Supplementary Material

1. Datasets

1.1 BaCelLo and BaCelLo IDS datasets

BaCelLo dataset (Pierlenoi, et al., 2006) and the BaCelLo independent dataset (IDS) (Casadio, et al., 2008) were used for the training and test sets, respectively. The BaCelLo dataset contains 2,597 animal proteins, 1,198 fungal proteins, and 491 plant proteins. This homology-reduced training dataset was extracted from Swiss-Prot release 48. By ignoring proteins annotated as 'membrane' or 'transmembrane', only globular proteins were considered. The animal and fungal proteins represent four localizations (nucleus, cytoplasm, mitochondrion, secretory pathway) and the plant proteins five localizations (with the addition of chloroplast).

The BaCelLo IDS was extracted from Swiss-Prot release 54. Only proteins newly added to the database from the release 49 and higher versions were collected. Furthermore, proteins that share 30% or lower identity (BLAST E-value>1E-3) with those from the release 48 (BaCelLo dataset) were extracted. This dataset contains 575 animal, 437 fungal, and 400 plant proteins. In order to avoid a bias towards the over-represented protein localizations, all sequences sharing the same localization and an alignment with an E-value lower than 1E-3 were clustered into 432 animal groups, 418 fungal groups, and 132 plant groups.

1.2 Höglund and Höglund IDS datasets

Höglund dataset (Höglund, et al., 2006) and the Höglund IDS (Blum et al., 2009) were used for the training and test sets, respectively. The Höglund dataset was extracted from Swiss-Prot release 42 and contains 5,959 eukaryotic proteins. Plant proteins have the following ten localizations: the chloroplast, cytoplasm, endoplasmic reticulum, extracellular space, Golgi apparatus, mitochondrion, nucleus, peroxisome, plasma membrane, and vacuole. Fungal proteins share the same subcellular localizations as plant proteins except for the chloroplast. Finally, animal proteins share all localizations with fungal cell with lysosomes in place of vacuoles.

The Höglund IDS was extracted from Swiss-Prot release 55.3 and clustered in the same way as the BaCelLo IDS, resulting in 158 animal groups, 106 fungal groups, and 30 plant groups. However, we used the 40% sequence identity threshold for plant proteins to increase the data size, because the number of plant proteins is relatively small. The Höglund IDS covers six localizations for animal proteins (extracellular region, plasma membrane, peroxisome, endoplasmic reticulum, Golgi apparatus, and lysosome), six localizations for fungal proteins (the same as those of animal proteins with vacuole replaced by lysosome) and seven localizations for plant proteins (chloroplast, in addition to the fungal localizations).

1.3 Human dataset

Human dataset (Shen and Chou, 2009) was collected from Swiss-Prot release 50.7. This dataset contains 3,106 human proteins covering 14 subcellular localizations (centriole, cytoplasm, cytoskeleton, endoplasmic reticulum, endosome, extracell, Golgi apparatus, lysosome, microsome, mitochondrion, nucleus, peroxisome, plasma membrane, and synapse), where 2,580 proteins belong to one subcellular location, 480 to two locations, 43 to three locations, and 3 to four locations. This dataset is quite stringent that none of the protein pairs has larger than 25% sequence identity.

2. Performance criteria

2.1 The Performance measures for BaCelLo IDS and Höglund IDS

The overall performance was measured by average sensitivity (AVG) and overall accuracy (ACC). Let $SE_i = T_i/n_i$ where n_i is the number of proteins at localization i and T_i is the number of correctly predicted proteins in localization i.

$$AVG = \frac{\sum_{i=1}^{C} SE_i}{C}, \quad ACC = \sum_{i=1}^{C} T_i \,/N,$$

where C is the number of subcellular localizations considered and N the total number of proteins in the test set.

The prediction performance for each subcellular location was measured by the sensitivity (*SE*), Specificity (*SP*), and the Matthews correlation coefficient (*MCC*), each of which is defined as follows:

$$SE = \frac{tp}{tp + fn}$$

$$SP = \frac{tp}{tp + fp}$$

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(\mathrm{tp} + \mathrm{fn})(\mathrm{tp} + \mathrm{fp})(\mathrm{tn} + \mathrm{fp})(\mathrm{tn} + \mathrm{fn})}}$$

where *tp*=true positive, *tn*=true negative, *fp*=false positive, and *fn*=false negative.

