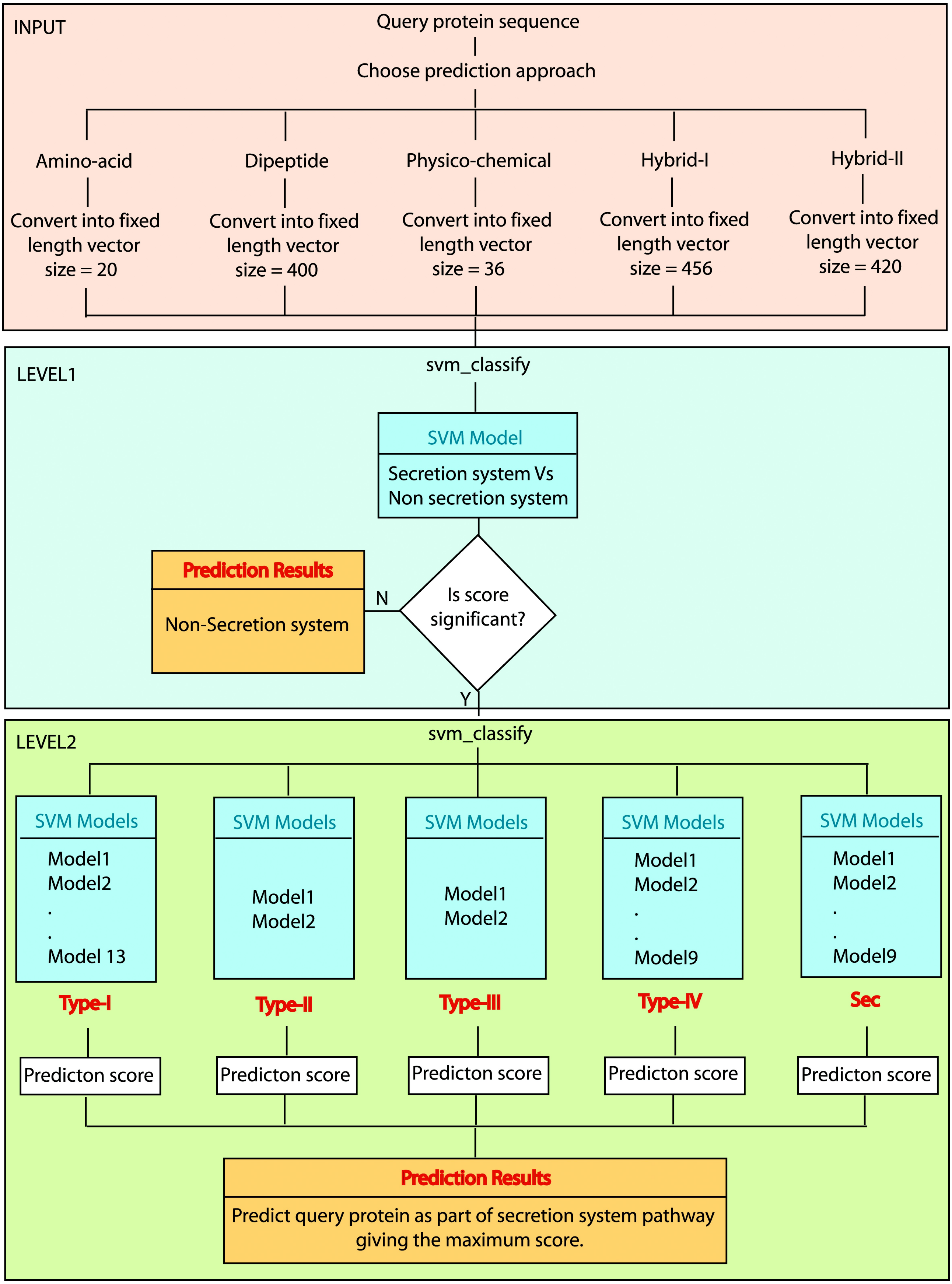
Supplementary data for SSPred: a prediction server based on SVM for the identification and classification of proteins involved in bacterial secretion systems

