Sequence Alignment

1. What are the applications of pairwise sequence similarity problem?
2. Why is local alignment more preferable than global alignment? What are the differences in dynamic programming formulation for these two problems?
3. Why do we need to address the affine gap penalties in sequence alignment problem? How does the dynamic programming formulation and runtime changes in this formulation?
4. Explain the application of multiple sequence alignment problem. Why we can not use dynamic programming to solve this problem?
5. What is BLAST? How does it save time while searching on a large database for sequence similarity?
6. Find the optimal global alignment of the following two sequences using -2 as the gap penalty, -1 as the mismatch penalty, and 2 as the score for a match. You have to use the Needleman-Wunsch algorithm, and show the corresponding dynamic programming (DP) table. In case there are multiple optimal alignments, find all of them. Please mark the paths, which correspond to the optimal alignments, in the DP table.
   ATCCTGT
   TCTCT
7. Consider the sequences and the scoring scheme mentioned in 6. Find the optimal local alignment(s) of these two sequences using Smith-Waterman algorithm. Show the DP table and mark the path(s) that corresponds to the alignment.
8. Now solve the global alignment version of the problem again for the affine gap penalty. Consider gap opening penalty -2 and gap extending penalty as -0.5.

Phylogeny Construction

1. Consider the following distance matrix on four taxa.

	a	b	c	d
a	0	4	5	4
b		0	3	4
c			0	5
d				0

Apply an additive phylogeny algorithm and find the tree.

2. Use the same matrix in the previous question and apply the UPGMA algorithm to find the tree.
3. What are the possible applications of a phylogenetic tree? How related are hierarchical clustering algorithms and phylogeny construction algorithms?

4. What is an ultrametric tree? How is that relevant in evolutionary understanding?

Hidden Markov Models

1. What is the log-odds ratio? How can it be used to find states from observations in HMM? What are the problems with using log-odds ratio to decode HMM states?
2. We have a 2-state HMM (Background B, Island I). Initial probabilities: P (B) = 0.5, P (I) = 0.5. Transition probabilities: P (B → B) = 0.85, P (B → I) = 0.15; P (I → I) = 0.85, P (I → B) = 0.15. Emission probabilities:
   E_B : P (A) = 0.25, P (C) = 0.25, P (G) = 0.25, P (T) = 0.25
   E_I : P (A) = 0.1, P (C) = 0.4, P (G) = 0.4, P (T) = 0.1
   Sequence: x = G C T
   Fill the Viterbi DP table and report the most likely state path.
3. Suppose you have the following sequence
   ACCGCCGT
   Considering all distributions uniform, run 2 iterations of the Baum-Welch Algorithm (use pseudo-count = +1).
4. What is the runtime complexity of the viterbi algorithm?
5. Can Baum-Welch algorithm guarantee optimal model and parameters? What is the solution to this problem?
6. Viterbi algorithm may result in underflow. What is the solution to this problem?
7. What is the space complexity of the Viterbi algorithm? Explain.

Supervised Learning

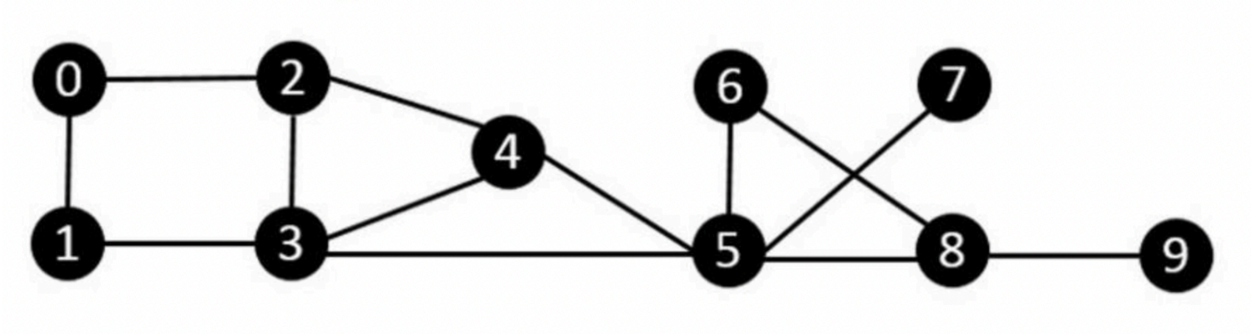
1. What is the difference between a classification and regression problem?
2. Is it possible to solve supervised classification using k-means clustering? How?
3. In the K-Nearest Neighbor algorithm, how to decide the value of K?
4. Are there any parameters of the K-NN algorithm? What is a hyperparameter? What are the hyperparameters of the K-NN algorithm?
5. What are the steps of developing machine learning algorithms for supervised learning?
    a. How to reduce redundancy while creating datasets? How is that ensured?
    b. What are the differences between training, validation and test sets? Why is it important to use random splits?
    c. What is cross validation?
    d. What types of features are used in DNA/RNA/Protein function prediction from sequences?
    e. What is the imbalance in supervised learning in bioinformatics? What are the ways to solve imbalance problems?
6. Suppose you have a nucleotide sequence ACCTGAC. Generate k-mer combination features for this using k=1,2,3.

