Neural Network-Based Discrimination of Golgi Type II Membrane Proteins with Better Accuracy

Tatsuki Kikegawa
Department of Electronics, Graduate School of Science and Technology, Meiji University, 1-1-1 Higashi-mita, Tama-ku, Kawasaki-shi, Kanagawa 214-8751, Japan
Email: ce61030@meiji.ac.jp

Kenji Etchuya and Yuri Mukai
Department of Electronics and Bioinformatics, School of Science and Technology, Meiji University, 1-1-1 Higashi-mita, Tama-ku, Kawasaki-shi, Kanagawa 214-8751, Japan
Email: {etchuya, yuri}@meiji.ac.jp

***Abstract*—Type II membrane proteins in the Golgi apparatus play important roles in biological functions, and predominantly exist as catalysts related to post-translational sugar modification. This study describes a new method for detecting Golgi-localized type II membrane proteins (GLs) from post-Golgi type II membrane proteins (PGs), which are mainly localized in the plasma membrane and endoplasmic reticulum (ER). The method is based on hydropathy profiles and the position-specific scoring matrix (PSSM) in combination with the back propagation artificial neural network (BP-ANN). The accuracy of discriminating GLs from PGs was evaluated in a 5-fold cross-validation test with 94.7% sensitivity and 93.5% specificity. This result shows that our method can predict GLs with high accuracy, and that the PSSM and BP-ANN combination can effectively discriminate GLs.**

***Index Terms*—Golgi, endoplasmic reticulum (ER), plasma membrane, type II membrane protein, discrimination, hydropathy analysis, position-specific scoring matrix (PSSM)**

I. Introduction

The Golgi apparatus has many important proteins that are mainly related to the modification of protein oligosaccharides and vesicular transport. Notably, Golgi-localized type II membrane proteins (GLs), including epimerases, nucleotidases, decarboxylases, oligosaccharide synthases, polysaccharide-degrading enzymes, and glycosyltransferases, are often found in the Golgi apparatus. The major members of the GLs are glycosyltransferases, which are related to protein sugar modification [1]–[3]. Sugars play important roles in many vital reactions such as cell adhesion, signal transfer, and subcellular localization, and their type and order variations depend on the localization of glycosyltransferases. Therefore, the identification and classification of GLs is essential in clarifying the mechanisms underlying sugar modification. Thus, a computational method for finding GLs in mammalian genomes is desired.

Manuscript received November 20, 2017; revised January 20, 2018.

The computational protein localization prediction systems like NNPSL [4], PSORT II [5], TargetP [6], SubLoc [7], iPSORT [8], LOC3D [9], PLOC [10], and other algorithms [11]–[17] are considered to be powerful tools as predictors of the subcellular localization of unknown proteins. However, these tools do not easily predict Golgi-localized membrane proteins because of the absence of clear signals or motifs for Golgi transport, and the low number of available membrane proteins in training datasets to develop the localization predictors. Therefore, an original GL detection system is required.

In our previous study, a detection algorithm of GLs was developed by a position-specific score matrix (PSSM) [18]. PSSM was calculated by the hydropathy alignment position-specific amino acid propensities of GLs and post-Golgi type II membrane proteins (PGs). Using this method, GLs were detected with 96.2% sensitivity and 93.5% specificity in a self-consistency test, and with 88.0% sensitivity and 85.5% specificity in a 5-fold cross-validation test. However, the method needed to be improved by combining machine learning approach.

In this study, GL and PG discrimination was achieved using a combination of PSSM and the back-propagation artificial neural network (BP-ANN). By applying BP-ANN, GL discrimination accuracy improved because the most appropriate alignment-position-specific connection strength could be adjusted. BP-ANN, with a three-layered structure, was trained by the PSSM of GLs and PGs. GLs could be detected with 94.7% sensitivity and 93.5% specificity in a 5-fold cross-validation test.

