

Improvement of PSORT II Protein Sorting Prediction for Mammalian Proteins

Mitsuteru C. Nakao

msn@ims.u-tokyo.ac.jp

Kenta Nakai

knakai@ims.u-tokyo.ac.jp

Human Genome Center, The Institute of Medical Science, The University of Tokyo,
4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

Keywords: protein sorting, subcellular localization, machine learning

1 Introduction

The PSORT system [8] is a unique tool for the prediction of protein subcellular localization in a sense that it can deal with proteins localized at almost all the subcellular compartments. In its several versions, PSORT II [5] was developed for the prediction of eukaryotic proteins using yeast sequences as its training data. The reason why the data from a single species were used was that training data were favored to reflect the subcellular proportion of a proteome. However, since the yeast is a unicellular organism, applying PSORT II to sequences of multicellular organisms can be problematic sometimes. For example, it has been pointed out that secretory proteins tend to be under-predicted. Since the first release of PSORT II, genome projects have been producing rich information of genes for many model organisms including yeasts, nematode, mouse and human. Amongst them, the information of mouse genes is managed in Mouse Genome Database [2] (MGD) and Mouse Genome Informatics (MGI). The information includes the data of the full length cDNAs [7] with the annotation of subcellular localization sites of their products. In this work, we report the improvement of PSORT II from three aspects: the employment of mammalian (murine) data, the optimization of the learning method, and the optimization of the sequence features used.

2 Methods and Results

Data set: From the MGI-2.8 database, amino acid sequences (1) that are annotated with the Gene Ontology [6] Cellular Component Terms (GO:C) verified experimentally (corresponding to the evidence codes, TAS, IDA, IMP, IPI, and IGI) and (2) that do not contain the word “fragment” in the definition line were selected for the training data. 13,527 genes had the corresponding amino acid sequence information and 1,317 genes were annotated by 1,624 GO:C Terms with the above evidence codes. Then, we translated the GO:C information into our classification of subcellular localization sites with a mapping table(`go2sc1`) we constructed. Although there were 1,140, 136 and 41 proteins that localize at single, dual and triple sites, respectively, we only used those with single-localization site for the training. The final non-redundant data set (`f1nr`) contained 599 protein sequences that were classified into the following nine classes: 19 cytoplasmic (`cyt`), 99 cytoskeletal (`csk`), 8 peroxisomal (`pox`), 121 plasma membrane (`pla`), 191 nuclear (`nuc`), 38 mitochondrial (`mit`), 42 golgi (`gol`), 42 extracellular (`exc`) and 19 endoplasmic reticulum (`end`) proteins.

Learning methods: The original PSORT II system was trained by the k -nearest neighbor (k NN) method. In this work, we assessed the prediction ability of two learning methods, the k NN rules and the support vector machine (SVM) [3], with the one-against-one method.

Feature selection: First, we added the prediction result of iPSORT [1] for signal peptides as one of the sequence features. The PSORT II system calculates 35 features in total from a query protein sequence. Since all features are treated equally, it is likely that using its subset yields better results. To this end, we investigated an optimized combination of the features with the forward method of feature selection, maximizing the Matthews' correlation coefficient (MCC) [4] in the leave-one-out cross-validation (LOOCV) test. The result of the feature selection is summarized in Table 1.

Table 1: Prediction performance of the models with selected features.

Model	Parameters	Total Accuracy (%)	Full Features Model (%)
kNN	$k = 1$	57.1	52.6
Weighted kNN	$k = 1$	57.8	51.6
SVM (RBF kernel)	$\gamma = 1/35, C = 810$	-	57.8

The new mammalian version of PSORT will be available at the PSORT WWW Server [8].

3 Discussion

In this work, we successfully improved PSORT II for the prediction of mammalian sequences. However, there still remain several issues for further study. One problem is that all proteins are assumed to be localized in a single subcellular compartment in the system. According to recent experiments, this assumption is not true; proteins can often be localized at multiple compartments. To avoid noises caused from this over-simplified assumption, a multi-class framework should be introduced. Similarly, it could be unrealistic also to assume that all query proteins are assigned to one of the known compartments. Finally, it is essential to improve the quality of each sequence feature used in the prediction.

References

- [1] Bannai, H., Tamada, Y., Maruyama, S., Nakai, N., and Miyano, S., Extensive feature detection of N-terminal protein sorting signals, *Bioinformatics*, 18:298–395, 2002.
- [2] Blake, J.A., Richardson, J.E., Bult, C.J., Kadin, J.A., Eppig, J.T., and the Mouse Genome Database Group, The Mouse Genome Database (MGD): The model organism database for the laboratory mouse, *Nucleic Acids Res.*, 30:113–115, 2000.
- [3] Cortes, C. and Vapnik, V., Support-vector network, *Machine Learning*, 20:1–25, 1995.
- [4] Matthews, B.W., Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochimie. Biophys. Acta.*, 405:442–451, 1975.
- [5] Nakai, K. and Horton, P., PSORT: A program for detecting the sorting signals of proteins and predicting their subcellular localization, *Trends Biochem. Sci.*, 24(1):34–35, 1999.
- [6] The Gene Ontology Consortium, Gene Ontology: Tool for the unification of biology, *Nature Genetics*, 25:25–29, 2000.
- [7] The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium, Functional annotation of a full-length mouse cDNA collection, *Nature*, 409:685–690, 2001.
- [8] <http://psort.ims.u-tokyo.ac.jp/>