Electronic Supplementary Material (ESI) for Molecular BioSystems.

Sup-A. The results of the other ensemble learning methods

Win-size	Residue	Accu	Prec	Sens	Spec	F1sc	MCC
7	K	0.94	0.94	0.94	0.94	0.94	0.88
	P	0.97	0.98	0.96	0.98	0.97	0.95
9	K	0.97	0.95	0.99	0.96	0.97	0.94
	P	0.96	0.96	0.97	0.95	0.96	0.92
11	K	0.95	0.93	0.98	0.93	0.95	0.91
	P	0.96	0.95	0.98	0.94	0.96	0.92
13	K	0.95	0.94	0.96	0.94	0.95	0.91
	P	0.97	0.96	0.98	0.95	0.97	0.93
15	K	0.95	0.93	0.98	0.93	0.95	0.91
	P	0.96	0.96	0.95	0.97	0.96	0.92
17	K	0.96	0.99	0.94	0.99	0.96	0.93
	P	0.96	0.97	0.94	0.98	0.96	0.92
19	K	0.97	0.94	1.00	0.94	0.97	0.94
	P	0.96	0.97	0.95	0.97	0.96	0.92

Table A-1. Jackknife cross validation results of adaptive boosting (AdaBoost) model for hydroxylation site prediction

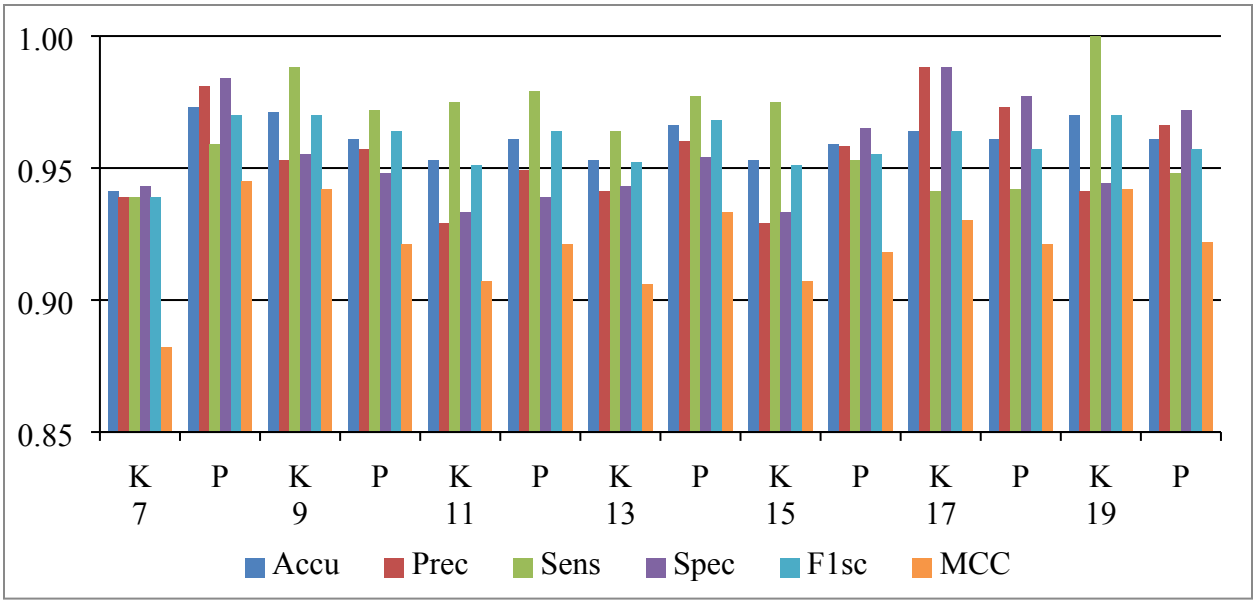


Figure A-1. Chart of the evaluation metrics for AdaBoost based model for hydroxylation site prediction

Win-size	Residue	Accu	Prec	Sens	Spec	F1sc	MCC
7	K	0.96	1.00	0.92	1.00	0.96	0.92
	P	0.94	0.96	0.92	0.97	0.94	0.89
9	K	0.95	0.92	0.99	0.92	0.95	0.91
	P	0.93	0.91	0.97	0.90	0.94	0.87
11	K	0.96	0.93	0.99	0.93	0.96	0.92
	P	0.94	0.92	0.97	0.91	0.94	0.88
13	K	0.97	0.95	0.98	0.95	0.96	0.93
	P	0.93	0.92	0.95	0.91	0.94	0.86
15	K	0.96	0.94	0.98	0.94	0.96	0.92
	P	0.94	0.97	0.91	0.97	0.94	0.89
17	K	0.95	0.99	0.92	0.99	0.95	0.91
	P	0.95	0.97	0.92	0.98	0.94	0.90
19	K	0.96	0.94	0.99	0.94	0.96	0.93
	P	0.94	0.96	0.92	0.97	0.94	0.88

