

MSLVP: prediction of multiple subcellular localization of viral proteins using support vector machine

Anamika Thakur, Akanksha Rajput and Manoj Kumar*

Bioinformatics Centre, Institute of Microbial Technology, Council of Scientific and Industrial Research, Sector 39-A, Chandigarh-160036, India

* To whom correspondence should be addressed. Tel, 91-172-6665453; Fax, 91-172-2690585; 91-172-2690632; Email, manojk@imtech.res.in

Supplementary Figures

Figure S1: Flowchart for extracting physicochemical properties for example id “ZH0H040101” (*Calculate stability scale from knowledge-based atom-atom potential*) from AAindex database for SVM input

Figure S2: Receiver operating characteristic (ROC) curves of hybrid-II SVM predictive model of first-tier algorithm using *one-versus-other* classification approach at 90% sequence identity during 10-fold cross validation

Supplementary Tables

Table S1. List of 544 Amino Acid Indices in AAindex ver9.1

Table S2 (a). Performance of SVM based first-tier (single, double and multiple viral SCLs) predictive models using *one-versus-other* classification approach at 90% sequence identity using physicochemical properties during 10-fold cross validation

Table S2 (b). Performance of SVM based first-tier (single, double and multiple viral SCLs) predictive models using *one-versus-other* classification approach at 90% sequence identity using physicochemical properties on independent data set

Table S3. List of top-10 physicochemical ids extracted from 544 physicochemical properties of AAindex database for twenty-one predictive models at 90% sequence identity

Table S4. Performance of SVM based first-tier (single, double and multiple viral SCLs) predictive models using *one-versus-other* classification approach at 90% sequence identity on independent data set

Table S5. Performance of SVM based first-tier (single, double and multiple viral SCLs) predictive models using *one-versus-other* classification approach at 30% sequence identity on independent data set

Table S6. Performance of SVM based second-tier (single SCLs -8) predictive models using *one-versus-other* classification approach at 90% sequence identity on independent data set

Table S7. Performance of SVM based second-tier (single SCLs -8) predictive models using *one-versus-other* classification approach at 30% sequence identity on independent data set

Table S8. Performance of SVM based second-tier (double SCLs -4) predictive models using *one-versus-other* classification approach at 90% sequence identity on independent data set

Table S9. Performance of SVM based second-tier (multiple SCLs -6) predictive models using *one-versus-other* classification approach at 90% sequence identity on independent data set

Table S10. Performance of SVM based first-tier (single, double and multiple viral SCLs) predictive models using *one-versus-one* classification approach at 90% sequence identity on independent data set

Table S11. Performance of SVM based first-tier (single, double and multiple viral SCLs) predictive models using *one-versus-one* classification approach at 30% sequence identity on independent data set

Table S12. Performance of SVM based second-tier (single SCLs -8) predictive models using *one-versus-one* classification approach at 90% sequence identity on independent data set

Table S13. Performance of SVM based second-tier (single SCLs -8) predictive models using *one-versus-one* classification approach at 30% sequence identity on independent data set

Table S14. Performance of SVM based second-tier (double SCLs -4) predictive models using *one-versus-one* classification approach at 90% sequence identity on independent data set

Table S15. Performance of SVM based second-tier (multiple SCLs -6) predictive models using *one-versus-one* classification approach at 90% sequence identity on independent data set

Table S16. Performance of SVM based first-tier and second-tier (single, double and multiple viral SCLs) predictive models using *one-versus-one* classification approach at 90% and 30% sequence identity using **feature selection** on independent data set

Table S17. Comparison of MSLVP and other servers on independent dataset (90% sequence identity) using *one-versus-other* approach. Column 2 has number of viral protein entries for subcellular locations for independent validation and columns 3-6 have correctly predicted entries by respective method.

