Protein Subcellular Location Prediction Using Principal Component Analysis

Daichi Nogami[1] Yuuichi Nakano[1] Y-h. Taguchi[1,a)]

Abstract: Inference of subcellular location only using sequence information is important. In this paper, we proposed the application of principal component analysis-based linear discriminant analysis for subcellular location information. It achieved the performance of Area Under the Curve under Receiver Operating Characteristic curve mostly more than 0.95 if less than 80% sequence identity non-redundancy was applied.

Keywords: protein, subcellular location, principal component analysis, linear discriminant analysis

1. Introduction

Inference of subcellular location of protein is important. There are many researches that aimed this goal[1],[2],[3],[4]. Especially, inference based upon only amino acid sequence was mostly useful [5],[6]. Yet if protein annotation can be used for inference, it can outperform the performances achieved by sequence-based method. Li et al [7] recently reported that their GO term based method could achieve high performance. Their method mostly achieved more than 0.9 Matthews Correlation Coefficient (MCC) for 12 subcellular locations with 7579 proteins having less than 80% pairwise sequence similarity[1].

In this paper, we try to compete with GO term based method using sequence based method. Then we found that our sequence based method can compete with GO term based method if the numbers of proteins in positive and negative sets do not differ from each other so much. We also found that generally AUC is as large as 0.90 by our methods.

2. Methods

2.1 Data set

Four data sets were used for this study. iLoc8897[8], Euk7579[1] and Hum3681[9] were employed for GO term based methods[7]. Multiloc dataset [10] was employed for recently proposed sequence based method, Boosting Class Association Rules (BCAR)[11].

2.2 semi-supervised principal component analysis based linear discriminant analysis with voting

Supposed we have M positive proteins and $N(>M)$ negative proteins. Negative proteins are divided into $[N/M]+1$ sets including equal to or less than M negative proteins, where $[x]$ means the integer which is the largest and not more than real number x. Using every negative proteins set and the positive proteins set, performance was measured by leave one out cross validation. The remaining negative proteins are discriminated by LDA trained. This trial was repeated employing every negative proteins set as training set. Label of individual protein is decided by voting over $[N/M]+1$ trials. For each discrimination, semi-supervised learning was used;principal components (PCs) was computed by using all of proteins. Then, discrimination was performed using obtained PCs. Optimal number of PCs was decided so as to have maximum performances. For AUC calculation, geometric means of P-values computed for each trial are used.

2.3 Performance measure

In additon to MCC, Area Under the Curve (AUC) under Receiver Operating Characteristic (ROC) curve was employed because of imbalance data set[12], although very few researches report AUC about protein subcellular location prediction[13],[14].

3. Results

Tables 1 and 2 summarized the performance comparisons between our method and other previous methods. Our method is clearly comparative with GO term based method only when restriction of sequence similarity is modest ($\leq$ 80%) and there are enough number of proteins in targeted category, e.g., Cytoplasm, Extra cell, Mitochondrion, Nucleus and Plasma membrane for Euk7579 data set (underlined in Table 1). There are no other cases where our methods compete with GO term based method. Nevertheless, since some GO terms were assigned using sequence similarity, we are not sure if it is meaningful to apply GO term based discrimination to non-redundant data set where no pairwise sequence similarity more than 25% is allowed. On the other hand, our method clearly outperforms BCAR[11]. Since our methods achieved generally high AUC values, we can conclude that our method can infer protein subcellular location with high accuracy.

Acknowledgments This study was supported by KAKENHI, 23300357 and Chuo University Joint Research Grant.

[1] Department of Physics, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112–8551, Japan
[a)] tag@granular.com

Table 1 Comparison between GO term based method[7] and our method. For MCC values underlined, our method is competitive with GO term based method.

Sequence identity	Euk7579 ≤ 80%				Human3681 ≤ 25%				iLoc8897 ≤ 25%			
		GO Term based	our method			GO Term based	our method			GO Term based	our method	
	N	MCC		AUC	N	MCC		AUC	N	MCC		AUC
Acrosome									14	0.87	0.23	0.78
Cell membrane									657	0.91	0.71	0.85
Cell wall									49	0.88	0.60	0.90
Centrosome									96	0.87	0.67	0.92
Chloroplast	671	0.99	0.64	0.89					385	0.99	0.67	0.91
Centriole					77	0.92	0.66	0.92				
Cyanelle									79	1.00	0.83	0.97
Cytoplasm	1241	0.90	0.86	0.96	817	0.90	0.47	0.81	2186	0.94	0.33	0.82
Cytoskeleton	40	0.96	0.84	0.95	79	0.82	0.51	0.82	139	0.76	0.55	0.82
Endosome					24	0.74	0.65	0.84	41	0.75	0.17	0.62
Endoplasmic reticulum	114	0.97	0.66	0.91	229	0.90	0.26	0.78	457	0.95	0.68	0.88
Extracell	861	0.97	0.98	0.99	385	0.97	0.45	0.94	1048	0.98	0.77	0.96
Golgi apparatus	47	0.97	0.58	0.93	161	0.89	0.52	0.79	254	0.91	0.45	0.76
Hydrogenosome									10	1.00	0.11	0.88
Lysosome	93	0.97	0.77	0.98	77	0.94	0.74	0.90	57	0.88	0.56	0.92
Melanosome									47	0.95	0.50	0.91
Microsome					24	0.80	0.55	0.91	13	0.86	0.21	0.89
Mitochondrion	727	0.95	0.96	0.99	364	0.96	0.78	0.90	610	0.98	0.66	0.88
Nucleus	1932	0.87	0.83	0.96	1021	0.89	0.57	0.86	2320	0.89	0.54	0.87
Peroxisome	125	0.96	0.25	0.81	45	0.95	0.51	0.82	110	0.97	0.36	0.80
Spindle pole body									68	0.92	0.68	0.97
Plasma membrane	1674	0.96	0.95	0.99	354	0.89	0.70	0.83				
Synapse					22	0.85	0.30	0.58	47	0.79	0.50	0.83
Vacular	54	0.93	0.56	0.92					170	0.96	0.65	0.88

