PLOC: Prediction of Subcellular Location of Proteins

Keun-Joon Park[1] **Minoru Kanehisa**[2]
park-kj@aist.go.jp kanehisa@kuicr.kyoto-u.ac.jp

Yutaka Akiyama[1]
akiyama@cbrc.jp

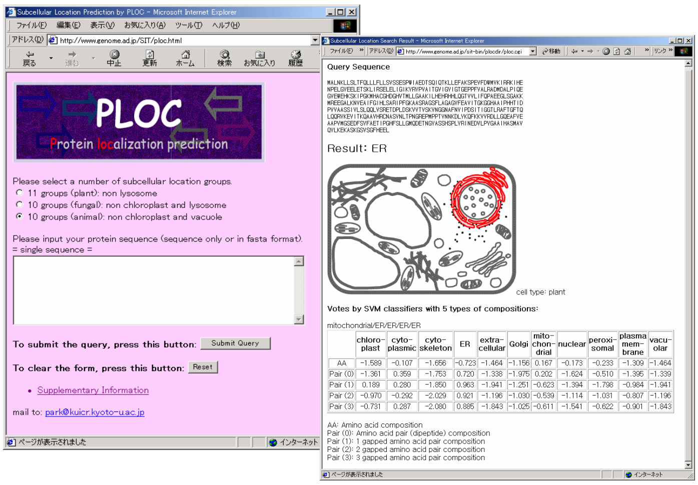
[1] Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Aomi Frontier Bldg. 17F, 2-43 Aomi, Koto-ku, Tokyo 135-0064, Japan
[2] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Keywords: sequence analysis, support vector machines, subcellular localization

1 Introduction

To understand the functions of various proteins, it would be helpful to obtain information about their subcellular location. This is because the subcellular location of proteins is closely related to their function. Thus, it would be worthwhile to develop a fast computational prediction method to identify protein's subcellular location.

In this research, we have developed a prediction tool named PLOC (Protein LOCalization prediction) using the compositions of amino acids and (gapped) amino acid pairs by support vector machines [2].


Subcellular Location Prediction by PLOC - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://www.genome.ad.jp/SIT/ploc.html
戻る 進む 中止 更新 ホーム 検索 お気に入り 履歴
PLOC
Protein localization prediction
Submit Query
Reset
ページが表示されました
インターネット
Subcellular Location Search Result - Microsoft Internet Explorer
ファイル(F) アドレス(D) http://www.genome.ad.jp/sit-bin/plocdir/ploc.cgi 移動 リンク

Please select a number of subcellular location groups.
○ 11 groups (plant): non lysosome
○ 10 groups (fungal): non chloroplast and lysosome
● 10 groups (animal): non chloroplast and vacuole

Please input your protein sequence (sequence only or in fasta format).
= single sequence =

To submit the query, press this button: Submit Query

To clear the form, press this button: Reset

- Supplementary Information

mail to: park@kuicr.kyoto-u.ac.jp

Query Sequence

```
MALNKLLSLTFQLLLFLLSVSSESPWIAEDTSQIQTKLLEFAKSPEVFDWMVKIRRKIHE
NPELGYEELETSKLIRSELELIGIKYRYPVAITGVIGVIGTGEPPFVALRADMDALPIQE
GVEWEHKSKIPGKMHACGHDGHVTMLLGAAKILHEHRHHLQGTVVLIFQPAEEGLSGAKK
MREEGALKNVEAIFGIHLSARIPFGKAASRAGSFLAGAGVFEAVITGKGGHAAIPHHTID
PVVAASSIVLSLQQLVSRETDPLDSKVVTVSKVNGGNAFNVIPDSITIGGTLRAFTGFTQ
LQQRVKEVITKQAAVHRCNASVNLTPNGREPMPPTVNNKDLYKQFKKVVRDLLGQEAFVE
AAPVMGSEDFSYFAETIPGHFSLLGMQDETNGYASSHSPLYRINEDVLPYGAAIHASMAV
QYLKEKASKGSVSGFHEEL
```

Result: ER

cell type: plant

Votes by SVM classifiers with 5 types of compositions:

mitochondrial/ER/ER/ER/ER

	chloro-plast	cyto-plasmic	cyto-skeleton	ER	extra-cellular	Golgi	mito-chon-drial	nuclear	peroxi-somal	plasma mem-brane	vacu-olar
AA	−1.589	−0.107	−1.656	−0.723	−1.464	−1.156	0.167	−0.173	−0.233	−1.309	−1.464
Pair (0)	−1.361	0.359	−1.753	0.720	−1.338	−1.975	0.202	−1.624	−0.510	−1.395	−1.339
Pair (1)	0.189	0.280	−1.850	0.963	−1.941	−1.251	−0.623	−1.394	−1.798	−0.984	−1.941
Pair (2)	−0.970	−0.292	−2.029	0.921	−1.196	−1.030	−0.539	−1.114	−1.031	−0.807	−1.196
Pair (3)	−0.731	0.287	−2.080	0.885	−1.843	−1.025	−0.611	−1.541	−0.622	−0.901	−1.843

AA: Amino acid composition
Pair (0): Amino acid pair (dipeptide) composition
Pair (1): 1 gapped amino acid pair composition
Pair (2): 2 gapped amino acid pair composition
Pair (3): 3 gapped amino acid pair composition

Figure 1: Screen shot of the PLOC and the example of result.

This prediction method is available at `http://www.genome.ad.jp/SIT/ploc.html` (Figure 1).

2 Method and Results

We considered 12 subcellular locations in eukaryotic cells: chloroplast, cytoplasm, cytoskeleton, endoplasmic reticulum, extracellular medium, Golgi apparatus, lysosome, mitochondrion, nucleus, peroxisome, plasma membrane, and vacuole. For the construction of the data set, 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. After homology check (80.0%), the total number of proteins in the final data set was 7589.

From the protein sequence data set, a set of Support Vector Machines (SVMs) was trained based on its amino acid, amino acid pair, and from one to three gapped amino acid pair compositions. The 12 SVMs were prepared at five different composition data. The feature vector has 20 coordinates 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 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:

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

where:

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

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.

We analyzed the change of prediction rates using amino acid composition by feature subset selection (backward selection) analysis. The result shows that the most important feature vector of amino acid composition is the value of cysteine (Cys). The composition of alanine (Ala) shows the opposite tendency.

Table 1: Prediction results from different numbers of subcellular locations.

Accuracy	12 locations	11 locations	10 locations	10 locations
		for plant cells	for fungal cells	for animal cells
TA(%)	78.2±0.9	78.5±0.9	79.5±0.9	79.6±0.9
LA(%)	57.9±2.1	57.9±1.3	56.8±1.9	59.9±3.3

3 Discussion

Results obtained through five-fold cross validation tests showed an improvement in prediction accuracy over the algorithm based on the amino acid composition mainly [1].

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 also want prediction system for some specific locations only. We have to 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] Cai, Y.D., Liu, X.J., Xu, X.B., and Chou, K.C., Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect, *J. Cell. Biochem.*, 84:343–348. 2002.

[2] Park, K.-J. and Kanehisa, M., Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs, *Bioinformatics*, 19(13):1656–1663, 2003.

[3] PLOC, `http://www.genome.ad.jp/SIT/ploc.html`