# Machine Learning in Computational Biology (Fall 2023)

**Assignment #2**

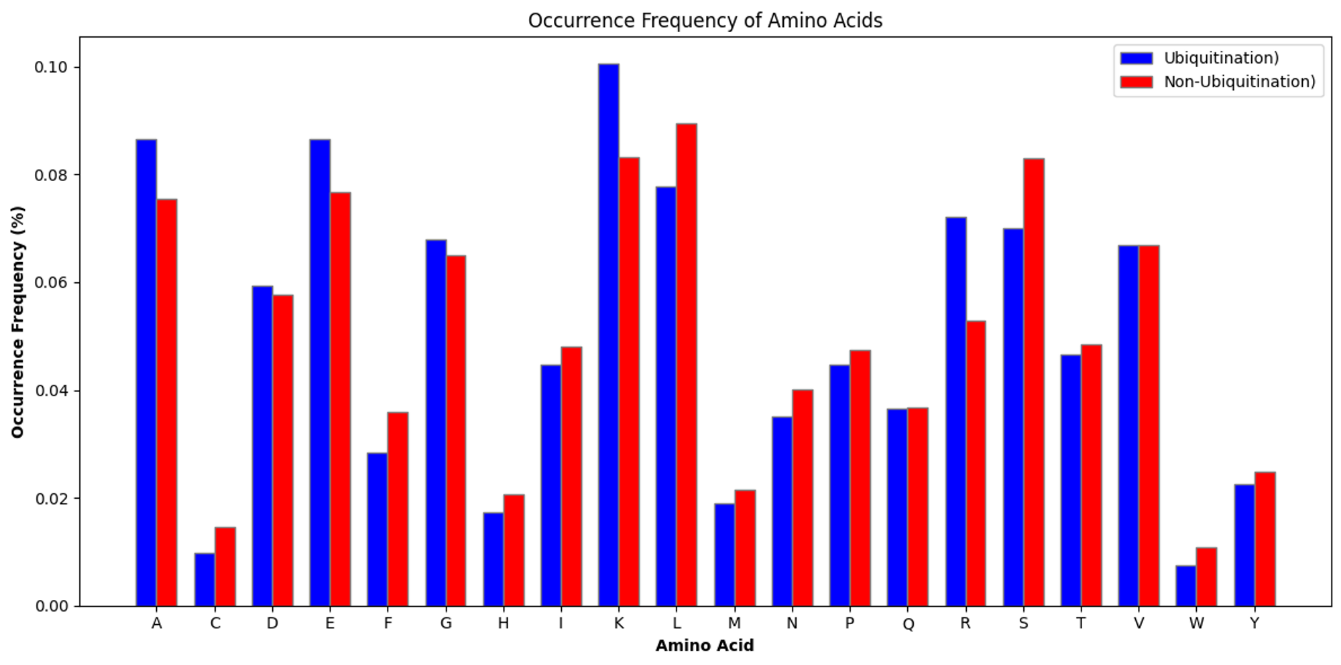
**Deadline: 23:59:59 17th November**, 2023 (Delayed submission is not allowed for any reason)

**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.

# [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

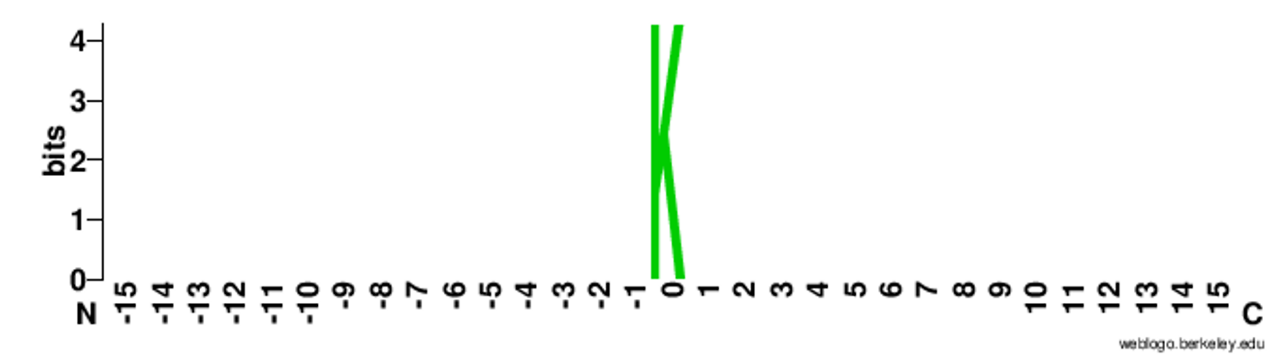
50  similarity.

