Appendix I: LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST

Dan Xie[1], Ao Li[1], Minghui Wang[1], Zhewen Fan[2], Huanqing Feng[1*],

[1]*Department of Electronic Science and Technology, University of Science and Technology of China, Hefei, P R, China*

[2]*Department of Biomedical Engineering, City University of New York, New York, U.S.A*

Confusion matrices for all prediction analyses in original paper

Each row in the confusion matrix is the distribution of the predicted location given the corresponding true subcellular location, and each column is the distribution of the true location given the corresponding predicted subcellular location. “Total” denotes the total number of proteins belonging to one of the four subcellular locations in each row or column of the confusion matrix.

Table S1. Confusion matrix of LOCSVMPSI on the RH-2427 data set*

	Cytoplasmic	Extra-cellular	Mitochondrial	Nuclear	Total
Cytoplasmic	592	6	22	64	684
Extra-cellular	6	301	4	14	325
Mitochondrial	25	4	258	34	321
Nuclear	40	4	16	1037	1097
Total	663	315	300	1149	2427

*Results in this table were obtained by using a jackknife test on the RH-2427 data set

Table S2. Confusion matrix of LOCSVMPSI on the RK-7579 data set*

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	Total
C1	513	52	0	0	3	0	0	48	47	1	6	1	671
C2	30	948	1	2	17	1	2	55	162	3	18	2	1241
C3	0	3	24	0	0	0	0	0	12	1	0	0	40
C4	1	16	0	70	10	0	2	4	6	0	5	0	114
C5	5	13	0	2	772	0	9	9	35	2	10	4	861
C6	1	11	0	0	2	22	0	1	10	0	0	0	47
C7	0	2	0	0	22	0	58	2	2	2	4	1	93
C8	49	69	0	0	13	0	0	496	84	5	10	1	727
C9	11	94	0	1	20	3	0	17	1767	0	19	0	1932
C10	4	29	0	0	2	0	0	21	5	52	12	0	125
C11	7	17	0	1	14	0	2	12	33	1	1586	1	1674
C12	1	5	0	0	13	0	1	3	8	0	1	22	54
Total	622	1259	25	76	888	26	74	668	2171	67	1671	32	7579

*Results in this table were obtained by using 5-fold cross validation test on the RK-759 data set. Abbreviations used in this table are: C1: chloroplast proteins; C2: cytoplasmic proteins; C3: cytoskeleton proteins; C4: ER proteins; C5: extracellular proteins; C6: golgi apparatus proteins; C7: lysosomal proteins; C8: mitochondrial proteins; C9: nuclear proteins; C10: peroxisomal proteins; C11: plasma membrane proteins; C12: vacuolar proteins. More details about this data set can be found at http://web.kuicr.kyoto-u.ac.jp/~park/Seqdata/

Table S3. Confusion matrix of LOCSVMPSI on the SWISS-PROT new-unique data set with the model trained with the RH-2427 data set*

	Cytoplasmic	Extra-cellular	Mitochondrial	Nuclear	Total
Cytoplasmic	101	0	8	37	146
Extra-cellular	7	82	4	35	128
Mitochondrial	5	0	51	4	60
Nuclear	28	3	6	141	178
Total	141	85	69	217	512

* Results in this table were obtained by using the whole RH-2427 data set as the training set and evaluating the prediction performance of LOCSVMPSI on the SWISS-PROT new-unique data set.

Table S4. Confusion matrix of LOCSVMPSI on the SWISS-PROT new-unique data set with the model trained with the PK-7579 data set*

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	Total
C1	0	0	0	0	0	0	0	0	0	0	0	0	0
C2	4	108	0	0	1	0	0	1	30	1	1	0	146
C3	0	0	0	0	0	0	0	0	0	0	0	0	0
C4	0	0	0	0	0	0	0	0	0	0	0	0	0
C5	0	5	0	0	98	0	2	2	19	0	1	1	128
C6	0	0	0	0	0	0	0	0	0	0	0	0	0
C7	0	0	0	0	0	0	0	0	0	0	0	0	0
C8	1	3	0	1	0	0	0	53	2	0	0	0	60
C9	0	11	0	0	4	1	0	7	150	0	5	0	178
C10	0	0	0	0	0	0	0	0	0	0	0	0	0
C11	0	0	0	0	0	0	0	0	0	0	0	0	0
C12	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	5	127	0	1	103	1	2	63	201	1	7	1	512

* Results in this table were obtained by using the whole PK-7579 data set as the training set and evaluating the prediction performance of LOCSVMPSI on the SWISS-PROT new-unique data set. Abbreviations used in this table are: C1: chloroplast proteins; C2: cytoplasmic proteins; C3: cytoskeleton proteins; C4: ER proteins; C5: extracellular proteins; C6: golgi apparatus proteins; C7: lysosomal proteins; C8: mitochondrial proteins; C9: nuclear proteins; C10: peroxisomal proteins; C11: plasma membrane proteins; C12: vacuolar proteins. More details about this data set can be found at http://web.kuicr.kyoto-u.ac.jp/~park/Seqdata/