Supporting Information to SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins

August 31, 2009

1 Cross-validation test

SherLoc2 was tested via 5-fold cross-validation on the MultiLoc datasets [1]. Table 1 shows a detailed statistic of the cross-validation. The confusion matrices for the animal, fungi, and plant model are shown in Table 2, Table 3, and Table 4, respectively.

Loc	**No.**	**Animals**		**Fungi**		**Plants**	
		SE	**MCC**	**SE**	**MCC**	**SE**	**MCC**
nu	837	0.96	0.92	0.93	0.91	0.93	0.91
cy	1411	0.90	0.90	0.91	0.91	0.86	0.90
mi	510	0.96	0.93	0.96	0.95	0.95	0.93
ch	449					0.97	0.95
ex	843	0.91	0.92	0.90	0.91	0.91	0.92
pm	1238	0.94	0.95	0.95	0.95	0.95	0.95
pe	157	0.98	0.87	0.96	0.88	0.97	0.87
er	198	0.92	0.90	0.93	0.88	0.92	0.87
go	150	0.93	0.84	0.90	0.84	0.91	0.82
ly	103	0.96	0.84				
va	63			0.97	0.74	0.95	0.76

Table 1: 5-fold cross-validation performance on MultiLoc datasets. For each subcellular location (Loc) (abbreviations from the paper are used) the number of proteins (No.) is shown. For each organism and subcellular location the sensitivity (SE) and Matthews Correlation Coefficient (MCC) are also shown.

Loc	**nu**	**cy**	**mi**	**ex**	**pm**	**pe**	**er**	**go**	**ly**
nu	783	46	3	0	0	2	1	2	0
cy	58	1271	33	8	4	30	3	4	0
mi	3	9	488	1	0	7	0	2	0
ex	1	5	1	770	26	3	9	6	22
pm	2	5	2	24	1169	0	5	22	9
pe	0	1	1	0	0	154	0	1	0
er	0	0	0	5	2	0	182	7	2
go	0	2	0	0	0	0	6	140	2
ly	0	0	0	1	3	0	0	0	99

Table 2: Confusion matrix of 5-fold cross-validation of SherLoc2 on the MultiLoc animal dataset.

Loc	nu	cy	mi	ex	pm	pe	er	go	va
nu	778	52	5	0	0	1	0	1	0
cy	59	1288	19	8	2	24	6	3	2
mi	3	6	490	1	2	6	0	2	0
ex	0	7	0	757	25	2	12	9	31
pm	2	4	0	25	1177	1	9	13	7
pe	0	3	2	0	0	151	1	0	0
er	0	0	0	2	1	0	185	8	2
go	0	0	0	2	1	1	8	135	3
va	0	1	0	0	0	0	0	1	61

Table 3: Confusion matrix of 5-fold cross-validation of SherLoc2 on the MultiLoc fungi dataset.

Loc	nu	cy	mi	ch	ex	pm	pe	er	go	va
nu	780	48	3	2	1	0	1	0	0	2
cy	64	1263	19	14	6	3	30	6	5	1
mi	3	8	482	9	0	1	7	0	0	0
ch	0	1	9	437	0	0	1	0	0	1
ex	0	7	0	0	764	23	1	10	11	27
pm	2	6	0	0	27	1170	0	10	21	2
pe	0	1	2	1	0	0	152	0	1	0
er	0	0	0	0	2	2	0	182	10	2
go	0	1	0	0	1	1	0	8	137	2
va	0	0	0	1	0	1	0	1	0	60

Table 4: Confusion matrix of 5-fold cross-validation of SherLoc2 on the MultiLoc plant dataset.

2 The BLAST predictor

To asses the performance of SherLoc2 on a novel set of proteins we make comparisons with predictions based on BLAST homology search.

First, all proteins from Swiss-Prot release 42.0 with known subcellular localization are extracted. For all proteins we checked whether a corresponding keyword occurs as substring in the "CC" or "KW" field of its Swiss-Prot entry. The keywords for the localizations are shown in Table 5. If so, we assigned the corresponding localizations. Consequently, we allow proteins to have multiple localizations.

The BLAST predictor itself works very straightforward. Given a protein sequence, BLAST is used to find proteins with high local similarity. The BLAST predictor assigns the location of the top-ranked annotated BLAST hit that has only one known localization. If no annotated protein is found with an E-value cutoff of 10, the protein is assigned to the class "None".

