arrows) of the HMM.



b) (2pt) How many emission parameters are needed to describe this model?

20 emission parameters for 20 amino acids that are emitted and observable. Since each of the k number of states can emit each of the m observations, we would have $km = 3 * 20 = 60$total possibilities for the emission probability distribution.

c) (2pt) How many transition parameters are needed to describe this model?

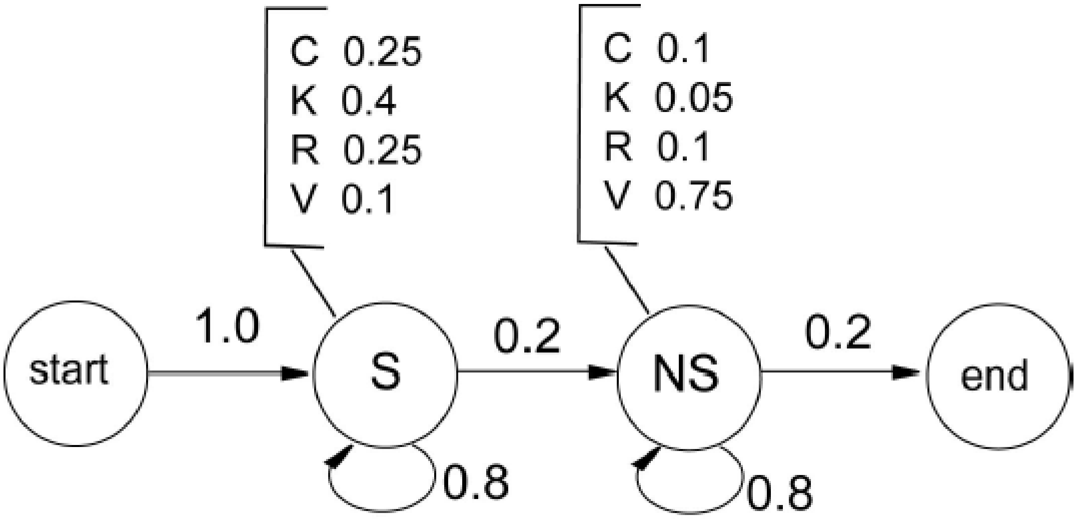
We can transition from any one of k total states to another (including staying in the same one), there are $k^2 = 3^2 = 9$ parameters for transitions.

d) (2pt) What is hidden in this hidden Markov model?

The 3 secondary structure confirmations: “α-helix”, “β-strand”, and “turn.”

2. You suspect that there is a signal peptide in PepY and you will use an HMM to predict its position. The model and parameters are given in the graph below. Note in the figure ‘S’ stands for “signal peptide” state and ‘N’ (marked NS in the diagram) for “Non-signal peptide” state.

Emission of PepY (s) : KKRKVRR
State Path of PepY (π) : SSSSNNN



a) (4pt) You are given a sequence s and a path π (above), what is $P(s, \pi)$?

$$P(s,\pi) = \prod_{t=1}^{T} P(\pi_t \mid \pi_{t-1}) P(s_t \mid \pi_t)$$

$$= P(\pi_1)P(s_1 \mid \pi_1) \ * \ P(\pi_2 \mid \pi_1)P(s_2 \mid \pi_2) \ * \ \dots \ * \ P(\pi_T \mid \pi_{T-1})P(s_T \mid \pi_T)$$

$$= (1)(0.4) * (0.8)(0.4) * (0.8)(0.25) * (0.8)(0.4) * (0.2)(0.75) * (0.8)(0.1) * (0.2)(0.1)$$

$$= 0.00000196608$$

b) (6pt) Name the algorithm used for each of the following questions:

(i) Given a sequence, what is the most likely path through the model?

Viterbi algorithm.

(ii) Given a sequence, how likely did it come from this model?

