

CSC / BIO 310

Bioinformatics

Instructor: Dr. Laurie J. Heyer

Review II

Due Saturday, April 12, at 12:00 noon.

Instructions:

VERY IMPORTANT: Once you read beyond page 1 in this file, you may not get help from ANYONE. I will answer questions up until the time you read page 2, but not afterward. Read these instructions carefully before you continue.

You may use any openly available inanimate resources to help you with this review.

However, you may not use code from the internet, you may not request a copy of already written code from anyone, and you may not discuss any aspect of the review with anyone other than me.

Take special care to protect your work so others do not accidentally find it laying around in the lab, up on a computer screen somewhere, or on the whiteboard.

You have unlimited time to work on this review, in any number of sittings.

In addition to the accuracy of your solutions, you will be graded on the readability of your code and your output, and the use of good programming practices as discussed in the text and in class.

Page two will specify exactly what you need to turn in to me by email. Failure to follow those instructions will incur a penalty. Late penalty of 10 points per hour.

Email to me the following three files:

(1) A Perl file called *name_r2_logo.pl*, where *name* is replaced with your name, containing your solutions to question 1

(2) A Word document containing your answer to question 3

(3) A Perl file called *name_r2_jc.pl*, where *name* is replaced with your name, containing the programming part of your solution to question 4

Also turn in a hard copy of your answers to questions 2 and 4(b).

QUESTIONS:

1. Go to JASPAR and find the transcription factor binding site frequency matrix for **ELK-1** (ID: MA0028).
 - a. Copy and paste the matrix, using the JASPAR format (including labels, brackets, etc) into a plain text file called **ma0028.txt**.
 - b. Read the sequence logo paper linked to from the course web page to see how to compute the heights of each letter at each position in a logo.
 - c. Write a Perl script to determine the heights, printing them to the screen in a well-organized table. Ignore the correction factor $e(n)$ for small number of sequences, and use the convention that $f(b,l)\log_2(b,l) = 0$ whenever $f(b,l) = 0$. You can verify your answer is correct by comparing your heights to those in the logo at the JASPAR site.

2. Consider the genome rearrangement problem given in #3 of the handout, under “Exercises” and “Presentation Problems”.
 - a. Draw the reality and desire diagram for this permutation.
 - b. Identify the good and bad cycles, good and bad components, hurdles and superhurdles for this permutation.
 - c. Use the information from part (b) to determine the reversal distance between these two gene orders.
 - d. Submit the starting permutation to GRIMM to determine a sequence of reversals that sorts the permutation in the number of steps you found in (c). Print the output of GRIMM to include with your solution.

3. One of the following sequences is a randomly generated DNA sequence, the other is part of a real DNA sequence that I modified in a few places. Which do you think is which? Justify your answer.

```
atgacttgactgactgacgacctatggcgtatagac  
caagccgttcttggtaccgatggtgggagtgtctaat
```

4. Write a Perl script to read a set of FASTA-formatted sequences from a file, and compute all the pairwise Jukes-Cantor distances. You may want to look at code you wrote for the project and/or the first review to help you with this task.
 - a. Use your program to read the sequences in the file phyloseq.txt and compute the distances needed to build a phylogenetic tree of these 5 sequences. Your program should print a nice table of distances like those we have used in class to start the process of building a tree.

You can use fdnadist program at EMBOSS to check your answers, or generate distances for you if you cannot get your Perl code to work.

- b. Using the distances found in part (a), build a phylogenetic tree using the UPGMA algorithm. Show your work.