50 % CD-HIT has uploaded

# [Sequence log and TwoSampleLogo] (20 points)

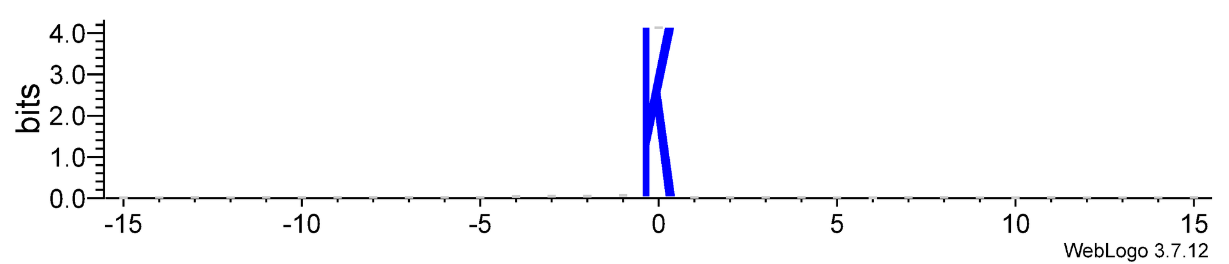
* 1. In order to observe the position-specific AAC for protein ubiquitination sites, the **WebLogo** tool can be utilized to create **frequency** and **entropy plots** of sequence logos for visualizing the potential amino acid motifs surrounding the modification sites (centered as position 0). Please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