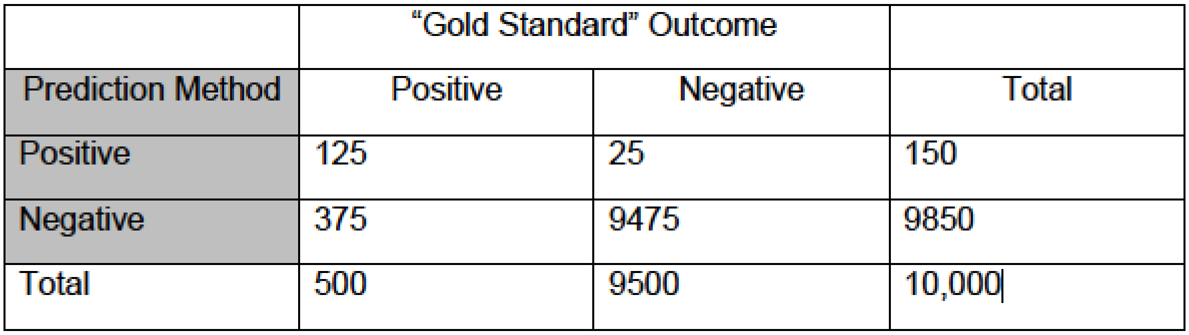
The forward algorithm.

(iii) Given unlabeled training data, how do I determine the emission and transition parameters?

Baum-Welch algorithm.

3. Consider a new algorithm for predicting whether a particular RNA binding protein binds to an exon. 10,000 exons are evaluated by the prediction method and a cutoff of 2 was selected. Everything scoring above a 2 was considered positive for the RNA binding protein whereas

everything below this score was classified as negative. These results were then compared to a gold standard method of determining whether the RNA binding protein associates with the exon. The results are shown in the following table:



	“Gold Standard” Outcome		
Prediction Method	Positive	Negative	Total
Positive	125	25	150
Negative	375	9475	9850
Total	500	9500	10,000

Calculate:

a) (4pt) Sensitivity

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{125}{125 + 375} = 0.250$$

b) (3pt) Specificity

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{9475}{9475 + 25} \approx 0.997$$

c) (3pt) Positive predictive value

$$\text{Positive predictive value} = \frac{TP}{FP + TP} = \frac{125}{25 + 125} \approx 0.833$$

4. Consider the following multiple sequence alignment (spaces included for ease of reading) for the proto-insulin gene:

Human: ATGGCCCTGT GGATGCGCCT CCTGCCCCTG CTGGCGCTGC TGGCCCTCTG
Sheep: ATGGCCATGT GGACACGCCT GGTGCCCCTG CTGGCCCTGC TGGCACTCTG
Chick: ATGGCTCTAT GGACACGCCT TCTGCCTCTA CTGGCCCTGC TAGCCCTCTG

a) (4pt) You are considering the Jukes-Cantor model of sequence evolution, which is a single parameter model of evolution (typically described simply as α). Given only the comparison between Human and Sheep as training data, what is your best estimate of α?

α denotes the rate of observable substitutions over 1 time step. From Human to Sheep, 5/50 have undergone mutations. So α is estimated to be 5/50 or 1/10.

