

Homework 5

Due Nov 11

1. Write a script using the 'csv' package to parse [this *E.coli* vs *Salmonella enterica* BLAST report](#) and [this *E.coli* vs *Yersinia pestis* BLAST report](#) and make a new report which prints out the *E.coli* sequence ID, and a column for the best *Salmonella* hit and the best *Yersinia* hit. best hit for each sequence to a file call 'Ecoli_report.tab'. Also print to STDOUT a summary report with the total number of sequences with a good hit in *E.coli* to each Salmonella and Yersina (e.g. 1000 proteins in Ecoli had a Salmonella hit, 700 had a hit to Yersinia). Screen the search results - don't take a best hit if something is less than 40% identical at the protein level.
2. Using BioPython SeqIO module and any others you need, write a script to parse [the transcript sequences](#) from the Tick *Ixodes scapularis* genome and print out.
 - the max, minimum, and average length of these transcripts.
 - Number of transcripts which start with ATG
 - GC content of the transcripts