Discrimination of Mammalian GPI-Anchored Proteins by Hydropathy and Amino Acid Propensities

Yuri Mukai,1,2,† Masami Ikeda,3 Hirotaka Tanaka,2 Tatsuya Konishi,2 Osamu Oura,1 and Takanori Sasaki1,2

1Department of Electronics and Bioinformatics, School of Science and Technology, Meiji University, 1-1-1 Higashi-mita, Tama-ku, Kawasaki, Kanagawa 214-8571, Japan
2Department of Electronics, Graduate School of Science and Technology, Meiji University, 1-1-1 Higashi-mita, Tama-ku, Kawasaki, Kanagawa 214-8571, Japan
3Department of Chemistry and Biological Science, School of Science and Technology, Aoyama Gakuin University, 5-10-1 Fuchinobe, Sagamihara, Kanagawa 252-5258, Japan

Received October 9, 2012; Accepted December 3, 2012; Online Publication, March 7, 2013 [doi:10.1271/bbb.120781]

The glycosylphosphatidylinositol (GPI) attachment is a most important post-translational modification of proteins that plays essential post-translational roles in promoting the biochemical activities of eukaryotic cells. Described here is an analysis of the amino acid properties of mammalian GPI-anchored proteins (GPI-APs) and the development of an innovative method of detecting them. GPI-APs are characterized by two high-hydropathy regions: the signal peptide, located inside the Endoplasmic Reticulum (ER), and the GPI attachment signal, a sequence adjacent to the GPI-anchoring site (the ω-site). Especially in sequence analysis of known GPI-APs, there were some distinct aspects of the amino acid propensities around the ω-sites. Therefore, a method of detecting GPI-APs was developed based on hydropathy profiles and a position-specific scoring matrix (PSSM) calculated by position-specific amino acid propensities. First, sequences of GPI-APs and negative controls, determined by screening based on hydropathy and residue volume profiles, were aligned based on residue volume profiles in the C-terminal region, and the position-specific amino acid propensities of each group were calculated according to their alignment positions. Then, a PSSM was devised using the amino acid propensities of GPI-APs and negative controls, and discrimination scores were estimated for each dataset. Based on these scores at a threshold was fixed for each dataset. GPI-APs were detected with 81.1% sensitivity and a 0.818 success rate in an optimized calculation region determined by adjusting the window size of this region using a 5-fold dataset. The results indicate that a PSSM around the ω-site can effectively discriminate GPI-APs.

Key words: computational discrimination; glycosylphosphatidylinositol (GPI)-anchored protein; hydropathy screening; position-specific scoring matrix; post-translational modification

Post-translational modifications strictly control protein functions and subcellular locations. Among them, glycosylphosphatidylinositol (GPI) is an important biomolecule related to post-translational modification that binds soluble proteins and translocates them to the surface of the plasma membrane. GPI-APs act as core molecules in Eukaryotic cells and are responsible for a diversity of functions, including immune recognition, complement regulation, transmembrane signaling, cell adhesion, and embryogenesis. Several GPI-APs are known to show correlations with serious human disorders, including bovine spongiform encephalopathy (BSE), Creutzfeld-Jacob disease (CJD), paroxysmal nocturnal hemoglobinuria (PNH), and deep vein thrombosis (DVT). Additionally, GPI-APs in Plasmodium falciparum (the human malaria parasite) are vaccine candidates. Thus the identification and functional analysis of GPI-APs is believed to be crucial to an understanding of the vital activity of Eukaryote cells, resolution of the molecular mechanisms of incurable human disorders, and the medical treatment of those disorders. Bioinformatics approaches are anticipated to be powerful tools to discover novel genes comprehensively from an immense amount of genome information. Therefore, the development of computational methods to predict GPI-APs with high accuracy from genome/protein sequences is of the utmost importance.

GPI-AP translocation to the plasma membrane occurs in the following steps. First, the pre-mature form GPI-AP is localized in the ER lumen by the hydrophobic signal peptide at the N-terminus immediately after the completion of biosynthesis. Then, the GPI attachment signal, located at the C-terminus of the pre-maturely formed proteins, also has high hydropathy and interacts with the ER membrane and the GPI modification enzyme, transamidase. In the mature form GPI-AP, the signal peptide and the GPI attachment signal are cleaved off from the protein functional domain, and GPI is modified at the ω-site. As can be seen by the
Computational Discrimination of GPI-Anchored Proteins

Materials and Methods

Dataset preparation. Datasets of mammalian GPI-APs (83 entries) were obtained from the UniProt Knowledgebase/Swiss-Prot protein sequence database release 54.0 (July 2007) by conducting a search with keywords “mammalia” in the OC lines and “GPI” in the FT LIPID lines. Entries that had “fragment” in the DE lines and “potential” or “probable” annotations in the FT LIPID lines were excluded from the dataset. Non-GPI-APs, used as negative control (48,974 entries), were extracted by retrieval with the keyword “mammalia” in the OC lines, and entries containing the keyword “GPI” in the FT LIPID lines and/or the KW lines were removed.

