Electronic Supplementary Material (ESI) for Molecular BioSystems.

SUPPLEMENTARY MATERIALS

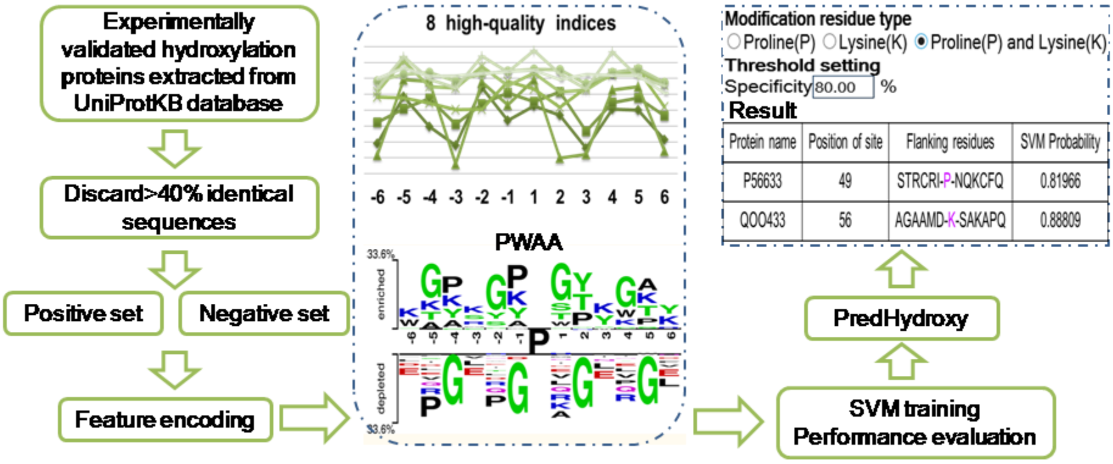
PredHydroxy: Computational prediction of protein hydroxylation site locations based on the primary structure

Shao-Ping Shi[a], Xiang Chen[b], Hao-Dong Xu[b], Jian-Ding Qiu[b,*]

[a] Department of Mathematics, Nanchang University, Nanchang, 330031, China

[b] Department of Chemistry, Nanchang University, Nanchang, 330031, China

* Corresponding Author: Jian-Ding Qiu, Department of Chemistry, Nanchang University, Nanchang 330031, China. Tel: + 86-791-83969518, e-mail: jdqiu@ncu.edu.cn



Modification residue type

○ Proline(P) ○ Lysine(K) ● Proline(P) and Lysine(K)

Threshold setting

Specificity 80.00 %

Result

Protein name	Position of site	Flanking residues	SVM Probability
P56633	49	STRCRI-P-NQKCFQ	0.81966
Q00433	56	AGAAMD-K-SAKAPQ	0.88809

Fig. S1 The system flow chart of the proposed method.

Table S1. The number of positive and negative sites in the training data and independent test dataset.

Dataset	Residue type	Number of positive sites	Number of negative sites
Training data	Proline	481	481
	Lysine	57	57
Independent test data	Proline	54	54
	Lysine	6	6

Table S2. The values corresponding to each amino acid for 8 high-quality indices (HQI).