To evaluate the performance, cluster-based evaluations were performed (Casadio, et al., 2008): (1) similar proteins in each localization are clustered, (2) the rates of correct and incorrect predictions are computed for each cluster, (3) the rates for all of the clusters with the same localizations were averaged for each localization, and (4) using these averaged rates of true and false prediction, SE, SP, MCC, AVG, and ACC are then evaluated.

2.2 The Performance measure for Human dataset

For the Human dataset, some proteins may occur in two or more locations and the 3,106 different proteins actually correspond to 3,681 locative proteins (given a protein coexisting at two different subcellular locations, it will be counted as two 'locative proteins'). For studying proteins with multiple subcellular location sites, a quality control function and new success rate was introduced (Shen and Chou, 2007).

Let $C(\theta) = \{C_1(\theta),\ C_2(\theta), \ldots, C_{m(\theta)}(\theta)\}$ be the predicted subcellular locations for a query protein when a threshold θ is used, while the real subcellular locations of the protein are $R = \{R_1,\ R_2, \ldots, R_r\}$. WegoLoc uses θ as a multiplex parameter that allows multiple localization prediction, meaning that any predicted localization with probability higher than $[\theta \times highest\ probability\ of\ location]$ will be assigned to the query protein.

A quality control function is defined by

$$Q(\theta) = H(\theta) - |S(\theta)|$$

where a hit function $H(\theta) = \sum_{i=1}^{m(\theta)} \Delta_i\ (C_i(\theta),\ \mathrm{R})$,

$$\Delta_i(C_i(\theta), R) = \begin{cases} 1, & if\ C_i(\theta)\ \in R \\ 0, & otherwise \end{cases}$$

and $|S(\theta)|$ represents the number of elements in the set $S(\theta) = [R \cup C(\theta)] - [R \cap C(\theta)]$. The sum of $Q(\theta)$ for all query proteins is dubbed the *quality function value*, which gives the quality of the threshold θ.

Let U_k^μ be a set of predicted subcellular localizations of the k th protein in the μ th localization. The overall success rate (ACC) is then defined by

$$\frac{1}{\widetilde{N}}\sum_{\mu=1}^{14}\sum_{k=1}^{\tilde{n}_\mu}\Delta[U_k^\mu,\mu]$$

where $\tilde{n}_\mu$ is the number of locative proteins in the μ th localization, $\widetilde{N}=\widetilde{n}_1+\widetilde{n}_2+\cdots+\widetilde{n}_{14}$, and the Δ function is defined by

$$\Delta\left[U_k^\mu,\mu\right]=\begin{cases}1, if\ \mu\ \in U_k^\mu\\ 0, otherwise\end{cases}$$

3. Training and test procedure

WegoLoc was trained using LIBSVM software (Chang and Lin, 2001) with the radial basis kernel function. The parameters of LIBSVM were determined as follows.

- Gamma parameter in the kernel function was optimized by a grid search.
- Fixed value of parameter $C=1$ was used.
- In order to reduce the over-prediction effect when using unbalanced training datasets, the weight for each subcellular location i was assigned as follows:

$$wi=\frac{the\ number\ of\ proteins\ in\ training\ data}{the\ number\ of\ proteins\ belonging\ to\ subcellular\ location\ i\ in\ training\ data}$$

3.1 Training and test for BaCelLo IDS and Höglund IDS

To determine the gamma parameters in LIBSVM, ten-fold cross-validations were performed using only training datasets, BaCelLo and Höglund datasets, respectively. These gamma parameters were used for the tests on BaCelLo IDS and Höglund IDS, respectively.

3.2 Training and test for Human dataset

For a comparison, we performed a Jackknife test for WegoLoc as was done for Hum-mPLoc 2.0 (Shen and Chou, 2009) using the same Human dataset. The jackknife test for the 3,681 locative proteins requires the same number of training and test procedures, which causes heavy computational costs. Thus, we used a simplified test. We first performed a 10-fold cross-validation using the entire Human dataset to find the optimal parameters of LIBSVM. Using these parameters, LIBSVM was trained for each of 3,681 training sets and used for test.

