Membrane proteins in the Golgi apparatus play important roles in biological functions, predominantly as catalysts related to post-translational modification of protein oligosaccharides. We succeeded in extracting the characteristics of Golgi type II membrane proteins computationally by comparison with those of Golgi no retention proteins, which are mainly localized in the plasma membrane. Golgi type II membrane proteins were detected by combining hydropathy alignment and a position-specific score matrix of the amino acid propensities around the transmembrane region. We achieved 96.2% sensitivity, 93.5% specificity, and a 0.949 success rate in a self-consistency test. In a 5-fold cross-validation test, 88.0% sensitivity, 85.5% specificity, and a 0.867 success rate were achieved.

Key words: Golgi; type II membrane protein; discrimination; hydropathy analysis; position-specific score matrix

The Golgi apparatus is an important organelle in eukaryotic cells. It is involved mainly in the modification of protein oligosaccharides and vesicular transport. There are many proteins in the Golgi apparatus with important roles in its function, notably Golgi-localized type II membrane proteins (GLs), including epimerases, nucleotidases, decarboxylases, oligosaccharide synthases, polysaccharide-degrading enzymes, and glycosyltransferases. Glycosyltransferases are major members of the GL and are one of the most important types of proteins involved in the modification of protein oligosaccharides.1–3) Oligosaccharides play important roles in many vital reactions, including cell adhesion, signal transfer, and subcellular localization. Variations in oligosaccharide modification depend on the localization of the protein and families of oligosaccharide-related proteins in the Golgi apparatus, including the glycosyltransferases. Hence identification and classification of GLs is essential in clarifying the mechanisms underlying oligosaccharide modification. To this end, the development of a computational method to distinguish the genes encoding GLs in mammalian genomes is desired.

A sequence similarity search is the usual method used to find useful genes from open reading frames in mammalian genomes, but these similarities are low among the families of GL genes. Even when a sequence similarity search is conducted in protein sequence databases, only known GLs with strong similarity to each other are found, and novel ones are rarely recognized. Hence the development of a GL detection method that does not rely on sequence similarity is needed to identify unannotated GLs.

The computational approaches used in subcellular localization prediction systems, such as NNPSL,4) PSORT II,5) TargetP,6) SubLoc,7) iPSORT,8) LOC3D,9) PLOC,10) WoLF PSORT,11) and other algorithms,12–17) are considered to be useful for the selection of Golgi-localized membrane proteins. Although these subcellular localization prediction systems predict soluble proteins well, they cannot predict Golgi-localized membrane proteins well. The reasons are: (i) database entries with sequences closely similar to those of known Golgi-localized proteins are required by prediction systems that use sequence similarity searches; (ii) there is no clear signal or motif for Golgi transport; and (iii) few membrane proteins are available for the construction of training datasets for developing localization prediction system. Therefore, the development of an original computational detection system for GLs is required.

In this work, we focused on the transmembrane region (TMR) at the N-terminus of GLs. First, we scanned each sequence for average hydrophobicity, and aligned the regions spanning the N-terminal 100 amino acid residues according to the position that showed the highest average hydrophobicity in the TMR. Next, we analyzed position-specific amino acid propensities, based on the alignment positions of GLs and post-Golgi type II membrane proteins (PG), except for the Golgi retention proteins, which are mainly localized in the
plasma membrane. We created a position-specific score matrix (PSSM)\(^1\) using the various amino acid propensities and estimated the discrimination score for each dataset.

Applying this score, 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. PGs were also detected with 93.5% sensitivity and 96.2% specificity in a self-consistency test, and with 85.3% sensitivity and 88.2% specificity in a 5-fold cross-validation test. Thus, we succeeded in detecting the TMRs in the GLs by combining hydropathy alignment and the PSSM. Recent studies have reported prediction methods for transmembrane topology (i.e., TMR position at the sequence and N-tail location) of membrane proteins, which can be correctly predicted at 75–80% for one TMR (1-TMR) using only amino acid sequence information.\(^2\) Here we propose a GL discrimination method with high accuracy from the sequence data predicted to be type II membrane proteins (i.e., 1-TMR and the N-tail loop localized in the cytoplasm/lumen). This method, which is based on the characteristics of the amino acid sequences around the Golgi TMR, can be used to detect the genes of type II membrane proteins localized at the Golgi membrane in mammalian genomes.