Table A-2. Jackknife cross validation results of Bootstrap Aggregating (Bagging) meta-estimator model for hydroxylation site prediction

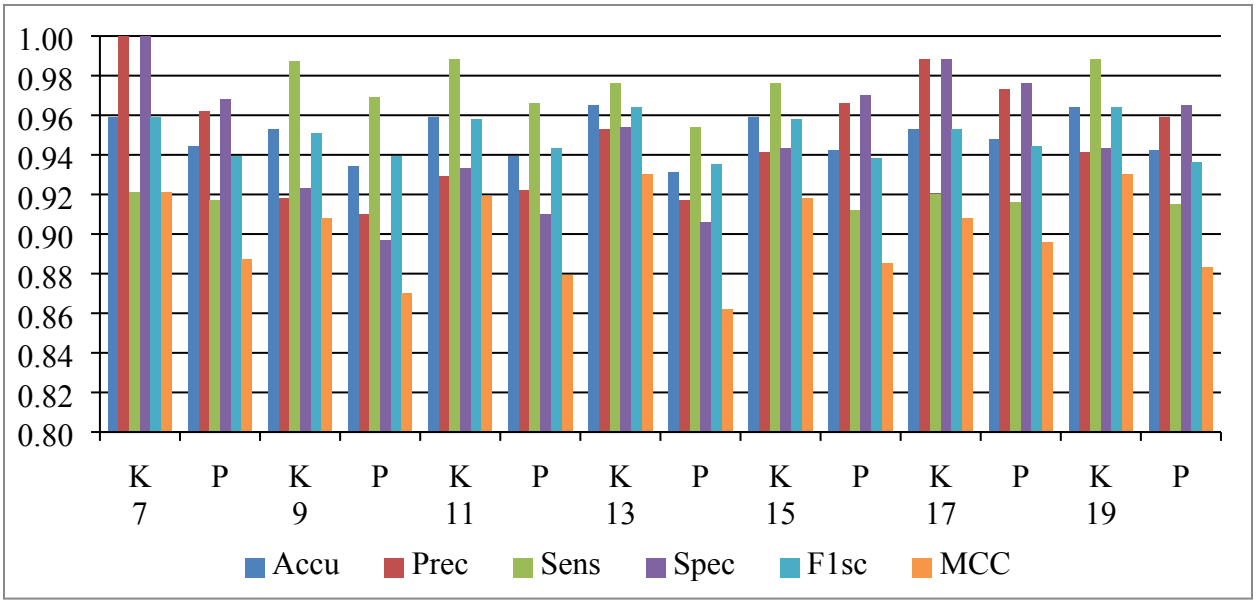


Figure A-2. Chart of the evaluation metrics for Bagging based model for hydroxylation site prediction

Win-size	Residue	Accu	Prec	Sens	Spec	F1sc	MCC
7	K	0.94	0.94	0.93	0.94	0.93	0.87
	P	0.96	0.97	0.94	0.98	0.96	0.92
9	K	0.95	0.93	0.96	0.93	0.95	0.90
	P	0.95	0.94	0.98	0.93	0.96	0.91
11	K	0.95	0.93	0.98	0.93	0.95	0.91
	P	0.95	0.93	0.98	0.92	0.96	0.91
13	K	0.97	0.98	0.97	0.98	0.97	0.94
	P	0.96	0.95	0.98	0.94	0.96	0.92
15	K	0.96	0.95	0.96	0.95	0.96	0.92
	P	0.96	0.97	0.94	0.97	0.95	0.91
17	K	0.97	1.00	0.94	1.00	0.97	0.94
	P	0.95	0.97	0.93	0.97	0.95	0.91
19	K	0.95	0.93	0.96	0.93	0.95	0.89
	P	0.95	0.97	0.92	0.98	0.95	0.90

Table A-3. Jackknife cross validation results of Gradient Boosting model for hydroxylation site prediction

