SherLoc: Comprehensive Prediction of Protein Subcellular Localization by Integrating Clues from Sequence Data and the Literature

H. Shatkay[1], A. Höglund[2], S. Brady[1], T. Blum[2], P. Dönnes[2] and O. Kohlbacher[2]

[1] School of Computing, Queen's University, Kingston, Ontario, Canada
[2] Div. for Simulation of Biological Systems, ZBIT/WSI, University of Tübingen, Germany

Knowing the subcellular localization of a protein can elucidate its function, its role in both healthy processes and in the onset of disease, and its potential use as a drug target. Experimental methods determining subcellular localization are reliable and accurate but slow and labor-intensive. In contrast, high-throughput computational prediction tools enable proteome-wide initial "triage" and provide information that is not otherwise attainable (e.g. for proteins whose composition is deduced from a genomic sequence but are hard to isolate, produce, or locate experimentally).

Much progress in computational prediction of protein subcellular localization using sequence-based information has been reported since the early 1990's. PSORT, a rule-based expert system was introduced by Nakai and Kanehisa [1], and later improved upon using classifiers based on machine learning [2]. Other prominent systems, ChloroP [3] and TargetP [4], based on artificial neural networks, have demonstrated a high accuracy when applied to a limited set of subcellular localizations in either plant (ChloroP) or animal cells (TargetP). Other recent methods use a variety of machine learning techniques. Most of them focus on a few subcellular localizations and improve prediction accuracy, while others support many localizations, albeit with reduced performance. The best performing comprehensive systems reported so far are PLOC [5] and more recently MultiLoc [6].

We present **SherLoc**, a prediction system integrating several types of sequence-based features (using MultiLoc) with information derived from text. We test SherLoc on a variety of previously used data sets, as well as on a new set devised specifically to test its predictive power. In all these tests *SherLoc performs significantly better than any previously reported method*. From a bio-text mining perspective, SherLoc forms a landmark, as it is the first reported system to demonstrate a *significant quantitative* improvement in the performance of a biological task by introducing a text mining element.

The following tables demonstrate the overall improved performance of SherLoc compared with TargetP and PLOC on their own data sets.

SherLoc is available online at: ***http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc/***.

System	**Plant** (accuracy, sens., spec.)	**Non-Plant** (accuracy, sens., spec.)
TargetP	0.85, 0. 86, 0.83	0.90, 0.91, 0.85
SherLoc	**0.95, 0.94, 0.93**	**0.96, 0.97, 0.94**

Table 1: *Overall accuracy and average sensitivity and specificity on the TargetP data set.*

System	**Plant** (accuracy, sens.)	**Animal** (accuracy, sens.)	**Fungal** (accuracy, sens.)
PLOC	0.78, 0. 58,	0.80, 0.60	0.80, 0.57
SherLoc	**0.85, 0.84**	**0.86, 0.85**	**0.85, 0.84**

Table 2: *Overall accuracy and average sensitivity on the PLOC data set (PLOC does not report specificity).*

References

[1] Nakai K. and Kanehisa M.A. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 14(4):897-911.

[2] Horton P. and Nakai K. (1997). Better prediction of protein cellular localization sites with the k nearest neighbors classifiers. ISMB'97:147-52.

[3] Emanuelsson O., Nielsen H., von Heijne G. (1999). ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science* 8(5), 978-84.

[4] Emanuelsson O., Nielsen H., Brunak S., and von Heijne G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* 300(4):1005-16.

[5] Park K.J. and Kanehisa M.A. (2003). Prediction of protein subcellular location by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics* 19(13), 1656-63.

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