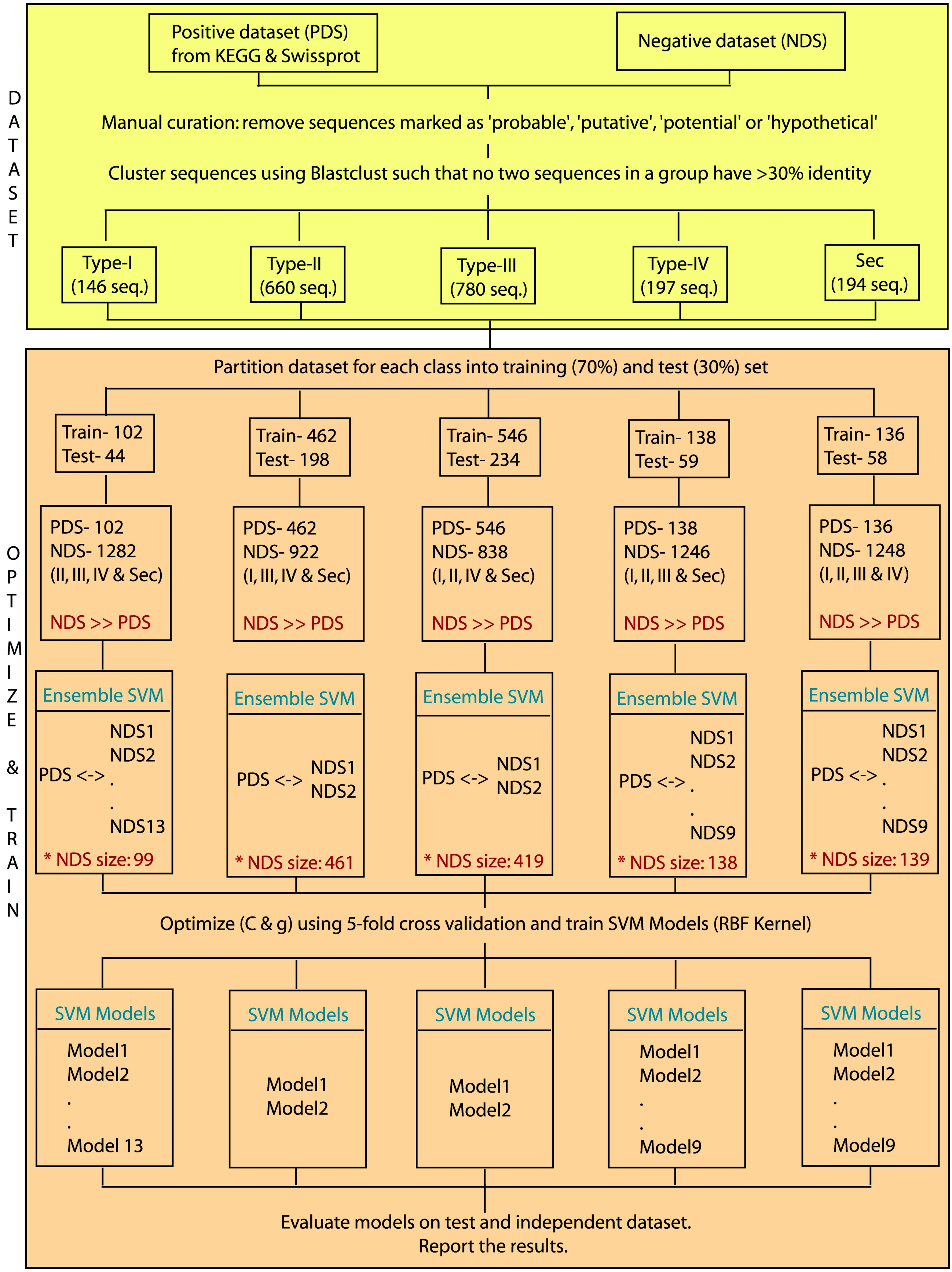
Sachin Pundhir and Anil Kumar

Supplementary Table 1: Dataset used for training and performance evaluation of SVM models in SSPred.

