MultiLoc2 and SherLoc2: improved prediction of subcellular protein localization

Torsten Blum[1], Sebastian Briesemeister[1], Scott Brady[2], Yin Lam[2], Oliver Kohlbacher[1], Hagit Shatkay[2]

1 Introduction

The function of a protein is highly correlated with its subcellular localization. However, determining the subcellular localization of a protein experimentally can be difficult and time-consuming. Computational methods for the prediction of subcellular locations of proteins from the sequence alone are an attractive alternative.

MultiLoc2 [1] and SherLoc2 [3] both significantly extend and improve upon previous high-accuracy location prediction methods. In addition to information about N-terminal sorting signals, amino acid composition, and location specific domains, both predictors integrate phylogenetic profiles and Gene Ontology (GO) terms. Moreover, SherLoc2 uses textual information from PubMed abstracts. MultiLoc2 and SherLoc2 predict all 11 main eukaryotic locations. In addition, we provide a low-resolution version of MultiLoc2, which is specialized on discriminating globular proteins from secreted proteins and, consequently, predicts only up to five key locations. MultiLoc2 and SherLoc2 perform significantly better than their previous versions [5, 9] and other state-of-the-art subcellular localization predictors.

MultiLoc2 and SherLoc2 are available as free webservices:
http://www-bs.informatik.uni-tuebingen.de/Services/MultiLoc2/
http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc2/

2 Materials and Methods

For training MultiLoc2 and SherLoc2 we used the MultiLoc dataset, extracted from Swiss-Prot release 42 [5]. It contains proteins from all 11 main eukaryotic locations: nucleus (nu), cytoplasm (cy), mitochondrion (mi), chloroplast (ch), plasma membrane (pm), extracellular space (ex), endoplasmic reticulum (er), Golgi apparatus (go), lysosome (ly), and vacuole (va). The low-resolution version of MultiLoc2 was trained on the BaCelLo datasets [8], extracted from Swiss-Prot release 48, which consist of globular proteins from five localizations (nu, cy, mi, ch, and Secretory Pathway (SP)). Both datasets are homology reduced. Along with these two datasets, two independent datasets (IDS) have been created, which consist of novel proteins added after Swiss-Prot release 48. The BaCelLo IDS covers five localizations (nu, cy, mi, ch, SP) [8], whereas the MultiLoc IDS covers the six remaining localizations (pm, ex, er, go, ly, va) [1].

MultiLoc2 and SherLoc2 are support vector machine-based prediction systems. The output of six subclassifiers (seven in the case of SherLoc2) is collected into a protein profile vector that again forms the input for a final SVM classifier. The subclassifiers differ in their input: SVMTarget uses partial amino acid composition to detect N-terminal targeting peptides. SVMaac uses overall amino acid composition. SVMSA scans for signal anchors, present in membrane proteins. MotifSearch searches for known localization signals and relevant domains. PhyloLoc utilizes the fact that proteins sharing location tend to also share a similar phylogenetic distribution of their homologs in organisms with known genome [6]. It creates a phylogenetic profile by searching for homologous proteins in 78 known genomes. GOLoc uses terms from the GO vocabulary for its prediction. These terms describe biological processes, cellular components, or a molecular function. GOLoc scans for domains using InterproScan and transfers GO annotations from the domains to the protein. It explicitly does not use annotations from the protein itself. In fact, transferred GO terms are properties of all proteins that contain the same InterPro domain. EpiLoc [2], the seventh subclassifier in SherLoc2, is a text-based subclassifier that represents PubMed abstracts linked to the protein's Swiss-Prot entry as weighted term-vectors. A set of distinguishing terms, are probabilistically weighted and used as features in the term vector. SherLoc2 provides

[1] Div. for Simulation of Biological Systems, ZBIT/WSI, Eberhard-Karls-Universität Tübingen, Germany
E-mail: {blum,briese,kohlbacher}@informatik.uni-tuebingen.de

[2] School of Computing, Queen's University, Kingston, Ontario, Canada.
E-mail: sbrady@atum.com, {3ypl,shatkay}@cs.queensu.ca

several novel ways of handling proteins that do not have text associated with them; moreover, it allows users to interactively describe a protein in case no Swiss-Prot entry for the protein exists or no textual information can be transferred from a close homolog.