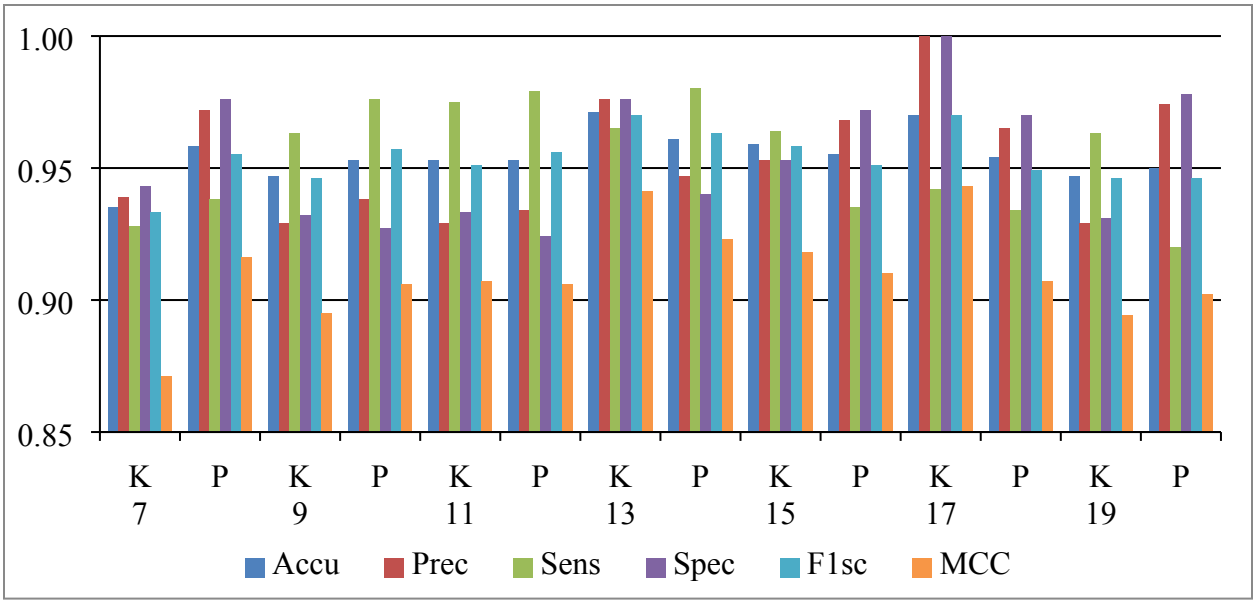


Figure A-3. Chart of the evaluation metrics for Gradient Boosting based model for hydroxylation site prediction

Win-size	Residue	Accu	Prec	Sens	Spec	F1sc	MCC
7	K	0.94	0.92	0.95	0.92	0.93	0.87
	P	0.88	0.86	0.87	0.88	0.86	0.75
9	K	0.91	0.91	0.92	0.91	0.91	0.82
	P	0.85	0.87	0.87	0.84	0.87	0.70
11	K	0.89	0.92	0.88	0.91	0.90	0.79
	P	0.86	0.87	0.88	0.84	0.87	0.72
13	K	0.91	0.89	0.93	0.90	0.91	0.82
	P	0.88	0.89	0.88	0.87	0.89	0.75
15	K	0.90	0.89	0.91	0.90	0.90	0.80
	P	0.87	0.84	0.86	0.87	0.85	0.73
17	K	0.89	0.90	0.88	0.91	0.89	0.79
	P	0.86	0.81	0.86	0.85	0.84	0.71
19	K	0.93	0.93	0.93	0.93	0.93	0.86
	P	0.85	0.82	0.83	0.86	0.83	0.69

Table A-4. Jackknife cross validation results of Extra-Trees Classifier model for hydroxylation site prediction

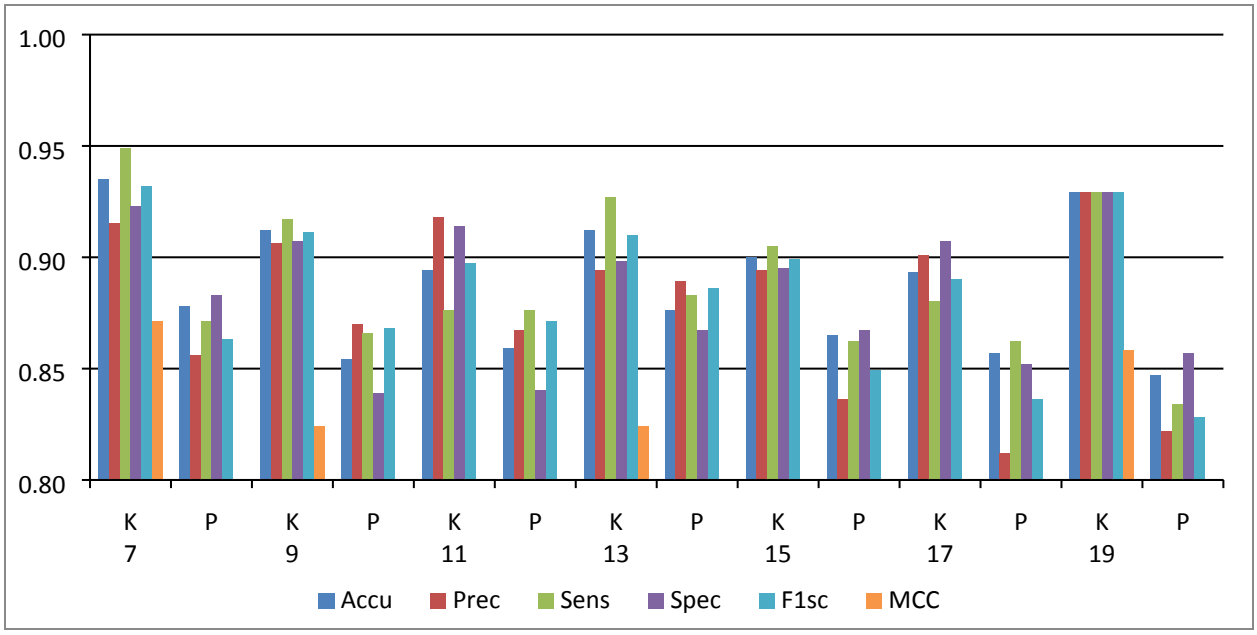


Figure A-4. Chart of the evaluation metrics for Extra-Trees Classifier based model for hydroxylation site prediction

