Supplementary Information

Prediction of S-nitrosylation Sites by Integrating Support Vector Machine and Random Forest

Md. Mehedi Hasan[a], Balachandran Manavalan[b] , Mst. Shamima Khatun[a], and Hiroyuki Kurata[a,c*]
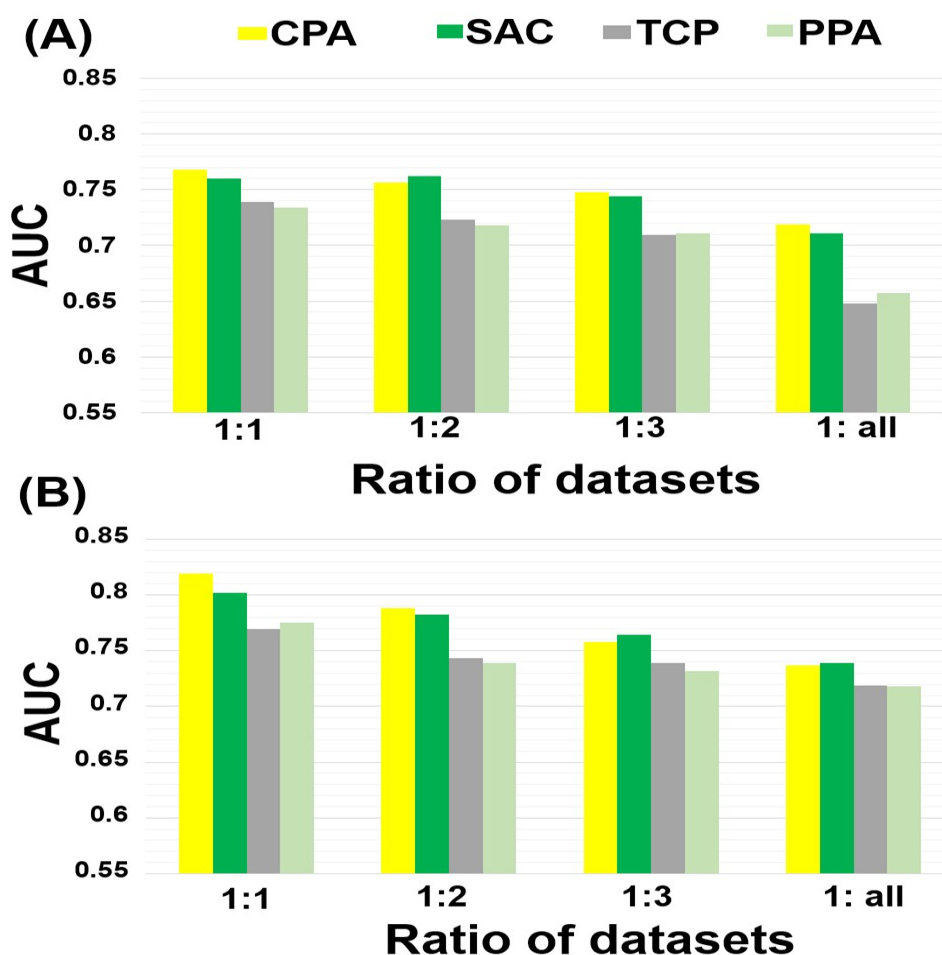
Figure S1. Effect of the ratio of positive vs negative training datasets on prediction performances by the ML models with a single encoding scheme of the CPA, SAC, TCP, or PPA.