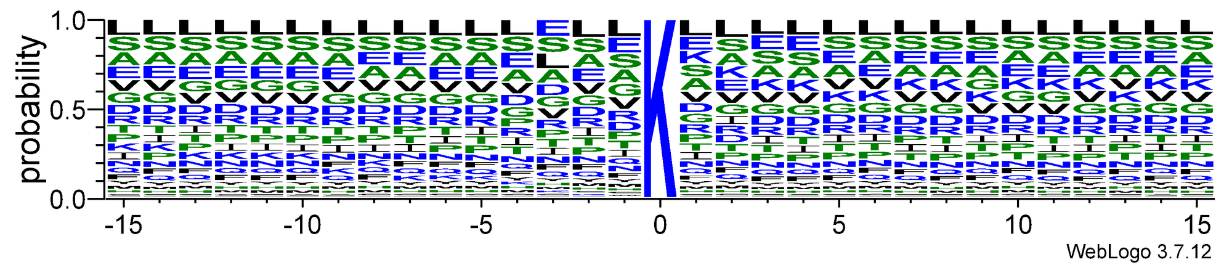
Positive Dataset



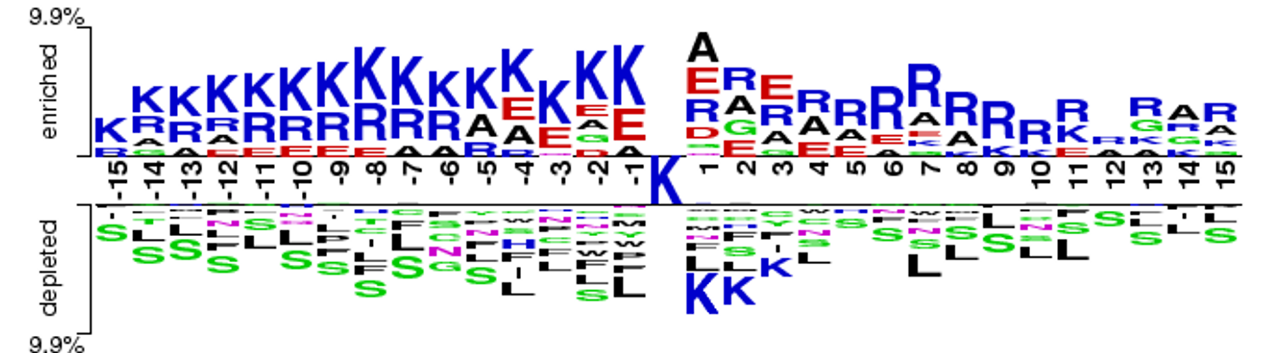


Negative Dataset





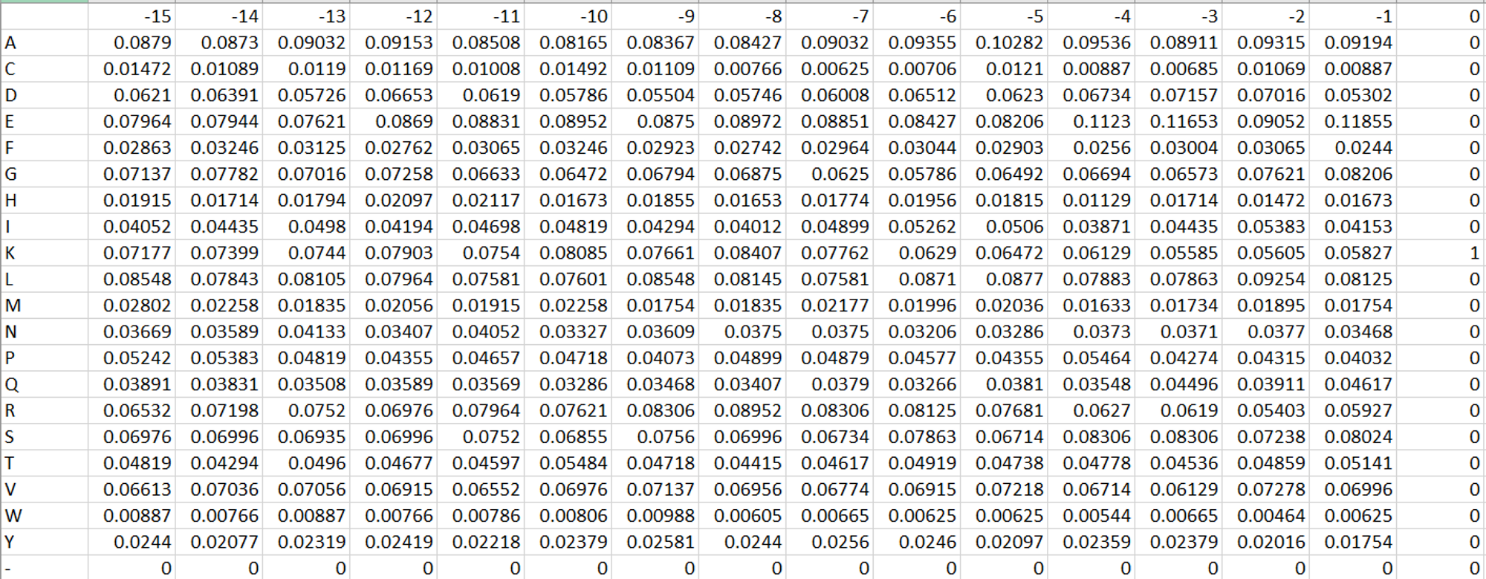
* 1. In order to investigate the difference of position-specific AAC between ubiquitination and non-ubiquitination sites, please use the **TwoSampleLogo** to visualize potential amino acid motifs surrounding the modification sites as shown below.



# [Positional Weighted Matrix] (60 points)

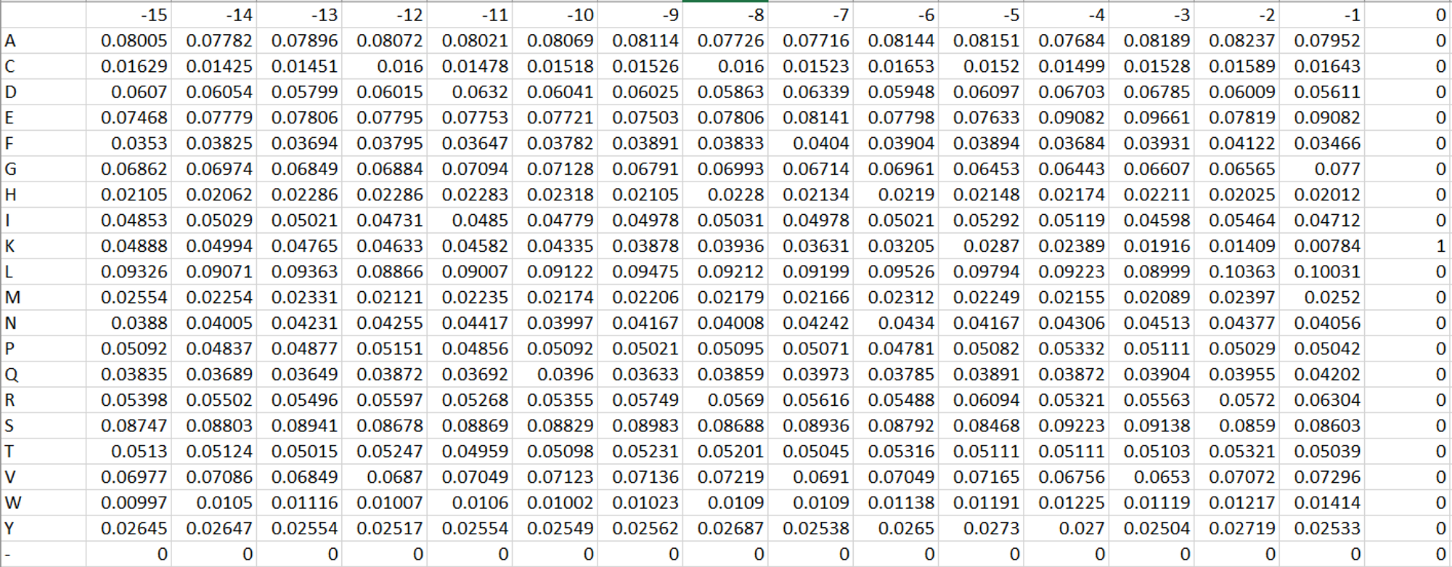
* 1. Please create the positional weighted matrix (PWM) for both positive and negative datasets, as shown below.