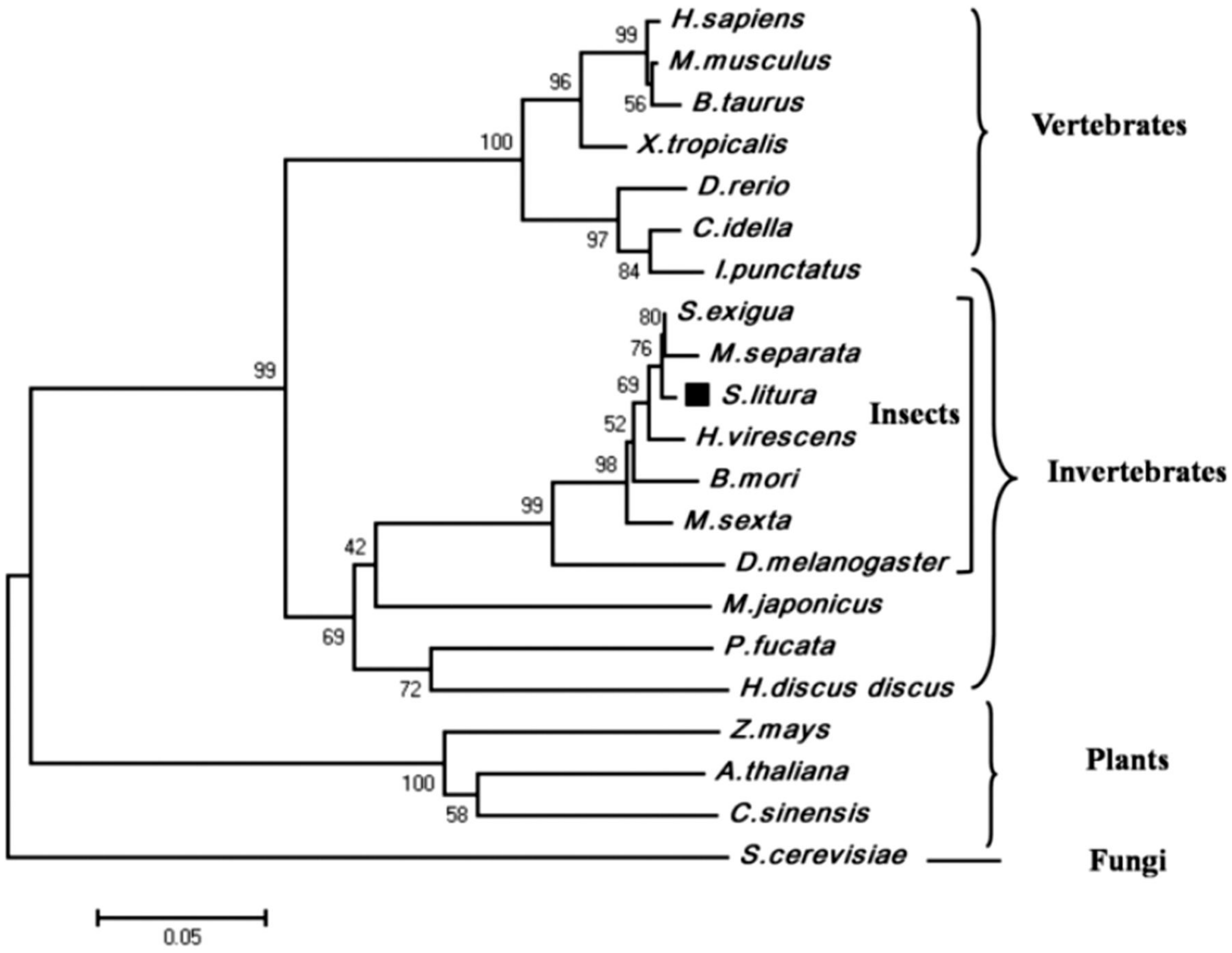
b) (3pt) Would the mutation rate be greater or less than the observed substitution rate for mammals? Why?

Greater. The mutation rate reflects the probability of a change in genome sequence between a parent and its offspring. It's the result of unrepaired DNA damage, polymerase errors, movements of transposable elements, or other molecular processeses that introduce errors during transmission of genetic info. Only the mutations in lineages that persist contribute to the substitution rate measued by whole-genome sequencing. Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4239992/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4239992/)

c) (3pt) From the standpoint of constructing a phylogenetic tree, how many positions (columns) in this alignment are informative?

We'd be interested in the positions that do not align in all 3 species. In the first block, there are 3 such positions; 2nd block: 2; 3rd block: 4; 4th block: none; 5th block: 2. A total of 11 positions that are informative of evolutionary relationship/ancestry.

5. Consider the following phylogenetic tree:



a) (2pt) Is this a cladogram or a phylogram?

