Table S3. Performance of MetaP[1] in predicting non-membrane proteins[2]

Database	Cytoplasmic			Periplasmic			Extracellular		
	Precision[3]	Recall[4]	Accuracy[5]	Precision	Recall	Accuracy	Precision	Recall	Accuracy
db2	0.82	0.97	0.91	0.85	0.77	0.87	0.92	0.79	0.92
db3	0.82	0.98	0.92	0.84	0.77	0.87	0.92	0.79	0.92

[1] This version of MetaP comprises from the following element algorithms: CELLO, SUBLOC, and LOCTREE. This is the earlier version of MetaP used in a previous study (Luo, H., R. Benner, R. A. Long, and J. Hu. 2009. Subcellular localization of marine bacterial alkaline phosphatases. Proc. Natl. Acad. Sci. U.S.A. 106:21219–21223).

[2] The first 200 amino acids at the N-terminal were removed in the testing sequences, and the sequences with no less than 30 amino acids were used in the analysis.

[3] Precison = TP / (TP+FP).

[4] Recall = TP / (TP+FN).

[5] Accuracy = (TP+TN) / (TP+TN+FP+FN)

Here, only extracellular and periplasmic proteins from database 2 (db2) (19) and database 3 (db3) (15) were used in order to balance the number of true positive and true negative cases. The database 1 (db1) (18) was not used because many secretory sequences in db1 are not labeled with finer subcellular localizations (e.g. extracellular or periplasmic).

TP is true positive; TN is true negative; FP is false positive; FN is false negative. For instance, in the case of extracellular proteins, TP is the number of proteins predicted to be extracellular which are indeed extracellular; TN is the number of proteins predicted to be non-extracellular which are indeed non-extracellular; FP is the number of proteins predicted to be extracellular which are indeed non-extracellular; FN is the number of proteins predicted to be non-extracellular which are indeed extracellular.

Table S4. Performance of new MetaP[1] in predicting non-membrane proteins[2]

Database	Cytoplasmic			Periplasmic			Extracellular		
	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy
db2	0.90	0.93	0.94	0.87	0.88	0.91	0.98	0.77	0.93
db3	0.89	0.92	0.93	0.87	0.88	0.92	0.98	0.81	0.94

[1] MetaP 2.0 comprises from the following element algorithms: CELLO, SUBLOC, and PSLDOC. MetaP 2.0 is used in the present study.

[2] The first 200 amino acids at the N-terminal were removed in the testing sequences, and the sequences with no less than 30 amino acids were used in the analysis.

Table S5. Performance of PSLDOC in predicting inner- and outer-membrane proteins[1]

Database	InnerMembrane			OuterMembrane		
	Precision	Recall	Accuracy	Precision	Recall	Accuracy
db2	0.98	0.84	0.96	0.92	0.97	0.97
db3	0.99	0.84	0.96	0.92	0.96	0.96

[1] The first 200 amino acids at the N-terminal were removed in the testing sequences, and the sequences with no less than 30 amino acids were used in the analysis.

Table S6. Performance of CELLO in predicting inner- and outer-membrane proteins[1]

Database	InnerMembrane			OuterMembrane		
	Precision	Recall	Accuracy	Precision	Recall	Accuracy
db2	0.99	0.74	0.94	0.97	0.86	0.95
db3	1	0.76	0.94	0.97	0.90	0.96

[1] The first 200 amino acids at the N-terminal were removed in the testing sequences, and the sequences with no less than 30 amino acids were used in the analysis.

Table S10. Pearson's correlation coefficient between subcellular localization and axis coordinates in two-dimensional space

		Dimension 1	Dimension 2
Bacteroidetes	Cytoplasm	0.28	0.24
	Inner Membrane	-0.05	-0.81
	Periplasm	0.63	-0.04
	Outer Membrane	-0.9	0.01
	Extracellular	-0.62	0.58
OMG	Cytoplasm	0.64	-0.56
	Inner Membrane	-0.87	-0.4
	Periplasm	0.56	0.68
	Outer Membrane	-0.86	0.36
	Extracellular	-0.04	0.73
Roseobacter	Cytoplasm	-0.83	0.16
	Inner Membrane	-0.92	-0.13
	Periplasm	0.98	-0.03
	Outer Membrane	-0.89	0.1
	Extracellular	0.19	0.97
SAR11	Cytoplasm	-0.84	0.14
	Inner Membrane	0.88	-0.19
	Periplasm	0.31	-0.03
	Outer Membrane	-0.55	0.01
	Extracellular	0.61	0.6
Synechococcus	Cytoplasm	-0.88	-0.2
	Inner Membrane	0.97	0.01
	Periplasm	-0.85	0.1
	Outer Membrane	-0.72	-0.28
	Extracellular	-0.25	0.85