Table 2 Performance comparison between a recently proposed sequence based method (BCAR)[11] and our method.

Sequence identity		plant ≤ 80%			animal ≤ 80%		
		BCAR	our method		BCAR	our method	
	N	MCC		AUC	MCC		AUC
Chloroplast	981	0.76	0.82	0.96			
Cytoplasm	3133	0.76	0.81	0.97	0.67	0.77	0.96
Endoplasmic reticulum	442	0.85	0.85	0.96	0.68	0.87	0.96
Extracellular space	3243	0.46	0.91	0.99	0.71	0.91	0.99
Golgi apparatus	563	0.81	0.79	0.96	0.82	0.82	0.96
Lysosome	147				0.71	0.86	0.98
Mitochondrion	1728	0.65	0.79	0.96	0.61	0.83	0.97
Nucleus	820	0.84	0.87	0.98	0.62	0.87	0.98
Peroxisome	262	0.66	0.71	0.92	0.68	0.69	0.93
Plasma membrane	2218	0.67	0.86	0.97	0.83	0.86	0.97
Vacuole	54	0.75	0.79	0.92			

References

[1] 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*, Vol. 19, No. 13, pp. 1656–1663 (2003).

[2] Nakashima, H. and Nishikawa, K.: Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies, *J. Mol. Biol.*, Vol. 238, No. 1, pp. 54–61 (1994).

[3] Imai, K. and Nakai, K.: Prediction of subcellular locations of proteins: where to proceed?, *Proteomics*, Vol. 10, No. 22, pp. 3970–3983 (2010).

[4] Park, K.-J., Kanehisa, M. and Akiyama, Y.: PLOC: Prediction of Subcellular Location of Proteins, *GENOME INFORMATICS SERIES*, pp. 559–560 (2003).

[5] Matsuda, S., Vert, J. P., Saigo, H., Ueda, N., Toh, H. and Akutsu, T.: A novel representation of protein sequences for prediction of subcellular location using support vector machines, *Protein Sci.*, Vol. 14, No. 11, pp. 2804–2813 (2005).

[6] Tamura, T. and Akutsu, T.: Subcellular location prediction of proteins using support vector machines with alignment of block sequences utilizing amino acid composition, *BMC Bioinformatics*, Vol. 8, p. 466 (2007).

[7] Li, L., Zhang, Y., Zou, L., Li, C., Yu, B., Zheng, X. and Zhou, Y.: An ensemble classifier for eukaryotic protein subcellular location prediction using gene ontology categories and amino acid hydrophobicity, *PLoS ONE*, Vol. 7, No. 1, p. e31057 (2012).

[8] Chou, K. C., Wu, Z. C. and Xiao, X.: iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins, *PLoS ONE*, Vol. 6, No. 3, p. e18258 (2011).

[9] Shen, H. B. and Chou, K. C.: A top-down approach to enhance the power of predicting human protein subcellular localization: HumPLoc 2.0, *Anal. Biochem.*, Vol. 394, No. 2, pp. 269–274 (2009).

[10] Hoglund, A., Donnes, P., Blum, T., Adolph, H. W. and Kohlbacher, O.: MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition, *Bioinformatics*, Vol. 22, No. 10, pp. 1158–1165 (2006).

[11] Yoon, Y. and Lee, G. G.: Subcellular Localization Prediction through Boosting Association Rules, *IEEE/ACM Trans Comput Biol Bioinform* (2011).

[12] Huang, J. and Ling, C.: Using AUC and accuracy in evaluating learning algorithms, *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 17, No. 3, pp. 299–310 (online), DOI: 10.1109/TKDE.2005.50 (2005).

[13] Kaundal, R., Saini, R. and Zhao, P. X.: Combining machine learning and homology-based approaches to accurately predict subcellular localization in Arabidopsis, *Plant Physiol.*, Vol. 154, No. 1, pp. 36–54 (2010).

[14] King, B. R., Latham, L. and Guda, C.: Estimation of Subcellular Proteomes in Bacterial Species, *The Open Applied Informatics Journal*, Vol. 3, pp. 1–11 (2009).