II. Materials and Methods

A. Dataset Preparation

Datasets of mammalian GLs and PGs were obtained from the UniProt Knowledgebase/Swiss-Prot protein sequence database release 2017_04 (April 2017) [19] by conducting a search with the keywords “Mammalia” in OC lines and “type II membrane protein” in CC lines. Entries that had a “Fragment” annotation in the DE lines were excluded from the dataset. The GLs were also

doi: 10.18178/ijpmbs.7.1.16-19

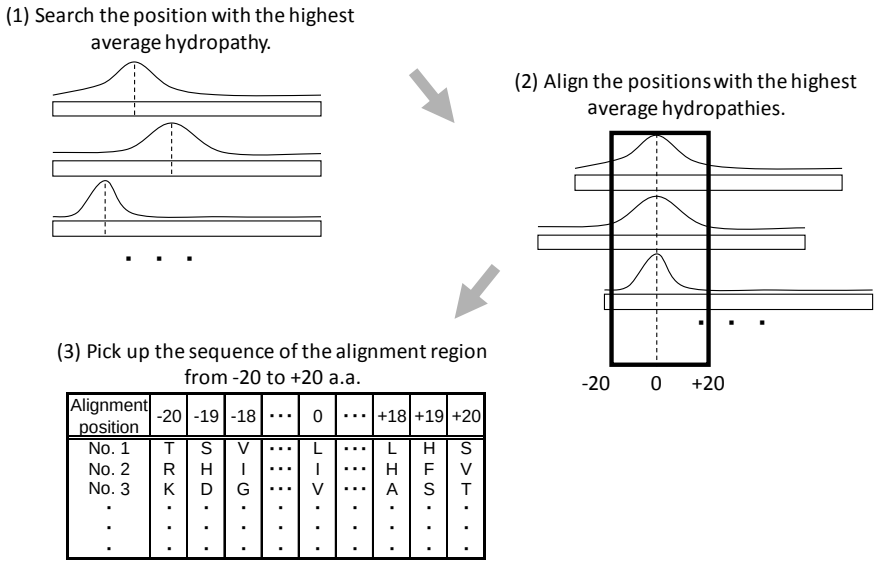
distinguished from the negative control (PGs) based on the "Golgi" annotation for subcellular localization. Data with the annotation "Potential" or "Probable" were eliminated. Representative sequences were extracted from the groups that clustered based on 100% sequence identity using the CD-HIT program [20].

B. Hydropathy Alignment and Dataset Extraction for Position-Specific Scoring Matrix

The average hydrophobicity of each protein was estimated using the moving average method with a sliding window of a certain size. The Kyte-Doolittle (K-D) hydropathy index [21] was used to calculate amino acid hydrophobicity. Accordingly, the average hydrophobicity (H_i) of the 100 amino acid residues at the N-terminus was expressed as follows:

$$\overline{H_i} = \frac{1}{w}\sum_{k=i-m}^{i+m} H(k), \quad \left(m = \frac{w-1}{2}\right) \tag{1}$$

Where, $H(k)$ is the K-D hydropathy index at the sequence position k and w is the sliding window size for average calculation. As shown in Fig. 1, the most hydrophobic position for each sequence was determined by the moving average method, and the hydrophobicity profiles of these sequences were aligned by superpositioning the most hydrophobic positions (standard points). The sequences in the -20 to +20 amino acid region were extracted based on the standard point.



Alignment position	-20	-19	-18	···	0	···	+18	+19	+20
No. 1	T	S	V	···	L	···	L	H	S
No. 2	R	H	I	···	I	···	H	F	V
No. 3	K	D	G	···	V	···	A	S	T
·	·	·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·	·	·

Figure 1. Hydropathy alignment and sequence extraction.

The following equation was used to calculate the position-specific amino acid propensity (f_{jp}) of each protein:

$$f_{jp} = \frac{n_{jp}}{\sum_{j=1}^{20} n_{jp}} \tag{2}$$

Where, p represents the alignment position determined from the position with the highest average hydrophobicity. To avoid setting the denominator at zero in the case of the PSSM calculation, a constant mode of the pseudo-count was introduced [22] as follows:

$$f_{jp} = \frac{n_{jp} + \frac{\varepsilon}{20}}{\sum_{j=1}^{20} n_{jp} + \varepsilon} \tag{3}$$

Where, ε is the pseudo-count (= 1). The position-specific score s_{jp} was computed based on the following equation:

$$s_{jp} = \ln\left(\frac{f_{jp}^{GL}}{f_{jp}^{PG}}\right) \tag{4}$$

Where, the superscript GL and PG represents the amino acid propensity in the GL and PG dataset, respectively.