Materials and Methods

Dataset preparation. Two known nonredundant protein datasets were used, containing mammalian GLs (422 sequences) and PGs (417 sequences). They were obtained from UniProt Knowledgebase/Swiss-Prot protein sequence database release 57.0 (24-Mar-2009)\(^3\) by searching with the keywords “Mammalia” in the OC lines and “type II membrane protein” in the CC lines for subcellular localization. Entries that had a “Fragment” annotation in the DE lines were deleted from the datasets. The GLs were also distinguished from the negative control, PGs, based on “GolgI” annotation for subcellular localization. Data with the annotation “Potential” or “Probable” were eliminated. Representative sequences were extracted from groups that clustered on the basis of 98% sequence identity in the 100 amino acid residues at the N-terminus using the single-linkage clustering program BLASTClust (unpublished, a document available at http://www.ncbi.nlm.nih.gov/BLAST/docs/blastclust.html) in the BLAST package. The numbers of entries in the nonredundant GL and PG datasets for evaluating the discrimination accuracy in self-consistency and 5-fold cross-validation tests were 344 and 356 sequences, respectively.

To evaluate discrimination performance objectively, GL and PG sequences not included in Swiss-Prot release 57.0 were extracted from release 2010_09 (10-Aug-2010) by a procedure similar to that described above. This dataset comprises 4 GLs and 10 PGs with lower than 50% full-length sequence identity to above training dataset.

Hydropathy alignment. 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\(^4\) 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:

\[
H_i = \frac{1}{w} \sum_{k=1}^{w} H(k), \quad (m = w - 1/2)
\]

where \(H(k)\) is the K–D hydropathy index at sequence position \(k\) and \(w\) is the sliding window size for average calculation. The most hydrophobic position was determined for each sequence by moving average method, and the hydrophobicity profiles of these sequences were aligned by superpositioning the most hydrophobic positions (standard points).

Amino acid propensities and position-specific score matrix (PSSM).

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

\[
f_{p,j} = \frac{n_{jp}}{20} \sum_{j=1}^{20} n_{jp}
\]

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,\(^5\) as follows:

\[
f_{p,j} = \frac{n_{jp} + \epsilon}{20} \sum_{j=1}^{20} n_{jp} + \epsilon
\]

where \(\epsilon\) is the pseudo-count (= 1). The position-specific score \(s_{jp}\) is computed generally by dividing \(f_{p,j}\) by the background propensity was calculated from all the representative sequences of GL and PG. Thus, the same background propensity was used for both GLs and PGs as an original calculation method in this study. Thus \(s_{jp}\) was computed based on the following equation:

\[
s_{jp} = \ln \left( \frac{f_{p,j}}{f_{p,j}} \right)
\]

where the superscript GL represents the amino acid propensity in the GL dataset and the superscript PG represents the propensity calculated from the PG dataset. The discrimination score \(S\) was estimated by taking the sum of the position-specific scores at the various alignment positions and normalizing them by the number of added amino acids for positions \(M\) to \(N\), as shown in the following equation:

\[
S = \frac{1}{L} \sum_{j=1}^{N} s_{jp}, \quad (L = N - M + 1, M < N)
\]

Accuracy test. Self-consistency and 5-fold cross-validation tests, often used for the discriminant analysis of protein sequences,\(^6\) were performed to estimate prediction accuracy using the sensitivity, the specificity, and the success rate, as shown below:

\[
\text{Sensitivity} = \frac{\text{Correctly predicted true}}{\text{True}} \times 100
\]

\[
\text{Specificity} = \frac{\text{Correctly predicted true + Incorrectly predicted false}}{\text{Correctly predicted true}} \times 100
\]

\[
\text{Success rate} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \times \frac{1}{100}
\]

The threshold for discriminating GL from PG was fixed as follows: (i) if the frequency distributions of the discrimination scores for GLs and PGs showed no overlap, then the threshold was the average of the minimum score of the GLs and the maximum score of the PGs; (ii) if there was an overlap, then the threshold was determined by scanning within the overlapping area so that the success rate (Eq. (8)) was maximized. If there were a number of candidates for the threshold, the average of the highest and lowest candidates was used for discrimination. In the self-consistency test, the nonredundant datasets (GL: 344 and PG: 356; see “Materials and Methods”) were used for PSSM creation and accuracy estimation. In the 5-fold cross-validation test, four-fifths of the nonredundant datasets were randomly selected to create the PSSM, and the other one-fifth was used to test the discrimination. The average sensitivity, specificity, and success rate of an average of 1,000 random selections were calculated.

