Machine Learning in Computational Biology (Fall 2023)

Assignment #2

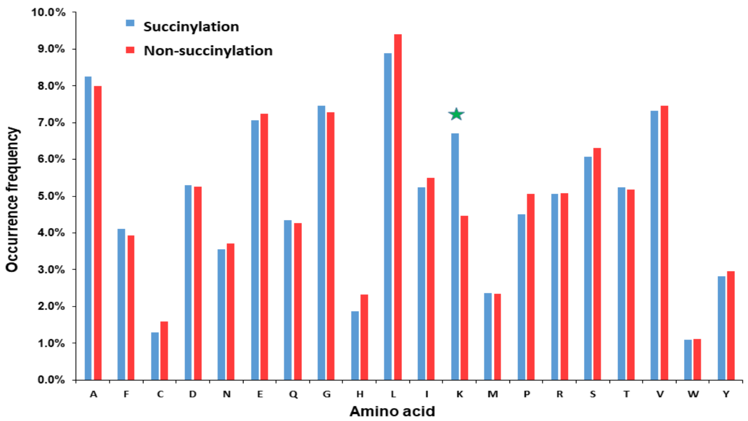
Deadline: 23:59:59 17th November, 2023 (Delayed submission is not allowed for any reason)

Purpose: to enhance the learning outcomes for the topics in “**Features Encoding and Investigation**” and “**Protein Sequence Analysis**”.

After the removal of homologous sequences in both positive and negative dataset using CD-Hit, in order to carry out a binary classification between ubiquitination and non-ubiquitination sites, please accomplish the tasks described as follows.

1. [Amino acid composition] (20 points)

Amino acid composition (AAC) is a common method used to transform protein sequences into 20-dimensional numeric vectors. As you learned from Chapter 2 in this class, please calculate AAC for each sequences in both positive and negative datasets; then, the comparison of AAC between positive (ubiquitination) and negative data (non-ubiquitination) can be displayed as the histogram shown below.

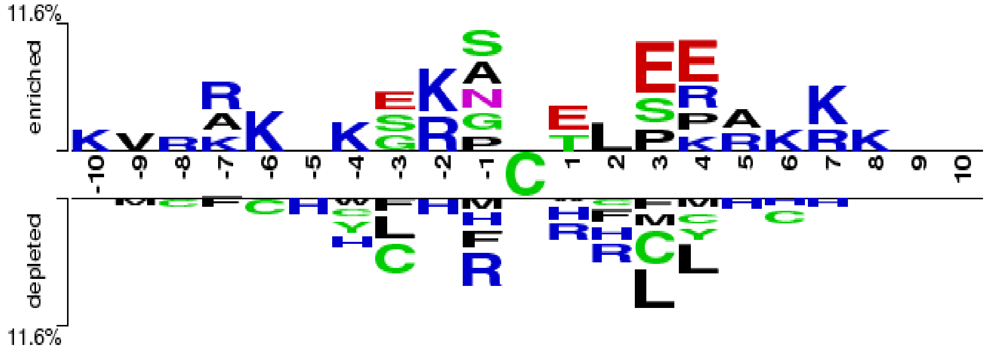


Please provide the histogram plot of AAC based on your training dataset using CD-Hit with 50% similarity.

2. [Sequence log and TwoSampleLogo] (20 points)

