Introduction to Molecular Biology

1. Use the Codon Table to find the mRNA and gene that might translate to the following protein: VKLFPWFNQY
    a. How many possibilities are there for this gene/mRNA?
2. What is the reverse complement of the following dna sequence: CTGGTCAGGTCA? It is a standard to write the DNA sequence from 5' → 3'. Write your reverse complement following that.
3. In general, the genes or coding regions of the DNA to be transcribed to mRNA and then translated into proteins. The number of genes in an organism is small compared to the number of proteins. How are these large numbers of proteins then translated?
4. What are the fundamental tasks of the following entities?
    a. DNA Polymerase
    b. RNA polymerase
    c. mRNA
    d. tRNA
    e. sRNA
    f. Ribosome
    g. Proteins
    h. Transcription Factors
5. How many k-mers are possible for k=4 for DNA, RNA and Protein sequences?

Finding Origin of Replication

1. Draw the skew diagram for the following genome: CGGCAGCGTCGCCACACGATCGATCGATCGCGTCGTCGATGCATGCTA
2. What is the runtime of the these methods shown in class:
    a. PatternCount(Text, Pattern)
    b. FrequentWords(Text, k)
3. Find out all k-mers with 2 or less hamming distance from the following k-mer: ACGT
4. What role does the directionality of DNA play in finding origin or replication?

Finding Regulatory Motifs

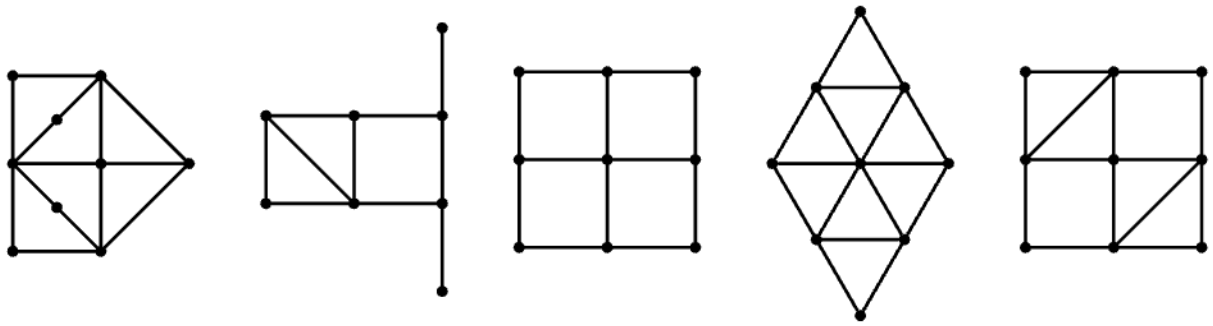
1. Consider the following set of motifs: ACGT, ACCT,AGGT, ATAC, ACGG. Now answer the following:
    a. What is the score of this motif set if we use the hamming distance version?
    b. Calculate the profile matrix for this set.
    c. What is the score of this motif set if we use the entropy version? (use laplacian correction, constant = 1)
    d. If we have another k-mer, ACCC, what is the probability, P(motif=ACCC|profile)=?

2. Both of the problems, *regulatory motif finding* and *clump finding for Oric* uses frequent words with approximate matches. Why is this algorithm more suitable for the latter problem compared to the first one?
3. Suppose you have to find a k-mer which is a suspect motif in the following strings:
   ACCAGTCG
   TCGGTACG
   CAGTCGAT
   CAGTCAGT
   CGATCGAT

   Now simulate the following algorithm up to 2 iterations:
   a) Greedy Motif Search
   b) Monte Carlo or Randomized Motif Search
   c) Gibbs Sampler (for randomized generation, find out the probabilities and take the most probable one from the normalized probabilities)
4. In Gibbs Sampling, after creating the profile matrix, the k-mer probability distribution does not add up to 1. Why?
5. There is a parameter in Gibbs Sampler which is k. How can we fix this k? What happens with the increase and decrease of k?
6. What problems might arise in Motif Finding with randomized algorithms if we have gene upstream regions with skewed distribution of nucleotides?