Results and Discussion

Hydrophobicity profiles and alignments of GL sequences

Figure 1 shows the average hydrophobicity profiles of the first 100 amino acids from the N-terminus of human
(Fig. 1a) and mouse (Fig. 1b) proteins as representatives of mammal GLs (solid line) and PGs (dotted line), calculated using a sliding window of 15 residues. The arrows marked “TMR” show the transmembrane region annotated in Swiss-Prot. In both cases, the most hydrophobic position within the 100 amino acid residues at the N-terminus was included in the annotation region for the TMR. As a consequence, the TMR was found in the region that included the most hydrophobic position. Moreover, the hydrophobic peak width of the GLs (solid line) was smaller than that of the PGs (dotted line). In this calculation, sliding windows of several sizes were evaluated, all of which clearly showed characteristics of GLs and PGs, and all the proteins evaluated with a sliding window of 15 residues showed the same tendencies. Almost all the GLs had highest average hydrophobicity profiles within the region 50 residues from the N-terminus. By contrast, the most hydrophobic positions of the PGs covered a wide area spanning the 100 amino acid residues from the N-terminus. That the most hydrophobic peaks of the PGs were wider and higher than those of the GLs indicates that PGs have a larger hydrophobic domain than GLs.

The GL and PG sequences of the nonredundant datasets (GL, 344, and PG, 356) were aligned by superpositioning the positions that showed the highest average hydrophobicity within the 100 amino acid residues from the N-terminus. Hydropathy alignment was used because the reciprocal similarity of the GL sequences was too low to allow alignment based on amino acid similarities. Figure 2 shows hydrophobicity profiles that were averaged over all the entries in the nonredundant datasets after alignment of the profiles of which had been smoothed with a sliding window of 15 residues. This hydropathy alignment indicates that the TMRs of the GLs are shorter and less hydrophobic than those of the PGs. These results are consistent with the hypothesis of the Munro group, who reported that the TMR of vertebrate Golgi membrane tended to be of shorter length (with an average of 20.6 residues) and less hydrophobic than the post-ER organelles (trans-Golgi network and endosome) and plasma membrane.

The distance from the N-terminus to the TMR is much shorter in GLs than in PGs, as mentioned above. Hence we estimated the distance from the N-terminus to the most hydrophobic position. Figure 3 shows the percentage occurrences of the amino acids plotted as a function of the hydropathy alignment position. The percentage was estimated from the number of entries in which the amino acid residue occurs in that alignment position across all entries. A position that does not reach 100% indicates
that an entry that does not have any amino acid at that position because the distance from the N-terminus to the standard point (0) is short. This distance in the PGs appeared to be greater than that in the GLs, because the percentage occurrence of the amino acids remained at a high level. The percentage amino acid occurrence also corresponded to the reliability of the average hydrophobicity profiles and the amino acid propensities. In the creation of the PSSM and the estimation of the score for discrimination, the data for the area before $-14$ from the standard point (shown by a gray zone) were not used, because the percentage amino acid occurrences of the GLs and PGs in this area were lower than 90%.

