

WoLF PSORT: Protein Localization Prediction Software

Paul Horton¹

horton-p@aist.go.jp

Keun-Joon Park²

park-kj@aist.go.jp

Takeshi Obayashi³

toobayas@bio.titech.ac.jp

Kenta Nakai²

knakai@ims.u-tokyo.ac.jp

- ¹ Computational Biology Resarch Center, National Institute of Advance Industrial Science and Technology, 135-0064 Tokyo, Koutou, Aomi 2-43.
- ² Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan.
- ³ Graduate School of Pharmaceutical Sciences, Chiba University, Japan.

Keywords: Protein Localization, k Nearest Neighbors, Feature Selection, PSORT

1 Introduction

Intracellular localization is an important clue to the function of proteins and aberrant localization has been implicated in human disease. Fortunately the one dimensional amino acid sequences of proteins, can yield much accessible information regarding localization. WoLF PSORT draws heavily from the PSORT [2] and PSORTII [3] programs, but unlike those older programs, uses feature selection and a flexible scoring model to increase accuracy and handle multiply localized proteins.

2 Method

Like PSORTII, WoLF PSORT uses the k nearest neighbors classifier (k NN) to predict localization sites. However it differs from PSORTII in several ways. **Dataset:** WoLF PSORT has been trained with a newly constructed dataset based on SWISS PROT (Uniprot) 45.0 and Gene Ontology. **Multiple Localization:** WoLF PSORT is designed to handle multiply localized proteins. **Candidate Features:** WoLF PSORT uses amino acid composition and four features taken from iPSORT[4] in addition to PSORT [2] features.

WoLF PSORT uses the WoLF feature selection and weighting program to increase prediction accuracy. WoLF is a feature selection and weighting program which heuristically searches over the space of integer weights (including zero – thus generalizing feature selection) using k NN for classification and a jackknife test to evaluate each vector of feature weights. The flow of information in WoLF PSORT is shown figure 1.

3 Results

Cross-validation results suggest that WoLF PSORT is significantly more accurate than PSORTII. The results indicate that WoLF PSORT gives useful predictions for many sequences that are not highly similar to any sequence in the training data. We have constructed a public web server wolfpsort.org to provide WoLF PSORT predictions. The server has been accessed by over 7,200 unique URLs and served more than 73,000 predictions. We have recently finished a minimal packaging of the software which is available (via free academic or non-free commercial license) upon request.

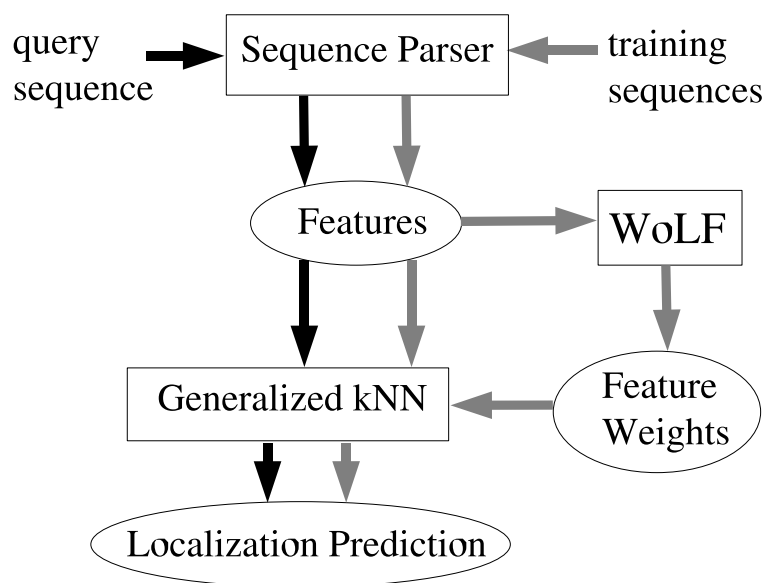


Figure 1: WoLF Schematic: Rectangles represent program modules. Ovals represent computed quantities. The flow of information from the training sequences and query sequences are shown in gray and black respectively.

4 Discussion

k NN is well known as a simple, yet effective classifier. Unfortunately it is also known to be relatively susceptible to the so called “curse of dimensionality”, *i.e.* the tendency to over-generalize when the number of features shown to a classifier increases. WoLF PSORT alleviates this problem by using WoLF feature weighting, while maintaining the simplicity and transparency of k NN classification. The candidate features contain much *ab initio* information about sorting signals and thus the method is a powerful complement to sequence similarity methods which rely almost entirely on non-causal information.

References

- [1] Paul Horton, Keun-Joon Park, Takeshi Obayashi, Kenta Nakai Protein Subcellular Localization Prediction with WoLF PSORT to appear in *Proceedings of APBC06*, Taiwan. 2006.
- [2] Kenta Nakai and Minoru Kanehisa, A Knowledge Base for Predicting Protein Localization Sites in Eukaryotic Cells, *Genomics*, 14:897-911, 1992.
- [3] Paul Horton and Kenta Nakai, Better Prediction of Protein Cellular Localization Sites with the k Nearest Neighbors Classifier, *Proc. 5th Intelligent Systems for Molecular Biology*, 147-152, 1997.
- [4] Hideo Bannai, Yoshinori Tamada, Osamu Maruyama, Kenta Nakai and Satoru Miyano, Extensive feature detection of N-terminal protein sorting signals, *Bioinformatics*, 18:298-305, 2002.