Supplementary Figures

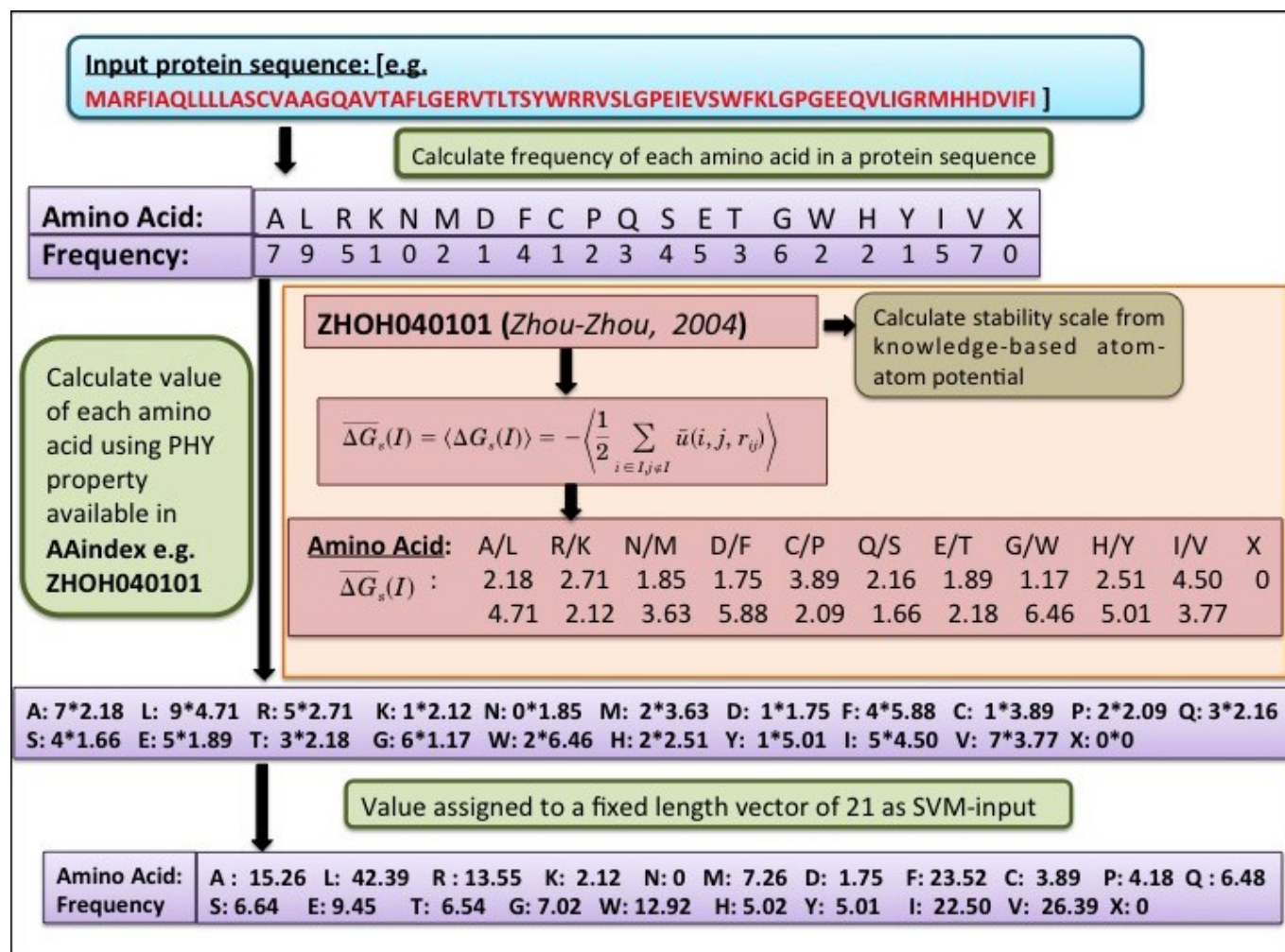


Figure S1: Flowchart for extracting physicochemical properties for example id "ZHOH040101" (Calculate stability scale from knowledge-based atom-atom potential) from AAindex database for SVM input

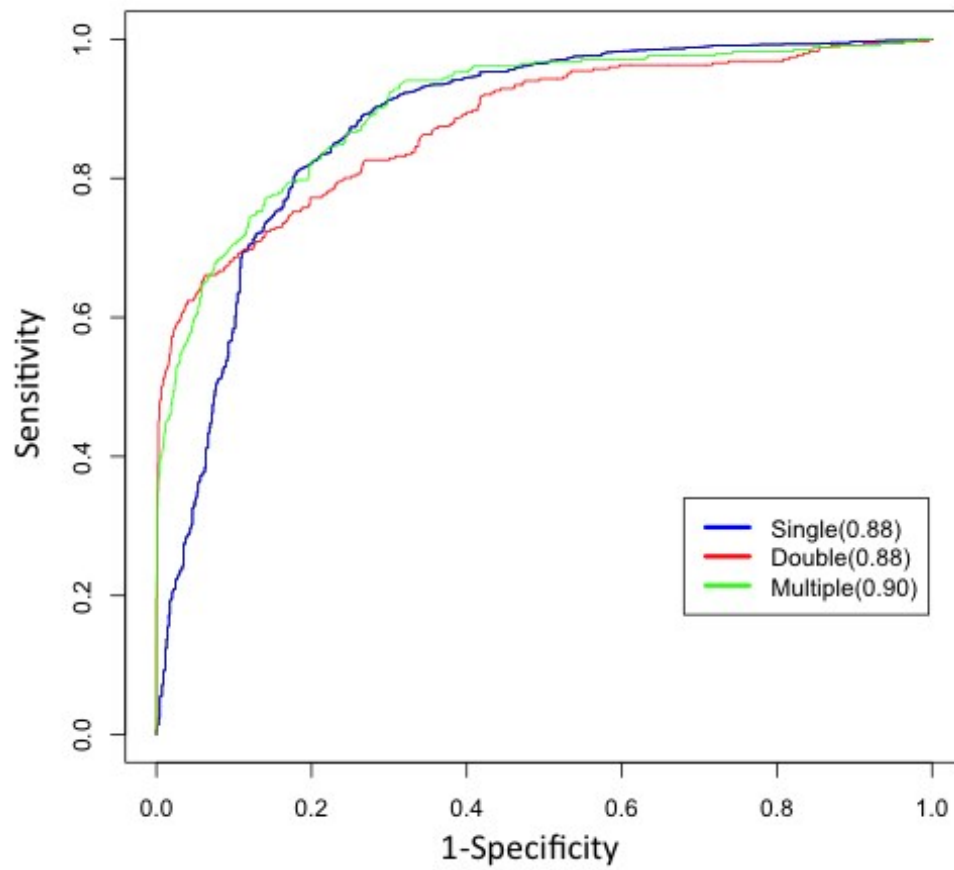


Figure S2: Receiver operating characteristics (ROC) curves of hybrid-II SVM predictive model of first-tier algorithm using *one-versus-other* classification approach at 90% sequence identity during 10-fold cross validation

Supplementary Tables

Table S1. List of 544 Amino Acid Indices in AAindex ver9.1

S. No.	Physicochemical IDs	Property
1.	ANDN920101	alpha-CH chemical shifts (Andersen et al., 1992)
2.	ARGP820101	Hydrophobicity index (Argos et al., 1982)
3.	ARGP820102	Signal sequence helical potential (Argos et al., 1982)
4.	ARGP820103	Membrane-buried preference parameters (Argos et al., 1982)
5.	BEGF750101	Conformational parameter of inner helix (Beghin-Dirkx, 1975)
6.	BEGF750102	Conformational parameter of beta-structure (Beghin-Dirkx, 1975)
7.	BEGF750103	Conformational parameter of beta-turn (Beghin-Dirkx, 1975)
8.	BHAR880101	Average flexibility indices (Bhaskaran-Ponnuswamy, 1988)
9.	BIGC670101	Residue volume (Bigelow, 1967)
10.	BIOV880101	Information value for accessibility; average fraction 35% (Biou et al., 1988)
11.	BIOV880102	Information value for accessibility; average fraction 23% (Biou et al., 1988)
12.	BROC820101	Retention coefficient in TFA (Browne et al., 1982)
13.	BROC820102	Retention coefficient in HFBA (Browne et al., 1982)
14.	BULH740101	Transfer free energy to surface (Bull-Breese, 1974)
15*.	BULH740102	Apparent partial specific volume (Bull-Breese, 1974)

* For complete list of 544 Physicochemical properties visit url : http://www.genome.jp/aaindex/AAindex/list_of_indices

Table S2 (a). Performance of SVM based first-tier (single, double and multiple viral SCLs) predictive models using *one-versus-other* classification approach at 90% sequence identity using physicochemical properties during 10-fold cross validation

