CS313 Exercise 3 Cover Page

Name(s): ________________________________

In the *TIME* column, please estimate the time you spent on the parts of this exercise. Please try to be as accurate as possible; this information will help us to design future exercises.

PART	TIME	SCORE
Exercise		

Task 1: Global Sequence Alignment with Linear Gap Penalty

For the two nucleotide sequences `CGGCTTG` and `AGGTTC`, fill in the table below to determine (a) the optimal *global* alignment score **and** (b) the optimal global pairwise alignment corresponding to this score. You should assume a match score of +5, a mismatch score of -4, and a linear gap score of -6.

		A	G	G	T	T	C
C							
G							
G							
C							
T							
T							
G							

Task 2: Local Sequence Alignment with Linear Gap Penalty

For the two nucleotide sequences `CGGCTTG` and `AGGTTC`, fill in the table below to determine (a) the optimal *local* alignment score **and** (b) the optimal local pairwise alignment corresponding to this score. You should assume a match score of +5, a mismatch score of -4, and a linear gap score of -6.

		A	G	G	T	T	C
C							
G							
G							
C							
T							
T							
G							

Task 3: Local Sequence Alignment with Affine Gap Penalty

For the two nucleotide sequences `CCTGCAATG` and `CCTAAT`, fill in the table below to determine (a) the optimal *local* alignment score **and** (b) the optimal local pairwise alignment corresponding to this score. You should assume a match score of +5, a mismatch score of -4, and affine gap scores of *alpha*=-7 (the gap open penalty) and *beta*=-2 (the gap extension penalty).

		C	C	T	A	A	T
C							
C							
T							
G							
C							
A							
A							
T							
G							

Task 4: Pairwise Sequence Analysis

A linear gap penalty can be viewed as a special case of an affine gap penalty. What values for *alpha* (the gap open penalty) and *beta* (the gap extension penalty) are equivalent to using a linear gap score of -6?

For two sequences, what is the relationship, if any, between their optimal global alignment score and their optimal local alignment score, assuming the same scoring model is used for both global and local alignment? In other words, might the optimal global score be larger than the optimal local score, might the optimal global score be smaller than the optimal local score, might the two scores be equal?

Suppose we use the following scoring model for aligning two sequences: match score of +5, mismatch score of -4, linear gap score of -6. Do you expect the optimal local alignment score of two *short random* sequences to be greater than, less than, or equal to the optimal local alignment score of two *long random* sequences? Why?

Suppose we use the following scoring model for aligning two sequences: match score of +5, mismatch score of -4, linear gap score of -6. Do you expect the optimal global alignment score of two *short random* sequences to be greater than, less than, or equal to the optimal global alignment score of two *long random* sequences? Why?

Suppose we use the following scoring model for aligning two sequences: match score of +8, mismatch score of 0, linear gap score of -1. Do you expect the optimal global alignment score of two *short random* sequences to be greater than, less than, or equal to the optimal global alignment score of two *long random* sequences? Why?

Task 5: Normal distribution

The *normal distribution*, also known as the Gaussian distribution or the bell curve, occurs frequently in many different settings. For example,

- The height of women is approximately normally distributed
- The measure of LDL (bad) cholesterol appears normally distributed in adults
- The width of stripes on a zebra is said to be normally distributed
- Most measurement errors are assumed to be normally distributed

In the figure below, the width of stripes in zebras is illustrated. In Figure 1a, the histogram shows the stripe widths of 1,000 randomly chosen mountain zebras. In Figure 1b, the graph reflects the *percent* of the 1,000 mountain zebras with a given stripe width (rather than the *number* of zebras with a given stripe width). Figure 1b is a normalized version of Figure 1a, i.e., Figure 1b was obtained by dividing each bar in Figure 1a by the total number of zebras sampled, 1,000. Figure 1c is the same as Figure 1b, except that a line graph has been added. The line corresponds to a *normal distribution* that approximates the stripe width data for mountain zebras. From the data in