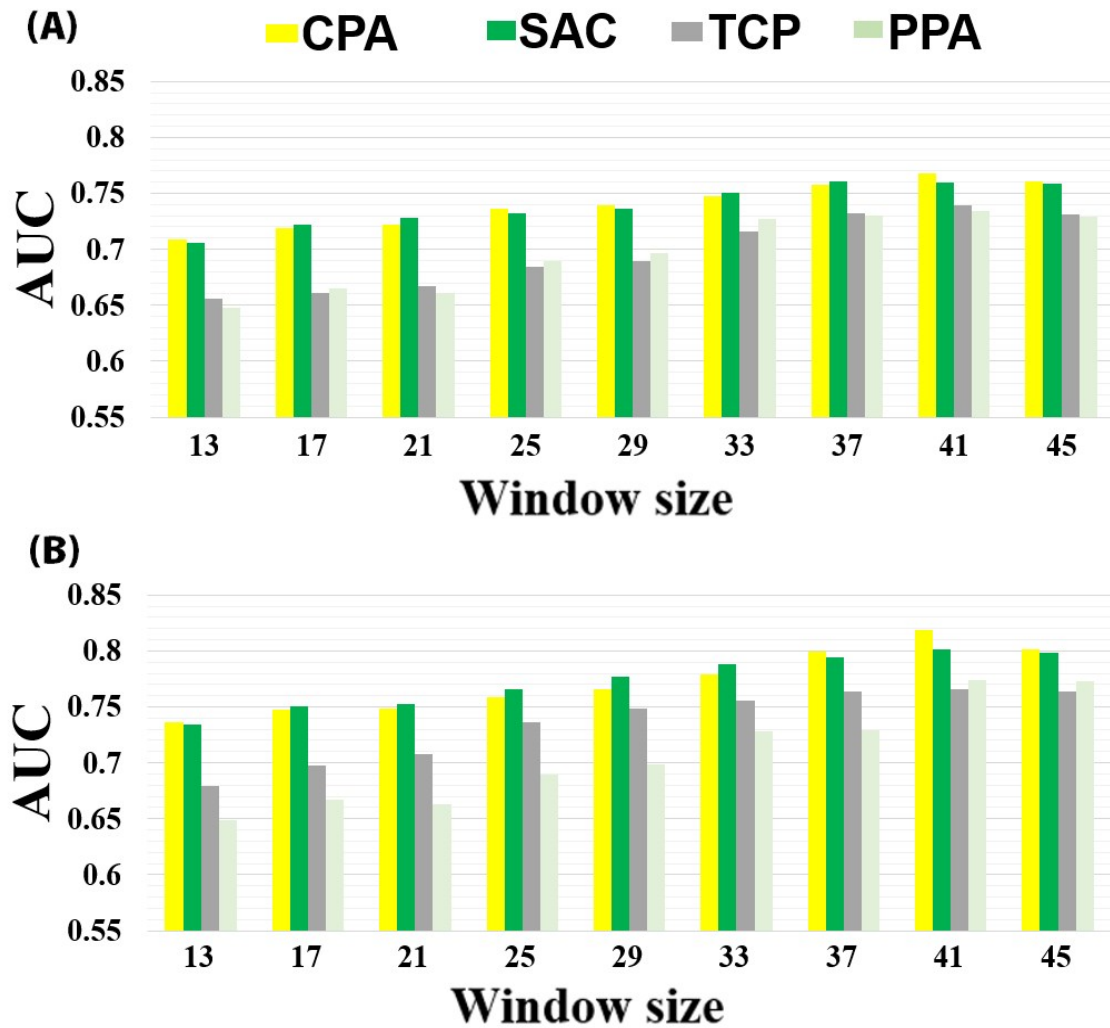 (A) AUC values provided by the SVM. (B) AUC values provided by the RF

Figure S2. Effect of window sizes on the AUC values by a single encoding scheme of the CPA, SAC, TCP, or PPA on the training datasets by 5-fold CV test.
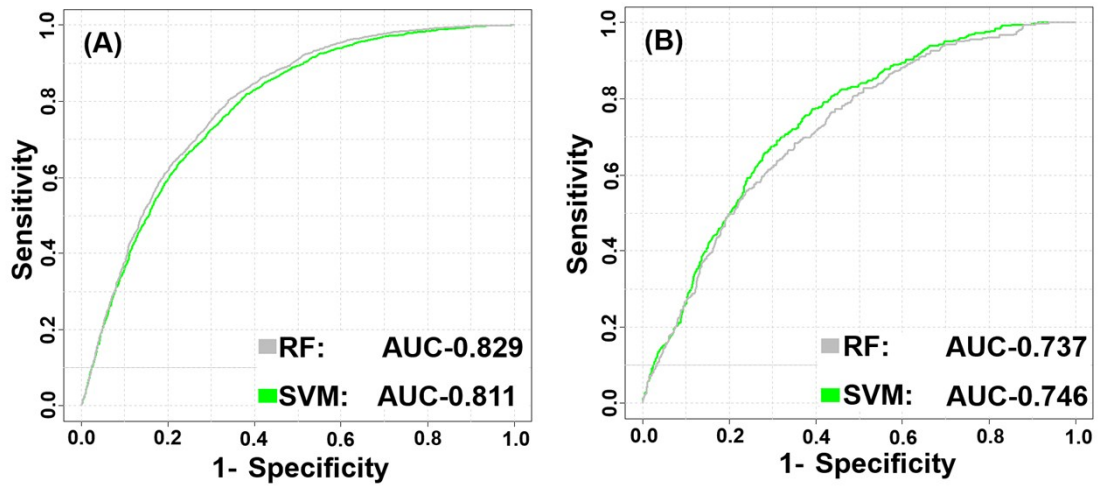(A) SVM and (B) RF algorithms.