Subcellular Location	Vector	Prec	Sen	Spec	Acc	MCC	AUC	GC
Single-Location	420 (Top-20)	89.41	90.25	73.58	85.45	0.64	0.89	$g:1e-05$ c:10
	630 (Top-30)	89.70	90.55	74.31	85.87	0.65	0.89	$g:1e-05$ c:50
	840 (Top-40)	90.17	88.31	76.20	84.82	0.64	0.89	$g:1e-05$ c:50
	1050 (Top-50)	89.59	89.49	74.31	85.12	0.64	0.90	$g:1e-05$ c:10
Double-Location	420 (Top-20)	69.88	64.10	95.25	90.68	0.62	0.87	$g:1e-05$ c:10
	630 (Top-30)	69.60	65.24	95.10	90.72	0.62	0.87	$g:1e-05$ c:50

	840 (Top-40)	72.17	63.53	95.79	91.05	0.63	0.87	g:1e-05 c:50
	1050 (Top-50)	72.64	63.53	95.88	91.14	0.63	0.87	g:1e-05 c:50
Multiple-Location	420 (Top-20)	75.00	74.11	95.96	92.89	0.70	0.93	g:1e-05 c:50
	630 (Top-30)	69.44	77.08	94.46	92.02	0.69	0.93	g:1e-05 c:10
	840 (Top-40)	70.46	77.38	94.70	92.27	0.69	0.93	g:1e-05 c:10
	1050 (Top-50)	70.88	76.79	94.84	92.31	0.69	0.93	g:1e-05 c:10

Prec, Precision; Sen, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Matthew's Correlation Coefficient; AUC, Area Under Curve;

Table S2 (b). Performance of SVM based first-tier (single, double and multiple viral SCLs) predictive models using *one-versus-other* classification approach at 90% sequence identity using physicochemical properties on independent data set

Subcellular Location	Vector	Prec	Sen	Spec	Acc	MCC	AUC	GC
Single-Location	420 (Top-20)	89.89	88.95	75.00	84.96	0.63	0.88	g:1e-05 c:10
	630 (Top-30)	90.91	89.47	77.63	86.09	0.66	0.90	g:1e-05 c:50
	840 (Top-40)	90.91	89.47	77.63	86.09	0.66	0.90	g:1e-05 c:50
	1050 (Top-50)	91.94	90.00	80.26	87.22	0.69	0.91	g:1e-05 c:10
Double-Location	420 (Top-20)	68.75	82.50	93.36	91.73	0.71	0.92	g:1e-05 c:10
	630 (Top-30)	73.91	85.00	94.69	93.23	0.75	0.90	g:1e-05 c:50
	840 (Top-40)	73.91	85.00	94.69	93.23	0.75	0.90	g:1e-05 c:50
	1050 (Top-50)	62.50	87.50	90.71	90.23	0.69	0.90	g:1e-05 c:50
Multiple-Location	420 (Top-20)	66.67	78.95	93.42	91.35	0.68	0.93	g:1e-05 c:50
	630 (Top-30)	63.27	81.58	92.11	90.60	0.67	0.94	g:1e-05 c:10
	840 (Top-40)	65.22	78.95	92.98	90.98	0.67	0.94	g:1e-05 c:10
	1050 (Top-50)	66.67	78.95	93.42	91.35	0.68	0.94	g:1e-05 c:10

Prec, Precision; Sen, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Matthew's Correlation Coefficient; AUC, Area Under Curve;

Table S3. List of top-10 physicochemical ids extracted from 544 physicochemical properties of AAIndex database for twenty-one predictive models at 90% sequence identity

Model name	Class	Physicochemical IDs*
	First-tier (Single, Double and Multiple viral SCLs)	
Model_1	First-Tier (Single Location)	ZHOH040102, LEVM780104, WOLS870103, ZHOH040101, LEVM780101, VINM940102, WERD780101, WOLS870101, YANJ020101, ZIMJ680104
Model_2	First-Tier (Double Location)	ZHOH040101, ZHOH040102, NAKH900106, RADA880108, YUTK870104, ZASB820101, ZHOH040103, ZIMJ680101, ZIMJ680102, ZIMJ680104
Model_3	First-Tier (Multiple Location)	PALJ810114, ZASB820101, ZHOH040101, ZHOH040103, ZIMJ680101, ZIMJ680104, YUTK870103, YUTK870104, ZHOH040102, ZIMJ680102
	Second-tier (Single SCLs-8)	
Model_4	Cytoplasm	ZHOH040102, PTIO830101, PALJ810102, NAKH900102, MONM990101, LEVM780104, ZHOH040101, RACS820111, RACS820110, RACS820108
Model_5	Nucleus	ZHOH040101, ZHOH040102, WOLS870101, PALJ810104, NISK860101, MIYS990105, MIYS990104, MIYS990103, MIYS850101, LEVM780101
Model_6	Extracell	ZIMJ680104, MAXF760106, WOLS870103, LEVM760106, LEVM760105, WOLR810101, MEIH800102, YANJ020101, LEVM780104, LEVM780101
Model_7	Single-Pass Membrane	SUEM840101, RICJ880108, VINM940102, RACS820112, PRAM900102, PARJ860101, PALJ810109, PALJ810107, PALJ810101, LEVM780104
Model_8	Multi-Pass Membrane	PONP800103, NISK800101, OOBM770103, OOBM770101, LEVM760102, WERD780101, VINM940104, VINM940102, SNEP660101, ROSM880105
Model_9	Endoplasmic Reticulum	ZIMJ680104, ZIMJ680101, ZHOH040102, ZHOH040101, YUTK870102, YUTK870101, YANJ020101, WOLS870103, WOLS870102, WOLS870101
Model_10	Others Single	RICJ880117, TANS770106, FODM020101, FAUJ880106, ANDN920101, ZHOH040102, WOLS870103, WERD780101, VINM940101, TANS770103
Model_11	Capsid	QIAN880120, PONP800106, RACS770103, RACS770102, RACS770101, QIAN880139, QIAN880137, QIAN880136, QIAN880134, QIAN880133
	Second-tier (Double SCLs-4)	
Model_12	Host nucleus. Host cytoplasm. [N.C.]	ZHOH040102, ZHOH040101, WILM950102, RADA880108, PARJ860101, MIYS990102, MEEJ810102, WERD780101, ROBB790101, NAKH900106
Model_13	Host cytoplasm. Host cell membrane; Peripheral membrane protein; Cytoplasmic side. [C.CM.]	ZIMJ680105, ZIMJ680104, ZIMJ680103, ZIMJ680102, ZIMJ680101, ZHOH040103, ZHOH040102, ZHOH040101, ZASB820101, YUTK870104
Model_14	Host nucleus. Extracell [N.E.]	ZIMJ680105, ZIMJ680104, ZIMJ680102, ZIMJ680101, ZHOH040103, ZHOH040102, ZHOH040101, ZASB820101, YUTK870104, YUTK870103
Model_15	Others double	NAKH900104, RICJ880107, RADA880108, NAKH900106, WOLS870101, FAUJ830101, FASG890101, BROCC820102, ZHOH040101, ROSM880102