	Dataset		
Class	**Training (70%)**	**Test (30%)**	**Independent**
Secretion system	1384	593	112
Non-secretion system	1352	580	88
Type-I	102	44	4
Type-II	462	198	25
Type-III	546	234	30
Type-IV	138	59	33
Sec	136	58	20



Supplementary Figure 1: Flow diagram showing various steps involved in prediction analysis with SSPred. Also shown is the sequential arrangement of trained SVM models in two levels (Level1 and Level2).


DATASET
Positive dataset (PDS) from KEGG & Swissprot
Negative dataset (NDS)
Manual curation: remove sequences marked as 'probable', 'putative', 'potential' or 'hypothetical'
Cluster sequences using Blastclust such that no two sequences in a group have >30% identity
Type-I (146 seq.)
Type-II (660 seq.)
Type-III (780 seq.)
Type-IV (197 seq.)
Sec (194 seq.)
OPTIMIZE & TRAIN
Partition dataset for each class into training (70%) and test (30%) set
Train- 102 Test- 44
Train- 462 Test- 198
Train- 546 Test- 234
Train- 138 Test- 59
Train- 136 Test- 58
PDS- 102 NDS- 1282 (II, III, IV & Sec) NDS >> PDS
PDS- 462 NDS- 922 (I, III, IV & Sec) NDS >> PDS
PDS- 546 NDS- 838 (I, II, IV & Sec) NDS >> PDS
PDS- 138 NDS- 1246 (I, II, III & Sec) NDS >> PDS
PDS- 136 NDS- 1248 (I, II, III & IV) NDS >> PDS
Ensemble SVM PDS <-> NDS1 NDS2 . . NDS13 * NDS size: 99
Ensemble SVM PDS <-> NDS1 NDS2 * NDS size: 461
Ensemble SVM PDS <-> NDS1 NDS2 * NDS size: 419
Ensemble SVM PDS <-> NDS1 NDS2 . . NDS9 * NDS size: 138
Ensemble SVM PDS <-> NDS1 NDS2 . . NDS9 * NDS size: 139
Optimize (C & g) using 5-fold cross validation and train SVM Models (RBF Kernel)
SVM Models Model1 Model2 . . Model 13
SVM Models Model1 Model2
SVM Models Model1 Model2
SVM Models Model1 Model2 . . Model9
SVM Models Model1 Model2 . . Model9
Evaluate models on test and independent dataset. Report the results.

Supplementary Figure 2: Flow diagram showing various steps involved during data set creation, optimization and training of SVM modules. Also shown are the number of SVM modules trained for each secretion system class designated as '*Ensemble of SVM classifiers*'.

Supplementary Table 2: Prediction performance of various SVM models developed using different features of protein sequence. The performance was evaluated on Training dataset using 5-fold cross validation.

Class		Approach Used [a)]				
		Amino-acid (A)	Dipeptide (B)	Physico-chem (C)	Hybrid-I (A+B+C)	Hybrid-II (A + PSSM)
Sec. Sys.	*Sen*	87.93	86.26	85.18	87.49	**90.96**
	Spe	85.06	84.39	84.17	87.28	**88.46**
	ACC	86.51	85.34	84.68	87.39	**89.73**
	MCC	0.73	0.71	0.69	0.75	**0.80**
Type-I	*Sen*	92.61	95.91	91.57	93.52	**97.52**
	Spe	92.06	94.37	91.37	93.51	**97.10**
	ACC	92.35	95.15	91.47	93.51	**97.32**
	MCC	0.85	0.90	0.83	0.87	**0.95**
Type-II	*Sen*	76.63	82.68	73.48	79.44	**85.72**
	Spe	75.11	72.08	71.74	79.96	**84.69**
	ACC	75.88	77.41	72.62	79.70	**85.21**
	MCC	0.52	0.55	0.46	0.60	**0.71**
Type-III	*Sen*	84.90	87.37	86.62	89.84	**92.03**
	Spe	84.99	88.13	80.75	87.52	**92.72**
	ACC	84.94	87.70	84.10	88.84	**92.33**
	MCC	0.70	0.76	0.68	0.78	**0.85**
Type-IV	*Sen*	84.67	83.17	83.63	85.63	**88.36**
	Spe	79.55	78.55	79.38	81.08	**83.39**
	ACC	82.10	80.85	81.49	83.35	**85.86**
	MCC	0.65	0.62	0.63	0.67	**0.72**
Sec	*Sen*	81.07	84.61	81.32	83.13	**87.00**
	Spe	83.29	84.59	81.04	84.73	**88.83**
	ACC	82.19	84.58	81.18	83.93	**87.92**
	MCC	0. 65	0.70	0.63	0.68	**0.76**
Average	*Sen*	84.64	86.67	83.63	86.51	**90.27**
	Spe	83.34	83.68	81.41	85.68	**89.20**
	ACC	84.00	85.17	82.59	86.12	**89.73**
	MCC	0.68	0.71	0.65	0.73	**0.80**