3 Results and Discussion

The prediction performance was measured using average sensitivity (SEN) and average accuracy (ACC). In a 5-fold cross-validation setting MultiLoc2 improves by $10-13\%$ in SEN and in ACC over the original MultiLoc. Similarly, SherLoc2 shows improvement of $7-9\%$ in both SEN and ACC compared to its previous version, SherLoc [9]. Details are shown in Table 1.

Version	**MultiLoc**	**MultiLoc2**	**SherLoc**	**SherLoc2**
ML Animals	79% (76%)	89% (89%)	87% (86%)	94% (93%)
ML Fungi	78% (77%)	89% (89%)	85% (85%)	94% (93%)
ML Plants	78% (76%)	89% (89%)	86% (85%)	94% (93%)

Table 1: Performance comparison of MultiLoc, the high resolution version of MultiLoc2, SherLoc, and SherLoc2 based on a 5-fold cross-validation on the MultiLoc (ML) dataset with respect to SEN and ACC (in parentheses).

In a second benchmark study we compared MultiLoc2 and SherLoc2 against other state-of-the-art predictors: the BaCelLo predictor [8], WoLF PSORT [4] and LOCTree [7]. Table 2 shows the results of all predictors on the BaCelLo IDS. MultiLoc2 and SherLoc2 perform better for animal and for plant proteins, and comparable for fungal proteins. The low-resolution version of MultiLoc2 performs best since it is specialized for globular proteins. On the second benchmark set, the MultiLoc IDS, MultiLoc2 and SherLoc2 perform significantly better than WoLF PSORT (see Table 2). Most interestingly, MultiLoc2 and SherLoc2 show very high robustness throughout all datasets, even for locations where only little training data is available.

Version	**SherLoc2**	**MultiLoc2**	**MultiLoc2-LowRes**	**WoLF PSORT**	**BaCelLo**	**LOCTree**
BA Animals	76% (71%)	75% (68%)	**80%** (**73%**)	69% (71%)	69% (64%)	61% (62%)
BA Fungi	61% (59%)	59% (52%)	66% (**60%**)	62% (51%)	**71%** (57%)	55% (47%)
BA Plants	69% (69%)	65% (62%)	**72%** (**76%**)	46% (57%)	61% (69%)	65% (70%)
ML Animals	**39%** (54%)	38% (57%)	-	24% (**58%**)	-	-
ML Fungi	27% (**32%**)	**30%** (31%)	-	17% (22%)	-	-
ML Plants	**37%** (**33%**)	30% (30%)	-	17% (20%)	-	-

Table 2: Performance of SherLoc2, MultiLoc2, the low-resolution version of MultiLoc2 (MultiLoc2-LowRes), WoLF PSORT, the BaCelLo predictor, and LOCTree on the BaCelLo (BA) IDSs and the MultiLoc (ML) IDSs with respect to SEN and ACC (in parentheses). The top-scoring SEN and ACC are highlighted in bold. No results could be obtained for MultiLoc2-LowRes, the BaCelLo predictor, and LOCTree on the ML IDS since they predict only five locations.

References

[1] Blum, T., Briesemeister, S., Kohlbacher, O. 2008. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *Bioinformatics* submitted.

[2] Brady, S., Shatkay, H. 2008. EpiLoc: a (working) text-based system for predicting protein subcellular location. In: *Pac. Symp. Biocomput.*, pp. 604-615.

[3] Briesemeister, S., Blum, T., Brady, S., Lam, Y., Kohlbacher, O., Shatkay, H. 2008. SherLoc2: a high-accuracy hybrid method for predicting protein subcellular localization. *Bioinformatics* to be submitted.

[4] Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C .J., Nakai, K. 2007. WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 35:W585-W587.

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

[6] Marcotte, E. M., Xenarios, I., van der Bliek, A. M., Eisenberg, D. 2000. Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl Acad. Sci. USA* 97:12115-12120.

[7] Nair, R., Rost, B. 2005. Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.* 348:85-100.

[8] Pierleoni, A., Martelli, P. L., Fariselli, P. L., Casadio, R. 2006. BaCelLo: a balanced subcellular localization predictor. *Bioinformatics* 22(14):408-416.

[9] Shatkay, H., Höglund, A., Brady, S., Blum, T., Dönnes, P., Kohlbacher, O. 2007. SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics* 23(11):1410-1417.