Figure S3. Prediction performance provided by the RF and SVM with the sequential combination of the CPA, SAC, TCP, and PPA.
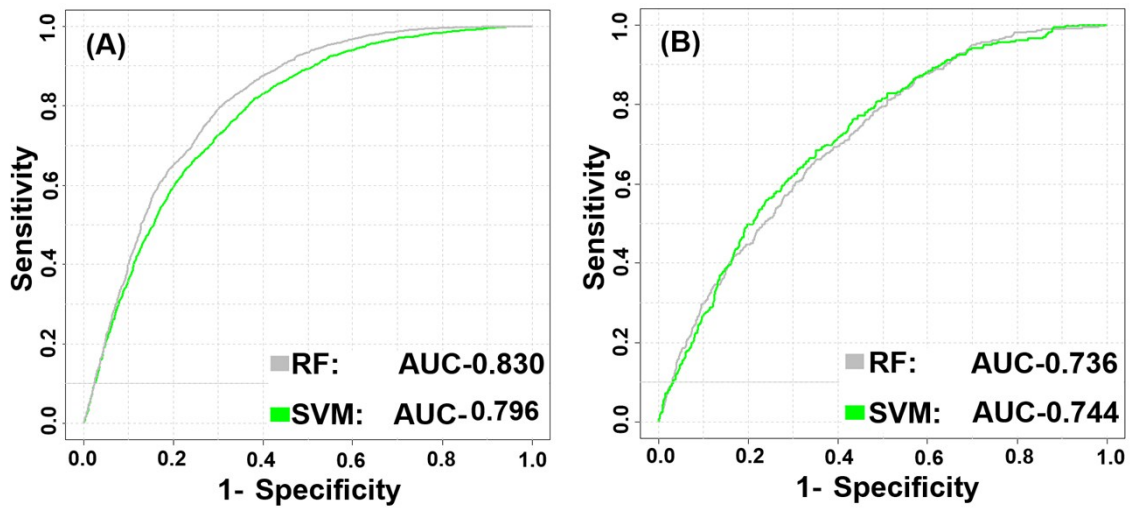(A) Training data. (B) Independent data.

Figure S4. Prediction performance by the meta-classifiers of the SVM and RF.
(A) Training data. (B) Independent data.

Table S1. Selected amino acid index properties for the PPA encoding scheme.

PPA property	Amino acid index properties						
BLAM930101	0.96	0.77	0.39	0.42	0.42	0.80	0.53
	0.00	0.57	0.84	0.92	0.73	0.86	0.59
	-2.50	0.53	0.54	0.58	0.72	0.63	
MAXF760101	1.43	1.18	0.64	0.92	0.94	1.22	1.67
	0.46	0.98	1.04	1.36	1.27	1.53	1.19
	0.49	0.70	0.78	1.01	0.69	0.98	
TSAJ990101	89.3	190.3	122.4	114.4	102.5	146.9	138.8
	63.8	157.5	163.0	163.1	165.1	165.8	190.8
	121.6	94.2	119.6	226.4	194.6	138.2	
NAKH920108	9.36	0.27	2.31	0.94	2.56	1.14	0.94
	6.17	0.47	13.73	16.64	0.58	3.93	10.99
	1.96	5.58	4.68	2.20	3.13	12.43	
CEDJ970104	7.9	4.9	4.0	5.5	1.9	4.4	7.1
	7.1	2.1	5.2	8.6	6.7	2.4	3.9
	5.3	6.6	5.3	1.2	3.1	6.8	
LIFS790101	0.92	0.93	0.60	0.48	1.16	0.95	0.61
	0.61	0.93	1.81	1.30	0.70	1.19	1.25
	0.40	0.82	1.12	1.54	1.53	1.81	
NOZY710101	0.5	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.5	1.8	1.8	0.0	1.3	2.5
	0.0	0.0	0.4	3.4	2.3	1.5	
HUTJ700103	154.33	341.01	207.90	194.91	219.79	235.51	223.16
	127.90	242.54	233.21	232.30	300.46	202.65	204.74
	179.93	174.06	205.80	237.01	229.15	207.60	
NAKH900109	9.25	3.96	3.71	3.89	1.07	3.17	4.80
	8.51	1.88	6.47	10.94	3.50	3.14	6.36
	4.36	6.26	5.66	2.22	3.28	7.55	
BIOV880101	16.	-70.	-74.	-78.	168.	-73.	-106.
	-13.	50.	151.	145.	-141.	124.	189.
	-20.	-70.	-38.	145.	53.	123.	
MIYS990104	-0.04	0.07	0.13	0.19	-0.38	0.14	0.23
	0.09	-0.04	-0.34	-0.37	0.33	-0.30	-0.38
	0.19	0.12	0.03	-0.33	-0.29	-0.29	