3.3 Prediction of multiple localizations

For predicting the subcellular localizations of multiplex(locative) proteins that exist at multiple locations, we modified a functionality of LIBSVM. We use the "one-against-one" approach where $k(k-1)/2$ binary classifiers are devised for k subcellular locations. Such binary classifiers are trained using data from the corresponding pair of subcellular locations, and then each binary classifier predicts the localization of a query protein. Using these $k(k-1)/2$ results, we calculate the probability of location $\mathrm{P}(m)$ by the ratio of the number of predicted location m to the total number of predicted locations $k(k-1)/2$. Using the user-defined multiplex threshold θ, any predicted localization with probability higher than $[\theta\times highest\ probability\ of\ location]$ will be assigned to the query protein as well.

3.4 GO with protein sequence file

We collected protein sequences that are represented by GO terms before WegoLoc prediction. To map a protein sequence into GO terms, the following procedure is used.

1. UniProtKB-GOA (version 81) was downloaded from http://www.ebi.ac.uk/GOA, which contains associations between gene products and GO terms.
2. We choose gene products with amino acid sequences from Swiss-Prot release 57 (downloaded from http://www.uniprot.org/downloads) which are simultaneously present in the above UniProtKB-GOA file. We call this file 'SEQ' which is a subset of Swiss-Prot release 57.
3. For each input protein, BLAST searches for the most similar protein in 'SEQ', and the corresponding GO terms of the most similar protein is fetched from UniProtKB-GOA.

4. Accuracy of WegoLoc for each subcellular location

4.1 Results on BaCelLo IDS and Höglund IDS

In Supplementary Table 1 and 2, BLAST E-value threshold was set to 1E0, multiplex threshold was set to 1 and No. is the number of groups per location.

Supplementary Table 1. Prediction results using BaCelLo IDS

	Animals				Fungi				Plants			
Location	No.	SE	SP	MCC	No.	SE	SP	MCC	No.	SE	SP	MCC
Secretory pathway	75	100.0	100.0	100.0	9	100.0	100.0	100.0	6	100.0	100.0	100.0
Mitochondrion	48	97.9	100.0	98.8	77	98.7	98.7	98.4	6	83.3	82.8	82.3
Chloroplast	-	-	-	-	-	-	-	-	72	99.0	99.0	100.0
Nucleus	224	98.7	98.2	96.8	152	99.3	99.3	99.0	36	93.5	100.0	95.6
Cytoplasm	85	95.3	95.3	94.1	180	98.9	98.9	98.0	17	100.0	87.9	92.9
	AVG=98.0, ACC=98.2				AVG = 99.2, ACC=99.0				AVG=95.4, ACC=97.5			

Supplementary Table 2. Prediction results using Höglund IDS

	Animals				Fungi				Plants			
Location	No.	SE	SP	MCC	No.	SE	SP	MCC	No.	SE	SP	MCC
Extracellular region	78	96.2	96.2	100.0	7	100.0	100.0	100.0	1	100.0	100.0	100.0
Plasma membrane	34	100.0	91.9	94.7	29	86.2	96.2	88.0	6	100.0	85.7	90.6
Peroxisome	3	100.0	100.0	100.0	5	100.0	100.0	100.0	2	100.0	100.0	100.0
Endoplasmic reticulum	25	96.0	100.0	97.6	46	97.8	91.8	90.6	6	100.0	75.0	82.9
Golgi apparatus	14	71.4	90.9	79.0	8	87.5	100.0	93.1	6	89.3	100.0	89.4
Lysosome	4	100.0	100.0	100.0	-				-			
Vacuole	-				11	81.8	100.0	89.5	9	77.8	100.0	84.3
	AVG=93.9, ACC=94.9				AVG = 92.2, ACC=92.5				AVG=93.5, ACC=90.0			