Hydropathy alignment and PSSM for GL detection
Position-specific amino acid propensities based on the hydrophathy alignments are shown in Fig. 4. This graph represents the physicochemical characteristics of the amino acid residues as determined by Granseth et al.\textsuperscript{30} Equation (2), which does not take the pseudo-count into consideration, was used to calculate amino acid frequencies. These values were evaluated between positions $-14$ and $+20$, except in the gray zone, where a low percentage occurrence of amino acids was observed, as shown in Fig. 3. In VAFILM, a group of hydrophobic residues, the area from $-7$ to $+7$ is thought to be the core of the TMR, because the propensity of hydrophobic residues rose dramatically in this area. No marked differences were found between the GLs (solid line) and PGs (dotted line) in the TMR core, whereas the PG hydrophobicity was higher than the GL hydrophobicity on both sides of the TMR core ($-12$ to $-8$ and $+8$ to $+11$). In CGPHNQST, a group of polar residues, no marked differences were observed between the GLs and PGs, except at $+1$, $+2$, and $+4$. The propensity for polar residues was lower in the area from $-7$ to $+7$, in contrast to the propensity for hydrophobic residues. The minimum propensities were found at the edges of the TMR ($-7$ and $+7$). In WY, a group of aromatic residues, the propensity was low near the TMR core in the GLs and PGs. However, the propensity was remarkably high at $+8$ in the GLs, which is thought to be a boundary of the TMR core, and decreased thereafter. In KR, a group of positively charged residues, the propensities in both GLs and PGs were very low in the TMR core ($-7$ to $+7$), and were completely absent in GLs and PGs in the areas from $-1$ to $+3$ and $-2$ to $-1$ respectively. Nevertheless, the propensities were remarkably high in the area before $-8$, especially in GLs. The fact that the propensities were higher before the TMR core than after it in both GLs and PGs is thought to be attributable to the effect of the positive-inside rule,\textsuperscript{31} in which a positively charged residue behaves as a clasp when a biosynthesized polypeptide enters the membrane. In DE, a group of negatively charged residues, the propensities were lower than those of any other residue groups before the TMR core, and were zero at $-4$, $-3$, $+1$, $+2$, and $+4$ in PGs, and from $-7$ to $-2$ in GLs. However, the propensities gradually recovered and increased after the TMR core in both GLs and PGs.

Sharpe et al.\textsuperscript{29} recently reported a comprehensive comparison of the TMRs of vertebrate and fungi proteins localized in the ER, Golgi, and plasma membranes.\textsuperscript{29} Especially, our results for amino acid propensities are close to their observed amino acid composition around TMR with regard as follows: (i) LF showed significantly lower propensities on the cytoplasmic side in TMR than on the lumenal side, and (ii) KR appeared to show markedly higher propensities before the cytosolic edge of TMR.

The PSSM was developed to observe the characteristics of individual amino acids and to discriminate GLs from PGs using these characteristics (Table 1). The position-specific score for each amino acid at each position (Eq. (4)) was calculated using 20 position-specific amino acid propensities, taking the pseudo-count into consideration (Eq. (3)). In Table 1, a positive score is given when the appearance frequency in the GLs exceeds that in the PGs. A score with an absolute value near 0 indicates that there is little difference in propensity between the GLs and the PGs. The characteristics of the amino acid propensities, as suggested by the PSSM, were as follows: For charged residues, remarkable appearance frequencies were found before and after the TMR core. Table 1 shows the different propensities for each amino acid, and Fig. 4 shows those of the

![Fig. 4. Position-Specific Amino Acid Propensities of GLs (solid line) and PGs (dotted line) Plotted for (a) Hydrophobic, (b) Polar, (c) Aromatic, (d) Positively Charged, and (e) Negatively Charged Residues.](image-url)

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Propensity (%)</th>
<th>Propensity (%)</th>
<th>Propensity (%)</th>
<th>Propensity (%)</th>
<th>Propensity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydrophobic</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Polar</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Aromatic</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Positively Charged</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Negatively Charged</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 1. PSSM Calculated on the Basis of the Hydropathy Alignment and Position-Specific Amino Acid Propensities for GL and PG