Genome Assembly

1. Why is de bruijn graphs preferred over overlap graphs in genome assembly problems?
2. Construct Overlap graph and de-bruijn graphs for the following DNA string:
   CGATCGATGTGTCGAT
   Assume k=3 and enumerate all kmers first. Also try to draw the de-bruijn graph for paired reads considering (k=3,d=1) reads.
3. Assume you have the following reads (k=10)
   GTGTGGGGCA
   GGGCAGCGAG
   GCGAGGTTTT
   You don't have all possible reads generated. Assume there is a full coverage with these 3 reads. Now, i) create a overlap graph ii) considering shorter k-mers (k=3) create a de-bruijn graph from these three reads. For both i) and ii) construct the superstring.
4. What happens to the de-bruijn graph if read length increases? Why we can not increase read length?
5. Does de-bruijn graph guarantee unique euler paths? How does that effect genome assembly? What are the solutions to this problem?
6. How does creating shorter reads from original reads help in genome assembly?
7. Explain contigs and bubbles and reason behind them.
8. For each of the following graphs, find out if they have a euler path/cycle and hamiltonian path/cycle and if yes, find that.



9.

Sequencing and Beyond

1. You are given a data file with many low-quality reads (mean Q < 15). Describe how you would preprocess this data before assembly or variant calling.
2. What is the purpose of the Phred+33 encoding scheme in FASTQ files?
3. If the quality score of a base is 30, what is the estimated probability that the base is incorrect?
4. Why do most sequencing reads have lower base quality toward the end?
5. Why is it necessary to check the reverse complement of a read during genome mapping?
6. A 100 bp read has high quality in the first 70 bases (Phred > 30) and drops to Phred < 15 in the last 30 bases. Would you prefer to align the full read or a trimmed read to the genome? Justify.

String Matching

1. “Spaced K-mer indices can save space and time in read alignment” - do you support this statement? Please justify your answer.
2. How exact matching algorithms can be used for approximate matching in an efficient way?
3. For the following pattern and text, illustrate the working principle of the Boyer-Moore algorithm and show how much is saved using the Boyer-Moore algorithm.
   Text: ACCGCGAGCGACGAGC
   Pattern: CGAGC

4. There are some pros and cons using Hash Tables and Multi-Maps for index data-structure. Explain each with examples.
5. What are the ways to reduce the size of the index table? How does that affect the search efficiency?

Clustering Gene Expression Data

1. Consider these data points, (1,2), (5,6), (4,3) and (6,4). Now initialize any two of them as cluster centers for a clustering with k=2. Now iterate 1 step for K-Means and Soft K-Means clustering (using beta=0.25)
2. For the previous data, if initially the first point is selected as a cluster center, what are the probabilities of the other points being selected as the second center?
3. What are the limitations of a Hard K-Means Clustering Algorithm? How can each of these be mitigated or handled to improve the performance of the algorithm?
4. Why is clustering genes important in Biology?
5. What are the biological motivations for Soft-Kmeans clustering and hierarchical clustering?
6. You are given expression values for 4 genes across 4 experimental conditions:

   GeneA: [2.0, 3.0, 2.5, 3.5]
   GeneB: [2.2, 3.1, 2.7, 3.6]
   GeneC: [5.0, 5.5, 5.1, 5.4]
   GeneD: [8.0, 7.5, 7.8, 7.7]

   Perform Agglomerative Hierarchical Clustering using, distance metric: Euclidean distance, linkage method: single linkage (minimum distance between any two points in clusters) Approach: Bottom-up (merge closest clusters).
7. Repeat the previous problem using average distances.
8. What are the possible disadvantages of hierarchical clustering?

Dimensionality Reduction

1. Why do we need dimensionality reduction for gene expression data?
2. Among three methods shown in class, PCA, MDS and SNE, which do you prefer and why? Explain pros and cons of each method. What do each of these methods try to minimize?
3. What is KL-divergence? Explain how SNE works with an example of input space of two dimensions and output space of one dimension. Illustrate using 3 sample data points: (2,1), (3,4), (0,3).
4. Explain MDS and PCA with the same data points. Explain the working principles only.
5. Why is distance based metric like euclidean distance / L1 metric/L3 metric not suitable for SNE?
6. Explain Soft Neighborhood. Why is the squared distance used in the soft-neighborhood formula?
7. You want to reduce dimensionality using PCA. How do you select the output dimensionality?
8. In gene expression data, suppose you are using PCA to reduce dimensions, is it possible to know which input dimensions or genes are important using PCA?

Nature of Question

1. Origin of replication + Motif Finding
2. Genome Assembly
3. Read Mapping, Alignment
4. Clustering + Dimensionality Reduction