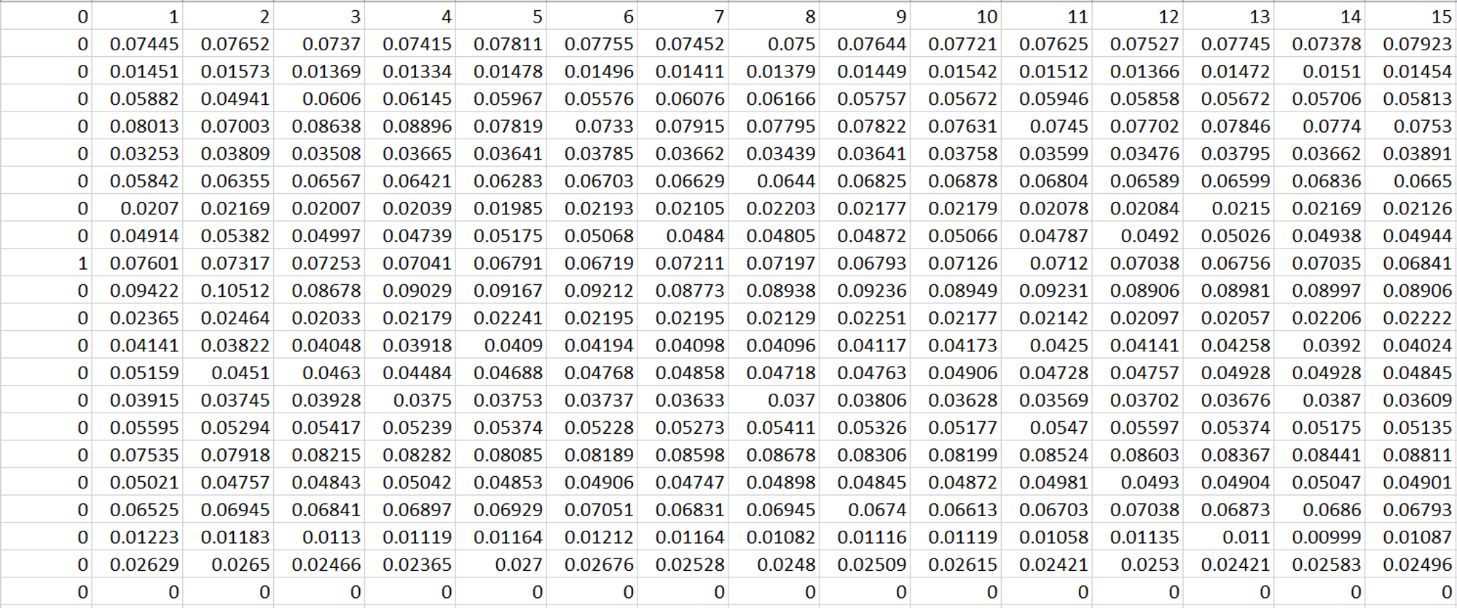
Positive Dataset

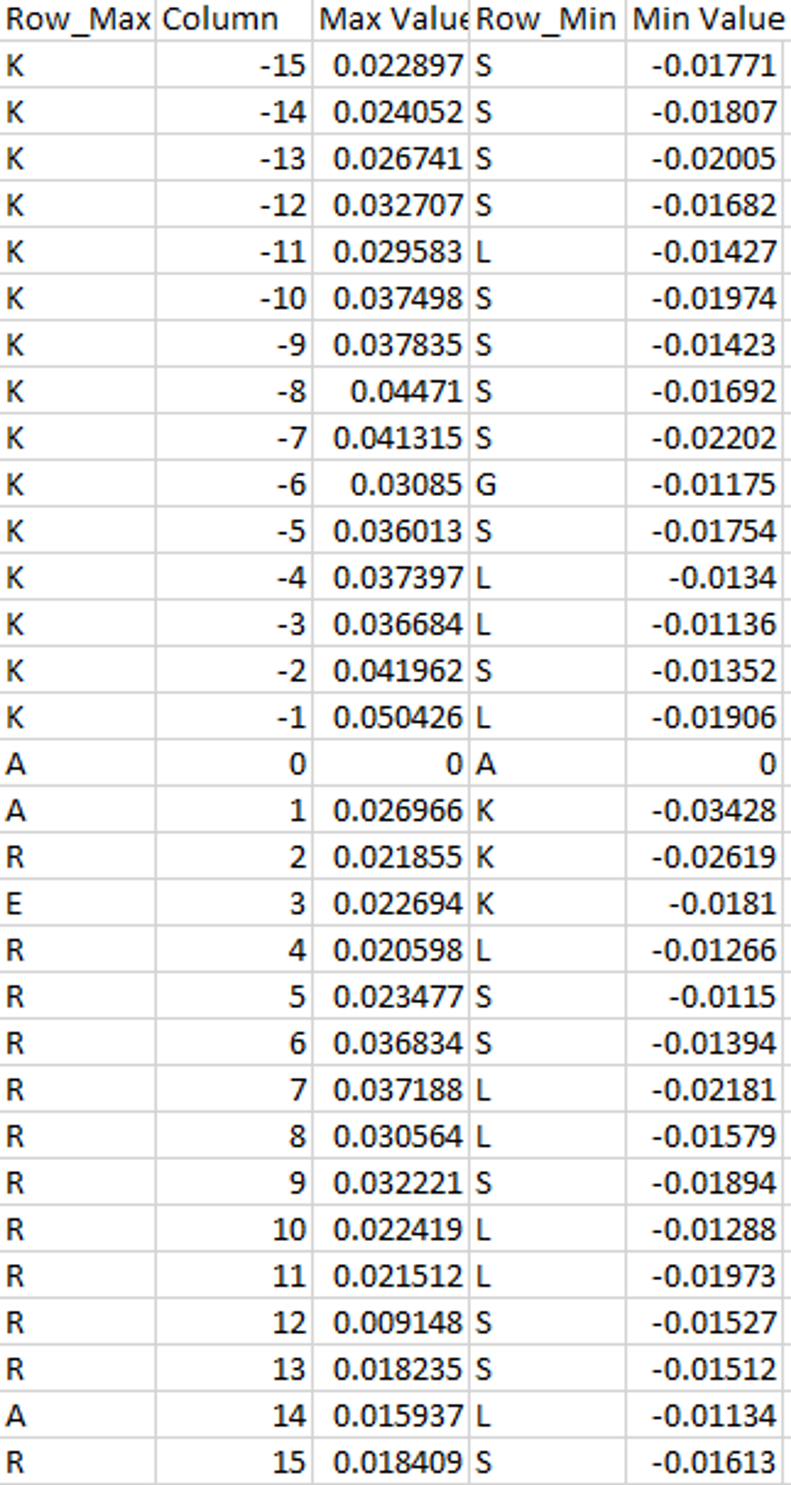


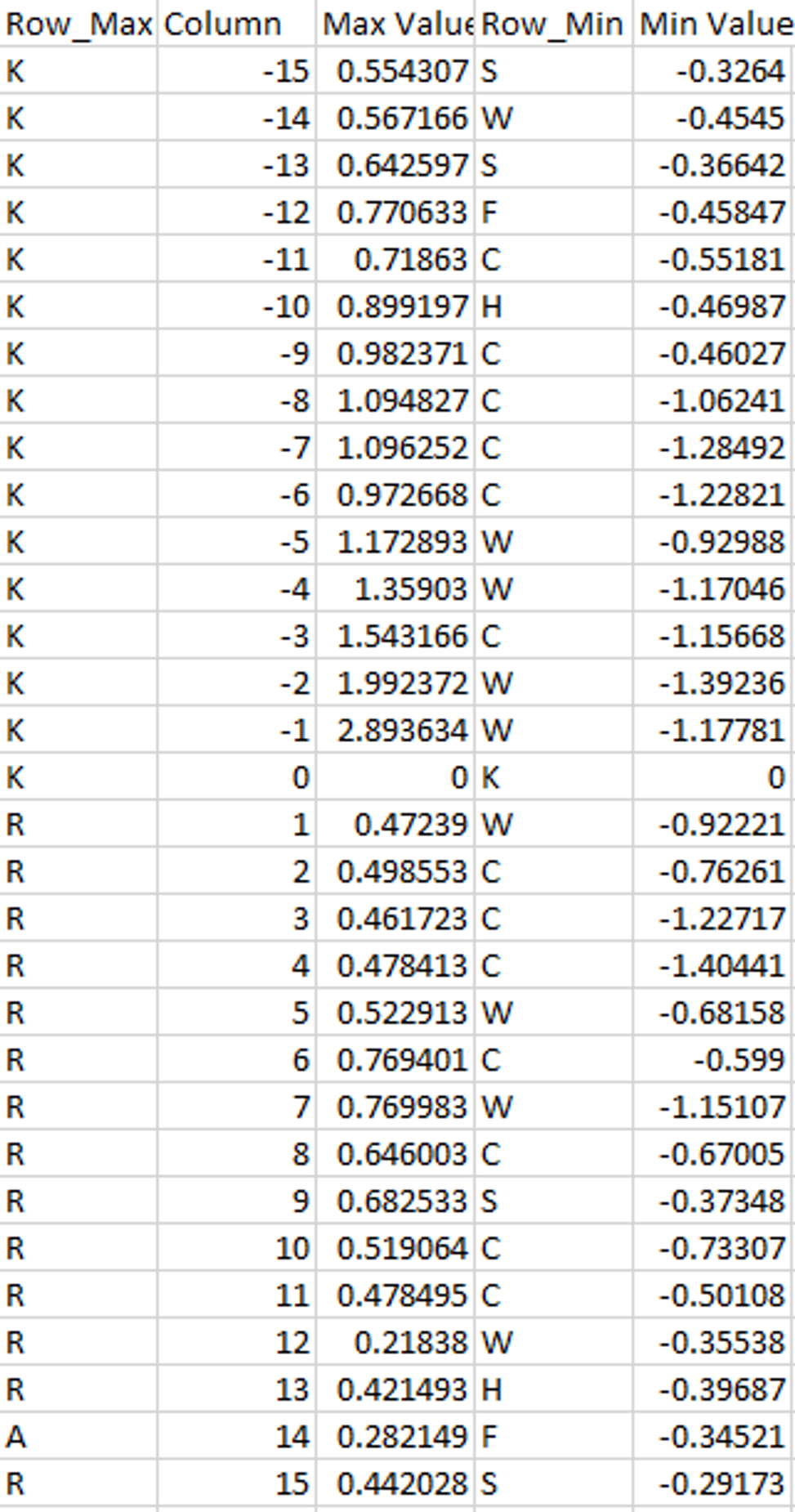
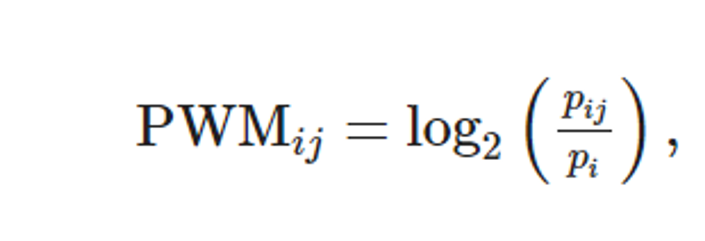


Negative Dataset





* 1. Please compare the PWM of positive dataset to that of negative dataset to identify the significantly differential represented amino acids in each position.
* Difference 比較：將兩個 PWM table 互減 (Positive - Negative)
* 取 log2 的方法



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

可以拿去預測，只是結果好壞不可知。

因為透過兩個方法去觀察 positive data 和 negative data 的差異性

1. Difference
2. Log2 Transformation

發現，確實在特定位點會有胺基酸特異性 (就以這筆資料來說)。

比如，做完 log2 的 table 中，位置 -1 的氨基酸位點，最大和最小值是整筆資料中相差較大的。在這個位點，可以 Positive 的資料是 K 比較重要，而 Negative 則是 W 貢獻度比較大。但如果有一個統計上的顯著性衡量標準去評判何謂顯著會更好。