Machine Learning in Computational Biology (Fall 2023)

Assignment #2

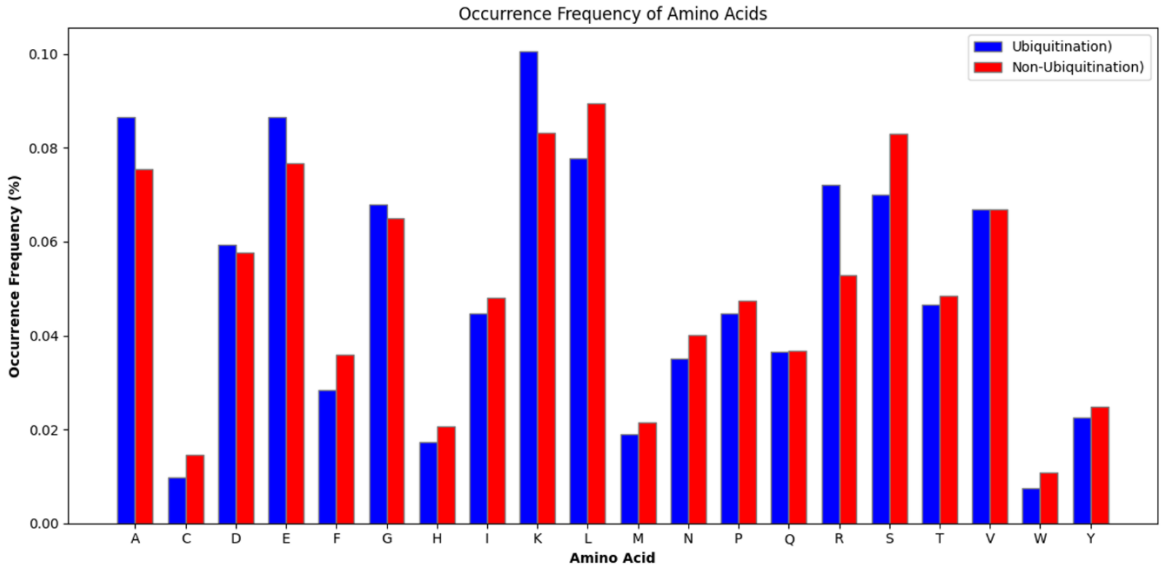
Deadline: 23:59:59 17th November, 2023 (Delayed submission is not allowed for any reason)

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.