We experimented with other keywords for localizations as well as different E-value cutoffs and found this setting to be well performing.

3 Independent dataset test

SherLoc2 was evaluated on two independent datasets (IDSs), the BaCelLo IDS [2] and the animal Höglund IDS [3]. Table 6 shows a detailed statistic for the BaCelLo IDS. The performance of all predictors is worse on the Höglund IDS compared to the BaCelLo IDS. Due to this fact, we evaluated all predictors using top-three sensitivity, which measures the ratio of correctly predicted proteins within the top three ranked locations. The top-three sensitivity allows to draw conclusions on the correctness of the returned location distribution. Table 7 shows a detailed statistic for the Höglund IDS.

Subcellular localization	Corresponding keywords
Cytoplasm	"cy"
Nucleus	"nucle"
Lysosome	"ly"
Endoplasmic Reticulum	"endopl"
Plasma membrane	"plasma"
Perxosisome	"peroxi"
Mitochondrion	"mitochon"
Golgi apparatus	"golgi"
Extracellular	"secret", "extracell"
Chloroplast	"chloroplast"
Vacuole	"vacuol"

Table 5: Subcellular localizations are assigned to a protein from Swiss-Prot if at least on corresponding keyword occurs as substring in the "CC" or "KW" field of its Swiss-Prot entry.

Version	Loc	No.	SherLoc2		MultiLoc2		WoLF PSORT		Euk-mPloc		BLAST	
			SE	MCC	SE	MCC	SE	MCC	SE	MCC	SE	MCC
Animals	SP	75	0.87	0.83	0.87	0.79	0.92	0.80	0.44	0.48	0.38	0.39
	mi	48	0.79	0.75	0.83	0.75	0.71	0.63	0.25	0.41	0.25	0.37
	nu	224	0.62	0.58	0.58	0.54	0.77	0.58	0.73	0.52	0.39	0.21
	cy	85	0.77	0.43	0.71	0.39	0.34	0.23	0.49	0.29	0.37	0.22
Fungi	SP	9	0.78	0.57	0.78	0.63	0.89	0.73	0.89	0.57	0.22	0.06
	mi	77	0.45	0.54	0.51	0.52	0.53	0.44	0.37	0.48	0.31	0.39
	nu	152	0.52	0.35	0.50	0.32	0.93	0.35	0.66	0.39	0.44	0.23
	cy	180	0.69	0.29	0.56	0.22	0.11	0.19	0.56	0.39	0.38	0.22
Plants	SP	6	0.83	0.71	0.83	0.50	0.33	0.24	0.83	0.50	0.50	0.48
	mi	6	0.67	0.48	0.67	0.40	0.42	0.52	0.00	-0.04	0.51	0.44
	ch	72	0.64	0.63	0.53	0.51	0.61	0.43	0.33	0.30	0.58	0.62
	nu	36	0.89	0.74	0.86	0.74	0.72	0.52	0.49	0.39	0.70	0.55
	cy	17	0.44	0.30	0.37	0.20	0.24	0.28	0.55	0.26	0.62	0.41

Table 6: Performance of SherLoc2, MultiLoc2, WoLF PSORT, Euk-mPloc, and the BLAST predictor on the BaCelLo IDS. For each organism version and subcellular location (Loc) (abbreviations from the paper are used) the number of proteins (No.) and the performance of the predictors concerning the sensitivity (SE) and Matthews Correlation Coefficient (MCC) are shown.

Loc	No.	SherLoc2		MultiLoc2		WoLF PSORT		Euk-mPloc		BLAST	
		SE	SE3	SE	SE3	SE	SE3	SE	SE3	SE	SE3
ex	78	0.61	0.85	0.78	0.91	0.93	0.97	0.20	-	0.20	-
pm	34	0.53	0.83	0.55	0.78	0.41	0.59	0.32	-	0.03	-
pe	3	0.00	0.33	0.33	1.00	0.00	0.00	0.00	-	0.00	-
er	25	0.56	0.80	0.28	0.70	0.08	0.40	0.28	-	0.12	-
go	14	0.36	0.71	0.07	0.57	0.00	0.07	0.00	-	0.00	-
ly	4	0.25	0.50	0.25	0.75	0.00	0.25	0.25	-	0.00	-

Table 7: Performance of SherLoc2, MultiLoc2, WoLF PSORT, Euk-mPloc, and the BLAST predictor on the Höglund IDS. For each organism version and subcellular location (Loc) (abbreviations from the paper are used) the number of proteins (No.) and the performance of the predictors concerning the sensitivity (SE) and top-three sensitivity (SE3) are shown. For Euk-mPloc and the BLAST predictor no SE3 values could be obtained due to the fact that both predictors return no location distribution.

