**OH-PRED: prediction of protein hydroxylation sites by
incorporating adapted normal distribution bi-profile Bayes feature
extraction and physicochemical properties of amino acids**

Cang-Zhi Jia^{1,*}, Wen-Ying He¹, Yu-Hua Yao^{2,*}

¹Department of Mathematics, Dalian Maritime University, No. 1 Linghai Road,
Dalian 116026, China;

²College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China

*To whom correspondence should be addressed
(email:cangzhijia@dlmu.edu.cn;yaoyuhua2288@163.com).

The BLOSUM62 scoring matrix used in the study

	C	S	I	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
I	-1	1	5	-1	0	-2	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	0	-2	-2	-2
P	-3	-1	-1	7	-1	-2	-2	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	0	-1	4	0	-2	-2	-1	-1	-2	-1	-1	-1	-1	-1	0	-2	-2	-3
G	-3	0	-2	-2	0	6	0	-1	-2	-2	-2	-2	-2	-3	-4	-4	-3	-3	-3	-2
N	-3	1	0	-2	-2	0	6	1	0	0	1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	-1	-1	-2	-1	1	6	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	-1	-1	-1	-2	0	2	5	2	0	0	1	-2	-3	-3	-2	-3	-2	-3
Q	-3	0	-1	-1	-1	-2	0	0	2	5	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8	0	-1	-2	-3	-3	-3	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5	-1	-3	-2	-2	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5	1	2	1	0	-1	-1
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4	2	3	0	-1	-3
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4	1	0	-1	-2
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	2
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

Table S1. The jackknife results of models using ANBPB on different window sizes.

Residue type	Length	Sn(%)	Sp (%)	Acc(%)	MCC
Proline	17	8847	9118	8981	0.797
	15	88.62	91.49	90.04	0.801
	13	89.07	89.47	89.27	0.785
Lysine	17	91.75	92.78	92.27	0.845
	15	92.78	98.97	95.88	0.919
	13	90.72	1	95.36	0.911

Table S2. The 13 relevant physicochemical properties selected in this paper.

Property	AAindex	Abbreviation
Refractivity	MCMT640101	AA1
Flexibility	BHAR880101	AA2
Volume	CHOC750101	AA3
Transfer free energy to surface	BULH740101	AA4
Electron-ion interaction potential values	COSI940101	AA5
Hydrophilicity	HOPT810101	AA6
Polarity	CHAM820101	AA7
Hydrophobicity	EISD840101	AA8
Isoelectric point	ZIMJ680104	AA9
Optimized transfer energy parameter	OOBM850103	AA10
Optimized side chain interaction parameter	OOBM850105	AA11
Residue volume	BIGC670101	AA12
Normalized van der Waals volume	FAUJ880103	AA13

Table S3. The values corresponding to each amino acid for 13 physicochemical properties.

Name of the indices	AA1	AA2	AA3	AA4	AA5	AA6	AA7	AA8	AA9	AA10	AA11	AA12	AA13
Amino acids and their values	4.34	0.357	91.5	-0.2	0.0373	-0.5	0.046	0.25	6	0.46	1.16	52.6	1
	26.66	0.529	202	-0.12	0.0959	3	0.291	-1.76	10.76	-1.54	1.72	109.1	6.13
	13.28	0.463	135.2	0.08	0.0036	0.2	0.134	-0.64	5.41	1.31	1.97	75.7	2.95
	12	0.511	124.5	-0.2	0.1263	3	0.105	-0.72	2.77	-0.33	2.66	68.4	2.78
	35.77	0.346	117.7	-0.45	0.0829	-1	0.128	0.04	5.05	0.2	0.5	68.3	2.43
	17.56	0.493	161.1	0.16	0.0761	0.2	0.18	-0.69	5.65	-1.12	3.87	89.7	3.95
	17.26	0.497	155.1	-0.3	0.0058	3	0.151	-0.62	3.22	0.48	2.4	84.7	3.78
	0	0.544	66.4	0	0.005	0	0	0.16	5.97	0.64	1.63	36.3	0
	21.81	0.323	167.3	-0.12	0.0242	-0.5	0.23	-0.4	7.59	-1.31	0.86	91.9	4.66
	19.06	0.462	168.8	-2.26	0	-1.8	0.186	0.73	6.02	3.28	0.57	102	4
	18.78	0.365	167.9	-2.46	0	-1.8	0.186	0.53	5.98	0.43	0.51	102	4
	21.29	0.466	171.3	-0.35	0.0371	3	0.219	-1.1	9.74	-1.71	3.9	105.1	4.77
	21.64	0.295	170.8	-1.47	0.0823	-1.3	0.221	0.26	5.74	0.15	0.4	97.7	4.43
	29.4	0.314	203.4	-2.33	0.0946	-2.5	0.29	0.61	5.48	0.52	0.43	113.9	5.89
	10.93	0.509	129.3	-0.98	0.0198	0	0.131	-0.07	6.3	-0.58	2.04	73.6	2.72
	6.35	0.507	99.1	-0.39	0.0829	0.3	0.062	-0.26	5.68	-0.83	1.61	54.9	1.6
	11.01	0.444	122.1	-0.52	0.0941	-0.4	0.108	-0.18	5.66	-1.52	1.48	71.2	2.6
	42.53	0.305	237.6	-2.01	0.0548	-3.4	0.409	0.37	5.89	1.25	0.75	135.4	8.08
	31.53	0.42	203.6	-2.24	0.0516	-2.3	0.298	0.02	5.66	-2.21	1.72	116.2	6.47
	13.92	0.386	141.7	-1.56	0.0057	-1.5	0.14	0.54	5.96	0.54	0.59	85.1	3

Table S4. The performance of models trained with various features combination for hydroxyproline in jackknife test.