Sup-B. Feature vectors for the different window sizes

Win-size	Selected features for hydroxyproline final model								Feature vector length
	HQI1	HQI3	HQI4	HQI5	HQI7	HQI8	ENT1	ACH	
7	6	6	6	6	6	6	3	3	**42**
9	8	8	8	8	8	8	3	4	**55**
11	10	10	10	10	10	10	3	6	**69**
13	12	12	12	12	12	12	3	7	**82**
15	14	14	14	14	14	14	3	8	**95**
17	16	16	16	16	16	16	3	9	**108**
19	18	18	18	18	18	18	3	10	**121**

Table B-1. The feature order and number of features in each selected feature type and feature vector size for hydroxyproline final model

Win-size	Selected features for hydroxylysine final model			Feature vector length
	HQI3	HQI4	ACH	
7	6	6	3	**15**
9	8	8	4	**20**
11	10	10	6	**26**

13	12	12	7	**31**
15	14	14	8	**36**
17	16	16	9	**41**
19	18	18	10	**46**

Table B-2. The feature order and number of features in each selected feature type and feature vector size for hydroxylysine final model

Sup-C. Feature profiles

Figure C-1 and C-2 show the HQI1 profiles of ten individual sequence windows of size of 15 residues and with positive or negative hydroxylation site in the center. The sequences are selected randomly from the benchmark sequences used for training. The figures are presented in 3-D. The x-axis shows the upstream and downstream flanking positions indexed from -7 to 7 based on their position from the hydroxylation site. The y-axis shows the scale of the amino acid property HQI1. The z-axis shows the individual windows numbered from 1 to 10.