| Residue | -14 | -13 | -12 | -11 | -10 | -9 | -8 | -7 | -6 | -5 | -4 | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---------|-----|-----|-----|-----|-----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A       | -0.26 | -0.76 | -0.63 | -0.68 | -1.82 | -0.57 | -0.21 | -0.41 | 0.31 | 0.50 | 0.14 | -0.25 | 0.09 | 0.38 | -0.25 | -0.18 | -0.52 | 0.14 |
| C       | -0.86 | -0.09 | 0.36 | -0.31 | -1.07 | 0.72 | 0.72 | -0.52 | 0.77 | 0.03 | 0.22 | -0.03 | 0.72 | -0.61 | 0.24 | 0.64 | 0.58 | -0.17 |
| D       | -2.58 | -4.82 | -1.70 | 1.58 | 3.74 | 0.03 | 1.10 | 0.03 | -3.01 | -3.68 | 0.03 | 0.03 | -3.01 | 1.10 | 0.03 | 0.03 | 0.03 | -5.67 |
| E       | -2.25 | -1.26 | 0.52 | -0.19 | -1.50 | 0.43 | 1.94 | -3.68 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 3.01 | 0.03 | 3.08 | 0.03 |
| F       | 0.01 | 0.15 | -0.34 | -0.36 | 0.52 | -0.66 | -0.94 | 0.52 | 1.02 | -0.16 | 0.14 | 0.36 | 1.60 | 0.54 | 0.31 | 0.58 | 0.70 | -0.35 |
| G       | -0.39 | -0.31 | -1.39 | -0.21 | -0.60 | -0.01 | -1.10 | -1.46 | -0.77 | 0.59 | -0.17 | 0.09 | -0.28 | -0.28 | 0.15 | 0.03 | 0.03 | -1.16 | -0.34 |
| H       | -0.68 | 0.09 | 2.25 | 0.26 | 1.02 | 1.05 | 0.81 | 0.03 | 0.03 | 4.65 | 0.03 | 0.03 | 0.03 | 3.08 | 3.75 | 0.03 | 3.03 | 3.75 |
| I       | 0.16 | -0.50 | -1.18 | -0.87 | -0.98 | -1.54 | 0.03 | -0.48 | -0.22 | -0.29 | -0.28 | 0.03 | -0.98 | -0.66 | -0.22 | -0.33 | -0.34 | -0.26 |
| K       | 0.22 | 0.02 | 0.19 | 0.14 | 0.67 | -0.03 | 0.22 | 0.03 | -0.08 | -0.01 | -0.08 | 0.03 | 3.75 | 0.03 | 0.03 | 3.03 | -3.68 | 0.03 |
| L       | -0.14 | 0.09 | 0.03 | -0.23 | -0.34 | -0.63 | -0.49 | 0.24 | 0.16 | 0.27 | 0.02 | 0.10 | 0.02 | -0.16 | -0.08 | -0.10 | 0.14 |
| M       | 2.48 | 0.82 | 0.23 | -0.28 | 0.02 | -0.38 | -0.94 | 0.17 | 0.54 | -1.06 | 1.00 | 1.89 | -0.87 | -0.99 | 0.62 | 0.37 | 0.72 | -0.94 |
| N       | -1.89 | 0.44 | -0.01 | 0.96 | -1.16 | -1.56 | -0.26 | 0.03 | 0.03 | -3.01 | -0.36 | 0.70 | 0.03 | 1.10 | 3.08 | 3.08 | 3.08 |
| P       | -0.79 | 0.26 | 0.25 | 0.08 | 0.57 | -0.66 | 0.03 | 4.15 | -0.63 | 3.75 | 0.54 | 3.75 | 0.44 | 0.03 | -0.03 | -0.05 | 0.03 | -0.63 |
| Q       | 1.07 | 0.82 | 0.68 | 0.01 | 0.19 | 1.05 | 1.12 | 0.03 | 4.15 | -0.25 | 3.08 | -3.01 | 0.03 | 5.81 | 4.83 | -3.68 | 0.03 | 0.94 |
| R       | 1.00 | 0.86 | 1.07 | 1.07 | 0.91 | 1.29 | 1.01 | 3.08 | 0.03 | 0.03 | 3.08 | 0.03 | 3.08 | 0.03 | 0.03 | 0.03 | -3.01 | 0.03 | -3.68 | 0.03 |
| S       | -0.36 | -0.38 | -0.01 | 0.42 | -0.40 | -0.03 | -0.29 | 1.40 | -1.16 | -0.30 | 0.03 | 0.66 | -0.75 | -0.13 | -0.02 | 1.37 | 0.48 | -0.15 |
| T       | -0.63 | -0.46 | -0.98 | 0.56 | -0.40 | -0.60 | 0.33 | 0.44 | 0.23 | -0.10 | -0.88 | 0.32 | 0.03 | 0.99 | 0.56 | 0.09 | 0.91 | -0.43 |
| V       | 0.49 | -0.80 | -1.14 | 0.78 | 0.10 | -0.77 | -0.21 | -0.29 | -0.88 | -0.72 | -0.06 | -0.57 | -0.42 | -0.06 | 0.00 | -0.45 | -0.07 | 0.34 |
| W       | -0.12 | -0.03 | -0.92 | 0.19 | -0.67 | 0.48 | 0.99 | 0.70 | 1.10 | 0.03 | 0.03 | 0.13 | 1.72 | 0.72 | -0.43 | 4.15 | 4.43 |
| Y       | -2.17 | -2.93 | -2.74 | 0.55 | 0.99 | -0.60 | -1.10 | 5.23 | 1.60 | -0.58 | -1.72 | -1.42 | -4.08 | 3.18 | 4.15 | -5.16 | 0.72 | -0.36 |

Residue groups. For example, for Y and P, the position-specific score at -7 showed large absolute and positive values (Table 1). This means that the propensity in GLs is remark high as compared with that in PGs at -7, which is the start position of the TMR core, whereas the propensities of the hydrophobic residues were high in the TMR core for both GLs and PGs. For another example, the W propensity in GLs was higher than that in the PGs around the TMR core (-9 to +9). Therefore, the characteristics of the position-specific amino acid propensities are fully comprehensible with the use of the PSSM.

