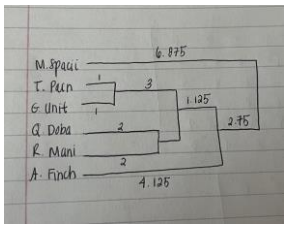
Homework 4: Trees and Models
(100 points total)

Assignment guidelines

1) Submit your assignment files on canvas under module 9: Trees and Models
2) Please submit your code in file(s) called [name].py. Your code should be easy to open in a text editor so that someone can download and use the function you write.
3) Please submit a pdf with the answers to the questions at the bottom of the assignment (and your visualization)
4) Please submit a text file with the output of your UPGMA code called [name]_UPGMA.txt
5) Please submit a text file with the output of your classification code called [name]_Model.txt

Complete the class assignment: UPGMA (50 points)

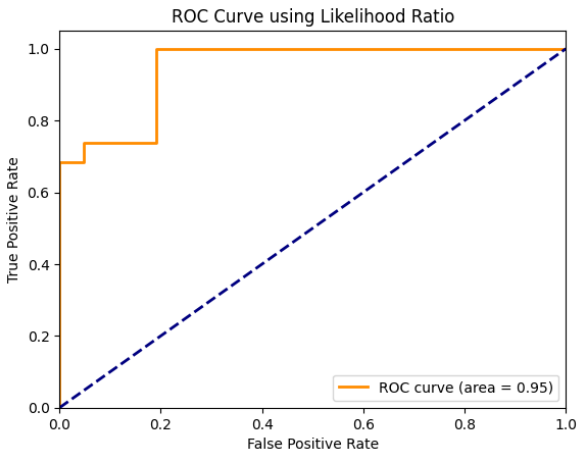
1) Complete the functions find smallest and update matrix (30 points)
    a. Code meets specifications – the function exists and makes correct input and output
        i. Code includes findSmallest() function which takes in a distance matrix and identifies the row and column of the smallest entry – chooses randomly if multiples ones are equal. (5 points)
        ii. Code includes updateMatrix() function which takes in a distance matrix row and column and removes those species from the matrix replacing them with the average distances to all the remaining points (15 points)
    b. Correct tree is returned (10 points)
2) Edit the program so that distances are returned (20 points)
    a. Output is correct with distances being added to the species tree in bracket notation like: ((species_1,species_2:distance), species_3:distance) (10 points)
    b. Draw the correct tree (by hand is fine). Remember that it should be ultrametric! (10 points)
        i. It is a little unclear, but the 3 includes the 1 and is representative of the distance from the beginning to the grouping of T. Pain and G. Unit as well as Q. Doba and R. Mani.



Complete the class assignment: Models (50 points)

1) We have two models (Coding and Non-coding) that represent the likelihood of a codon changing from a given codon to any different codon including itself. They represent the probabilities at time **t** where t is the distance of divergence between the ancestor and M.Spacii. Complete the functions to calculate the log likelihood of the coding model describing the sequences (20 points).
    a. Code meets specifications:
        i. scoreModels function is completed (5 points)
        ii. Program outputs a list of sequences by ID with an associated label "likely coding" or "likely non-coding" (5 points)
        iii. Resulting classification is correct (10 points)
2) Evaluate your results (20 points)
    a. For this section use the data file Spacii_2100.fa – a more divergent set of sequences.
    b. Consider that sequences marked with _n_ are truly non-coding and _c_ are truly coding. Conceptually, explain why some are misclassified (10 points)
        i. There are a few biological reasons as to why sequences might get misclassified. For instance, there are some non-coding sequences that may have functional elements such as enhancers and promoters. These may have similar patterns to coding sequences based on motifs. Also, if sequences are shorter, there may not be enough context to correctly classify them as coding or non-coding.
        ii. Since we are also using a more divergent set of sequences, the likelihood of codon transitions from A to B is greater than codon transitions from A to A. Coding matrices do not represent these differences, and so there are lower scores for the coding sequences. As a result, the model may have trouble distinguishing the scores between coding and non-coding.
        iii. However, misclassification may have some implications in downstream analysis of drug discovery or understanding gene regulation properties.
    c. Create an ROC curve with the results using the likelihood ratio as a threshold (what is the true and false positive rate for various threshold values). You are not required to code this. Based on the ROC curve, what cutoff would you use? Justify your answer – be sure to explain whether you might care more about specificity or sensitivity! (10 points)
        i. The best value for the cutoff would be the point that is furthest away from the curve that optimizes the true positive rate and minimizes the false positive rate. When I use the Spacii_2100.fa file with more divergent sequences, the model classified all the sequences as non-coding when some could be coding. This is a case in which sensitivity was prioritized.