C. Score Estimation Using Back-Propagation Artificial Neural Network

BP-ANN, with a three-layered structure (Fig. 2), was used to discriminate the GLs from PGs. Position-specific scores, calculated using equation (4), were assigned to a 41-residue amino acid sequence of the extracted GL and PG entries. After 41 of the position-specific scores of each entry were input into the first layer, BP-ANN was trained 150 times.

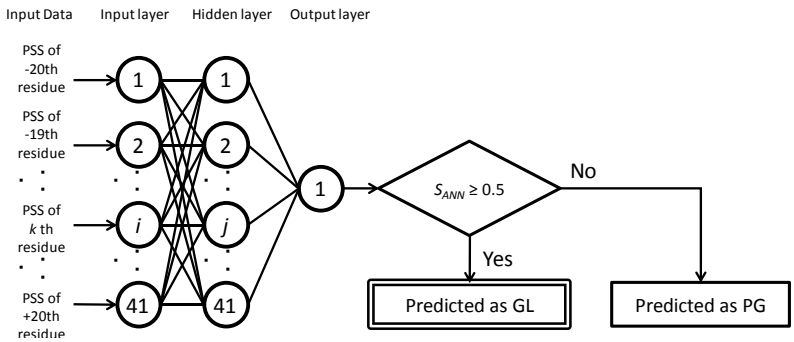


Figure 2. Flow chart for the GL discrimination method using BP-ANN.

D. Evaluation of the GL Discrimination Accuracy

The n-fold cross-validation test is often used for the discriminant analysis of protein sequences [23]–[26] to estimate the prediction accuracy based on sensitivity (*Sn*), specificity (*Sp*), and success rate (*Sr*) as shown below:

$$Sn = TP / (TP + FN) \quad (5)$$

$$Sp = TN / (TN + FP) \quad (6)$$

$$Sr = (Sn \times Sp)^{1/2} \quad (7)$$

In the above equations, *TP*, *TN*, *FP*, and *FN* indicate true positive, true negative, false positive, and false negative, respectively. In the 5-fold cross-validation test, four folds of the non-redundant datasets were randomly selected to create the PSSM and the remaining fold was used to test the discrimination. The average sensitivity, specificity, and success rate was calculated for 1000 random selections.

III. Results and Discussion

The number of entries in the non-redundant GL and PG datasets from the UniProt Knowledgebase/Swiss-Prot protein sequence database release 2017_04 for evaluating the discrimination accuracy in the 5-fold cross-validation tests were 258 and 213 sequences, respectively. The position-specific amino acid propensities of the GLs and PGs based on the hydropathy alignments were calculated using equation 3, which has a pseudo-count. PSSM was created to identify the characteristics of individual amino acids and to discriminate GLs from PGs using those characteristics.

Table I shows the accuracy of GL and PG discrimination using the original non-redundant datasets (258 GLs and 213 PGs) with/without BP-ANN as evaluated in the 5-fold cross-validation test. In the test without BP-ANN, the discrimination score (*S*) was estimated by taking the sum of the position-specific scores at each alignment position and normalizing them to the number of amino acids added from positions *M* to *N* (*Ws*), as shown in the equation below:

$$S = \frac{1}{L}\sum_{p=M}^{N} s_{jp}, \quad (L = N - M + 1, \quad M < N) \quad (9)$$

The lower boundary position (*M*), used to calculate the discrimination score, varied in the range of -20 to -5, and the upper boundary position (*N*) varied from +5 to +20. Maximum accuracy was obtained by adding the scores of the 35 residues located within the -18 to +16 range.

TABLE I. The Accuracy of GL and PG Discrimination with/without BP-ANN Using the 5-Fold Cross-Validation Test

BP-ANN	Ws	*Sn* [%]	*Sp* [%]	*Sr*
+	41	94.7	93.5	0.94
—	35	89.5	84.0	0.86
— (previous study)	23	88.0	85.3	0.87

BP-ANN is a network of non-linear processing units that have adjustable connection strengths between each layer. The BP-ANN in our method has only one hidden layer and therefore, the connection strengths directly provide information regarding the amino acids and the positions that are important for the GL characterization. Discrimination accuracy improved because the most appropriate alignment-position-specific connection strengths could be adjusted by training BP-ANN with PSSM.