a) Amino-acid, amino acid composition is used as input; Dipeptide, dipeptide composition is used as input; Physico-chem, 36 physico-chemical properties of amino-acids is used as input; PSSM + A, a 420 dimension feature vector is used as training vector (400 dimensions generated from PSI-BLAST profiles and 20 dimensions from amino acid composition).

Supplementary Table 3: Prediction performance of various SVM models developed using different features of protein sequence. The performance was evaluated on Test dataset using validation test.

Class		**Approach Used** [a]				
		Amino-acid (A)	Dipeptide (B)	Physico-chem (C)	Hybrid-I (A+B+C)	Hybrid-II (A +PSSM)
Secretion System	*Sen*	83.16	84.01	83.16	86.03	**90.91**
	Spe	85.34	86.72	84.14	87.76	**90.52**
	ACC	84.24	85.35	83.65	86.88	**90.72**
	MCC	0.69	0.71	0.67	0.74	**0.81**
Type-I	*Sen*	90.22	95.80	90.21	92.31	**95.28**
	Spe	91.04	92.46	87.96	91.38	**96.40**
	ACC	90.62	94.15	89.10	91.85	**95.78**
	MCC	0.82	0.88	0.78	0.84	**0.92**
Type-II	*Sen*	71.97	81.06	70.71	77.02	**87.88**
	Spe	76.28	73.54	69.98	79.65	**83.59**
	ACC	74.12	77.31	70.34	78.33	**85.73**
	MCC	0.49	0.55	0.41	0.57	**0.72**
Type-III	*Sen*	88.04	86.54	85.26	89.74	**94.23**
	Spe	85.92	88.74	80.28	87.62	**93.06**
	ACC	87.12	87.49	83.11	88.83	**93.72**
	MCC	0.74	0.75	0.66	0.78	**0.88**
Type-IV	*Sen*	82.22	83.15	81.11	83.52	**84.08**
	Spe	82.04	73.15	74.26	77.22	**79.81**
	ACC	82.13	78.15	77.69	80.37	**81.65**
	MCC	0.64	0.57	0.56	0.61	**0.64**
Sec	*Sen*	79.28	85.31	78.53	83.24	**85.50**
	Spe	80.19	82.78	78.33	82.78	**87.78**
	ACC	79.74	84.03	78.43	83.00	**86.65**
	MCC	0.60	0.68	0.57	0.66	**0.73**
Average	*Sen*	82.48	86.00	81.50	85.31	**89.65**
	Spe	83.49	83.00	79.16	84.40	**88.53**
	ACC	83.00	84.41	80.38	84.88	**89.04**
	MCC	0.66	0.69	0.61	0.70	**0.78**

[a] Amino-acid, amino acid composition is used as input; Dipeptide, dipeptide composition is used as input; Physico-chem, 36 physico-chemical properties of amino-acids is used as input; PSSM + A, a 420 dimension feature vector is used as training vector (400 dimensions generated from PSI-BLAST profiles and 20 dimensions generated from amino acid composition).