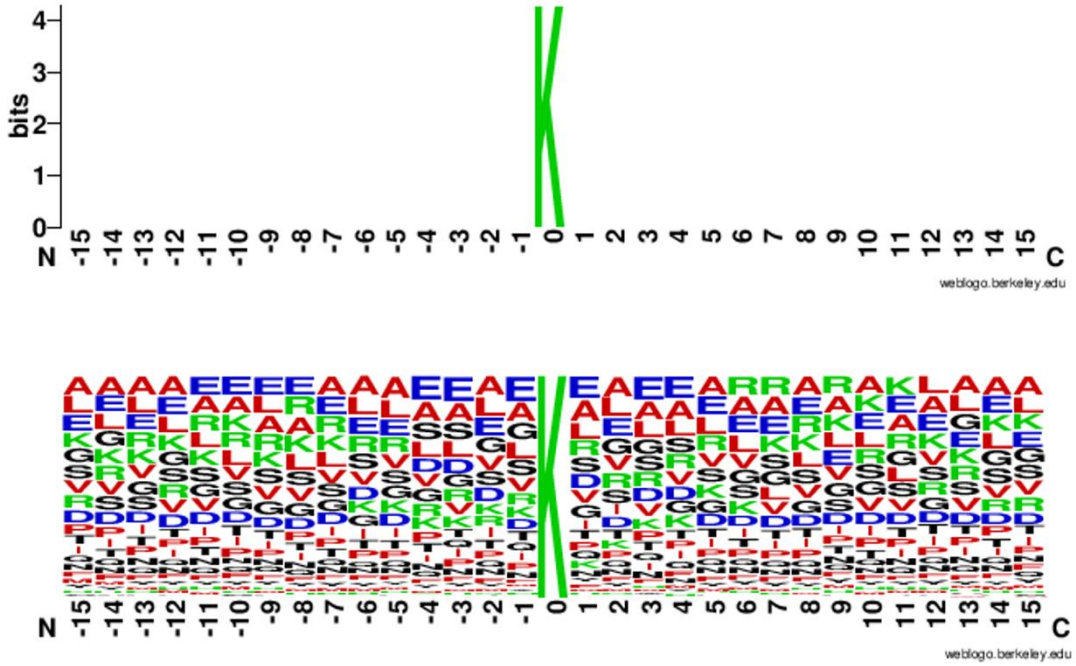
50 % CD-HIT has uploaded

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

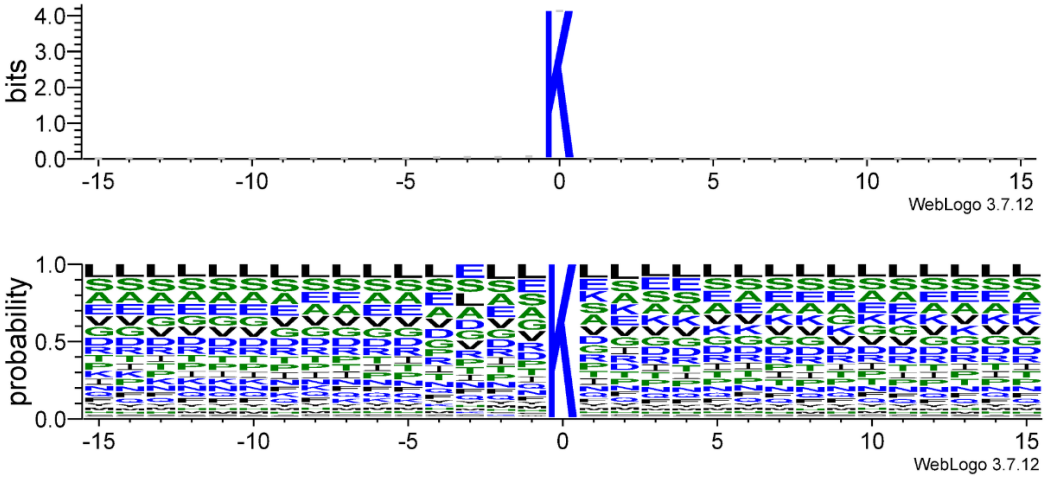
2.1 In order to observe the position-specific AAC for protein ubiquitination sites, the **WebLogo** tool can be utilized to create **frequency** and **entropy plots** of sequence logos for visualizing the potential amino acid motifs surrounding the modification sites (centered as

position 0). Please provide the **frequency** and **entropy plots** of sequence logos on both positive and negative datasets.

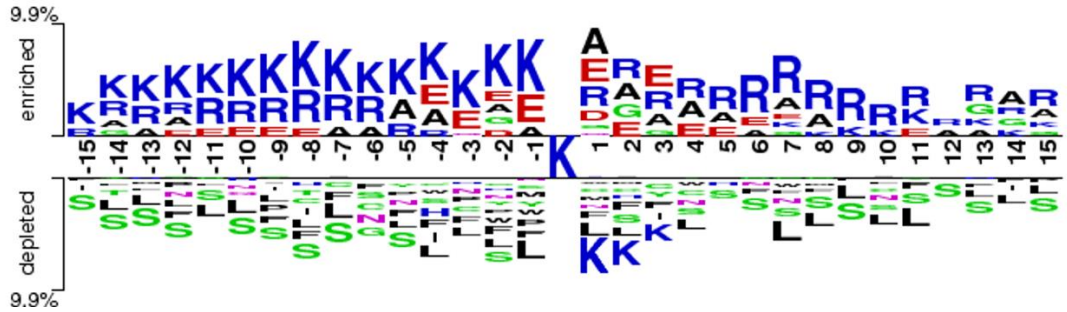
Positive Dataset



Negative Dataset



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.

