Supplementary File 1

PERFORMANCE MEASUREMENTS

We let *C* represent all of the possible subcellular localization classes and *D* represent the data being predicted. We maintain the following counts for each subcellular localization class $c_j \in \mathcal{C}$:

- True positives (TP_j) – the number of sequences correctly predicted to localize into c_j.
- True negatives (TN_j) – the number of sequences correctly predicted to not localize into c_j.
- False positives (FP_j) – the number of sequences predicted to localize into c_j, but actually localize elsewhere.
- False negatives (FN_j) – the number of sequences that actually localize into c_j, but were predicted to localize elsewhere.

Using these four quantitative measures, we report a number of standard measurements in judging classifier performance:

- *Overall Accuracy* – a measure of the overall classifier performance. It is defined as the fraction of the data tested that is classified correctly. Though it is a poor measure to consider on highly unbalanced datasets, we still report it as a general overall comparative measure:

$$Accuracy = \frac{\sum_j TP_j}{|\mathcal{D}|}$$

- *Sensitivity* (a.k.a. *Recall*, *TP-rate*) – the fraction of data in class c_j that was correctly predicted. This gives a measure of individual class accuracy:

$$Sens_j = \frac{TP_j}{TP_j + FN_j} = \frac{TP_j}{|\mathcal{D}_{c_j}|}$$

- *Precision* – the fraction of data predicted to be in class c_j that was actually correct:

$$Prec_j = \frac{TP_j}{TP_j + FP_j}$$

- *Specificity* – the fraction of data not in class c_j that was correctly predicted:

$$Spec_j = \frac{TN_j}{TN_j + FP_j} = \frac{TN_j}{\sum_{k \neq j} |\mathcal{D}_{c_k}|}$$

- *False Positive Rate* – the fraction of data not in class c_j that was incorrectly predicted to be in class c_j:

$$FPR_j = \frac{FP_j}{TN_j + FP_j} = \frac{FP_j}{\sum_{k \neq j} |\mathcal{D}_{c_k}|}$$

- *Matthews Correlation Coefficient* – a measure of overall predictive performance for class c_j. It has a range of (-1,+1), where 1 implies a perfect prediction, 0 implies random, and a value of less than 0 implies that the prediction is worse than random guessing:

$$MCC_j = \frac{TP_j TN_j - FP_j FN_j}{\sqrt{(TP_j + FP_j)(TP_j + FN_j)(TN_j + FP_j)(TN_j + FN_j)}}$$

- F_{1j} – an individual measure for class c_j that combines sensitivity (recall) and precision measured for that class:

$$F_{1_j} = \frac{2 \times Sens_j \times Prec_j}{Sens_j + Prec_j}$$

- *Mac-*F_1 – an average of all of the individual F_{1j} measures observed over all classes:

$$Mac\text{-}F_1 = \frac{\sum_{j=1}^{|\mathcal{C}|} F_{1j}}{|\mathcal{C}|}$$

Table S1. Coverage *vs* Precision for Gram-Negative Data

n-Gram	Coverage	Micro-Averaged Precision	Macro-Averaged Precision
4-gram	50%	99.7%	99.5%
	70%	99.2%	98.6%
	95%	92.2%	89.9%
	100%	90.3%	85.9%
5-gram	50%	100.0%	100.0%
	70%	99.3%	99.0%
	95%	91.8%	94.9%
	100%	89.7%	92.2%
6-gram	50%	100.0%	100.0%
	70%	99.7%	99.5%
	95%	91.8%	95.5%
	100%	89.8%	93.0%

This table shows how limiting the coverage of predictions generated by ngLOC through selecting various CS thresholds dramatically improves the precision. The coverage of predictions generated is increased by decreasing the CS threshold setting, thereby allowing lower confidence predictions to be generated. Both micro-averaged and macro-averaged precisions are shown in order to accurately convey performance across all classes. For 100% coverage, micro-averaged precision is equal to overall accuracy for the classifier.