	Second-tier (Multiple SCLs-6)	
Model_16	Virion tegument. Host cytoplasm. Host nucleus. [VT.C.N.]	ROBB760107, TANS770110, RACS820110, RACS820104, RACS820101, PRAM820101, PALJ810114, PALJ810106, MUNV940105, MUNV940104
Model_17	Host cell membrane; Lipid-anchor; Cytoplasmic side. Host cytoplasm › host perinuclear region. Secreted. [CM. C.S.]	ZIMJ680105, ZIMJ680104, ZIMJ680102, ZIMJ680101, ZHOH040103, ZHOH040102, ZHOH040101, ZASB820101, YUTK870104, YUTK870103
Model_18	Host nucleus › host nucleolus. Host cytoplasm. Secreted. [N.C.S.]	ZIMJ680104, ZIMJ680101, ZHOH040103, ZHOH040101, ZASB820101, YUTK870102, YUTK870101, YANJ020101, WOLS870103, WOLS870102
Model_19	Host cytoplasm. Host nucleus. Host mitochondrion. [C.N.M.]	ZIMJ680105, ZIMJ680104, ZIMJ680102, ZIMJ680101, ZHOH040103, ZHOH040102, ZHOH040101, ZASB820101, YUTK870104, YUTK870103
Model_20	Multiple	NAKH900106, RACS820107, PONP800107, NAKH900104, MIYS990102, SNEP660103, RICJ880117, RICJ880106, RACS820112, PALJ810114
Model_21	Others triple	TANS770101, QIAN880132, QIAN880104, PRAM820102, PALJ810114, PALJ810102, PALJ810101, NAGK730101, LEVM780104, VINM940104

* For further information of Physicochemical properties visit url: http://www.genome.jp/aaindex/AAindex/list_of_indices

Table S4. Performance of SVM based first-tier (single, double and multiple viral SCLs) predictive models using *one-versus-other* classification approach at 90% sequence identity on independent data set

Subcellular Location	Dataset	Properties	Prec	Sen	Spec	Acc	MCC	AUC	Parameters
Single-Location	V ^(190p+76n)	AAC	90.16	86.84	76.32	83.83	0.62	0.90	g:0.05 c:1
		DPC	91.43	84.21	80.26	83.08	0.61	0.89	g:0.01 c:10
		PHY	89.67	86.84	75.00	83.46	0.60	0.89	g:1e-05 c:50
		Hybrid-I	88.24	94.74	68.42	87.22	0.67	0.92	g:0.01 c:5
		Hybrid-II	87.76	90.53	68.42	84.21	0.60	0.85	g:1e-05 c:200
Double-Location	V ^(40p+226n)	AAC	77.78	52.50	97.35	90.60	0.59	0.84	g:0.01 c:5
		DPC	56.00	70.00	90.27	87.22	0.55	0.88	g:0.01 c:5
		PHY	77.78	70.00	96.46	92.48	0.69	0.92	g:1e-05 c:5
		Hybrid-I	92.00	57.50	99.25	93.77	0.70	0.87	g:0.01 c:5
		Hybrid-II	78.13	62.50	96.90	91.73	0.65	0.93	g:1e-05 c:5
Multiple-Location	V ^(38p+228n)	AAC	90.63	76.32	98.68	95.49	0.81	0.96	g:0.05 c:10
		DPC	70.27	68.42	95.18	91.35	0.64	0.92	g:0.05 c:10
		PHY	100.00	55.26	100.00	93.61	0.72	0.95	g:1e-05 c:10
		Hybrid-I	93.55	76.32	99.12	95.86	0.82	0.95	g:0.01 c:50
		Hybrid-II	86.96	52.63	98.68	92.11	0.64	0.95	g:1e-05 c:5
		Hybrid-I	89.23	86.57	95.08	92.35	0.82		

AVERAGE	V ^(268p+530n)	Hybrid-II	86.45	80.97	93.58	89.35	0.76		
---------	--------------------------	-----------	-------	-------	-------	-------	------	--	--

Prec, Precision; Sen, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Matthew's Correlation Coefficient; AUC, Area Under Curve; AAC, Amino Acid Composition; DPC, Dipeptide Composition; PHY, Physicochemical properties; Hybrid-I, AAC+DPC; Hybrid-II, AAC+DPC+PHY;

Table S5. Performance of SVM based first-tier (single, double and multiple viral SCLs) predictive models using *one-versus-other* classification approach at 30% sequence identity on independent data set

Subcellular Location	Dataset	Properties	Prec	Sen	Spec	Acc	MCC	AUC	Parameters
Single-Location	V ^(101p+32n)	AAC	81.01	63.37	53.12	60.90	0.14	0.57	g:0.05 c:1
		DPC	79.35	72.28	40.62	64.66	0.12	0.60	g:0.05 c:10
		Hybrid-I	82.93	67.33	56.25	64.66	0.21	0.66	g:0.01 c:5
		Hybrid-II	100.00	100.00	100.00	100.00	1.00	1.00	g:0.001 c:10
Double-Location	V ^(17p+116n)	AAC	16.67	11.76	91.38	81.20	0.04	0.60	g:0.05 c:1
		DPC	20.00	64.71	62.07	62.41	0.18	0.70	g:0.01 c:10
		Hybrid-I	19.30	64.71	60.34	60.90	0.17	0.67	g:0.005 c:100
		Hybrid-II	27.59	47.06	81.90	77.44	0.23	0.70	g:1e-05 c:10
Multiple-Location	V ^(16p+117n)	AAC	30.43	43.75	86.32	81.20	0.26	0.69	g:0.05 c:1
		DPC	15.63	62.50	53.85	54.89	0.11	0.60	g:0.05 c:1
		Hybrid-I	20.00	68.75	62.39	63.16	0.21	0.70	g:0.01 c:50
		Hybrid-II	33.33	25.00	93.16	84.96	0.21	0.72	g:1e-05 c:10
AVERAGE	V ^(134p+265n)	Hybrid-I	46.39	67.16	60.75	62.91	0.26		
		Hybrid-II	79.58	84.33	89.06	87.47	0.72		