Positive Dataset

	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0
A	0.0879	0.0873	0.09032	0.09153	0.08508	0.08165	0.08367	0.08427	0.09032	0.09355	0.10282	0.09536	0.08911	0.09315	0.09194	0
C	0.01472	0.01089	0.0119	0.01169	0.01008	0.01492	0.01109	0.00766	0.00625	0.00706	0.0121	0.00887	0.00685	0.01069	0.00887	0
D	0.0621	0.06391	0.05726	0.06653	0.0619	0.05786	0.05504	0.05746	0.06008	0.06512	0.0623	0.06734	0.07157	0.07016	0.05302	0
E	0.07964	0.07944	0.07621	0.0869	0.08831	0.08952	0.0875	0.08972	0.08851	0.08427	0.08206	0.1123	0.11653	0.09052	0.11855	0
F	0.02863	0.03246	0.03125	0.02762	0.03065	0.03246	0.02923	0.02742	0.02964	0.03044	0.02903	0.0256	0.03004	0.03065	0.0244	0
G	0.07137	0.07782	0.07016	0.07258	0.06633	0.06472	0.06794	0.06875	0.0625	0.05786	0.06492	0.06694	0.06573	0.07621	0.08206	0
H	0.01915	0.01714	0.01794	0.02097	0.02117	0.01673	0.01855	0.01653	0.01774	0.01956	0.01815	0.01129	0.01714	0.01472	0.01673	0
I	0.04052	0.04435	0.0498	0.04194	0.04698	0.04819	0.04294	0.04012	0.04899	0.05262	0.0506	0.03871	0.04435	0.05383	0.04153	0
K	0.07177	0.07399	0.0744	0.07903	0.0754	0.08085	0.07661	0.08407	0.07762	0.0629	0.06472	0.06129	0.05585	0.05605	0.05827	1
L	0.08548	0.07843	0.08105	0.07964	0.07581	0.07601	0.08548	0.08145	0.07581	0.0871	0.0877	0.07883	0.07863	0.09254	0.08125	0
M	0.02802	0.02258	0.01835	0.02056	0.01915	0.02258	0.01754	0.01835	0.02177	0.01996	0.02036	0.01633	0.01734	0.01895	0.01754	0
N	0.03669	0.03589	0.04133	0.03407	0.04052	0.03327	0.03609	0.0375	0.0375	0.03206	0.03286	0.0373	0.0371	0.0377	0.03468	0
P	0.05242	0.05383	0.04819	0.04355	0.04657	0.04718	0.04073	0.04899	0.04879	0.04577	0.04355	0.05464	0.04274	0.04315	0.04032	0
Q	0.03891	0.03831	0.03508	0.03589	0.03569	0.03286	0.03468	0.03407	0.0379	0.03266	0.0381	0.03548	0.04496	0.03911	0.04617	0
R	0.06532	0.07198	0.0752	0.06976	0.07964	0.07621	0.08306	0.08952	0.08306	0.08125	0.07681	0.0627	0.0619	0.05403	0.05927	0
S	0.06976	0.06996	0.06935	0.06996	0.0752	0.06855	0.0756	0.06996	0.06734	0.07863	0.06714	0.08306	0.08306	0.07238	0.08024	0
T	0.04819	0.04294	0.0496	0.04677	0.04597	0.05484	0.04718	0.04415	0.04617	0.04919	0.04738	0.04778	0.04536	0.04859	0.05141	0
V	0.06613	0.07036	0.07056	0.06915	0.06552	0.06976	0.07137	0.06956	0.06774	0.06915	0.07218	0.06714	0.06129	0.07278	0.06996	0
W	0.00887	0.00766	0.00887	0.00766	0.00786	0.00806	0.00988	0.00605	0.00665	0.00625	0.00625	0.00544	0.00665	0.00464	0.00625	0
Y	0.0244	0.02077	0.02319	0.02419	0.02218	0.02379	0.02581	0.0244	0.0256	0.0246	0.02097	0.02359	0.02379	0.02016	0.01754	0