Example 1) Specific comparisons.
Using 79 mitochondrion sequences in BaCelLo IDS fungal dataset, we compared MultiLoc2, BaCelLo, WoLF PSORT, SherLoc2 (stand-alone version without GoLoc), and WegoLoc. WegoLoc correctly predicted PSL of all the 79 sequences except for NCS6_YEAST. All the other methods also failed to correctly locate the sequence. Searching UniProtKB-GOA for the sequence, we found it had 100% identity with CTU1_YEAST whose localizations were cytoplasm and mitochondrion. WegoLoc predicted its localizations as cytoplasm (50%, best hit) and mitochondrion (33%), both of which are correct predictions actually.
There were 12 sequences that only WegoLoc correctly predicted their PSLs: YN92_SCHPO, MFB1_YEAST, TAR1_YEAST, YNT5_YEAST, OXR1_YEAST, MU167_SCHPO, YL091_YEAST, AI3_USTMA, YD185_YEAST, YO285_YEAST, YNU8_YEAST, and YHOA_SCHPO (UniProKB accession).

Example 2) Comparison for proteins with multiple localizations
Hum_mPLoc 2.0 also provides predictions for multiple localizations of input proteins. We compared WegoLoc with Hum-mPLoc 2.0 for predicting multiple localizations. Q9UHB9 is known to be localized at both cytoplasm and nucleus. Hum-mPLoc 2.0 predicted two locations endoplasmic reticulum, nucleus, while WegoLoc correctly predicted cytoplasm, nucleus, endoplasmic reticulum in order. Q53HL2 is also known to be localized at cytoplasm and nucleus. Hum-mPLoc 2.0 predicted only nucleus, while WegoLoc predicted centrosome, cytoskeleton, nucleus, each with 13.2% and cytoplasm with 11.0%.

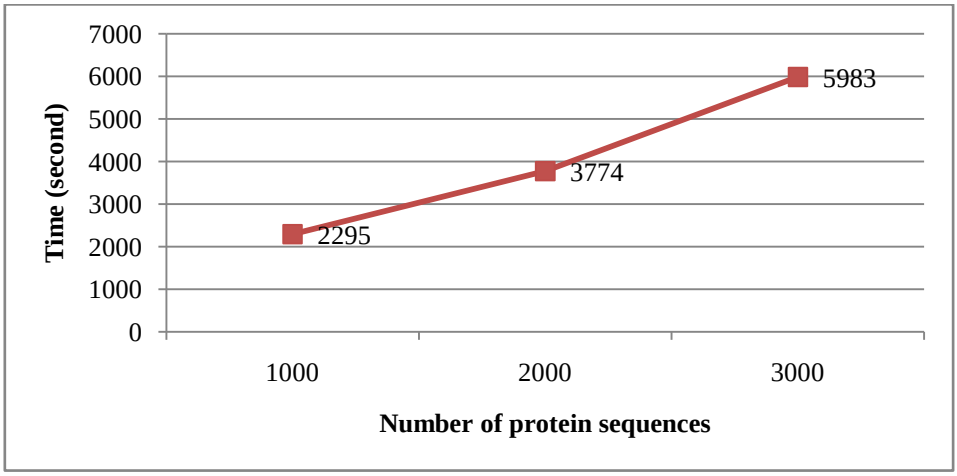
4.2 Results on Human dataset

In Supplementary Table 3, the E-value threshold is set to 1E0, the multiplex threshold is varied from 1 to 0.8, and No. is the number of locative proteins in each location. As the multiplex threshold decreases, the quality function value described in 2.2 decreases, because the false positive predictions increase.