Prec, Precision; Sen, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Matthew's Correlation Coefficient; AUC, Area Under Curve; AAC, Amino Acid Composition; DPC, Dipeptide Composition; Hybrid-I, AAC+DPC; Hybrid-II, AAC+DPC+PHY;

Table S6. Performance of SVM based second-tier (single SCLs -8) predictive models using *one-versus-other* classification approach at 90% sequence identity on independent data set

Subcellular Location	Dataset	Properties	Prec	Sen	Spec	Acc	MCC	AUC	Parameters
Cytoplasm	V ^(28p+162n)	AAC	83.33	71.43	97.53	93.68	0.74	0.89	g:0.05 c:10
		DPC	81.82	64.29	97.53	92.63	0.68	0.91	g:0.01 c:5
		PHY	95.00	67.86	99.38	94.74	0.78	0.94	g:1e-05 c:10
		Hybrid-I	70.97	78.57	94.44	92.11	0.70	0.93	g:0.01 c:10

		Hybrid-II	81.48	78.57	96.91	94.21	0.77	0.95	g:1e-05 c:50
Nucleus	V ^(42p+148n)	AAC	91.18	73.81	97.97	92.63	0.78	0.96	g:0.05 c:10
		DPC	77.50	73.81	93.92	89.47	0.69	0.92	g:0.01 c:10
		PHY	91.18	73.81	97.97	92.63	0.78	0.93	g:1e-05 c:5
		Hybrid-I	82.50	78.57	95.27	91.58	0.75	0.94	g:0.005 c:10
		Hybrid-II	84.21	76.19	95.95	91.58	0.75	0.95	g:1e-05 c:5
Endoplasmic Reticulum	T ^(3p+187n)	AAC	100.00	66.67	100.00	99.47	0.81	0.84	g:0.005 c:10
		DPC	100.00	66.67	100.00	99.47	0.81	0.97	g:0.005 c:5
		PHY	100.00	66.67	100.00	99.47	0.81	0.74	g:1e-05 c:0.05
		Hybrid-I	100.00	66.67	100.00	99.47	0.81	0.94	g:0.001 c:5
		Hybrid-II	100.00	66.67	100.00	99.47	0.81	0.70	g:1e-05 c:5
Single-pass Membrane	V ^(75p+115n)	AAC	66.67	90.67	70.43	78.42	0.60	0.90	g:0.05 c:10
		DPC	75.58	86.67	81.74	83.68	0.67	0.93	g:0.01 c:5
		PHY	77.27	90.67	82.61	85.79	0.72	0.92	g:1e-05 c:5
		Hybrid-I	78.16	90.67	83.48	86.32	0.73	0.94	g:0.01 c:5
		Hybrid-II	77.38	86.67	83.48	84.74	0.69	0.92	g:1e-05 c:50
Multi-pass Membrane	T ^(27p+163n)	AAC	81.82	66.67	97.55	93.16	0.70	0.95	g:0.01 c:0.5
		DPC	85.00	62.96	98.16	93.16	0.70	0.94	g:0.01 c:1
		PHY	75.00	77.78	95.71	93.16	0.72	0.91	g:1e-05 c:10
		Hybrid-I	69.70	85.19	93.87	92.63	0.73	0.96	g:0.005 c:50
		Hybrid-II	67.74	77.78	93.87	91.58	0.68	0.94	g:1e-05 c:100
Secreted	V ^(4p+186n)	AAC	50.00	50.00	98.92	97.89	0.49	0.96	g:0.01 c:5
		DPC	12.50	25.00	96.24	94.74	0.15	0.68	g:0.05 c:10
		PHY	37.50	75.00	97.31	96.84	0.52	0.98	g:1e-05 c:5
		Hybrid-I	100.00	50.00	100.00	99.47	0.71	0.89	g:0.005 c:5
		Hybrid-II	50.00	50.00	98.92	97.89	0.49	0.83	g:1e-05 c:5
Others-single	V ^(14p+176n)	AAC	54.55	42.86	97.16	93.16	0.45	0.81	g:0.01 c:10
		DPC	80.00	28.57	99.43	94.21	0.46	0.84	g:0.05 c:10
		PHY	80.00	57.14	98.86	95.79	0.66	0.86	g:1e-05 c:200
		Hybrid-I	75.00	42.86	98.86	94.74	0.54	0.88	g:0.01 c:5

		Hybrid-II	66.67	57.14	97.73	94.74	0.59	0.89	g:1e-05 c:50
Capsid	V ^(83p+189n)	AAC	67.65	55.42	88.36	78.31	0.47	0.82	g:0.01 c:1
		DPC	53.00	63.86	75.13	71.69	0.37	0.77	g:0.001 c:10
		PHY	59.74	55.42	83.60	75.00	0.40	0.77	g:1e-05 c:1
		Hybrid-I	54.81	68.67	75.13	73.16	0.42	0.80	g:0.01 c:1
		Hybrid-II	52.44	51.81	79.37	70.96	0.31	0.74	g:1e-05 c:1
AVERAGE	V ^(276p+1326n)	Hybrid-I	69.38	77.17	93.78	91.22	0.68		
		Hybrid-II	69.64	70.65	93.59	89.64	0.64		

Prec, Precision; Sen, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Matthew's Correlation Coefficient; AUC, Area Under Curve; AAC, Amino Acid Composition; DPC, Dipeptide Composition; Hybrid-I, AAC+DPC; Hybrid-II, AAC+DPC+PHY;

Table S7. Performance of SVM based second-tier (single SCLs -8) predictive models using *one-versus-other* classification approach at 30% sequence identity on independent data set