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0.10141	0.09476	0.0881	0.09234	0.09052	0.0873	0.08831	0.09032	0.08427	0.08468	0.08226	0.08427	0.08669	0.08972	0.09052
0	0.00786	0.00927	0.00585	0.00504	0.01169	0.00988	0.00867	0.00867	0.01149	0.00927	0.01069	0.01129	0.01391	0.0121	0.01492
0	0.07198	0.05484	0.06734	0.06552	0.0619	0.05444	0.06008	0.05766	0.05867	0.06069	0.05927	0.05847	0.05907	0.05746	0.05827
0	0.10444	0.08629	0.10907	0.10504	0.09012	0.08488	0.0879	0.08488	0.0748	0.08024	0.08508	0.08427	0.07782	0.08407	0.07641
0	0.02339	0.02742	0.0254	0.03327	0.03226	0.03024	0.03004	0.02823	0.02964	0.03165	0.02641	0.02923	0.03125	0.02883	0.03427
0	0.06573	0.07984	0.0748	0.06492	0.06734	0.07056	0.07379	0.06532	0.07097	0.07056	0.0748	0.06411	0.07823	0.07802	0.07399
0	0.01653	0.01512	0.01794	0.01855	0.0131	0.01855	0.01774	0.01774	0.02097	0.02056	0.02097	0.01976	0.01633	0.01935	0.01754
0	0.04839	0.05544	0.03972	0.04335	0.04597	0.04778	0.04516	0.04859	0.04677	0.05363	0.04234	0.05081	0.04435	0.04133	0.04738
1	0.04173	0.04698	0.05444	0.06431	0.06855	0.06734	0.07984	0.08044	0.08105	0.08044	0.0877	0.07641	0.07742	0.07964	0.07702
0	0.07923	0.09355	0.08105	0.07762	0.08468	0.08367	0.06593	0.07359	0.07681	0.07661	0.07258	0.08468	0.08065	0.07863	0.07823
0	0.01673	0.01915	0.01694	0.02056	0.02016	0.01875	0.01855	0.01593	0.02157	0.01956	0.01956	0.01915	0.01855	0.02238	0.02077
0	0.03407	0.03649	0.03931	0.03044	0.03669	0.03589	0.03266	0.03569	0.04012	0.03347	0.03931	0.0373	0.03992	0.03629	0.03488
0	0.04597	0.04254	0.0504	0.04516	0.04315	0.04234	0.04516	0.04839	0.04435	0.0494	0.04597	0.04456	0.04637	0.05141	0.04093
0	0.04516	0.03851	0.04234	0.04254	0.03851	0.0379	0.03629	0.03831	0.03589	0.03508	0.0373	0.0381	0.0373	0.0381	0.03367
0	0.07762	0.0748	0.0746	0.07298	0.07722	0.08911	0.08992	0.08468	0.08548	0.07419	0.07621	0.06512	0.07198	0.06169	0.06976
0	0.075	0.06815	0.0748	0.07359	0.06935	0.06794	0.07419	0.07218	0.06411	0.07339	0.07198	0.07077	0.06855	0.07782	0.07198
0	0.04798	0.0494	0.0496	0.0496	0.04778	0.04496	0.04919	0.0502	0.04294	0.04698	0.05222	0.05302	0.04839	0.04637	0.04879
0	0.06633	0.07621	0.06593	0.07137	0.06976	0.07399	0.06573	0.06734	0.07238	0.07016	0.06371	0.07258	0.06774	0.06512	0.07177
0	0.00645	0.00786	0.00605	0.00464	0.00726	0.00907	0.00524	0.00948	0.00927	0.00726	0.00948	0.00887	0.0121	0.00806	0.01069
0	0.02399	0.02339	0.01633	0.01915	0.02399	0.0254	0.0256	0.02238	0.02843	0.02218	0.02218	0.02722	0.02339	0.02359	0.02823
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Negative Dataset