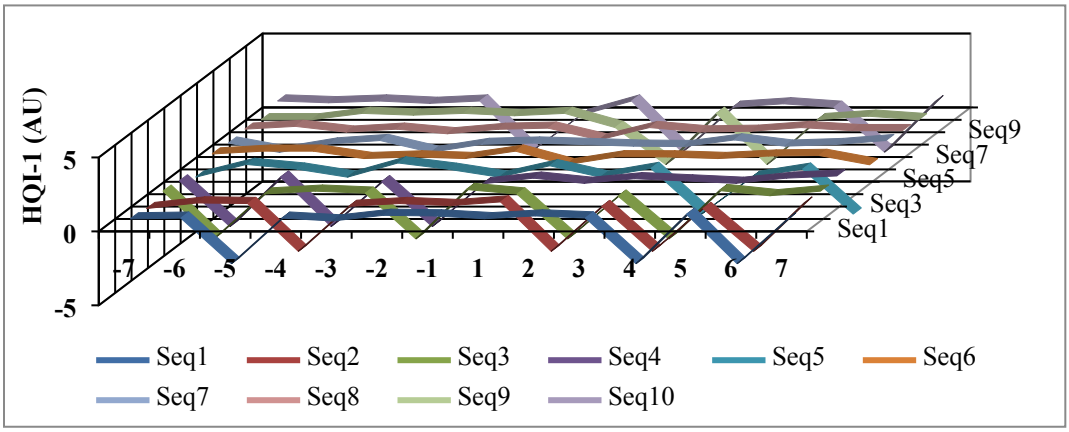


Figure C-1. High Quality indices-1 (HQI1) profile for positive hydroxyproline

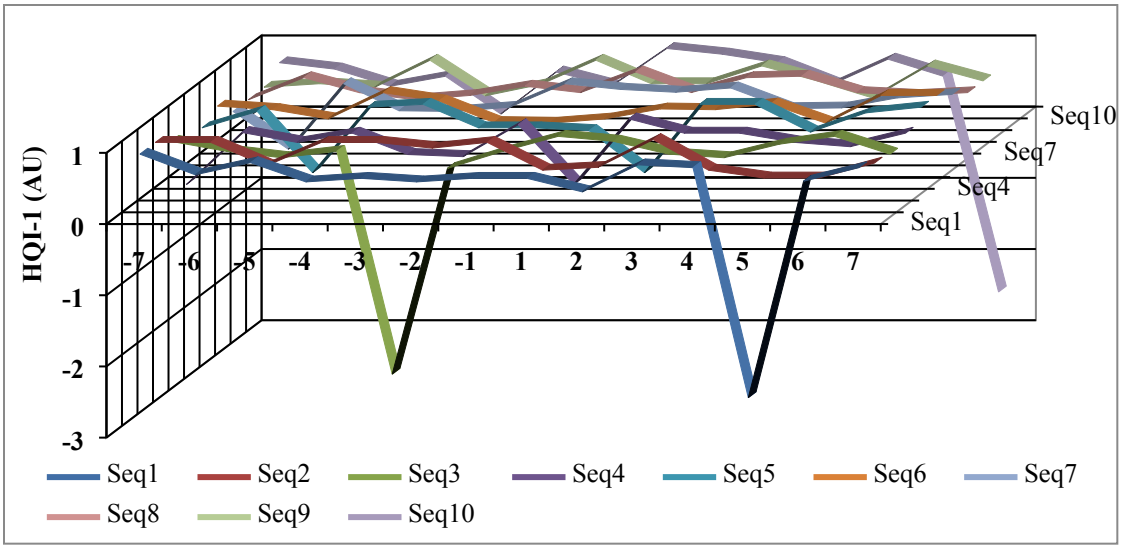


Figure C-2. High Quality indices-1 (HQI1) profile for negative hydroxyproline

Figure C-3 and C-4 show the ACH profiles of the same above ten individual sequence windows of size of 15 residues and with positive or negative hydroxylation site in the center. ACH is calculated by dividing the main sequence windows (here is 15) into sub-windows with the hydroxylation site in the center. The x-axis shows the sub-windows (3, 5, 7, 11, 13, and 15) while the y-axis shows the average hydrophobicity of sub-windows.

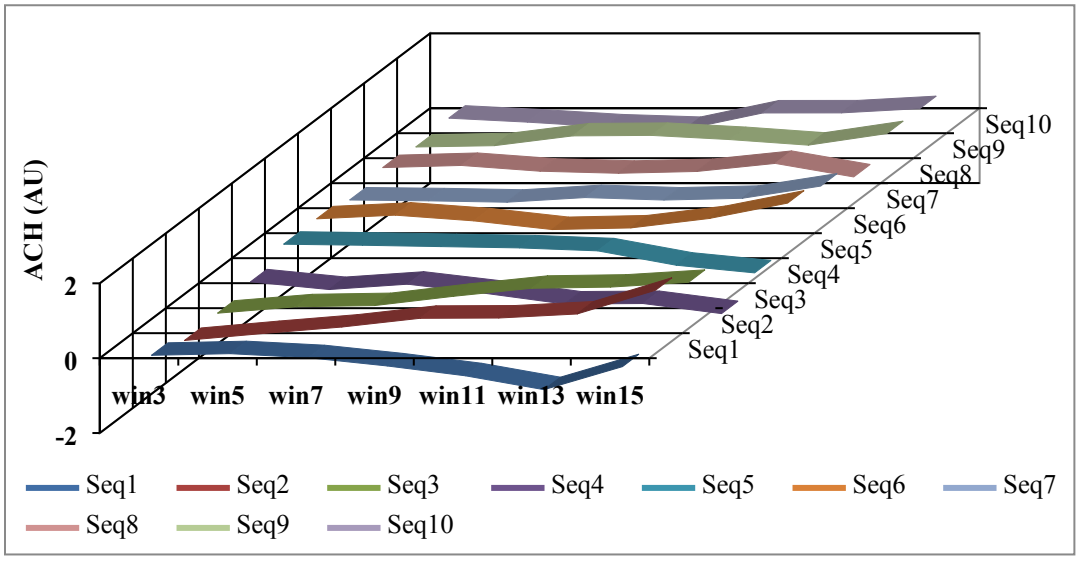


Figure C-3. Average cumulative hydrophobicity (ACH) profile for positive hydroxyproline (windows size is 15)

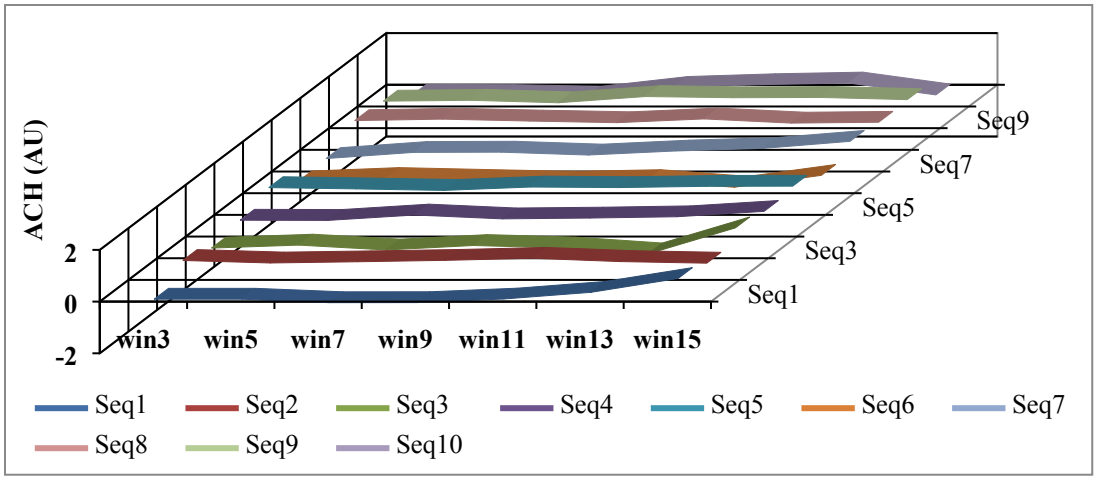


Figure C-4. Average cumulative hydrophobicity (ACH) profile for negative hydroxyproline (windows size is 15)

Figure C-5 and C-6 show Shannon entropy (Type II) profiles of the same above ten individual sequence windows of size of 15 residues and with positive or negative hydroxylation site in the center. The x-axis shows the upstream and downstream flanking positions indexed from -7 to 7 based on their position from the hydroxylation site. The y-axis shows the entropy based on position specific scoring matrix (PSSM).