Evaluation of detection methods

The discrimination score (Eq. (5)) for each dataset of GLs and PGs was calculated based on the PSSM. Figure 5 shows the frequency distributions of the discrimination scores for the GLs and PGs in the aligned region from -14 to +18, as calculated by the

![Fig. 5. Distribution of Discrimination Scores for GLs (black bar) and PGs (white bar) in the Aligned Region from -14 to +18 as Calculated by PSSM.](image-url)
PSSM. The discrimination scores calculated by the PSSM indicate that GLs could be clearly distinguished from PGs. The threshold for discrimination was set from the score distribution (see "Materials and Methods"). Table 2 shows the sensitivity, the specificity, and the success rate of our method for the discrimination of GLs from PGs using the original nonredundant dataset (GL, 344, and PG, 356), as evaluated by a self-consistency test and a 5-fold cross-validation test. The lower boundary position for the discrimination score calculation varied in a range from -14 to -4, and the upper boundary position varied from +10 to +20. Maximum accuracy was obtained by adding the scores of the 33 residues located within a range of -14 to +18 (M = -14 and N = +18, see Eq. (5)). This calculation was sufficient to quantify the characteristics of the TMR, because 90% of the TMR lengths in our datasets were within 33 residues (Fig. 3). The discriminatory ability of this method might be reduced when the calculation area exceeds 33 residues, because the discrimination score includes the effects of non-TMRs, such as the disordered region.

Discrimination accuracy is shown in Table 2. In a self-consistency test, GLs were detected with 96.2% sensitivity, 93.5% specificity, and 0.949 success rate when the discrimination threshold was set at 0.003. In a 5-fold cross-validation test, GLs were detected with 88.0% sensitivity, 85.5% specificity, and 0.867 success rate.

Performance was evaluated using an independent dataset (GL, 4, and PG, 10), which reduced redundancy lower than 50% sequence identity to the original non-redundant dataset by application of the PSSM and parameters (that is, discrimination threshold = 0.003, M = -14 and N = +18 obtained the highest success rate on a self-consistency test to the original nonredundant dataset). The proposed method can discriminate between these GLs and PGs with a sensitivity and specificity of 100% on an independent dataset (shown Table 3).

Yuan and Teasdale\(^\text{32}\) have reported an algorithm that discriminates eukaryotic Golgi retention type II membrane proteins from the no-retention type using hydrophobicity and a principal component analysis of amino acid propensities. Their algorithm also gave good results: 89.3% sensitivity in both a self-consistency and jack-knife tests. Our method differs from theirs in the following points: (i) our datasets were restricted to mammalian proteins because the Golgi membrane and its network vary widely among organisms;\(^\text{28}\) and (ii) the characteristics of amino acid propensities in GLs were identifiable because the PSSM is used in our method.

In this study, GLs were distinguished from PGs with sufficient accuracy using an algorithm that combines a hydropathy alignment based on the position that shows the highest average hydrophobicity, and the PSSM calculated from the ratio of position-specific amino acid propensities for GLs and PGs. The position-specific amino acid propensities around the TMR are an important parameter in the identification of GLs. The profiles are closely related to the characteristics of the TMRs of GLs, which are rather hydrophilic and short compared with those of PGs. This result was consistent with the characteristics of the Golgi membrane, which is rather hydrophilic and thin as compared with the post-ER organelles and the plasma membrane.\(^\text{28,29}\)

Table 3 shows the prediction results for GLs and PGs using the PSSM. The method can effectively distinguish GL from unknown protein sequences predicted as type II membrane proteins by prediction methods for transmembrane topology such as TMHMM,\(^\text{33}\) HMMTOP,\(^\text{34}\) SOSUI,\(^\text{35}\)

### Table 2. Prediction Accuracy of Discrimination between GLs and PGs in a Self-Consistency and 5-Fold Cross-Validation Tests