	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0
A	0.08005	0.07782	0.07896	0.08072	0.08021	0.08069	0.08114	0.07726	0.07716	0.08144	0.08151	0.07684	0.08189	0.08237	0.07952	0
C	0.01629	0.01425	0.01451	0.016	0.01478	0.01518	0.01526	0.016	0.01523	0.01653	0.0152	0.01499	0.01528	0.01589	0.01643	0
D	0.0607	0.06054	0.05799	0.06015	0.0632	0.06041	0.06025	0.05863	0.06339	0.05948	0.06097	0.06703	0.06785	0.06009	0.05611	0
E	0.07468	0.07779	0.07806	0.07795	0.07753	0.07721	0.07503	0.07806	0.08141	0.07798	0.07633	0.09082	0.09661	0.07819	0.09082	0
F	0.0353	0.03825	0.03694	0.03795	0.03647	0.03782	0.03891	0.03833	0.0404	0.03904	0.03894	0.03684	0.03931	0.04122	0.03466	0
G	0.06862	0.06974	0.06849	0.06884	0.07094	0.07128	0.06791	0.06993	0.06714	0.06961	0.06453	0.06443	0.06607	0.06565	0.077	0
H	0.02105	0.02062	0.02286	0.02286	0.02283	0.02318	0.02105	0.0228	0.02134	0.0219	0.02148	0.02174	0.02211	0.02025	0.02012	0
I	0.04853	0.05029	0.05021	0.04731	0.0485	0.04779	0.04978	0.05031	0.04978	0.05021	0.05292	0.05119	0.04598	0.05464	0.04712	0
K	0.04888	0.04994	0.04765	0.04633	0.04582	0.04335	0.03878	0.03936	0.03631	0.03205	0.0287	0.02389	0.01916	0.01409	0.00784	1
L	0.09326	0.09071	0.09363	0.08866	0.09007	0.09122	0.09475	0.09212	0.09199	0.09526	0.09794	0.09223	0.08999	0.10363	0.10031	0
M	0.02554	0.02254	0.02331	0.02121	0.02235	0.02174	0.02206	0.02179	0.02166	0.02312	0.02249	0.02155	0.02089	0.02397	0.0252	0
N	0.0388	0.04005	0.04231	0.04255	0.04417	0.03997	0.04167	0.04008	0.04242	0.0434	0.04167	0.04306	0.04513	0.04377	0.04056	0
P	0.05092	0.04837	0.04877	0.05151	0.04856	0.05092	0.05021	0.05095	0.05071	0.04781	0.05082	0.05332	0.05111	0.05029	0.05042	0
Q	0.03835	0.03689	0.03649	0.03872	0.03692	0.0396	0.03633	0.03859	0.03973	0.03785	0.03891	0.03872	0.03904	0.03955	0.04202	0
R	0.05398	0.05502	0.05496	0.05597	0.05268	0.05355	0.05749	0.0569	0.05616	0.05488	0.06094	0.05321	0.05563	0.0572	0.06304	0
S	0.08747	0.08803	0.08941	0.08678	0.08869	0.08829	0.08983	0.08688	0.08936	0.08792	0.08468	0.09223	0.09138	0.0859	0.08603	0
T	0.0513	0.05124	0.05015	0.05247	0.04959	0.05098	0.05231	0.05201	0.05045	0.05316	0.05111	0.05111	0.05103	0.05321	0.05039	0
V	0.06977	0.07086	0.06849	0.0687	0.07049	0.07123	0.07136	0.07219	0.0691	0.07049	0.07165	0.06756	0.0653	0.07072	0.07296	0
W	0.00997	0.0105	0.01116	0.01007	0.0106	0.01002	0.01023	0.0109	0.0109	0.01138	0.01191	0.01225	0.01119	0.01217	0.01414	0
Y	0.02645	0.02647	0.02554	0.02517	0.02554	0.02549	0.02562	0.02687	0.02538	0.0265	0.0273	0.027	0.02504	0.02719	0.02533	0
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0.07445	0.07652	0.0737	0.07415	0.07811	0.07755	0.07452	0.075	0.07644	0.07721	0.07625	0.07527	0.07745	0.07378	0.07923
0	0.01451	0.01573	0.01369	0.01334	0.01478	0.01496	0.01411	0.01379	0.01449	0.01542	0.01512	0.01366	0.01472	0.0151	0.01454
0	0.05882	0.04941	0.0606	0.06145	0.05967	0.05576	0.06076	0.06166	0.05757	0.05672	0.05946	0.05858	0.05672	0.05706	0.05813
0	0.08013	0.07003	0.08638	0.08896	0.07819	0.0733	0.07915	0.07795	0.07822	0.07631	0.0745	0.07702	0.07846	0.0774	0.0753
0	0.03253	0.03809	0.03508	0.03665	0.03641	0.03785	0.03662	0.03439	0.03641	0.03758	0.03599	0.03476	0.03795	0.03662	0.03891
0	0.05842	0.06355	0.06567	0.06421	0.06283	0.06703	0.06629	0.0644	0.06825	0.06878	0.06804	0.06589	0.06599	0.06836	0.0665
0	0.0207	0.02169	0.02007	0.02039	0.01985	0.02193	0.02105	0.02203	0.02177	0.02179	0.02078	0.02084	0.0215	0.02169	0.02126
0	0.04914	0.05382	0.04997	0.04739	0.05175	0.05068	0.0484	0.04805	0.04872	0.05066	0.04787	0.0492	0.05026	0.04938	0.04944
1	0.07601	0.07317	0.07253	0.07041	0.06791	0.06719	0.07211	0.07197	0.06793	0.07126	0.0712	0.07038	0.06756	0.07035	0.06841
0	0.09422	0.10512	0.08678	0.09029	0.09167	0.09212	0.08773	0.08938	0.09236	0.08949	0.09231	0.08906	0.08981	0.08997	0.08906
0	0.02365	0.02464	0.02033	0.02179	0.02241	0.02195	0.02195	0.02129	0.02251	0.02177	0.02142	0.02097	0.02057	0.02206	0.02222
0	0.04141	0.03822	0.04048	0.03918	0.0409	0.04194	0.04098	0.04096	0.04117	0.04173	0.0425	0.04141	0.04258	0.0392	0.04024
0	0.05159	0.0451	0.0463	0.04484	0.04688	0.04768	0.04858	0.04718	0.04763	0.04906	0.04728	0.04757	0.04928	0.04928	0.04845
0	0.03915	0.03745	0.03928	0.0375	0.03753	0.03737	0.03633	0.037	0.03806	0.03628	0.03569	0.03702	0.03676	0.0387	0.03609
0	0.05595	0.05294	0.05417	0.05239	0.05374	0.05228	0.05273	0.05411	0.05326	0.05177	0.0547	0.05597	0.05374	0.05175	0.05135
0	0.07535	0.07918	0.08215	0.08282	0.08085	0.08189	0.08598	0.08678	0.08306	0.08199	0.08524	0.08603	0.08367	0.08441	0.08811
0	0.05021	0.04757	0.04843	0.05042	0.04853	0.04906	0.04747	0.04898	0.04845	0.04872	0.04981	0.0493	0.04904	0.05047	0.04901
0	0.06525	0.06945	0.06841	0.06897	0.06929	0.07051	0.06831	0.06945	0.0674	0.06613	0.06703	0.07038	0.06873	0.0686	0.06793
0	0.01223	0.01183	0.0113	0.01119	0.01164	0.01212	0.01164	0.01082	0.01116	0.01119	0.01058	0.01135	0.011	0.00999	0.01087
0	0.02629	0.0265	0.02466	0.02365	0.027	0.02676	0.02528	0.0248	0.02509	0.02615	0.02421	0.0253	0.02421	0.02583	0.02496
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

