I573: Programming for Science Informatics

IUPUI

INDIANA UNIVERSITY
PURDUE UNIVERSITY
INDIANAPOLIS

Assignment 4

- **Plot a frequency distribution using R (with 100 AA intervals i.e, 0-100, 100-200 etc) of the protein lengths in E. coli using the ptt file available from the link below (ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Escherichia_coli_K_12_substr__MG1655_uid225/U00096.ptt)**

Data should be directly imported into R using the above file and the complete code from importing the data to plotting the figure should be submitted.

- **Download the genomic sequence of Escherichia coli K12 MG1655 in *.gbk format from the ftp site of NCBI and parse the file to print for each CDS the following information. The output of the script should be in the below tab-delimited format.**

190..255 thrL b0001 "thr operon leader peptide"
MKRISTTITTTTITITTGNGAG

(ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Escherichia_coli_K_12_substr__MG1655_uid225/U00096.gbk)