PLOC: Analysis of features for protein's subcellular localization prediction

Keun-Joon Park and Paul Horton[1]

Keywords: subcellular localization, amino acid composition, SVMs, sequence analysis

1 Introduction.

To understand the functions of proteins especially in genome sequencing projects, it is desirable to obtain a protein's subcellular locations automatically from its protein sequence. Park and Kanehisa have developed a prediction method PLOC (Protein LOCalization prediction) using the compositions of amino acid and (gapped) amino acid pairs by support vector machines [1]. PLOC is available at http://www.genome.ad.jp/SIT/ploc.html. In this research, we analyzed the relationship between prediction rate and feature subgroup selection in PLOC. We also have investigated some new features, such as the biological features from PSORT2 [2] (http://psort.ims.u-tokyo.ac.jp/).

2 Method.

We considered 12 subcellular locations in eukaryotic cells (maximum case): chloroplast, cytoplasm, cytoskeleton, endoplasmic reticulum, extracellular medium, Golgi apparatus, lysosome, mitochondrion, nucleus, peroxisome, plasma membrane, and vacuole. For the construction of our dataset, protein sequences were collected from the SWISS-PROT database. In the SWISS-PROT database, we checked keyword information about subcellular locations in the CC field, and also checked the OC field to remove prokaryotic proteins from the dataset. From the protein sequence data set, a set of Support Vector Machines (SVMs) for each subcellular location was trained based on its amino acid, amino acid pair, and from one to three gapped amino acid pair compositions. The case of 12 subcellular locations, 12 SVMs were prepared using five different kinds of composition data. The feature vector contains 20 elements for the amino acid composition, and 400 coordinates for the four kinds of amino acid pair compositions. The prediction methods based on these five different compositions information were then combined using a voting scheme.

The prediction performance was examined by the five-fold cross-validation test, in which the data set was divided into five subsets of approximately equal size (Table 1). In order to assess the accuracy of prediction methods we use two measures, the total accuracy (TA) and the location accuracy (LA) defined by:

$$TA = \frac{\sum_{i=1}^{k} T_i}{N}, \qquad LA = \frac{\sum_{i=1}^{k} P_i}{k},$$

where:

$$P_i = \frac{T_i}{n_i}.$$

[1] Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Aomi Frontier Bldg. 17F, 2-43 Aomi, Koto-ku, Tokyo 135-0064, Japan, E-mail: park-kj@aist.go.jp, horton-p@aist.go.jp

Here N is the total number of proteins in the data set, k is the number of subcellular locations, n_i is the number of proteins in each location i, and T_i is the number of correctly predicted proteins in each location i. In this method, the kernel of SVMs is RBF (Radial Basis Function). We also tested for more realistic repertoires of subcellular locations in different cell types, 11 subcellular locations excluding lysosome for a plant cell, 10 locations excluding chloroplast and lysosome for a fungal cell, and 10 locations excluding chloroplast and vacuole for an animal cell. Note that vacuoles in fungi or plants are thought to correspond to lysosomes in animals.

To investigate the importance of each amino acid pair compositions, forward feature selection analysis was done with the PLOC dataset. The forward selection procedure starts from the evaluation of each individual amino acid pair feature along with the base set of 20 amino acid composition features. The pair feature used in the best combination is then added to the base set and the procedure is repeated until no further improvement is obtained.

3 Results and Discussions.

Table 1 shows the first 10 amino acid pair features selected and each prediction rate from the forward feature selection analysis.

TA	LA	Amino acid pair features
0.727	0.569	GxxxP
0.728	0.572	PxxY
0.728	0.575	NF
0.729	0.577	PxxH
0.729	0.579	YF
0.730	0.580	GxQ
0.730	0.581	RxxY
0.732	0.581	SxxxF
0.733	0.581	LxxxP
0.734	0.582	HxP

Table 1: The first 10 informative amino acid pair compositions selected by forward subset feature selection in PLOC method.

We also acquired some informative feature set from PSORT2. PSORT2 contains 31 biological meaningful features. These features would be added as the new features to the new version of PLOC2 for yeast proteins. Further practical prediction method may be constructed by adding new subcellular locations or defining finer classifications, for example mitochondrial inner, outer membrane or matrix protein groups. And some researchers could also want prediction system for some specific locations only. We have to gather and consider these various needs from the users. Perhaps improvements of prediction rate can be obtained using additional new feature vectors for training of SVMs.

References

[1] Park, K. -J. and Kanehisa, M. 2003. Prediction of proteins subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics* 19:1656-1663.

[2] Nakai, K. and Horton, P. 1999. PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. *Trends Biochem. Sci.* 24:34-35.