Subcellular Location	Dataset	Properties	Prec	Sen	Spec	Acc	MCC	AUC	Parameters
Cytoplasm	V ^(11p+90n)	AAC	22.73	45.45	81.11	77.23	0.20	0.77	g:0.01 c:5
		DPC	33.33	72.73	82.22	81.19	0.40	0.73	g:0.01 c:50
		Hybrid-I	26.67	72.73	75.56	75.25	0.33	0.76	g:0.01 c:5
		Hybrid-II	21.05	36.36	83.33	78.22	0.16	0.67	g:1e-05 c:5
Nucleus	V ^(14p+87n)	AAC	29.63	57.14	78.16	75.25	0.28	0.83	g:0.01 c:5
		DPC	50.00	85.71	86.21	86.14	0.58	0.90	g:0.01 c:1
		Hybrid-I	54.55	85.71	88.51	88.12	0.62	0.92	g:0.0001 c:100
		Hybrid-II	41.38	85.71	80.46	81.19	0.51	0.88	g:1e-05 c:10
Endoplasmic Reticulum	V ^(1p+100n)	AAC	0	0	99.00	98.02	-0.01	0.69	g:0.01 c:50
		DPC	0	0	99.00	98.02	-0.01	0.74	g:0.01 c:50
		Hybrid-I	0	0	99.00	98.02	-0.01	0.82	g:0.01 c:5
		Hybrid-II	0	0	98.00	97.03	-0.01	0.25	g:1e-05 c:5
Single-pass Membrane	V ^(45p+56n)	AAC	64.81	77.78	66.07	71.29	0.44	0.79	g:0.01 c:1
		DPC	70.59	80.00	73.21	76.24	0.53	0.81	g:0.01 c:5
		Hybrid-I	72.55	82.22	75.00	78.22	0.57	0.83	g:0.005 c:5
		Hybrid-II	66.67	66.67	73.21	70.30	0.40	0.78	g:1e-05 c:5
Multi-pass		AAC	69.57	72.73	91.14	87.13	0.63	0.90	g:0.001 c:5

Membrane	V ^(22p+79n)	DPC	62.96	77.27	87.34	85.15	0.60	0.91	g:0.01 c:1
		Hybrid-I	60.00	81.82	84.81	84.16	0.60	0.90	g:0.0005 c:5
		Hybrid-II	72.00	81.82	91.14	89.11	0.70	0.92	g:1e-05 c:50
Secreted	V ^(3p+98n)	AAC	14.29	33.33	93.88	92.08	0.18	0.67	g:0.01 c:50
		DPC	14.29	33.33	93.88	92.08	0.18	0.62	g:0.001 c:50
		Hybrid-I	8.33	33.33	88.78	87.13	0.12	0.66	g:0.0005 c:50
		Hybrid-II	16.67	33.33	94.90	93.07	0.20	0.49	g:1e-05 c:5
Others-single	V ^(8p+93n)	AAC	80.00	50.00	98.92	95.05	0.61	0.72	g:0.05 c:10
		DPC	16.67	37.50	83.87	80.20	0.15	0.64	g:0.01 c:10
		Hybrid-I	12.50	50.00	69.89	68.32	0.12	0.67	g:0.01 c:5
		Hybrid-II	28.57	75.00	83.87	83.17	0.39	0.80	g:1e-05 c:100
Capsid	V ^(36p+101n)	AAC	45.10	63.89	72.28	70.07	0.33	0.72	g:0.0005 c:100
		DPC	45.83	30.56	87.13	72.26	0.20	0.62	g:0.0005 c:10
		Hybrid-I	35.85	52.78	66.34	62.77	0.17	0.65	g:0.0005 c:10
		Hybrid-II	35.00	58.33	61.39	60.58	0.17	0.60	g:1e-05 c:1
AVERAGE	V ^(140p+704n)	Hybrid-I	42.86	70.71	81.25	79.50	0.43		
		Hybrid-II	44.44	65.71	83.66	80.69	0.43		

Prec, Precision; Sen, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Matthew's Correlation Coefficient; AUC, Area Under Curve; AAC, Amino Acid Composition; DPC, Dipeptide Composition; Hybrid-I, AAC+DPC; Hybrid-II, AAC+DPC+PHY;

Table S8. Performance of SVM based second-tier (double SCLs -4) predictive models using *one-versus-other* classification approach at 90% sequence identity on independent data set

Subcellular Location	Dataset	Properties	Prec	Sen	Spec	Acc	MCC	AUC	Parameters
Host cytoplasm. Host cell membrane; Host cell membrane; Peripheral protein; Cytoplasmic side.	V ^(7p+33n)	AAC	85.71	85.71	96.97	95.00	0.83	0.94	g:0.01 c:10
		DPC	100.00	85.71	100.00	97.50	0.91	0.94	g:0.001 c:10
		PHY	100.00	85.71	100.00	97.50	0.91	0.99	g:1e-05 c:1
		Hybrid-I	100.00	85.71	100.00	97.50	0.91	0.94	g:0.0005 c:50
		Hybrid-II	100.00	85.71	100.00	97.50	0.91	0.98	g:1e-05 c:5
Host nucleus. Host cytoplasm.	V ^(15p+25n)	AAC	72.22	86.67	80.00	82.50	0.65	0.93	g:0.01 c:5
		DPC	85.71	80.00	92.00	87.50	0.73	0.91	g:0.01 c:5
		PHY	100.00	86.67	100.00	95.00	0.90	0.94	g:1e-05 c:200

		Hybrid-I	85.71	80.00	92.00	87.50	0.73	0.95	g:0.01 c:10
		Hybrid-II	85.71	80.00	92.00	87.50	0.73	0.91	g:1e-05 c:10
Host nucleus. Extracell	V ^(2p+38n)	AAC	66.67	100.00	97.37	97.50	0.81	0.99	g:0.001 c:100
		DPC	100.00	100.00	100.00	100.00	1.00	1.00	g:0.01 c:10
		PHY	100.00	50.00	100.00	97.50	0.70	1.00	g:1e-05 c:0.05
		Hybrid-I	100.00	50.00	100.00	97.50	0.70	0.83	g:1e-05 c:50
		Hybrid-II	100.00	50.00	100.00	97.50	0.70	1.00	g:1e-05 c:0.5
Others-double	T ^(17p+23n)	AAC	83.33	88.24	86.96	87.50	0.75	0.93	g:0.01 c:5
		DPC	83.33	88.24	86.96	87.50	0.75	0.96	g:0.01 c:10
		PHY	85.00	100.00	86.96	92.50	0.86	0.98	g:1e-05 c:5
		Hybrid-I	80.95	100.00	82.61	90.00	0.82	0.96	g:0.005 c:5
		Hybrid-II	72.73	94.12	73.91	82.50	0.68	0.95	g:0.0001 c:1
AVERAGE	V ^(41p+119n)	Hybrid-I	85.71	87.80	94.96	93.13	0.82		
		Hybrid-II	81.40	85.37	93.28	91.25	0.77		