Supplementary Table 3. Prediction results using Human dataset

Location	No.	Multiplex 1.00	Multiplex 0.95	Multiplex 0.90	Multiplex 0.85	Multiplex 0.80
Centrosome	77	92.2	92.2	98.7	98.7	98.7
Cytoplasm	817	51.7	51.8	85.2	85.2	97.3
Cytoskeleton	79	72.2	72.2	89.9	89.9	98.7
Endoplasmic reticulum	229	66.4	66.4	93.5	93.5	96.1
Endosome	24	62.5	62.5	95.8	95.8	95.8
Extracellular region	385	84.7	84.7	96.1	96.1	97.9
Golgi apparatus	161	78.9	78.9	96.9	96.9	98.8
Lysosome	77	92.2	92.2	98.7	98.7	98.7
Microsome	24	100.0	100.0	100.0	100.0	100.0
Mitochondrion	364	95.1	95.1	97.8	97.8	98.9
Nucleus	1021	71.1	71.1	89.8	89.8	94.2
Peroxisome	47	100.0	100.0	100.0	100.0	100.0
Plasma membrane	354	73.5	73.5	94.4	94.4	97.7
Synapse	22	77.3	77.3	90.9	90.9	90.9
AVG		79.8	79.8	94.8	94.8	97.4
ACC		72.3	72.3	91.8	91.8	96.8
Quality function value		1022	1022	873	872	-1637

4.3 processing time



Supplementary Figure 1. The processing time of WegoLoc for 1000, 2000, and 3000 sequences in one execution. The average processing time per sequence is two seconds.

The execution of the current WegoLoc can be further speeded up at least 10 times by using computational acceleration on graphics processing units (GPUs) and adopting accelerated versions of the BLASTP. Since WegoLoc will become fast enough to be used only via web.

5. Options for the high performance of WegoLoc

5.1 GOA versus InterPro Database

In order to fetch GO terms of a given protein sequence, GOA (Barrell, et al., 2009) and InterPro (Hunter, et al., 2009) database have been used. The rate of annotated protein sequences varies depending on the GO annotation methods. In Supplementary Table 4, we compare the annotation rates of BaCelLo IDS by a GOA-based method (Chi, 2010) and InterProScan (Hunter, et al., 2009). Since InterProScan is being updated, we showed the annotation rates which are obtained at august, 2011 as well as the rates from the previous version in the Blum's paper (Blum, et al., 2009). The results based on the two GOA versions (UniProt 70 and 81) are shown in the following table.

Supplementary Table 4. comparison of annotation rates (%) by GOA and InterProScan. (E-value threshold=0.001)

Dataset	InterProScan (Bum, et al., 2009)	InterProScan (august, 2011)	GOA (UniProtKB-GOA UniProt 70, March 10th, 2008)	GOA (UniProtKB-GOA UniProt 81, April 22th, 2010)
BaCelLo IDS animals	43	51	88	88
BaCelLo IDS fungi	34	47	100	100
BaCelLo IDS plants	79	82	99	99

As can be seen in Supplementary Table 4, the GOA-based method gives much higher annotation rates than any of the InterProScan results for BaCelLo IDS. Since the GO terms are highly correlated with the protein subcellular localizations, the prediction of subcellular localizations is improved with higher GO annotation rates. Thus, GOA-based method can be regarded as a better annotation tool than InterProScan for predicting protein subcellular localizations.

Supplementary Table 5. The sensitivity and accuracy of PSL prediction on GOA version 70 and 80 (%)

Dataset	GOA (UniProtKB-GOA UniProt 70, March 10th, 2008)	GOA (UniProtKB-GOA UniProt 81, April 22th, 2010)
BaCelLo IDS animals	97.15/98.15	97.97/98.15
BaCelLo IDS fungi	99.03/98.33	99.23/99.04
BaCelLo IDS plants	95.52/97.72	95.36/97.54

The numbers represent AVG/ ACC and are given in percentages.

We also performed PSL prediction using an older GOA version (version 70 as of Mar. 2008). The performance of WegoLoc was slightly degraded, but it still outperformed other GO-based methods that were published in 2009 or later (Table 1).

5.2 SVM versus KNN

We performed prediction tests using a nearest-neighbor classifier based on GOA annotation. Supplementary Table 6 shows the sensitivity (SE) for each subcellular localization, average sensitivity (AVG) and overall accuracy (ACC) using the same datasets, BaCelLo IDS animals and fungi tested in Supplementary Table 1. As can be seen in Supplementary Table 1 and 5, SVM shows far better accuracy than nearest-neighbor classifier. KNN (k nearest neighbor) method is a generalization of the nearest neighbor method. KNN is used in Hum-mPLoc 2.0 (Shen and Chou, 2009) and it showed a lower accuracy than WegoLoc as shown in Table 1.