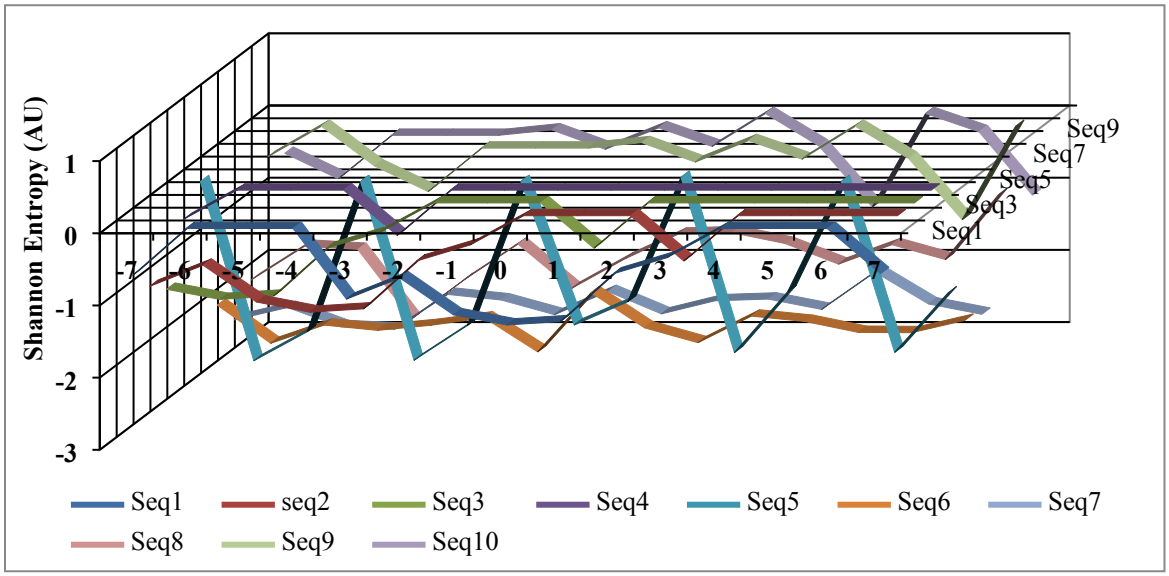


Figure C-5. Shannon entropy profile for positive hydroxyproline

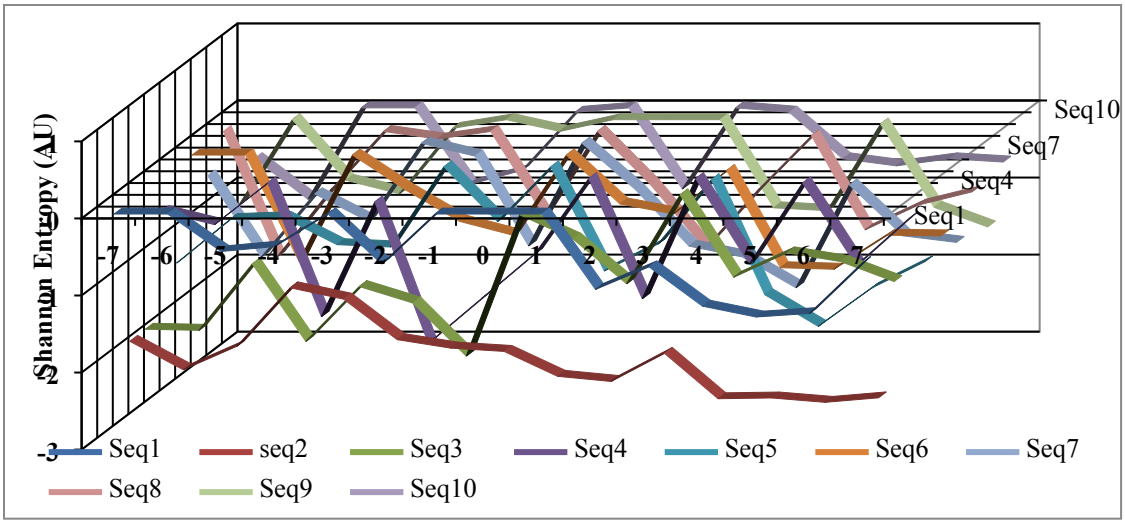
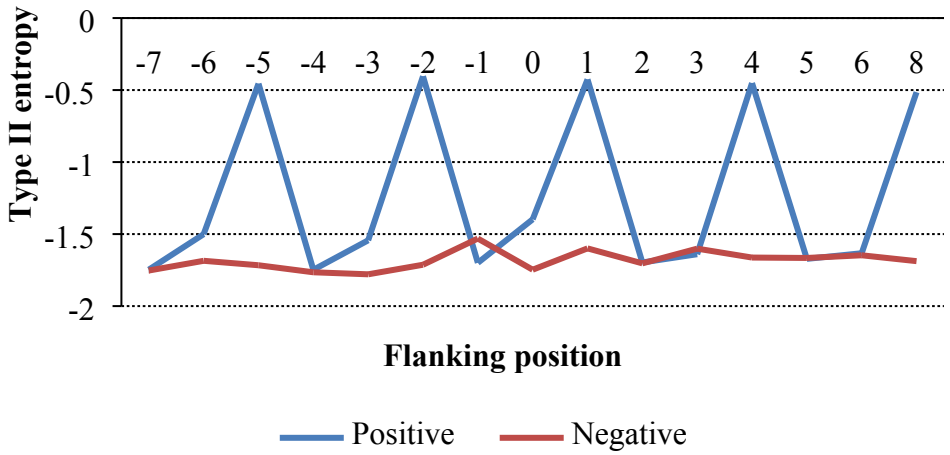