	Name of the Clusters	Electric properties	Hydrop-hobicity	Alpha and turn propensities	Physicochemical properties	Residue propensity	Compo-sition	Beta propensity	Intrinsic propensities
	Name of the Indices	BLAM 930101	BIOV 880101	MAXF 760101	TSAJ 990101	NAKH 920108	CEDJ 970104	LIFS 790101	MIYS 990104
Amino Acids and their values	Ala (A)	0.96	16	1.43	89.3	9.36	7.9	0.92	-0.04
	Cys (C)	0.42	168	0.94	102.5	2.56	1.9	1.16	-0.38
	Asp (D)	0.42	-78	0.92	114.4	0.94	5.5	0.48	0.19
	Glu (E)	0.53	-106	1.67	138.8	0.94	7.1	0.61	0.23
	Phe (F)	0.59	189	1.19	190.8	10.99	3.9	1.25	-0.38
	Gly (G)	0	-13	0.46	63.8	6.17	7.1	0.61	0.09
	His (H)	0.57	50	0.98	157.5	0.47	2.1	0.93	-0.04
	Ile (I)	0.84	151	1.04	163	13.73	5.2	1.81	-0.34
	Lys (K)	0.73	-141	1.27	165.1	0.58	6.7	0.7	0.33
	Leu (L)	0.92	145	1.36	163.1	16.64	8.6	1.3	-0.37
	Met (M)	0.86	124	1.53	165.8	3.93	2.4	1.19	-0.3
	Asn (N)	0.39	-74	0.64	122.4	2.31	4	0.6	0.13
	Pro (P)	-2.5	-20	0.49	121.6	1.96	5.3	0.4	0.19
	Gln (Q)	0.8	-73	1.22	146.9	1.14	4.4	0.95	0.14
	Arg (R)	0.77	-70	1.18	190.3	0.27	4.9	0.93	0.07
	Ser (S)	0.53	-70	0.7	94.2	5.58	6.6	0.82	0.12
	Thr (T)	0.54	-38	0.78	119.6	4.68	5.3	1.12	0.03
	Val (V)	0.63	123	0.98	138.2	12.43	6.8	1.81	-0.29
	Trp (W)	0.58	145	1.01	226.4	2.2	1.2	1.54	-0.33
	Tyr (Y)	0.72	53	0.69	194.6	3.13	3.1	1.53	-0.29
References		1	2	3	4	5	6	7	8

HQI1 denotes BLAM930101 which is the center of electric properties cluster; HQI2 denotes BIOV880101 which is the center of hydrophobicity cluster; HQI3 denotes MAXF760101 which is the center of alpha and turn propensities cluster; HQI4 denotes TSAJ990101 which is the center of physicochemical properties cluster; HQI5 denotes NAKH920108 which is the center of residue propensity cluster; HQI6 denotes CEDJ970104 which is the center of composition cluster; HQI7 denotes LIFS790101 which is the center of beta propensity cluster; HQI8 denotes MIYS990104 which is the center of intrinsic propensities cluster.

References

1 M. Blaber, X.J. Zhang and B.W. Matthews, *Science*, 1993, **260**, 1637-1640.

2 V. Biou, J.F. Gibrat, J.M. Levin, B. Robson and J. Garnier, *Protein Eng.*, 1988, **2**, 185-191.

3 F.R. Maxfield and H.A. Scheraga, *Biochemistry*, 1976, **15**, 5138-5153.

4 J. Tsai, R. Taylor, C. Chothia and M. Gerstein, *J. Mol. Biol.*, 1999, **290**, 253-266.

5 H. Nakashima and K. Nishikawa, *FEBS Lett.*, 1992, **303**, 141-146.

6 J. Cedano, P. Aloy, J.A. Perez-Pons and E. Querol, *J. Mol. Biol.*, 1997, **266**, 594-600.

7 S. Lifson and C. Sander, *Nature*, 1979, **282**, 109-111.

8 S. Miyazawa and R. L. Jernigan, *Proteins* 1999, **34**, 49-68.

Table S3. The performance of models trained with various features for hydroxyproline.

Training features	Performance（%）							
	TP	FP	TN	FN	Sn	Sp	Acc	MCC
HQI1(BLAM930101)	389	120	361	92	80.87	75.05	77.96	56.02
HQI2(BIOV880101)	383	85	396	98	79.63	82.33	80.98	61.98
HQI3(MAXF760101)	379	149	332	102	78.79	69.02	73.91	48.05
HQI4(TSAJ990101)	418	131	350	63	86.9	72.77	79.84	60.27
HQI5(NAKH920108)	416	125	356	65	86.49	74.01	80.25	60.98
HQI6(CEDJ970104)	399	106	375	82	82.95	77.96	80.46	60.99
HQI7(LIFS790101)	372	83	398	109	77.34	82.74	80.04	60.17
HQI8(MIYS990104)	395	101	380	86	82.12	79	80.56	61.15
PWAA	396	81	400	85	82.32	83.16	82.74	65.49
Combination of HQI + PWAA	403	71	410	78	83.78	85.24	84.51	69.03

PWAA: position weight amino acids composition; Combination of HQI: the combination of HQI1 (BLAM930101), HQI2 (BIOV880101), HQI3 (MAXF760101), HQI4 (TSAJ990101), HQI5 (NAKH920108), HQI6 (CEDJ970104), HQI7 (LIFS790101) and HQI8 (MIYS990104).