Supplementary Table 6. Prediction results by a nearest-neighbor classifier using fungal and animal BaCelLo IDS

Location	Animals	Fungi
Secretory pathway	95.9	100.0
Mitochondrion	91.7	94.8
Nucleus	78.8	92.8

Cytoplasm	74.7	77.2
	AVG=85.3, ACC=82.4	AVG=91.2, ACC=86.6

5.3 Weighted GO and non-weighted GO

Tests in a previous paper (Chi, 2010) shows that about 50% of the error rates are reduced for fungal BaCelLo IDS and fungal Höglund IDS by using weighted GO terms, while the errors for animal BaCelLo IDS and animal Höglund IDS do not change much. Thus, weighting GO terms needs to be considered to improve the prediction of subcellular localizations.

5.4 The process of WegoLoc prediction

As shown in Figure 1, each query protein is first BLASTed against GO annotated sequences to obtain the most similar protein with an E-value less than a given threshold. If such protein exits, then the corresponding GO terms of that protein are fetched from UniProtKB-GOA and used for PSL prediction. Otherwise, an amino acid composition-based prediction is applied as a backup method.

6. Description of the WegoLoc output (Figure 2)

1. The top table shows the input options selected by the user.
2. Prediction results are downloadable as a text file by clicking download.
3. Descriptions of each column in the lower table:
   A. Sequence name : the name of input amino acid sequence that is put after '>' notation in the FASTA input format.
   B. Predicted locations : the locations predicted by WegoLoc with higher scores than a given threshold.
   C. Probability of each location (%): the probability scores of each location predicted by WegoLoc. If the multiplex threshold θ is used, localizations with a probability score higher than [θ × the highest probability score in this column] will also be assigned to the query proteins. This probability threshold is represented in the bottom of the cell.
   D. Weight, GO terms (description, evidence codes): Weight means the value of the following GO term calculated by using WGO algorithm. The GO term links to the corresponding GO term information, lineage of GO terms and gene product association. Description shows the short biological explanation of the GO term. Evidence code means the attribution of the GO annotation such as a literature reference, another database or a computational analysis; detailed information on the meaning of these evidence codes can be found at http://www.geneontology.org/GO.evidence.shtml.
   E. Best BLAST hit (UniProtKB Accession: E-value): When a query sequence is entered, WegoLoc searches the proteins that has GO annotation(s) for the most similar protein by BLAST and makes use of all the corresponding GO terms. UniProtKB Accession provides the accession of the most similar protein and corresponding BLAST E-value.

REFERENCES

Barrell D., Dimmer E., Huntley R.P., Binns D., O'Donovan C., Apweiler R. (2009), The GOA database in 2009--an integrated Gene Ontology Annotation resource. Nucleic Acids Research 2009 37: D396-D403.

Blum, T., Briesemeister, S. and Kohlbacher, O. (2009) MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction, *BMC bioinformatics*, **10**, 274.

Casadio, R., Martelli, P.L. and Pierleoni, A. (2008) The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation, *Briefings in functional genomics & proteomics*, **7**, 63-73.
Chang, C.-C. and Lin, C.-J. (2001) LIBSVM: a library for support vector machines.
Höglund, A., *et al.* (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition, *Bioinformatics*, **22**, 1158-1165.
Chi, S.-M. (2010) Prediction of protein subcellular localization by weighted gene ontology terms, *Biochemical and biophysical research communications*, **399**, 402-405.
Hunter, S., *et al.*, InterPro: the integrative protein signature database (2009). *Nucleic Acids Res.* **37** (Database Issue), D211-215
Pierleoni, A., *et al.* (2006) BaCelLo: a balanced subcellular localization predictor, *Bioinformatic*, **22**, e408-e416.
Shen, H.B. and Chou, K.C. (2009) A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0, *Analytical biochemistry*, **394**, 269-274.
Shen, H.B. and Chou, K.C. (2007) Hum-mPloc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites, *Biochem. Biophy. Res. Comm.*, **355**, 1006-1011.