Homework 5

Skylar Chan

2023-03-03

1 Sed

1. Sed can be used to overwrite files after substituting their text. What option do you use to accomplish this? Also give a sed command that looks like it would overwrite the file with the edited text but gives an empty file instead. Hint: a similar problem appeared in an earlier homework.

Use sed (and only sed) to accomplish the following substitutions. Assume all letters are uppercase. All sed expressions should be enclosed using single quotes `'` to avoid shell expansions, which we will discuss later. Each command should look like `echo -e $your_sequence_here | sed $your_expr_here`, using options as needed, unless otherwise specified (note that `-r` and `-E` are equivalent).

2. Add a start codon (ATG) to the beginning of each line.

3. Add a poly A tail (AAAAA) to the end of each line.

4. Surround each line with a start codon and a poly A tail.

6. Restriction enzymes are enzymes that cut genetic sequences at specific locations. A common restriction enzyme is EcoRI, which cuts all occurences of GAATTC into G and AATTC, and all occurences of CTTAAG into CTTAA and G. Wikipedia has a good picture. Use sed to model EcoRI by inserting a newline in the places where EcoRI would cut.

2 Awk

Read about the GFF3 (click here) and BED (click here) file formats. Then provide a cut or awk pipeline that accomplishes the following. Example BED and GFF files are provided in the following directory:

`/afs/glue.umd.edu/class/spring2023/bsci/238g/0101/public/hw5/example-files/`

7. In a GFF file, count the number of sequences by phase. That is, give the number of sequences of phase 0, 1, and 2.

8. In a GFF file, print the genomic start and end columns.
9. In a BED file, print the score of positively-stranded features.
10. In a BED file, delete all optional columns.
11. Delete the 2nd row and 2nd column of a space-separated file.
12. Convert a tab separated file to a comma separated file.