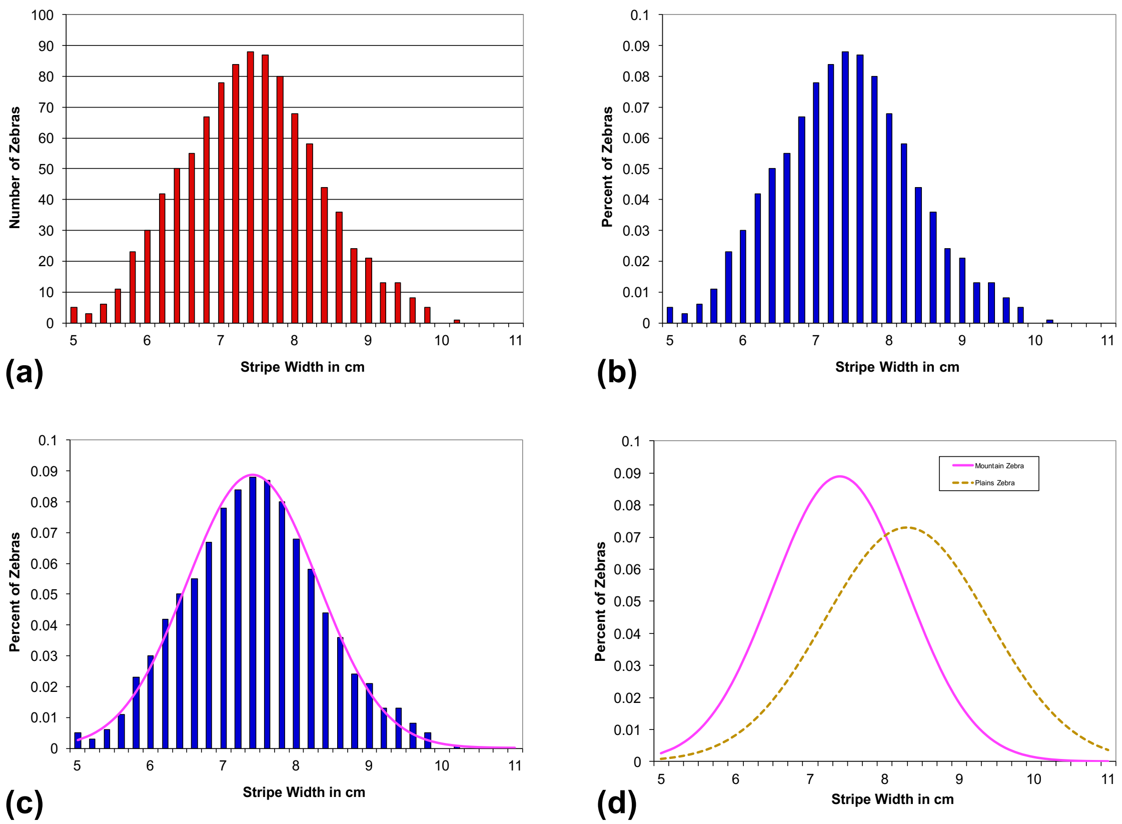

Figure 1. The figure illustrates the width of stripes in zebras. (a) The histogram shows the width of stripes for 1,000 randomly chosen mountain zebras. (b) The graph reflects the percent of the 1,000 mountain zebras with a given stripe width. (c) A normal distribution is shown that approximates the sample distribution of mountain zebra stripe widths. (d) A normal distribution approximating plains zebra stripe widths is shown in comparison to the normal distribution approximating mountain zebra stripe widths.

Figure 1a, the mean stripe width of the 1,000 mountain zebras was calculated (7.3 centimeters) and the standard deviation from the mean stripe width was calculated (0.9 centimeters). A normal distribution (the line graph) was then plotted in Figure 1c to approximate the sample distribution of 1,000 mountain zebra stripe widths. The normal distribution was determined based on a mean of μ=7.3 and a standard deviation of σ=0.9 according to the function $\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$. In Figure 1d, for comparison, a normal distribution approximating *plains* zebra stripe widths is shown along with the normal distribution approximating *mountain* zebra stripe widths.

Approximately how many of the 1,000 sampled mountain zebras have a stripe width greater than or equal to 9 centimeters?

Suppose we come across a mountain zebra while walking around campus. When we meet this zebra, the first thing that jumps into our minds, obviously, is that we should measure the width of its stripes. We determine that the zebra has stripes 9 centimeters wide. The p-value of this measurement is the chance that a mountain zebra has a stripe width greater than or equal to 9 centimeters. Approximately, what is the p-value of our result, i.e., what is the chance that a mountain zebra has a stripe width greater than or equal to 9 centimeters?

Based on the figure above, do you expect the standard deviation of plains zebra stripe widths to be greater than, less than, or equal to that of mountain zebra stripe widths? Why?

Suppose that we meet a plains zebra and observe that it has a stripe width of 9 centimeters. Based on the figure above, approximately what is the p-value of our observation?

Task 6: Alignment Score for a Yeast Gene

Download the `/home/cs313/download/TestAlignment` directory from the CS server. Study the `TestAlignment.java` file in the `TestAlignment` directory. The `TestAlignment` application (1) reads in two DNA sequences from files and computes their optimal global pairwise alignment and (2) reads in two protein sequences from files and computes their optimal global pairwise alignment. Note that the `TestAlignment` class utilizes two other classes, the `SequenceOps` class that we developed in Project #1 and the `Alignment` class whose contract can be found at the following URL: http://cs.wellesley.edu/~cs313/projects/project3/doc/Alignment.html

Execute the `TestAlignment` application and confirm that the optimal global alignment for the two DNA sequences has a score of 6 and the optimal global alignment for the two protein sequences has a score of -171. You can ignore the reported p-values of 0.0, which are inaccurate (you will be calculating the p-values in Project #3).

Using the `TestAlignment` application, calculate the optimal global alignment score for two genes that we studied in Exercise #1: the yeast hexokinase gene, *hxk1*, and its ortholog in *Drosophila melanogaster*, *hex-a*. You will need to download from the appropriate genome websites the DNA coding sequences for the two genes and the protein sequences for the two genes.

What is the optimal global alignment score when the DNA coding sequence of the yeast gene *hxk1* is aligned with the DNA coding sequence of the fly gene *hex-a* (for the fly gene *hex-a*, use the coding sequence, CDS, corresponding to Hex-A-RA rather than Hex-A-RB or Hex-A-RC)?

What is the optimal global alignment score when the protein sequence of the yeast gene *hxk1* is aligned with the protein sequence of the fly gene *hex-a* (for the fly gene *hex-a*, use the sequence corresponding to Hex-A-PA rather than Hex-A-PB or Hex-A-PC)?

Now generate two random DNA sequences. The first random DNA sequence should have the same length and expected GC-content as the DNA coding sequence of the yeast hexokinase gene. The second random DNA sequence should have the same length and expected GC-content as the DNA coding sequence of the fly hexokinase gene.

What is the optimal global alignment score for the two random DNA sequences?