A phylogram. It present branch lengths as proportional to some measure of divergence, seen in the given scale.

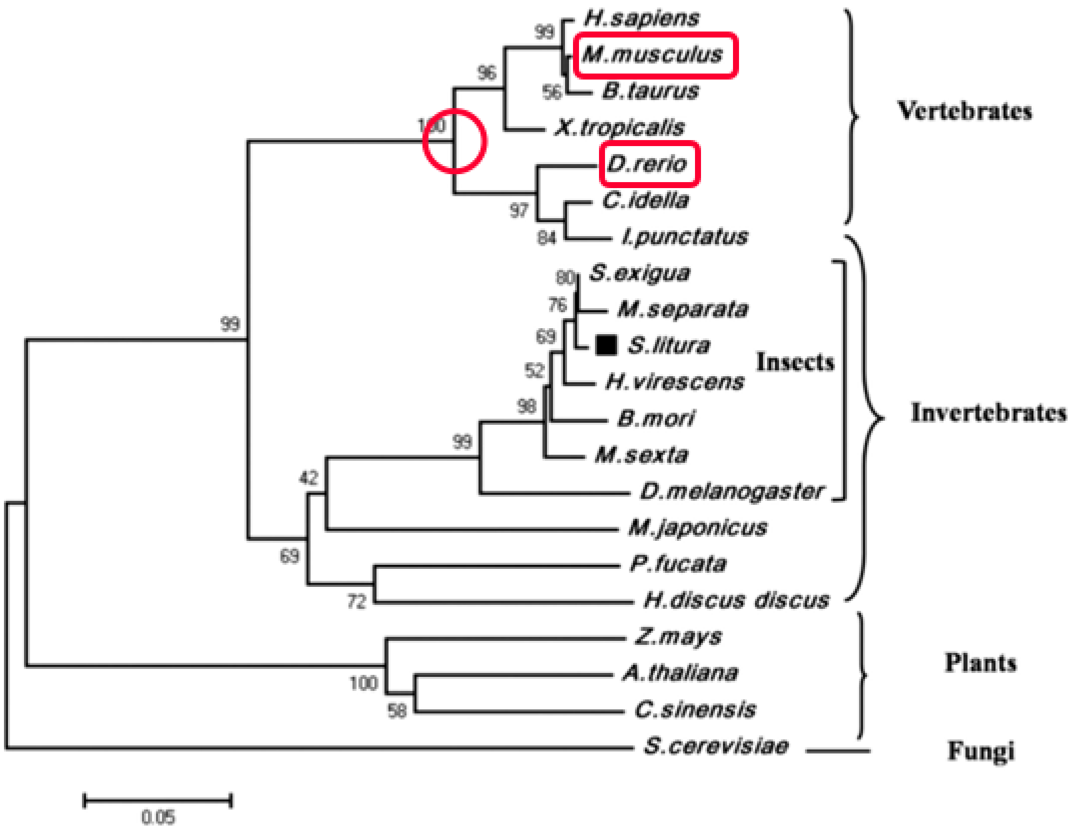
b) (2pt) Which sequence(s) is/are presumably the outgroup?

S. cerevisiae. From the proposed root on the left, this is the species that is distant from the rest of the species.

c) (2pt) Which sequence is most closely related to A.thaliana?

C. sinensis. It shares a common node/ancestor with *A.thaliana* at node near the number 58.

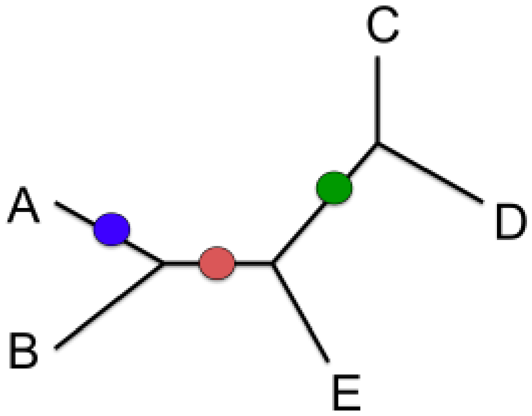
d) (2pt) Circle (on the tree above) the last common ancestor of M. musculus and D. rerio.



e) (2pt) Which branch(es) do you have the least confidence in? Why?