ii. Based on the ROC curve, we want the cutoff to be 0.7, which is the point in which the TPR value is the highest and the FPR still remains 0.



iii. In this case, sensitivity represents the true positive rate and is important for correctly identifying sequences that are coding as coding. It helps when identifying rare non-coding elements with functional regions. If there are sequences that are misclassified or missed, then we could have an incomplete analysis of whatever disease or gene being studied.

iv. Specificity is important when we want to avoid false positives, so it is indicative of the true negative rate. If we misclassify a coding sequence as non-coding, then we could be neglecting coding variations related to certain genes and diseases.

v. I believe that specificity is more important in this case, because we want to ensure we are minimizing misclassifications. If that means missing a few sequences but classifying the ones we have correctly, then the tradeoff is worth it. If there is a low specificity, then the model might incorrectly identify non-coding sequences as coding and lead to inaccurate gene predictions.

3) Consider the models at time **t'** > **t** : (10 points)

a. Describe what the coding matrix will look like at time 2**t**. Show how you would calculate probability of TTT staying TTT after time 2**t**. No need to write out the full equation – you may use (…) – just show me enough to make it clear you follow the calculation by listing a few terms. (5 points)

i. The likelihood of TTT remaining as the TTT codon at time 2t is lower than at time t, whereas the likelihood of TTT altering codons will be greater than at time t. As time progresses, there is a greater chance that TTT will turn into another codon because of mutations. Here are some of the equations that demonstrate this idea. We can conceptualize this idea by referencing the Jukes Cantor Model we learned earlier in the class.

ii. In this model, we are able to see the rates in which transitions and transversions occur for the 4 nucleotide bases. If we assume equal probability, then the rate of A staying A or G staying G (etc.) would be 25%.

iii. When we apply this idea to codons, we assume that at time t=0, TTT remaining TTT has a probability of 1, or 100% certainty. As time progresses, however, the probability decreases but never really becomes 0. We assume that the probability of TTT staying TTT at time infinity would be 1/64 because there are 64 codon possibilities. Therefore, at any of the intermediate steps, such as 2t, the likelihood has to be greater than 1/64 but not 1.

$$P(TTT, 2t) = P(TTT, 2t-1)\,P(TTT \mid TTT, t) + P(TTC, 2t-1)\,P(TTT \mid TTC, t) + \vdots + P(\chi, 2t-1)\,P(TTT \mid \chi, t)$$

where χ represents a non-TTT codon

$$P(TTT, 2t) = \sum_{\chi} P(\chi, 2t-1)\,P(TTT \mid \chi, t)$$

χ: 63 codons

b. Make a prediction about your classification success if the sequences had longer to diverge. Explain how you would expect the ROC curve to change as **t** increases. (5 points)

i. As sequences have more time to diverge, then there may be a greater accumulation of mutations as time progresses. When a greater number of mutations are introduced, it becomes harder to make the distinction between non-coding and coding. As a result, the ROC curve will shift and the area under the curve will start to decrease over time.

ii. As t increases, the ROC will start to shift downwards because of the lack of discrepancy between the two types of sequences. The true positive rate will also decrease and the false positive rate will likely increase.