Prec, Precision; Sen, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Matthew's Correlation Coefficient; AUC, Area Under Curve; AAC, Amino Acid Composition; DPC, Dipeptide Composition; PHY, Physicochemical properties; Hybrid-I, AAC+DPC; Hybrid-II, AAC+DPC+PHY;

Table S9. Performance of SVM based second-tier (multiple SCLs -6) predictive models using *one-versus-other* classification approach at 90% sequence identity on independent data set

Subcellular Location	Dataset	Properties	Prec	Sen	Spec	Acc	MCC	AUC	Parameters
Host cell membrane; Lipid-anchor; Cytoplasmic side. Host cytoplasm > host perinuclear region. Secreted.	V ^(3p+35n)	AAC	100.00	100.00	100.00	100.00	1.00	1.00	g:0.01 c:10
		DPC	100.00	100.00	100.00	100.00	1.00	1.00	g:0.001 c:10
		PHY	100.00	66.67	100.00	97.37	0.81	1.00	g:1e-05 c:0.05
		Hybrid-I	75.00	100.00	97.14	97.37	0.85	1.00	g:1e-05 c:50
		Hybrid-II	100.00	66.67	100.00	97.37	0.81	1.00	g:1e-05 c:0.5
Host cytoplasm. Host nucleus. Host mitochondrion.	V ^(2p+36n)	AAC	100.00	100.00	100.00	100.00	1.00	1.00	g:0.01 c:10
		DPC	100.00	100.00	100.00	100.00	1.00	1.00	g:0.001 c:10
		PHY	100.00	50.00	100.00	97.37	0.70	1.00	g:1e-05 c:0.5
		Hybrid-I	50.00	100.00	94.44	94.74	0.69	1.00	g:0.0001 c:10
		Hybrid-II	100.00	100.00	100.00	100.00	1.00	1.00	g:1e-05 c:5

Host nucleus > host nucleolus. Host cytoplasm. Secreted.	V ^(4p+34n)	AAC	100.00	75.00	100.00	97.37	0.85	1.00	g:0.0005 c:10
		DPC	100.00	100.00	100.00	100.00	1.00	1.00	g:0.001 c:50
		PHY	100.00	75.00	100.00	97.37	0.85	0.99	g:1e-05 c:0.01
		Hybrid-I	75.00	100.00	97.14	97.37	0.85	1.00	g:0.0001 c:100
		Hybrid-II	100.00	100.00	100.00	100.00	1.00	1.00	g:1e-05 c:10
Virion tegument. Host cytoplasm. Host nucleus.	V ^(4p+34n)	AAC	75.00	75.00	97.06	94.74	0.72	0.99	g:0.01 c:50
		DPC	75.00	75.00	97.06	94.74	0.72	0.96	g:0.005 c:50
		PHY	100.00	75.00	100.00	97.37	0.85	0.98	g:1e-05 c:5
		Hybrid-I	100.00	75.00	100.00	97.37	0.85	0.99	g:0.01 c:10
		Hybrid-II	100.00	75.00	100.00	97.37	0.85	0.88	g:1e-05 c:10
Others-triple	V ^(14p+24n)	AAC	81.25	92.86	87.50	89.47	0.79	0.97	g:0.05 c:10
		DPC	80.00	85.71	87.50	86.84	0.72	0.95	g:0.01 c:5
		PHY	100.00	85.71	100.00	94.74	0.89	0.97	g:0.0001 c:100
		Hybrid-I	77.78	100.00	83.33	89.47	0.81	0.98	g:0.01 c:100
		Hybrid-II	86.67	92.86	91.67	92.11	0.83	0.95	g:0.0001 c:10
Multiple	V ^(13p+25n)	AAC	73.33	84.62	84.00	84.21	0.67	0.97	g:0.001 c:200
		DPC	85.71	92.31	92.00	92.11	0.83	0.98	g:0.01 c:10
		PHY	68.42	100.00	76.00	84.21	0.72	0.94	g:0.0001 c:50
		Hybrid-I	86.67	100.00	92.00	94.74	0.89	0.98	g:0.005 c:50
		Hybrid-II	85.71	92.31	92.00	92.11	0.83	0.94	g:1e-05 c:200
AVERAGE	V ^(40p+188n)	Hybrid-I	79.17	97.44	94.71	95.18	0.85		
		Hybrid-II	90.00	90.00	97.87	96.49	0.88		

Prec, Precision; Sen, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Matthew's Correlation Coefficient; AUC, Area Under Curve; AAC, Amino Acid Composition; DPC, Dipeptide Composition; PHY, Physicochemical properties; Hybrid-I, AAC+DPC; Hybrid-II, AAC+DPC+PHY;

Table S10. Performance of SVM based first-tier (single, double and multiple viral SCLs) predictive models using *one-versus-one* classification approach at 90% sequence identity on independent data set

Subcellular Location	Dataset	Properties	Prec	Sen	Spec	Acc	MCC	Parameters
Single-Location	V(190p+76n)	Hybrid-II	98.96	100.00	97.44	99.25	0.98	g:0.0001 c:1
Double-Location	V(40p+226n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:1
Multiple-Location	V(38p+228n)	Hybrid-II	100.00	94.74	100.00	99.25	0.97	g:0.0001 c:1
AVERAGE	V(268p+530n)	Hybrid-II	99.25	99.25	99.63	99.50	0.99	g:0.0001 c:1

Prec, Precision; Sen, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Matthew's Correlation Coefficient; Hybrid-II, AAC+DPC+PHY;

Table S11. Performance of SVM based first-tier (single, double and multiple viral SCLs) predictive models using *one-versus-one* classification approach at 30% sequence identity on independent data set

Subcellular Location	Dataset	Properties	Prec	Sen	Spec	Acc	MCC	Parameters
Single-Location	V(101p+32n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:1
Double-Location	V(17p+116n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:1
Multiple-Location	V(16p+117n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:1
AVERAGE	V(134p+265n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:1

Prec, Precision; Sen, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Matthew's Correlation Coefficient; Hybrid-II, AAC+DPC+PHY;