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.

- Difference 比較：將兩個 PWM table 互減 (Positive - Negative)

Row_Max	Column	Max Value	Row_Min	Min Value
K	-15	0.022897	S	-0.01771
K	-14	0.024052	S	-0.01807
K	-13	0.026741	S	-0.02005
K	-12	0.032707	S	-0.01682
K	-11	0.029583	L	-0.01427
K	-10	0.037498	S	-0.01974
K	-9	0.037835	S	-0.01423
K	-8	0.04471	S	-0.01692
K	-7	0.041315	S	-0.02202
K	-6	0.03085	G	-0.01175
K	-5	0.036013	S	-0.01754
K	-4	0.037397	L	-0.0134
K	-3	0.036684	L	-0.01136
K	-2	0.041962	S	-0.01352
K	-1	0.050426	L	-0.01906
A	0	0	A	0
A	1	0.026966	K	-0.03428
R	2	0.021855	K	-0.02619
E	3	0.022694	K	-0.0181
R	4	0.020598	L	-0.01266
R	5	0.023477	S	-0.0115
R	6	0.036834	S	-0.01394
R	7	0.037188	L	-0.02181
R	8	0.030564	L	-0.01579
R	9	0.032221	S	-0.01894
R	10	0.022419	L	-0.01288
R	11	0.021512	L	-0.01973
R	12	0.009148	S	-0.01527
R	13	0.018235	S	-0.01512
A	14	0.015937	L	-0.01134
R	15	0.018409	S	-0.01613

- 取 log2 的方法

$$\mathrm{PWM}_{ij} = \log_2 \left(\frac{p_{ij}}{p_i} \right),$$

Row_Max	Column	Max Value	Row_Min	Min Value
K	-15	0.554307	S	-0.3264
K	-14	0.567166	W	-0.4545
K	-13	0.642597	S	-0.36642
K	-12	0.770633	F	-0.45847
K	-11	0.71863	C	-0.55181
K	-10	0.899197	H	-0.46987
K	-9	0.982371	C	-0.46027
K	-8	1.094827	C	-1.06241
K	-7	1.096252	C	-1.28492
K	-6	0.972668	C	-1.22821
K	-5	1.172893	W	-0.92988
K	-4	1.35903	W	-1.17046
K	-3	1.543166	C	-1.15668
K	-2	1.992372	W	-1.39236
K	-1	2.893634	W	-1.17781
K	0	0	K	0
R	1	0.47239	W	-0.92221
R	2	0.498553	C	-0.76261
R	3	0.461723	C	-1.22717
R	4	0.478413	C	-1.40441
R	5	0.522913	W	-0.68158
R	6	0.769401	C	-0.599
R	7	0.769983	W	-1.15107
R	8	0.646003	C	-0.67005
R	9	0.682533	S	-0.37348
R	10	0.519064	C	-0.73307
R	11	0.478495	C	-0.50108
R	12	0.21838	W	-0.35538
R	13	0.421493	H	-0.39687
A	14	0.282149	F	-0.34521
R	15	0.442028	S	-0.29173

3.3 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。