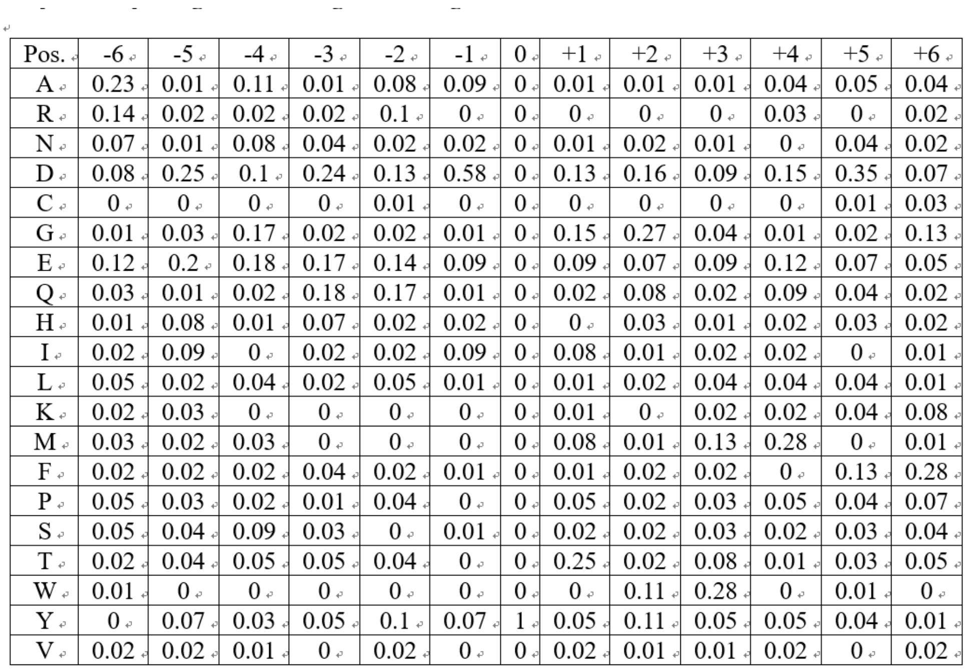
2.1 In order to observe the position-specific AAC for protein ubiquitination sites, the **WebLogo** tool can be utilized to create **frequency** and **entropy plots** of sequence logos for visualizing the potential amino acid motifs surrounding the modification sites (centered as position 0). Please provide the **frequency** and **entropy plots** of sequence logos on both positive and negative datasets.

2.2 In order to investigate the difference of position-specific AAC between ubiquitination and non-ubiquitination sites, please use the **TwoSampleLogo** to visualize potential amino acid motifs surrounding the modification sites as shown below.



3. [Positional Weighted Matrix] (60 points)

3.1 Please create the positional weighted matrix (PWM) for both positive and negative datasets, as shown below.



Pos.	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	0.23	0.01	0.11	0.01	0.08	0.09	0	0.01	0.01	0.01	0.04	0.05	0.04
R	0.14	0.02	0.02	0.02	0.1	0	0	0	0	0	0.03	0	0.02
N	0.07	0.01	0.08	0.04	0.02	0.02	0	0.01	0.02	0.01	0	0.04	0.02
D	0.08	0.25	0.1	0.24	0.13	0.58	0	0.13	0.16	0.09	0.15	0.35	0.07
C	0	0	0	0	0.01	0	0	0	0	0	0	0.01	0.03
G	0.01	0.03	0.17	0.02	0.02	0.01	0	0.15	0.27	0.04	0.01	0.02	0.13
E	0.12	0.2	0.18	0.17	0.14	0.09	0	0.09	0.07	0.09	0.12	0.07	0.05
Q	0.03	0.01	0.02	0.18	0.17	0.01	0	0.02	0.08	0.02	0.09	0.04	0.02
H	0.01	0.08	0.01	0.07	0.02	0.02	0	0	0.03	0.01	0.02	0.03	0.02
I	0.02	0.09	0	0.02	0.02	0.09	0	0.08	0.01	0.02	0.02	0	0.01
L	0.05	0.02	0.04	0.02	0.05	0.01	0	0.01	0.02	0.04	0.04	0.04	0.01
K	0.02	0.03	0	0	0	0	0	0.01	0	0.02	0.02	0.04	0.08
M	0.03	0.02	0.03	0	0	0	0	0.08	0.01	0.13	0.28	0	0.01
F	0.02	0.02	0.02	0.04	0.02	0.01	0	0.01	0.02	0.02	0	0.13	0.28
P	0.05	0.03	0.02	0.01	0.04	0	0	0.05	0.02	0.03	0.05	0.04	0.07
S	0.05	0.04	0.09	0.03	0	0.01	0	0.02	0.02	0.03	0.02	0.03	0.04
T	0.02	0.04	0.05	0.05	0.04	0	0	0.25	0.02	0.08	0.01	0.03	0.05
W	0.01	0	0	0	0	0	0	0	0.11	0.28	0	0.01	0
Y	0	0.07	0.03	0.05	0.1	0.07	1	0.05	0.11	0.05	0.05	0.04	0.01
V	0.02	0.02	0.01	0	0.02	0	0	0.02	0.01	0.01	0.02	0	0.02

3.2 Please compare the PWM of positive dataset to that of negative dataset to identify the significantly differential represented amino acids in each position.

3.3 Can we use the PWM to be a predictive model for the prediction of ubiquitination sites? Explain why or why not?