Based on the general principle of tree testing, long branches indicate stronger evolutionary signals but are more errone-prone. For this phylogram, I have the least confidence in the longer branches, such as the one for *S. cerevisiae* and generally the lower half of the phylogram consisting of plants and some inveterbrates.

6. Consider this unrooted tree:



a) (4pt) (Ignore the colored dots for this part.) How many unrooted and rooted trees are possible for this many operational taxonomic units (OTUs)?

Number of OTUs = n = 5

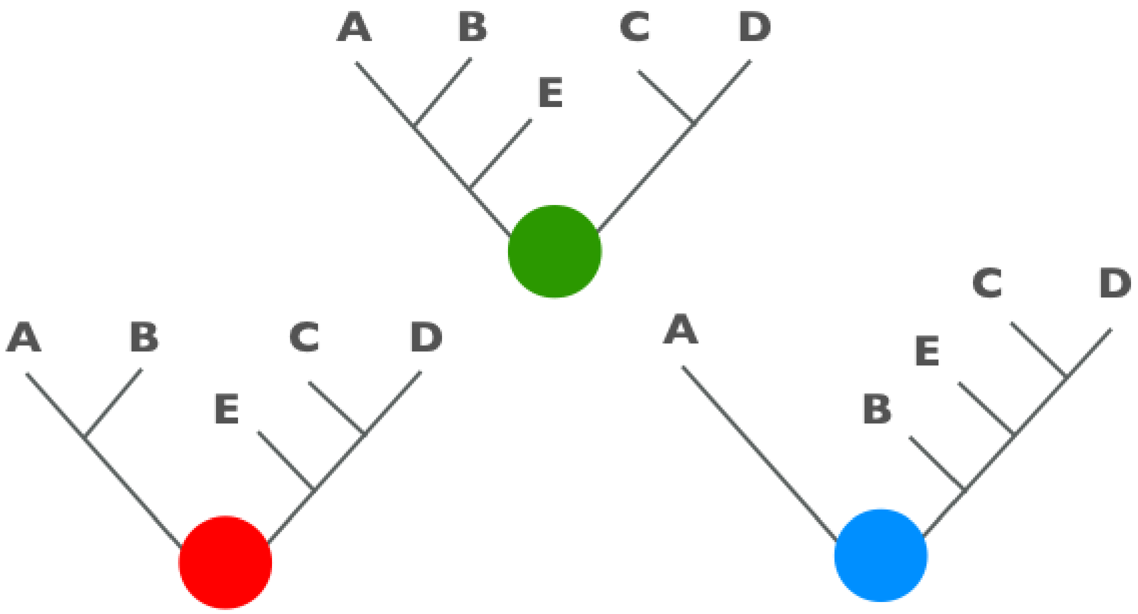
Rooted trees, given that n $\geq$ 2 OTUs:

$$\begin{aligned} N_R &= \frac{(2n-3)!}{2^{n-2}(n-2)!} \\ &= \frac{(2*5-3)!}{2^{5-2}(5-2)!} \\ &= 105 \end{aligned}$$

Unrooted trees, given that n $\geq$ 3 OTUs:

$$\begin{aligned} N_R &= \frac{(2n-5)!}{2^{n-3}(n-3)!} \\ &= \frac{(2*5-5)!}{2^{5-3}(5-3)!} \\ &= 15 \end{aligned}$$

b) (6pt; 2pt per node/tree) Draw the three rooted trees that arise by placing the root at each of the three labeled colored dots (blue, red, green).



7. (Advanced) Consider the two state HMM describing DNA sequence that was discussed in class. Namely where one state was GC-poor (we will call this state L) and one state is GC-rich (we will call this state H).