GLs were discriminated from PGs with high accuracy using an algorithm that involved hydropathy alignment, PSSM, and BP-ANN in this study. Hydropathy alignment can select the sequence around transmembrane regions of GLs and PGs as the first step [19]. The PSSM and BP-ANN combination can then find the region that has the characteristics of GLs with high accuracy. Our method can effectively identify GLs from unknown membrane protein sequences due to the high evaluation obtained from combining membrane protein prediction tools such as TMHMM [27], SOSUI [28], and other algorithms [29, 30]. This method is a promising tool that can contribute to the genome-wide screening of GLs. The comprehensive identification of GLs is expected to reveal the mechanisms of protein glycosylation.

IV. Conclusion

The combination of PSSM, based on position-specific amino acid propensities in sequences aligned with respect to residue size, and the BP-ANN technique allowed us to distinguish GLs from PGs with high accuracy. The position-specific amino acid propensities and the connection strengths in BP-ANN around the transmembrane regions are important parameters for the

identification of GLs.

ACKNOWLEDGMENT

This work was financially supported in part by Grant-in-Aid for Scientific Research (C) (KAKENHI).

REFERENCES

[1] C. R. Bertozzi and L. L. Kiessling, "Chemical glycobiology," *Science*, vol. 291, pp. 2357-2364, Mar. 2001.
[2] P. M. Rudd, T. Elliott, P. Cresswell, I. A. Wilson, and R. A. Dwek, "Glycosylation and the immune system," *Science*, vol. 291, pp. 2370-2376, Mar. 2001.
[3] L. Wells, K. Vosseller, and G. Hart, "Glycosylation of nucleocytoplasmic proteins: signal transduction and O-GlcNAc," *Science*, vol. 291, pp. 2376-2378, Mar. 2001.
[4] A. Reinhardt and T. Hubbard, "Using neural networks for prediction of the subcellular location of proteins," *Nucleic Acids Res.*, vol. 26, pp. 2230-2236, May 1998.
[5] K. Nakai, and P. Horton, "PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization," *Trends Biochem. Sci.*, vol. 24, pp. 34-36, Jan. 1999.
[6] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence," *J. Mol. Biol.*, vol. 300, pp. 1005-1016, July 2000.
[7] S. Hua and Z. Sun, "A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach," *Bioinformatics*, vol. 17, pp. 721-728, Apr. 2001.
[8] H. Bannai, Y. Tamada, O. Maruyama, K. Nakai, and S. Miyano, "Extensive feature detection of N-terminal protein sorting signals," *Bioinformatics*, vol. 18, pp. 298-305, Feb. 2002.
[9] R. Nair and B. Rost, "Sequence conserved for subcellular localization," *Proteins*, vol. 53, pp. 917-930, Dec. 2003.
[10] K. J. Park and M. Kanehisa, "Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs," *Bioinformatics*, vol. 19, pp. 1656-1663, Sep. 2003.
[11] K. C. Chou, "iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites," *Mol. Biosyst.*, vol. 8, pp. 629–641, Feb. 2012.
[12] W. Z. Lin, "iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins," *Mol. BioSyst.*, vol. 9, pp. 634–644, Apr. 2013.
[13] A. Adelfio, V. Volpato, and G. Pollastri, "SCLpredT: Ab initio and homology-based prediction of subcellular localization by N-to-1 neural networks," *SpringerPlus*, vol. 2, p. 502, Oct. 2013.
[14] T. Goldberg, *et al.*, "Loctree3 prediction of localization," *Nucl. Acids Res.*, vol. 42, pp. W350–W355, July 2014.
[15] A. M. Hasan, S. Ahmad, and K. I. Molla, "Protein subcellular localization prediction using multiple kernel learning based support vector machine," *Mol. BioSyst.*, vol. 13, pp. 785-795, Feb. 2017.
[16] X. Cheng, S. G. Zhao, W. Z. Lin, X. Xiao, and K. C. Chou, "pLoc-mAnimal: Predict subcellular localization of animal proteins with both single and multiple sites," *Bioinformatics*, Vol. 33, pp. 3524-3531, Nov. 2017.