7. What is 1-hot encoding? Find one hot encoding representation for the nucleotide sequence given in the previous question.
8. Consider the following data from a rna-seq experiments:
   Sample A: [2.0, 3.0, 2.5, 3.5]
   Sample B: [2.2, 3.1, 2.7, 3.6]
   Sample C: [5.0, 5.5, 5.1, 5.4]
   Sample D: [8.0, 5.5, 7.8, 3.7]
   Sample E: [10.0, 4.5, 7.8, 4.5]
   Sample F: [6.0, 2.5, 2.8, 9.7]
   Here sample A,B,C are from healthy individuals and sample D,E,F are from diseased individuals. Now Use k=3 and k=5 to find the class of a sample X =[5,4,3,2]
9. What effect does scaling have on the features of a classifier? How will that affect K-NN classification described in the previous question.

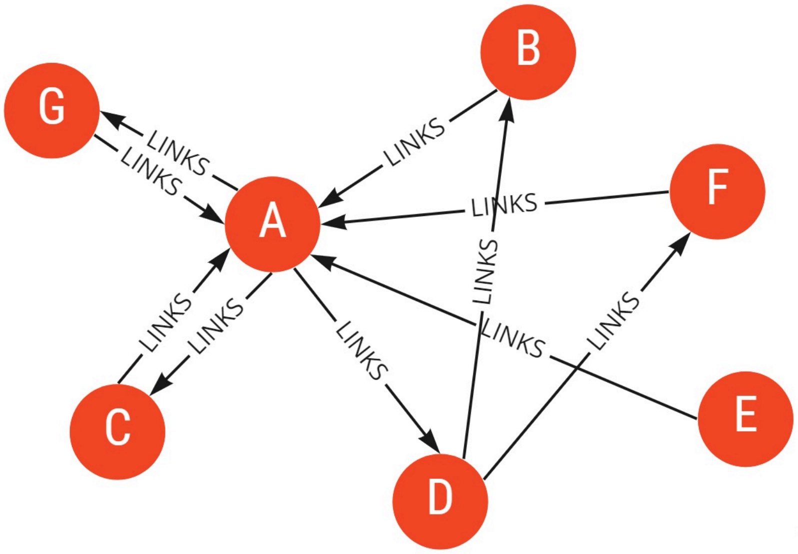
Neural Networks and Autoencoders

1. What is the working principle of a logistic regression classifier? How is it different from a linear regression model?
2. What is a regression problem?
3. What is the application of gradient descent algorithm in solving classification and regression problems? What is the effect of learning rate on gradient descent algorithm?
4. “Neural Network is an ensemble of logistic units” - explain this with an example.
5. What is the purpose of the sigmoid function?
6. Consider the classification problem in Question 8 of the previous section. You decided to use a single hidden layer neural network where the hidden layer will have 3 neurons and the output will be a single sigmoid unit. How many parameters gradient descent requires to learn for this problem?
7. What is a deep neural network? What does the output layer of a neural network look like? Does it depend on the taskt? Explain with examples.
8. What is an undercomplete autoencoder? Why is the bottleneck layer needed? What is the application of such a network?
9. What is a variational auto-encoder? How can it be used to generate synthetic data?
10. In RNA-Seq data, propose how denoising auto-encoders can play a role to improve data quality.

Network Analysis



1. For the above graph perform the following:
    a. Find the diameter
    b. Find closeness centrality of node 4
    c. Find betweenness centrality of node 4 and edge between 4-5
    d. Rank the nodes and edges (use any of the centrality measures)
2. Consider the following network:



   Now simulate random walk on this network to rank the nodes. Go up to 2 iterations. Initialize the probability distribution as 1/7 initially and after each iteration, show the updated values.
3. Provide examples of directed , undirected and weighted networks in Biology.
4. What are the types of analysis that are done on Biological Networks?
5. What are the advantages and disadvantages of a scale-free network?
6. Explain the small world effect and its implications for Protein-Protein Interaction Networks.