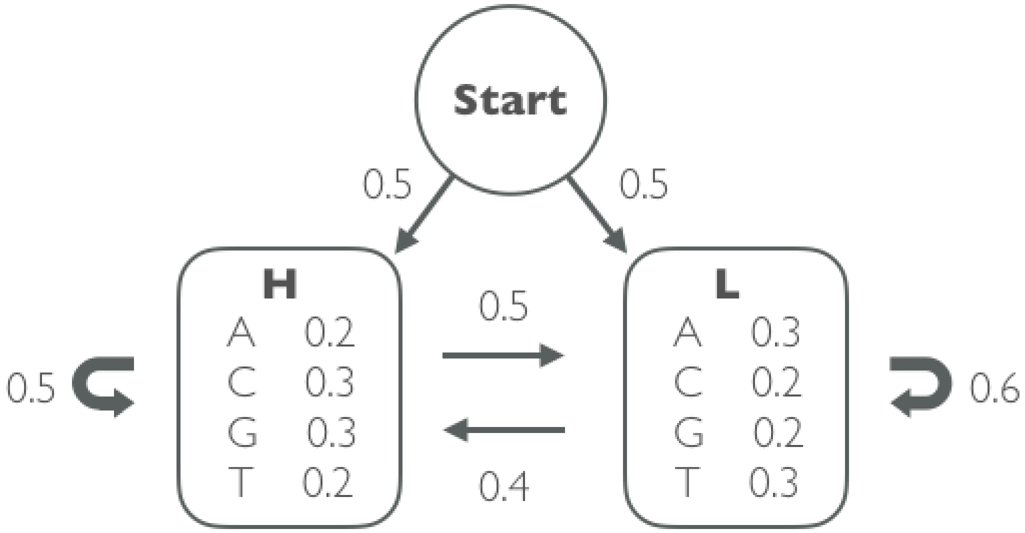
Consider the following parameters of the model:
T(H,H) = 0.5
T(H,L) = 0.5
T(L,H) = 0.4
T(L,L) = 0.6

Emissions:

	A	C	G	T
H	.2	.3	.3	.2
L	.3	.2	.2	.3

The probability of starting in H or L is 0.5 => T(0,L) = 0.5 T(0,H) = 0.5

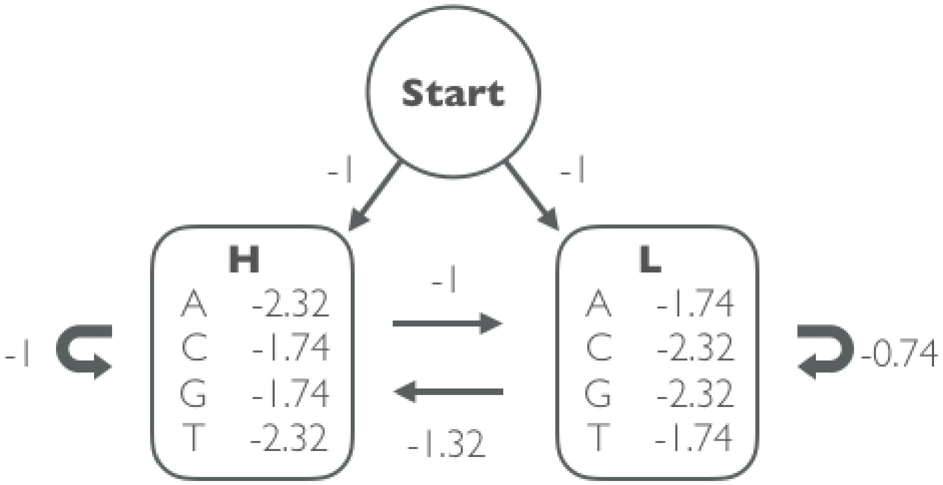
a) (2pt) Draw the HMM state diagram corresponding to this information.



b) (8pt) What is the most likely path for the sequence GGCACTGAA?