PUNT030101	-0.17	0.37	0.18	0.37	-0.06	0.26	0.15
	0.01	-0.02	-0.28	-0.28	0.32	-0.26	-0.41
	0.13	0.05	0.02	-0.15	-0.09	-0.17	
WOEC730101	7.0	9.1	10.0	13.0	5.5	8.6	12.5
	7.9	8.4	4.9	4.9	10.1	5.3	5.0
	6.6	7.5	6.6	5.3	5.7	5.6	
BASU050102	0.0728	0.0394	-0.0390	-0.0552	0.3557	0.0126	-0.0295
	-0.0589	0.0874	0.3805	0.3819	-0.0053	0.1613	0.4201
	-0.0492	-0.0282	0.0239	0.4114	0.3113	0.2947	
SUYM030101	-0.058	0.000	0.027	0.016	0.447	-0.073	-0.128
	0.331	0.195	0.060	0.138	-0.112	0.275	0.240
	-0.478	-0.177	-0.163	0.564	0.322	-0.052	

Table S2. Top 20 selected features based on the CPA, SAC, TCP, and PPA schemes by the WR test.

Sequential order	CPA		SAC		TCP		PPA	
	p-value	Selected pattern	*p*-value	Selected pattern	*p*-value	Selected pattern	*p*-value	Window position
1	6.03E-01	AL	3.06E-03	LL	2.54E-03	LKK	1.35E-02	(L, -9)
2	7.37E-06	LL	1.63E-04	L×E	6.52E-01	NRK	6.33E-04	(E, +11)
3	1.42E-04	K××A	1.70E-04	E×××R	2.49E-02	NLE	4.47E-02	(K, -2)
4	1.63E-05	Y×E	6.52E-02	E×L	9.95E-01	LIK	3.33E-06	(H, +6)
5	8.61E-01	R×××E	3.47E-02	L×G	2.56E-01	DER	5.44E-01	(D, -3)
6	6.90E-01	LA	8.41E-02	LA	9.37E-01	DAN	3.67E-02	(E, +6)
7	2.54E-03	A×××L	2.28E-03	AL	1.31E-02	DAV	7.57E-02	(V, -11)
8	9.17E-02	V×K	5.10E-03	L××××V	8.58E-01	ELL	1.14E-03	(S, -2)
9	6.16E-01	A××L	2.82E-01	V×K	9.38E-01	LLL	3.11E-03	(T, -12)
10	3.67E-01	KV	3.22E-02	E××R	4.25E-01	LAK	7.88E-02	(R, +6)
11	2.58E-01	E×××R	2.24E-07	K×L	9.84E-01	ELE	3.15E-02	(P, +8)
12	3.19E-01	W×Y	5.96E-04	E××××L	4.081E-04	DKL	1.02E-03	(V, -15)
13	8.25E-05	L××××A	2.68E-02	R×××E	7.25E-01	FKS	4.11E-02	(P, +2)
14	1.39E-03	L×××V	5.20E-01	L×××A	7.95E-05	GLK	3.49E-02	(K, +10)
15	2.61E-01	S×P	9.10E-02	R×L	8.29E-01	AAA	7.46E-02	(E, +7)
16	3.16E-02	KK	7.09E-04	KV	9.8E-03	EKK	5.44E-03	(N, +18)
17	1.15E-07	LS	1.34E-03	G×L	1.17E-01	ESV	3.11E-02	(Y, -13)
18	6.65E-01	E×L	5.88E-01	A××L	5.89E-08	AAL	3.45E-03	(R, +11)
19	7.42E-01	E×××D	3.86E-01	A×K	4.98E-02	ALL	2.22E-03	(G, +16)
20	5.57E-01	L×G	1.08E-03	A××××L	1.49E-02	DAK	2.54E-02	(L, +9)

A *p*-value is calculated by a two-sample *t*-test.

Table S3. Comparison of the amino acid residue patterns selected by different encodings.

Encoding	CPA		SAC		TCP		PPA	
Sample	SNO	non-SNO	SNO	non-SNO	SNO	non-SNO	SNO	non-SNO
Pattern	A×××L	LL	L×G	LL	NLE	LKK	(L, -9)	(E, +11)
		K××A	E××R	L×E	DKL	DA	(E, +6)	(K, -2)
		Y×E	K×L	E×××R	GLK	V	(V, -15)	(H, +6)
		L××××A	E××××L	AL	EKK		(R, +11)	(S, -2)
		L×××V	R×××E	L××××V	AAL		(K, +10)	(T, -12)
		KK	KV		ALL		(L, +9)	(P, +8)
		LS	G×L		DAK		(Y, -13)	(G, +16)
			A××××L					(P, +2)
								(N, 18)
#charged	0	3	3	2	4	2	3	2
#hydrophobic	2	3	3	3	3	3	2	2
#polar	0	1	0	0	1	0	1	4

Blue indicates charged amino acids; red hydrophobic amino acids; black polar amino acids.

#charged, the unique number of charged amino acid containing patterns;

#hydrophobic, the unique number of hydrophobic amino acid containing patterns;

#polar, the unique number of polar amino acid containing patterns.