Table S12. Performance of SVM based second-tier (single SCLs -8) predictive models using *one-versus-one* classification approach at 90% sequence identity on independent data set

Subcellular Location	Dataset	Properties	Prec	Sen	Spec	Acc	MCC	Parameters
Cytoplasm	V(28p+162n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:1
Nucleus	V(42p+148n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:1
Endoplasmic Reticulum	V(3p+187n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:1
Single-pass Membrane	V(75p+115n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:1
Multi-pass Membrane	V(27p+163n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:1
Secreted	V(4p+186n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:1
Others-single	V(14p+176n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:1
Capsid	V(83p+189n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:1
AVERAGE	V(276p+1326n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:1

Prec, Precision; Sen, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Matthew's Correlation Coefficient; Hybrid-II, AAC+DPC+PHY;

Table S13. Performance of SVM based second-tier (single SCLs -8) predictive models using *one-versus-one* classification approach at 30% sequence identity on independent data set

Subcellular Location	Dataset	Properties	Prec	Sen	Spec	Acc	MCC	Parameters
Cytoplasm	V ^(11p+90n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:0.05
Nucleus	V ^(14p+87n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:0.05
Endoplasmic Reticulum	V ^(1p+100n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:0.05
Single-pass Membrane	V ^(45p+56n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:0.05
Multi-pass Membrane	V ^(22p+79n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:0.05
Secreted	V ^(3p+98n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:0.05
Others-single	V ^(8p+93n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:0.05
Capsid	V ^(36p+101n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:0.05
<i>AVERAGE</i>	V ^(140p+704n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0005 c:0.05

Prec, Precision; Sen, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Matthew's Correlation Coefficient; Hybrid-II, AAC+DPC+PHY;

Table S14. Performance of SVM based second-tier (double SCLs -4) predictive models using *one-versus-one* classification approach at 90% sequence identity on independent data set

Subcellular Location	Dataset	Properties	Prec	Sen	Spec	Acc	MCC	Parameters
Host cytoplasm. Host cell membrane; Peripheral membrane protein; Cytoplasmic side.	V ^(7p+33n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:1
Host nucleus. Host cytoplasm.	V ^(15p+25n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:1
Host nucleus. Extracell	V ^(2p+38n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:1
Others-double	V ^(17p+23n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:1
<i>AVERAGE</i>	V ^(41p+119n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:1

Prec, Precision; Sen, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Matthew's Correlation Coefficient; Hybrid-II, AAC+DPC+PHY;

Table S15. Performance of SVM based second-tier (multiple SCLs -6) predictive models using *one-versus-one* classification approach at 90% sequence identity on independent data set

Subcellular Location	Dataset	Properties	Prec	Sen	Spec	Acc	MCC	Parameters
Host cell membrane; Lipid-anchor; Cytoplasmic side. Host cytoplasm › host perinuclear region. Secreted.	V(3p+35n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:10
Host cytoplasm. Host nucleus. Host mitochondrion.	V(2p+36n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:10
Host nucleus › host nucleolus. Host cytoplasm. Secreted.	V(4p+34n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:10
Virion tegument. Host cytoplasm. Host nucleus.	V(4p+34n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:10
Others-triple	V(14p+24n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:10
Multiple	V(13p+25n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:10
<i>AVERAGE</i>	V(40p+188n)	Hybrid-II	100.00	100.00	100.00	100.00	1.00	g:0.0001 c:10

Prec, Precision; Sen, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Matthew's Correlation Coefficient; Hybrid-II, AAC+DPC+PHY;

Table S16. Performance of SVM based first-tier and second-tier (single, double and multiple viral SCLs) predictive models using *one-versus-one* classification approach at 90% and 30% sequence identity using **feature selection** on independent data set

Class	Prec	Sen	Spec	Acc	MCC	AUC	Parameters
90%							
First-Tier Algorithm	90.30	90.30	95.15	95.53	0.85	0.76	g:0.005 c:0.01
Second-tier (Single Locations)	100.00	100.00	100.00	100.00	1.00	1.00	g:0.0005 c:50
Second-tier (Double Locations)	100.00	100.00	100.00	100.00	1.00	1.00	g:0.0001 c:0.01
Second-tier (Multiple Locations)	100.00	100.00	100.00	100.00	1.00	1.00	g:0.0005 c:0.01
30%							
First-Tier Algorithm	100.00	100.00	100.00	100.00	1.00	1.00	g:0.0005 c:0.01
Second-tier (Single Locations)	100.00	100.00	100.00	100.00	1.00	1.00	g:0.0005 c:0.05

Prec, Precision; Sen, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Matthew's Correlation Coefficient; AUC, Area Under Curve

Table S17. Comparison of MSLVP and other servers on independent dataset (90% sequence identity) using *one-versus-other* approach. Column 2 has number of viral protein entries for subcellular locations for independent validation and columns 3-6 have correctly predicted entries by respective method.

Subcellular Location	Entries	mGOASVM	mPLR-Loc	HybridGO-Loc	MSLVP
SINGLE					
Cytoplasm	28	3	28	28	28
Nucleus	42	3	40	41	38
Endoplasmic Reticulum	3	0	1	3	2
Single-pass Membrane	75	2	14	16	74
Multi-pass Membrane	27	0	3	3	27
Secreted	4	2	4	4	3
Others-single	14		0	0	6
Capsid	83	79	23	20	28
Total (SINGLE)	276	89	113	115	214
DOUBLE					
Host cytoplasm. Host cell membrane; Peripheral membrane protein; Cytoplasmic side.	7	0	0	0	5
Host nucleus. Host cytoplasm.	15	3	3	3	10
Host nucleus. Extracell	2	0	0	0	1
Others-double	17	0	2	0	16
Total (DOUBLE)	41	3	5	3	32
MULTIPLE					
Host cell membrane; Lipid-anchor; Cytoplasmic side. Host cytoplasm › host perinuclear region. Secreted.	3	0	0	0	2
Host cytoplasm. Host nucleus. Host mitochondrion.	2	0	0	0	2
Host nucleus › host nucleolus. Host cytoplasm. Secreted.	4	0	1	0	3
Virion tegument. Host cytoplasm. Host nucleus.	4	0	0	0	2
Others-triple	14	0	1	1	12
Multiple	13	0	0	0	10
Total (MULTIPLE)	40	0	2	1	31