Viterbi algorithm:
A = state 1; B = state 2; i = element 1; j = element 2
$P_A(i,x) = e_A(i)max_B(P_B(j,x-1).P_{AB})$

Using $log_2(P)$ for the probabilities for convenient calculations with sum instead of product.

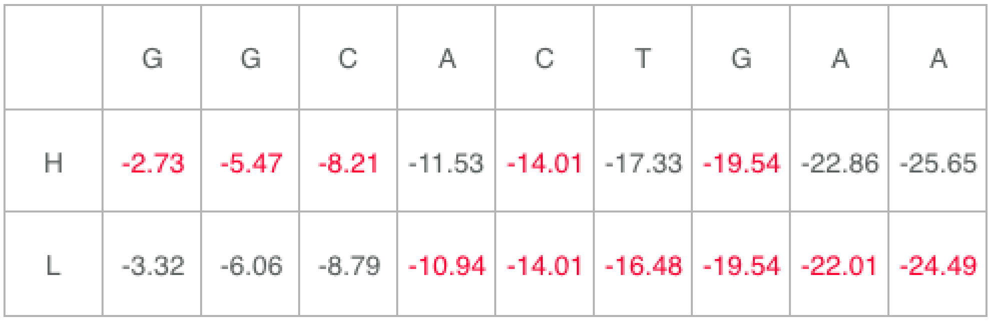


Probability that G in first position is emitted by state H: $P_H(G,1) = -1 + -1.74 = -2.74$
Probability that G in first position is emitted by state L: $P_L(G,1) = -1 + -2.32 = -3.32$

Probability that G in second position is emitted by state H:
$P_H(G,2) = P_H(G) + max(P_H(G,1) + P_{HH},\ P_L(G,1) + P_{LH})$
$= -1.74 + max(-2.74 - 1, -3.32 - 1.32)$
$= -5.47$

Probability that G in second position is emitted by state L:
$P_L(G,2) = P_H(G) + max(P_H(G,1) + P_{HL},\ P_L(G,1) + P_{LL})$
$= -2.32 + max(-2.74 - 1.32, -3.32 - 0.74)$
$= -6.06$

By this same calculation, we can end up with this matrix and traceback to find the path that corresponds to the highest probability at last position: -24.49:



	G	G	C	A	C	T	G	A	A
H	-2.73	-5.47	-8.21	-11.53	-14.01	-17.33	-19.54	-22.86	-25.65
L	-3.32	-6.06	-8.79	-10.94	-14.01	-16.48	-19.54	-22.01	-24.49

Most likely path: HHHLLLLLL. But since at position 5 and 7 we have equal values for states H and L, we may also have the path: HHHLHLHLL.

I implemented this in Python using its hmmlearn module and I get something close to the manual calculation above: HHLLLLLLL.

In [8]:

```
from __future__ import division
import numpy as np
from hmmlearn import hmm
```

In [41]:

```
states = ["H", "L"]
n_states = len(states)

observations = ["A", "C", "G", "T"]
n_observations = len(observations)

start_probability = np.array([0.5, 0.5])

transition_probability = np.array([
  [0.5, 0.5, 0.4],
  [0.5, 0.5, 0.6]
])

emission_probability = np.array([
    [0.2, 0.3, 0.3, 0.2],
    [0.3, 0.2, 0.2, 0.3],
])

model = hmm.MultinomialHMM(n_components=n_states)
model.startprob_=start_probability
model.transmat_=transition_probability
model.emissionprob_=emission_probability

# predict a sequence of hidden states based on visible states
sequence = np.array([[2,2,1,0,1,3,2,0,0]]).T

model = model.fit(sequence)
logprob, path = model.decode(sequence, algorithm="viterbi")
print("Observed sequence:", ", ".join(map(lambda x: observations[x], sequenc
print("Probable path:", ", ".join(map(lambda x: states[x], path)))
```

```
Observed sequence: G, G, C, A, C, T, G, A, A
Probable path: H, H, L, L, L, L, L, L, L
```