<table>
<thead>
<tr>
<th>Test method</th>
<th>GL</th>
<th>PG</th>
<th>GL</th>
<th>PG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity (%)</td>
<td>96.2</td>
<td>93.5</td>
<td>0.949</td>
<td>93.5</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>93.5</td>
<td>85.5</td>
<td>0.867</td>
<td>85.3</td>
</tr>
<tr>
<td>Success rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 3. Prediction Results of Discrimination among Individual Datasets of GLs and PGs

<table>
<thead>
<tr>
<th>Swiss-Prot ID (accession no.)</th>
<th>Annotation (GL/PG)</th>
<th>Sequence identity* (%)</th>
<th>Discrimination score</th>
<th>Prediction result</th>
</tr>
</thead>
<tbody>
<tr>
<td>GL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNT2A_HUMAN (Q8N0V5)</td>
<td>GL</td>
<td>43.5</td>
<td>0.129</td>
<td>GL</td>
</tr>
<tr>
<td>GNT2B_HUMAN (Q8WFS9)</td>
<td>GL</td>
<td>46.4</td>
<td>0.258</td>
<td>GL</td>
</tr>
<tr>
<td>XYLK_HUMAN (O75063)</td>
<td>GL</td>
<td>15.0</td>
<td>0.201</td>
<td>GL</td>
</tr>
<tr>
<td>XYLK_MOUSE (Q8VCS3)</td>
<td>GL</td>
<td>14.6</td>
<td>0.201</td>
<td>GL</td>
</tr>
<tr>
<td>EDEM1_HUMAN (Q92611)</td>
<td>PG</td>
<td>23.7</td>
<td>-0.030</td>
<td>PG</td>
</tr>
<tr>
<td>MGA_HUMAN (O43451)</td>
<td>PG</td>
<td>6.8</td>
<td>-0.050</td>
<td>PG</td>
</tr>
<tr>
<td>NALDL_MOUSE (Q7M758)</td>
<td>PG</td>
<td>37.1</td>
<td>-0.155</td>
<td>PG</td>
</tr>
<tr>
<td>NALDL_RAT (O54697)</td>
<td>PG</td>
<td>37.1</td>
<td>-0.215</td>
<td>PG</td>
</tr>
<tr>
<td>P4HTM_HUMAN (Q9NXG6)</td>
<td>PG</td>
<td>10.7</td>
<td>-0.013</td>
<td>PG</td>
</tr>
<tr>
<td>P4HTM_MOUSE (Q6BG58)</td>
<td>PG</td>
<td>11.1</td>
<td>-0.013</td>
<td>PG</td>
</tr>
<tr>
<td>SUS1_HUMAN (P14410)</td>
<td>PG</td>
<td>5.4</td>
<td>-0.343</td>
<td>PG</td>
</tr>
<tr>
<td>SUS1_RABIT (P7768)</td>
<td>PG</td>
<td>6.6</td>
<td>-0.252</td>
<td>PG</td>
</tr>
<tr>
<td>SUS1_RAT (P21739)</td>
<td>PG</td>
<td>1.0</td>
<td>-0.232</td>
<td>PG</td>
</tr>
<tr>
<td>SUS1_SUNMU (O62653)</td>
<td>PG</td>
<td>3.3</td>
<td>-0.254</td>
<td>PG</td>
</tr>
</tbody>
</table>

* The maximum sequence identity to an original nonredundant dataset was calculated using Align in the FASTA package (http://www.ebi.ac.uk/Tools/emboss/align/).
and ConPred II. This study is expected to contribute to further developments in the prediction of membrane protein localization. Furthermore, our method is expected to pave the way for comprehensive identification of GL genes not only in mammalian genomes, but also in other genomes by generating the pSSM for each phylum/class.

Acknowledgments

We are grateful to Professor Hisashi Narimatsu (AIST), Professor Kiyoshi Asai (University of Tokyo), and Professor Yutaka Akiyama (Tokyo Institute of Technology) for helpful discussion. This research was financially supported in part by the program Research Promoting Technological Seeds of the Japan Science and Technology Agency (JST).

References

23) Yuan Z and Teasdale RD, Bioinformatics, 18, 1109–1115 (2002).
41) Yuan Z and Teasdale RD, Bioinformatics, 18, 1109–1115 (2002).