Figure C-6. Shannon entropy profile for negative hydroxyproline

Figure C-7 shows the average Type II profiles for the positive (blue) and negative (red) windows for hydroxylysine used for training.



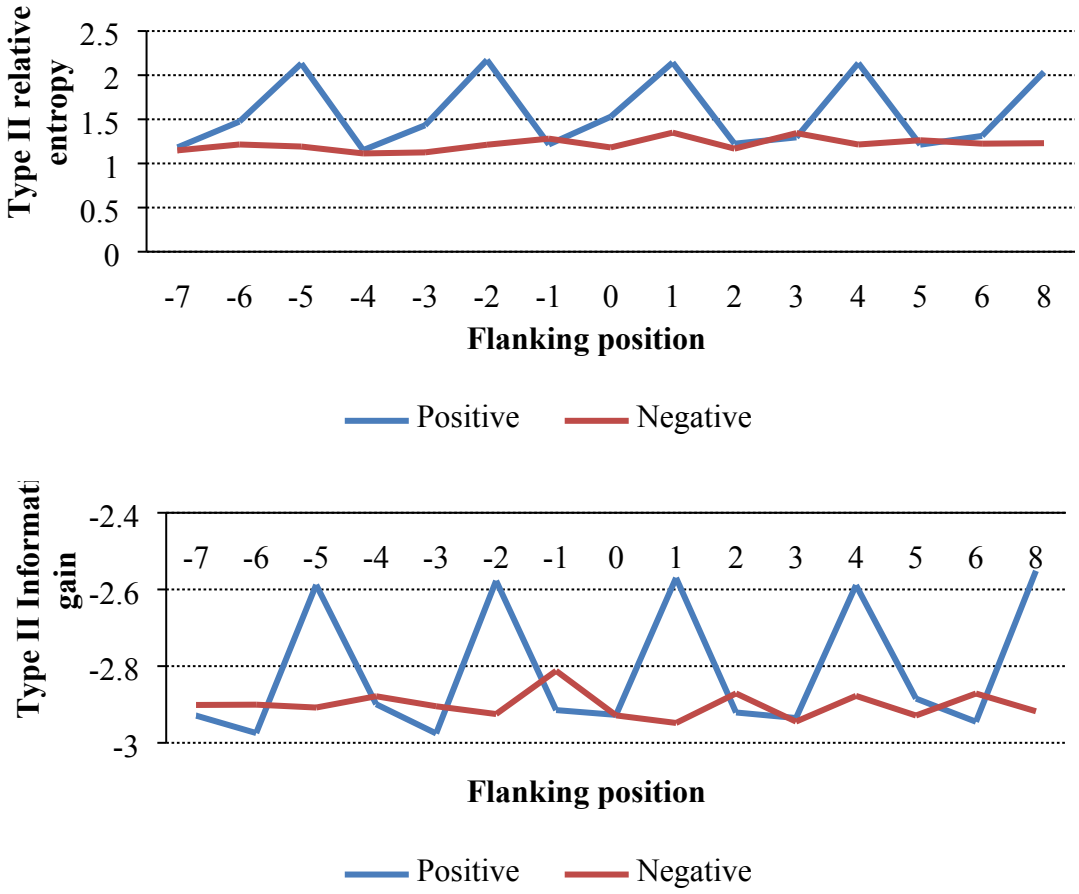


Figure C-7. Type II Entropy, relative entropy, and information gain for lysine

Sup-D. Independent sequences for testing

Table D

Independent test set for comparing RF-Hydroxysite with other methods

Table D-1: Dataset for Hydroxyproline sites

	Positive hydroxyproline		Negative hydroxyproline	
#	Uniprot ID	Position	Uniprot ID	Position
1	P02459	427	Q02388	668
2	Q7XAD0	72	Q9Y2N7	193
3	Q25460	598	Q02388	984
4	P30754	332	Q16665	229
5	Q05707	1531	P14618	295
6	P30754	266	P02459	231