8. (Advanced) (10 pt) Consider the following distance matrix:

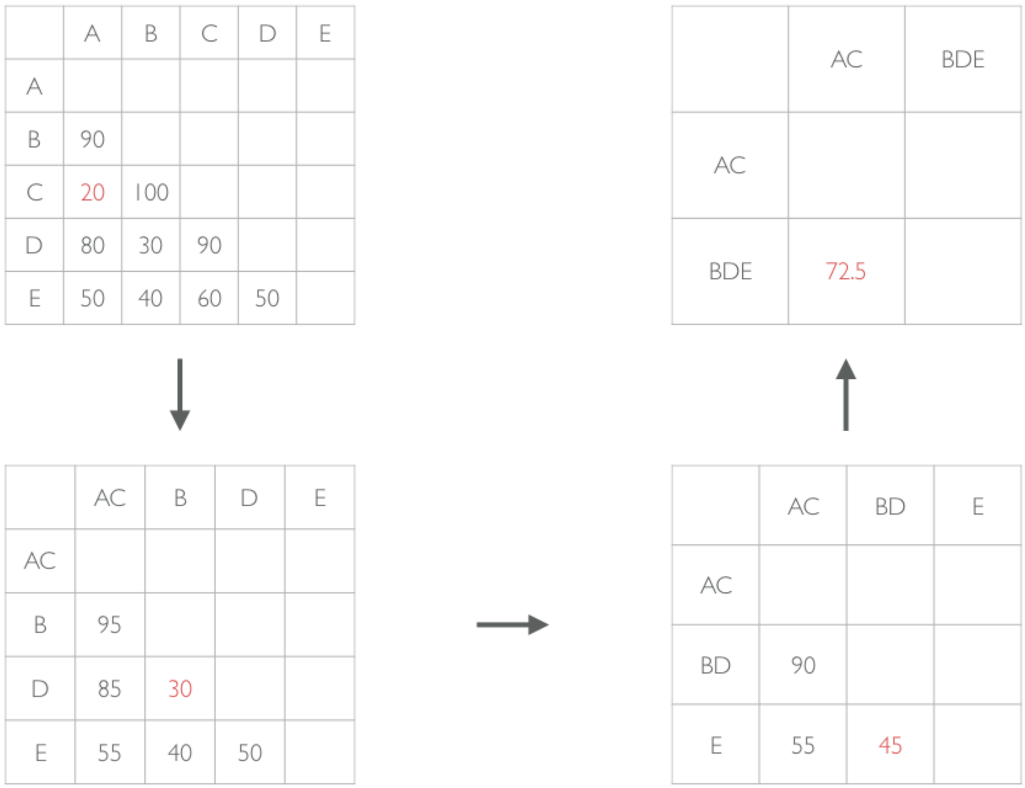
	A	B	C	D	E
A	-				
B	90	-			
C	20	100	-		
D	80	30	90	-	
E	50	40	60	50	-

Calculate a rooted tree using the UPGMA method of tree construction. For full credit you must show the final topology of the tree, the calculated branch lengths, the location of the root, and ALL intermediate matricies utilized in its construction.

UPGMA basic assumptions:

- Change in characteristics occurs in lineages over time

- Rate of mutations is contant ("molecular clock hypothesis"). All leaves have the same distance from the root.
- When a lineage splits, it divides into 2 groups.
- Tree is rooted: last common ancestor is known.



	A	B	C	D	E
A					
B	90				
C	20	100			
D	80	30	90		
E	50	40	60	50	

	AC	B	D	E
AC				
B	95			
D	85	30		
E	55	40	50	

	AC	BD	E
AC			
BD	90		
E	55	45	

	AC	BDE
AC		
BDE	72.5	

Average distance sample calculation:
dist(AC, B) = (dist(A, B) + dist(C, B)) / 2 = (90 + 100) / 2 = 95