[17] A. J. J. Almagro, C. K. Sønderby, S. K. Sønderby, H. Nielsen, and O. Winther, "DeepLoc: Prediction of protein subcellular localization using deep learning," *Bioinformatics*, vol. 33, pp. 3387-3395, Nov. 2017.
[18] Y. Mukai, *et al.*, "Discrimination of Golgi type II membrane proteins based on their hydropathy profiles and the amino acid propensities of their transmembrane regions," *Biosci. Biotechnol Biochem.*, vol. 75, pp. 82-88, Jan. 2011.
[19] C. O'Donovan, M. J. Martin, A. Gattiker, E. Gasteiger, A. Bairoch, and R. Apweiler, "High-quality protein knowledge resource: SWISS-PROT and TrEMBL," *Brief Bioinform.*, vol. 3, pp. 275-284, Sep. 2002.
[20] L. Fu, B. Niu, Z. Zhuy, S. Wu, and W. Li "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, pp. 3150–3152, Dec. 2012.
[21] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.*, vol. 157, pp. 105-132, May 1982.
[22] J. M. Claverie and S. Audic, "The statistical significance of nucleotide position-weight matrix matches," *Comput. Appl. Biosci.*, vol. 12, pp. 431-439, Oct. 1996.
[23] Y. Mukai, M. Ikeda, H. Tanaka, T. Konishi, O. Oura, and T. Sasaki, "Discrimination of mammalian GPI-anchored proteins by their hydropathy and amino acid propensities," *Biosci. Biotech. Biochem.*, vol. 77, pp. 526-533, Mar. 2013.
[24] M. C. Jespersen, B. Peters, M. Nielsen, and P. Marcatili, "BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes," *Nucl. Acids Res.*, vol. 45, pp. W24–W29, July 2017.
[25] K. Etchuya, R. Nambu, and Y. Mukai, "Discrimination of xylose O-glycosylation sites in mammalian proteins," *Chem. Lett.*, vol. 42, pp. 1043-1045., Sep. 2013.
[26] F. Pucci, R. Bourgeas, and M. Rooman, "Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC," *Sci. Rep.*, vol. 6, No. 23257, Mar. 2016.
[27] L. Käll, A. Krogh, and E. L. L. Sonnhammer, "Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server," *Nucleic Acids Res.*, vol. 35, pp. 429-432, July 2007.
[28] T. Hirokawa, S. Boon-Chieng, and S. Mitaku, "SOSUI: classification and secondary structure prediction system for membrane proteins," *Bioinformatics*, vol. 14, pp. 378-379, May 1998.
[29] G. E. Tusnády and I. Simon, "The HMMTOP transmembrane topology prediction server," *Bioinformatics*, vol. 17, pp. 849-850, Sep. 2001.
[30] M. Ikeda, M. Arai, D. M. Lao, and T. Shimizu, "Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies," *In. Silico. Biol.*, vol. 2, pp. 19-33, Jan. 2002.

Tatsuki Kikegawa received his bachelor degree from Meiji University in 2016. He is currently a master's course student of the Department of Electronics, Graduate School of Science and Technology, Meiji University, Japan. His research interests are subcellular localization mechanisms of transmembrane proteins using bioinformatics and experimental approaches. He is a member of the ISCB, JSBBA, MBSJ, BSJ, and PSSJ.

Kenji Etchuya received his Ph.D. degree in Electronics from Meiji University in 2017. He is currently a postdoctoral researcher in the Life-science Informatics laboratory, Department of Electronics and Bioinformatics, School of Science and Technology, Meiji University, Japan. His research interests are protein sugar modification using bioinformatics and experimental approaches. He is a member of the ISCB, JSBBA, MBSJ, BSJ, and PSSJ.

Yuri Mukai received her Ph.D. degree in Pharmaceutical Science from Hokkaido University in 2000. She is currently an associate professor in the Department of Electronics and Bioinformatics, School of Science and Technology, Meiji University, Japan. Her research interests are protein molecular biology and bioinformatics. She is a member of the ISCB, JSBBA, MBSJ, BSJ, PSSJ, and BBSJ.