4 Use of HomoLoc

As introduced in the paper, HomoLoc is used instead of EpiLoc in case no valid Swiss-Prot AC is given or no PubMed abstract is linked to the Swiss-Prot entry of the query protein. Since EpiLoc and HomoLoc were trained on Swiss-Prot 42.0 no protein from the IDSs has a valid Swiss-Prot AC.

If no homolog is found by BLAST, HomoLoc returns a uniform probability vector. In these cases, the SherLoc2 predictions equal the MultiLoc2 predictions. Table 8 shows how often this is the case.

For the BaCelLo plant IDS the number of non-homologous proteins is zero. This explains the performance gain of SherLoc2 compared to MultiLoc2 for the plant IDSs (see Table 2 in the paper).

Dataset	**No.**	**Ratio no homolog**
BaCelLo Animals IDS	576	61.9%
BaCelLo Fungi IDS	437	64.8%
BaCelLo Plants IDS	399	0.0%
Höglund Animals IDS	198	77.3%

Table 8: For each dataset the table shows the number of proteins (No.) and the ratio of proteins without a homolog in Swiss-Prot 42.0 (Ratio no homolog).

5 Use of DiaLoc

The performance of SherLoc2 can be improved by providing additional information. Especially, if no homolog exists and HomoLoc fails to give predictions, the use of DiaLoc can cover this missing source of information.

If no Swiss-Prot AC is known for a query protein, HomoLoc is used instead of EpiLoc. It might happen that no homolog is known for this protein in Swiss-Prot release 42.0. If so, HomoLoc returns a uniform probability vector. In the SherLoc2 webservice the user is asked whether he wants to accept the uniform probability vector or if he wants to add additional knowledge about the protein in a text field. This text is treated as an abstract that is linked to this protein. Depending on the amount of information the user provides, the prediction can quality improves.

The following example shows how prediction quality can improve by providing a description of the query protein. The protein with Swiss-Prot AC Q8R1W2 is present in the Endoplasmic Reticulum (ER). When running SherLoc2 with this protein as query, HomoLoc returns a uniform probability vector since no homologous protein is known in Swiss-Prot release 42.0. If we would not provide any information SherLoc2 return the plasma membrane as the most probable location with a probability of 33%. The ER is only the third most probable location with a probability of 16%. The scientist that studied the function of this protein [4], however, obtained additional information during their studies. Utilizing the information they annotated in the corresponding Swiss-Prot entry in the fields function, subunit structure and tissue specificity we provide the following protein description for DiaLoc:

May cause the redistribution of PAPOLB from the cytosol to the endoplasmic reticulum. It is integral to the membrane. It interacts with PAPOLB. It is expressed in germ cells within the testis from day 21 onwards.

Given this description SherLoc2 returns the ER as the most probable localization for this protein with a probability of 90%. One could argue that the endoplasmic reticulum is already contained in the description. However, the cytoplasm is mentioned in the description as well and still obtains 0% probability.

We also want to point out that a protein description makes the SherLoc2 prediction bias towards one or more locations. Thus, the user has to be careful when describing a protein. Making nonobjective conjectures about the protein might lead to the guessed location but not necessarily to the right one.

References

[1] A. Höglund, P. Dönnes, T. Blum, H.W. Adolph, and O. Kohlbacher. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, 22(10):1158–1165, 2006.

[2] R. Casadio, P.L. Martelli, and A. Pierleoni. The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Briefings in Functional Genomics and Proteomics*, 7(1):63, 2008.

[3] T. Blum, S. Briesemeister, and O. Kohlbacher. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular localization prediction. *BMC Bioinformatics*, accepted.

[4] H.S. Choi, S.H. Lee, H. Kim, and Y. Lee. Germ cell-specific gene 1 targets testis-specific poly (A) polymerase to the endoplasmic reticulum through protein–protein interactions. *FEBS letters*, 582(8):1203–1209, 2008.