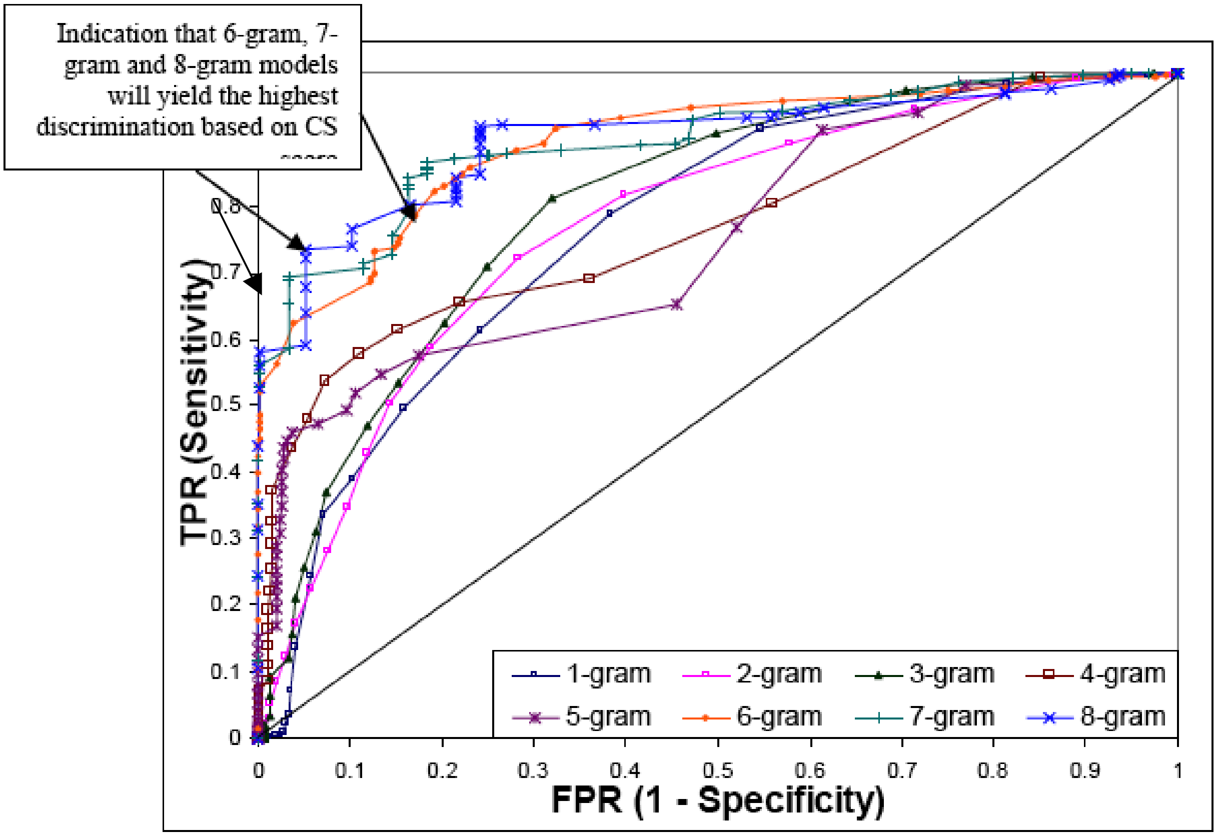

Fig. (S1). ROC curves for [1-8]-gram models on gram-positive data.

This figure depicts the ROC curves plotted for gram-positive data, where the sensitivity and specificity are based on macro-averaged calculations observed for distinct CS thresholds across each class. We point out that the 6-gram, 7-gram and 8-gram are likely to have the best discriminatory ability based on CS, and according to Table **1** in the main paper, they all have reasonable AUC calculations. However, the overall accuracy and Mac-F1 scores of the 7-gram and 8-gram model are significantly lower than the 6-gram, suggesting that the 6-gram is the best overall candidate model to use.

Table S2. Performance of ngLOC on *GP-Test* Dataset

	Sequences	*Gneg-PLoc**	*ngLOC*	*ngLOC(UO)*
Cytoplasm	210(62)	177/210 = 84.3%	205/210 = 97.6%	61/62 = 98.4%
Extracellular	20(6)	13/20 = 65.0%	20/20 = 100%	6/6 = 100%
Inner Membrane	345(49)	325/345 = 94.2%	345/345 = 100%	49/49 = 100%
Outer Membrane	13(1)	10/13 = 76.9%	13/13 = 100%	1/1 = 100%
Periplasmic	49(7)	43/49=87.8%	48/49 = 98%	6/7 = 85.7%
Overall Accuracy		89.1%	99.1%	123/125 = 98.4%

This table reports the individual sensitivity measures (TP / (TP + FN)) for each class resulting from predictions generated using the independent test dataset assembled by Chou and Shen [10], denoted as *GP-test*. The *Sequences* column indicates the number of sequences assigned to each localization class. The number in parenthesis represents the number of sequences in the test data that are not in the training data. The *ngLOC* column reports the results of the ngLOC method on the same test dataset. The *ngLOC*(*UO*) column reports the results of the ngLOC method on the sequences in GP-test that are not in the training data. *(The Gneg-PLoc results are taken from the results reported by Chou and Shen [10]).