Introduction to Computational Biology and Bioinformatics Bi/BE/CS183
Problem Set 1 Due: Wed, Feb 2/2017

1) What is the time complexity of the following sorting algorithm? Explain the reasoning behind this time complexity, and then write code (pseudocode is okay) for a sorting algorithm that runs in O(nlogn) time.

Given a 0-indexed array of n integers we will sort accordingly:

```
For i = 1 → length(array):
    j = i
    While j > 0 AND array[ j - 1] > array[ j ]:
        SWAP array[ j - 1 ] and array [ j ]
        j = j - 1
```

2) Samtools is a set of utilities that manipulate alignments in the BAM format. Now you are given a file of alignments named P1Q3.bam which can be downloaded from our course website. Read the manual and use Samtools (or any other tool you like) to find the alignments that overlap the region chr17:220-300.
(Hints: You may need to sort and index the .bam file first. The manual of samtools can be found here http://samtools.sourceforge.net/samtools.shtml and a detailed explanation of SAM and BAM format can be found here http://samtools.github.io/hts-specs/SAMv1.pdf.)

3) Given the sequences below

a. Calculate the observed and expected frequency of each possible 4-mer assuming each nucleotide appears at a probability of ¼
b. Compute the frequency of all possible 4-mers in this sequence
c. Are any 4-mers more common than you'd expect? If so, which ones?
    i. Provide a table of all 4-mers, the number of observed occurrences and the expected number of occurrences at random.
d. Highlight the unexpected 4-mer motifs that you discovered

AGTCGTACGTGAC
AGTAGACGTGCCG
ACGTGAGATACGT
GAACGGAGTACGT
TCGTGACGGTGAT

4) Given a file of sequences (sequences.txt, on our course website)

a. Implement a simple method that scores the Hamming distance for a pattern against each subsequence
b. Use your methods to score the following patterns using simple Hamming distance.

   TTGTAGG GAGGACC TATACGG CCGCAGG CAGCAGG
c. Which pattern is most likely to be the implanted motif?
d. What is the entropy of the discovered motif?

5) The randomized motif searching algorithm:

Pick a random position in each sequence to generate an initial k-mer. Create a profile matrix, score each sequence, and update the PWM accordingly.

What is the probability that you will select the correct motif using a randomized search algorithm? How many permutations would you need on average to be likely to approach a correct solution using a randomized search algorithm?

6) Now we will ask you to implement a randomized motif search. Using the file sequences.txt given in problem 5, find the motif using a randomized motif search. What profile matrix is obtained by running the algorithm once? 10 times? 1000 times?

This algorithm includes the following steps:

a. Pick a random position for each sequence to generate a random k-mer
b. Compute the profile matrix for the Motif using the k-mers selected in part a and update accordingly
c. Score each sequence using the final profile matrix

7) Continue using the file sequences.txt given in problem 5, identify the motif in the sequences using the Gibbs Sampling approach. Did you arrive at the same motif as problem 7, why?

8) In Southern California, Mount San Gorgonio is the highest peak, which doesn't have a lot of trees on top of it. Build a suffix trie and suffix tree for gorgonio$, by hand.

9) Suffix Arrays:

a) In any language of your choice, create a suffix array for Gorgonio$ and print out the results (index, element) of your array in ascending order.
b) Next, implement a "Query" method for your suffix array using the binary search method. Please include your well-documented code in your submission.
c) Using your implementation, create a suffix array for the portion of human chromosome 1 found in the .fasta file and query for the following sequences:
   i) atattaacaaagccaaaagtttcaaacttt
   ii) aaaattat

Report the locations (index) of all exact matches for these sequences, along with the code used the retrieve the locations.

10) When searching for a motif using k-mer enumeration, what’s the rationale for using the entropy metric rather than a simple difference score? Provide a simple example that illustrate this.