7	P30754	526	O75636	1
8	Q9M0S4	29	E7ENY8	84
9	Q25460	662	Q25460	849
10	P02747	98	Q02388	2859
11	Q7XAD0	115	P02459	44
12	P05997	283	P12111	3157
13	P69929	202	P04925	76
14	P15502	160	P02747	2
15	P25508	5	P08125	117
16	A1X158	65	P02459	1321
17	Q25460	108	P32121	118
18	P25508	2	P08123	412
19	Q25460	298	P08125	442
20	P02459	223	P02453	48
21	P30754	776	P04925	129
22	P02459	307	P86289	43
23	P0C8W0	46	Q6PSU2	150
24	P20908	668	Q9UKV8	520
25	P0C1X1	33	P20908	516
26	Q7XAD0	50	Q9Y4R8	775
27	Q805R9	40	P69928	22
28	Q9Y4R8	415	P05997	243
29	P69929	134	O75636	83
30	Q25460	292	P30754	73
31	P30754	848	P12111	1608
32	P69929	168	Q05707	653
33	Q05707	1463	P69929	133
34	Q25460	860	P86289	31
35	P29602	109	Q25460	96
36	Q805R9	34	P07550	316
37	P30754	380	E7ENY8	42
38	Q93WP8	35	Q05707	1266
39	P25508	20	Q9UKV8	583
40	Q24940	102	P58808	24
41	Q05707	1652	P69929	218
42	A1X158	276	Q05707	270
43	P30754	493	O97939	33
44	P20908	866	P08427	55
45	Q25460	714	O97939	1049
46	Q4ZJN1	144	Q24940	159
47	P02745	90	P69929	201
48	Q9C5S0	26	P0C8W0	2
49	P19999	66	Q9Y2N7	35

50	Q805R9	37	P12111	2225
51	Q93WP8	42	Q9M0S4	50
52	P0DKQ9	60	O97939	299
53	Q05707	1708	O97939	861
54	P02745	66	P05997	897
55	Q05707	1725	P12111	2390
56	P15502	458	P20908	370
57	A1X158	128	P08125	301
58	Q9GQV7	76	Q05707	1405
59	Q25460	658	P35248	99
60	A1X158	428	P0C1N5	50
61	P20908	899	P08125	196
62	P20908	563	P14618	95
63	P30754	656	P02747	124
64	A1X158	46	O75636	73
65	A1EC31	119	P12111	1915
66	P30754	506	Q25460	839
67	P0C1X1	62	B2KPN7	10
68	A1X158	218	Q24940	231
69	O75636	61	Q05707	1587
70	P02459	313	Q02388	613

Table D-2 Independent Dataset for Hydroxylysine sites

	Positive hydroxylysine		**Negative hydroxylysine**	
#	**Uniprot ID**	**Position**	**Uniprot ID**	**Position**
1	P12111	2330	P12111	2979
2	P30754	344	P12111	1484
3	P20908	857	P35248	315
4	Q60994	64	Q05707	997
5	P20908	812	P20908	1785
6	P02461	277	P12111	3028
7	P30754	926	P26368	455
8	P02747	68	Q02388	2159
9	P02461	1099	P19999	115

10	Q05707	1691	P12111	1924

Table E-1 and E-2 present the results of individual independent protein sequences, which were not used in the model learning. The whole sequences were used as input to demonstrate the ability of RF-hydroxysite to identify the already known sites. The third column shows the positions of the known positive sites in the sequence (the red is the position misclassified by RF-Hydroxysite).

Residue	UniprotID	Site positions	# of sites	# sites of predicted
Proline	O97939	547	1	1
	D2Y171	50, 58, 62, 64	4	4
	Q9GQV7	26, 59, 83	3	3
	Q05707	1467, 1470, 1482, 1497, 1503, 1517, 1520, 1532, 1538, 1544, 1550, 1556, 1565, 1568, 1574, 1577, 1580, 1595, 1598, **1643**, 1656, 1659, 1662, 1665, 1668, 1674, 1677, 1680, 1686, 1689, 1704, 1715, 1726, 1729, 1732, 1735, 1741, 1747	38	37
			Accuracy	**97.83**

Table E-1. A sample of independent sequences tested with RF-Hydroxysite. The sites are experimentally confirmed as hydroxyproline

Residue	UniprotID	Site positions	# of sites	# sites of predicted
Lysine	P02459	287, 299, 308, 374, 419, 452, 464, 470, 527, 542, 608, 620	12	12
	P0A6N4	34	1	1
			Accuracy	**100**

Table E-2. A sample of independent sequences tested with RF-Hydroxysite. The sites are experimentally confirmed as hydroxylysine