UCS654 - Predictive Analytics Using Statistics

Assignment02 - Feature Extraction

General Instructions – Must Read

- **Submission Due Date:** 06 Feb 2022 | 23:59:59
- **Marks:** 05 (Five)
- **Number of Questions:** 01
- **Submission Link:** **Click Here**
- **Submission Guidelines:** You need to submit single R file.
  - Single R (.r) file | File Name must be <YourRollNum>.r | Example: **10155.r**
- Your program must be run from **command line** only:
  - **Usages:** rscript <program.r> <InputDataFile>
  - **Example:** rscript 10155.r input.csv
- **Output File Name:** <output>-<YourRollNum>.csv
  - **Example:** output-10155.csv
- Your program must be capable to handle exception/error (if any) and write to **log file**:
  - Correct number of parameters (inputFileName).
  - Show the appropriate message for wrong inputs.
  - Handling of "File not Found" exception
  - Many sequences are missing in the input file, handle them (ignore them).
  - If any issue with the input record, it must be write to a log file
- **Note:**
  - Multiple submissions are allowed, but **latest submission** will be considered for the evaluation.
  - Submission link will open all the time, but only 50% marks will be awarded if you fail to submit with in the due date. No excuse will be consider for the submission.
  - **Zero marks** will be awarded for plagiarized code or result.

1. **Write a R program that extract the features (aliphatic index, Boman index, hmoment index, peptide charge, etc) of each "Peptide Sequence" given in the input file and create a feature matrix given below.** [Input file is available in "Input for Assignment02" folder]

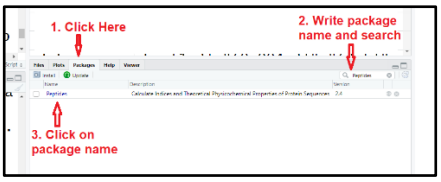
Peptide Sequence	len	Aliphatic_index	Boman_index	hmoment_index	peptide_chrage	molecular_weight	....	.....	.....	.....	Target
TLYGPQLSQKIVQIN	15	123.33	0.68	0.51	1.00	1701.98	....	.....	.....	.....	0
PSWGLVVTMFAWGYLI	24	101.25	-0.28	0.44	-0.91	2764.26	....	.....	.....	.....	1
TMIKTAVAVV	10	146.00	-1.23	0.40	1.00	1032.31	....	.....	.....	.....	1
AGISSLIIDPNPMFV	15	130.00	-0.64	0.23	-1.00	1573.87	....	.....	.....	.....	1
DPMIVGVLFIEIHMM	15	142.67	-1.24	0.13	-1.91	1745.19	....	.....	.....	.....	0
ELNNALQNLARTISE	15	117.33	2.44	0.73	-1.00	1685.85	....	.....	.....	.....	1
AGILLGLFYLVAVAR	15	188.67	-1.86	0.23	1.00	1575.96	....	.....	.....	.....	0
GKAGCQTYKWETFLTSI	20	44.00	1.02	0.30	-0.06	2216.49	....	.....	.....	.....	0
QLSAEYASTAAELSG	15	78.67	0.94	0.26	-2.00	1497.58	....	.....	.....	.....	1
AAVVRFQEAANKQKQ	15	65.33	2.52	0.52	2.00	1687.92	....	.....	.....	.....	1

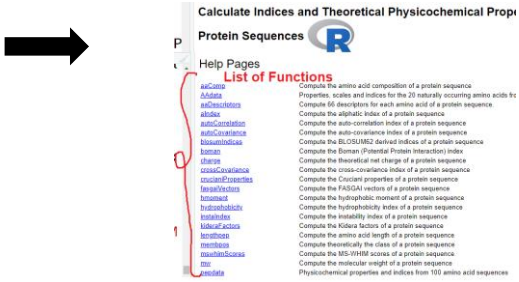
Please note: Steps for feature extraction

1. **First, install two packages: 'Peptides' and 'peptider'**

   install.packages('Peptides')
   install.packages('peptider')

2. **Explore all the functions of both the packages:**





3. **Give the sequence to these functions, get the value, merge them and write to the output file.**

4. **Sample Code**

```
library(Peptides)
library(peptider)
mydata = read.table("input.csv")   # Reading the sequence file
for (sequence in mydata$V1){
        F1 ← aIndex(sequence)            # F1: aliphaticIndex
        F2 ← boman(sequence)             # F2: bomanIndex
        F3 ← instaIndex(sequence)        # F3: instaIndex
        F4 ← ppeptide(sequence, libscheme = "NNK", N=10^8)  # F4: probabilityDetectionPeptide
        allFeatures = data.frame(F1,F2,F3,F4)   # Merging all features
        # Write and append allFeatures to the file
        write.table(allFeatures, "output-10155.csv", sep = ",", row.names=F, col.names = F, append = T)
}
cat("Done")
```