Features	Sn(%)	Sp (%)	Acc(%)	MCC
ANBPB	88.62	91.49	90.04	0.801
ANBPB+AA1	89.98	92.26	91.11	0.823
ANBPB+AA2	88.77	91.49	90.11	0.803
ANBPB+AA3	88.62	90.87	89.73	0.795
ANBPB+AA4	88.77	90.71	89.73	0.795
ANBPB+AA5	89.38	91.80	90.57	0.812
ANBPB+AA6	88.47	90.71	89.58	0.792
ANBPB+AA7	86.80	89.63	88.20	0.764
ANBPB+AA8	89.07	91.02	90.04	0.801
ANBPB+AA9	89.07	91.49	90.27	0.806
ANBPB+AA10	89.23	91.80	90.50	0.810
ANBPB+AA11	89.98	92.11	91.03	0.821
ANBPB+AA12	89.23	90.56	89.89	0.798
ANBPB+AA13	88.16	90.56	89.35	0.787
ANBPB +AA1+AA2	87.71	91.33	89.50	0.791
ANBPB +AA1+ AA5	90.74	92.11	91.42	0.829
ANBPB +AA1+ AA9	88.16	92.11	90.11	0.803
ANBPB +AA1+ AA10	88.77	91.64	90.19	0.804
ANBPB +AA1+ AA11	90.59	92.41	91.49	0.830
ANBPB +AA1+ AA5+ AA11	91.20	92.57	91.88	0.838

Table S5. The performance of models trained with various features combination for hydroxylysine in jackknife test.

Features	Sn(%)	Sp (%)	Acc(%)	MCC
ANBPB	92.78	98.97	95.88	0.919
ANBPB+AA1	93.81	98.97	96.39	0.929
ANBPB+AA2	94.85	96.91	95.88	0.918
ANBPB+AA3	94.85	98.97	96.91	0.939
ANBPB+AA4	93.81	98.97	96.39	0.929
ANBPB+AA5	92.78	96.91	94.85	0.898
ANBPB+AA6	94.85	98.97	96.91	0.939
ANBPB+AA7	94.85	98.97	96.91	0.939
ANBPB+AA8	92.78	97.94	95.36	0.908
ANBPB+AA9	91.75	95.88	93.81	0.877
ANBPB+AA10	94.85	96.91	95.88	0.918
ANBPB+AA11	94.85	95.88	95.36	0.907
ANBPB+AA12	95.88	98.97	97.42	0.949
ANBPB+AA13	94.85	97.94	96.39	0.928
ANBPB+AA12+AA1	93.81	98.97	96.39	0.929
ANBPB+AA12+AA3	93.81	98.97	96.39	0.929
ANBPB+AA12+AA4	9278	9897	9588	0.919
ANBPB+AA12+AA6	93.81	98.97	96.39	0.929
ANBPB+AA12+AA7	9485	98.97	96.91	0.939
ANBPB+AA12+AA13	9381	98.97	96.39	0.929

Table S6. The jackknife, 4-fold, 6-fold, 6-fold and 10-fold cross-validation results by the predictor on the unbalanced dataset.

Residue type	Cross-validation	Sn(%)	Sp (%)	Acc(%)	MCC
Proline	4-fold	80.95	96.73	92.56	0.804
	6-fold	80.40	96.58	92.41	0.798
	8-fold	80.80	96.98	92.72	0.808
	10-fold	81.23	97.32	93.04	0.817
	jackknife	83.46	95.26	92.17	0.796
Lysine	4-fold	90.04	98.58	96.34	0.905
	6-fold	92.03	99.29	97.39	0.931
	8-fold	90.05	98.21	96.08	0.901
	10-fold	90.48	98.32	95.81	0.897
	jackknife	90.72	98.60	96.60	0.909

Table S7. The jackknife test performances of OH-PRED and iHyd-PseAAC on the benchmark datasets in (Xu, 2014).

Residue type	Method	Sn(%)	Sp (%)	Acc(%)	MCC
Proline	iHyd-PseAAC [*]	70.68	89.03	78.42	0.52
	OH-PRED	75.82	82.32	80.93	0.52
Lysine	iHyd-PseAAC [*]	79.04	86.37	83.12	0.51
	OH-PRED	90.70	96.19	95.77	0.75

^{*}The data is directly obtained from (Xu, 2014).

Figure S1. The average values of F-score measurement on AA1, AA5 and AA11 properties for hydroxyproline

Figure S2. The average value of F-score measurement on AA12 property for hydroxylysine