Hydropathy analysis and dataset screening. The average hydrophathy for the GPI-AP and non-GPI-AP sequences was estimated by the moving average method with a sliding window of various sizes. The Kyte-Doolittle (K-D) hydropathy index was used to calculate the hydropathy of the amino acid sequences. Accordingly, the average hydropathy (H_k) of a sequence was expressed as:

\[ H_k = \frac{1}{w} \sum_{k=0}^{w-1} H(k), \]

where H(k) is the K-D hydropathy index at sequence position k, and w is the sliding window size used to calculate the average value.

According to the hydropathy profiles of GPI-APs (see “Hydrophathy profiles and screening”), three thresholds for training data selection of discrimination were defined. Entries that passed all three thresholds enumerated below, were selected as training datasets: (i) the maximum value of the average hydropathy of window size 11 for the 40 residues from the N-terminus was more than 50, (ii) the maximum value of the average hydropathy by window size 11 for the 40 residues from the C-terminus was more than 2.02, and (iii) the residue positions that had the maximum average hydropathy were less than 13 residues from the C-terminus. Thus 372 entries of non-GPI-APs were selected.

Non-GPI-AP entries showed a more than 30% similarity in full-length sequence identity to the sequences with “GPI” annotations in the FT LIPID lines were removed, and representative sequences were extracted from groups that clustered with a 30% similarity in identity as analyzed in the global alignment mode of single-linkage clustering program CD-HIT. The gap opening and elongation penalties of the alignment were (0, -1) for insertions and (-1, 0) for deletions in the query sequence. However, the GPI-AP entries were removed completely only when the sequences were identical, since the number of available GPI-AP entries was insufficient. The numbers of entries in the GPI-AP and non-GPI-AP datasets were 75 and 130 sequences respectively.

Residue volume analysis and sequence alignment. The volume score for the dataset was estimated by the moving average method with a sliding window of varying sizes. The DAWD720101 index was used to calculate the amino acid residue volume. Accordingly, the volume score (V_i) of a sequence was expressed as:

\[ V_i = \frac{1}{w} \sum_{k=0}^{w-1} V(k), \]

where V(k) is the DAWD720101 index at sequence position k, and w is the sliding window size used to calculate the average value.

The position of the minimum volume score was determined for each sequence by the moving average method. In regard to the volume profiles of GPI-APs (see “Amino acid residue volume analysis and sequence alignment”), after plotting the distribution of the minimum volume scores by window size 3 for the 20-30 residues from the C-terminus, a threshold was defined at 3.83 for selection of training data for discrimination. Thus 88 entries of non-GPI-APs were selected. Seventy-five GPI-AP and 88 non-GPI-AP sequences (Table 1) were aligned at the points of minimum volume scores (standard points) for the 20-30 residues from the C-terminus.

Calculation of position-specific amino acid propensity and position-specific scoring matrix. The following equation was used to calculate the position-specific amino acid propensity (f_p) of each position:

\[ f_p = \frac{n_p}{\sum_{j=1}^{20} n_j} \]

where p represents the standard points determined from the position with the lowest volume score, and n_p is the frequency of occurrence of the aforementioned information, the signal peptide and the GPI attachment signal are thought to be essential factors in the discrimination of GPI-APs.

Several means have been used to predict the C-terminus GPI attachment signal, each of which has a differing method including the following systems: Big-PI is based on scores that take into account the amino acid composition around the omega-site, whereas CHAPS uses the Kohonen Self-Organizing Map, a machine learning methods. Another method, PredGPI, can predict GPI attachment signals in two steps, a Neural Network (NN) and a Hidden Markov Model (HMM). Compared to these systems, PredGPI is a GPI attachment signal predictor of higher accuracy than the other methods, based on a Support Vector Machine (SVM) and an HMM. While the systems described above can discriminate GPI-APs with satisfactory performance from proteins already known to have a signal peptide, an additional signal peptide prediction tool e.g., SignalP, PredSi, Phobius, and SOSUIsignal, is necessary prior to GPI-AP detection. A more accurate method of Cao et al. used a SVM trained by GPI sequence data of many species, including mammals, to predict the GPI attachment signal and a majority voting strategy to detect signal peptides. It was evaluated by Aspergillus fungi ORFs, but the discrimination accuracy for mammalian GPI ORFs was unclear.

The following describes a method for the discrimination of GPI-APs specialized in mammals that can predict GPI-APs with high accuracy without signal peptide prediction tools. In this study, discrimination of GPI-APs from non-GPI-APs was achieved using a combination of hydropathy screening based on the hydropathy of N-terminus signal peptides and the C-terminus GPI attachment signals of GPI-APs and a position-specific scoring matrix (PSSM) using information gathered on amino acid composition around the omega-sites. In the first screening, the sequences of non-GPI-APs were selected according to their hydropathy and residue volume profiles, which had close similarities to those of GPI-APs. Then the sequences were aligned at the point of the minimum residue volume score of the 30 amino acids closest to the C-terminus. The PSSM alignment positions at −12 to +12 was created by calculating the position-specific amino acid propensities of GPI-APs and non-GPI-APs. When applying the PSSM to each dataset, discrimination scores were estimated. Using these estimations, GPI-APs were detected with 81.1% sensitivity at a 0.818 success rate in an optimized calculation region determined by adjusting the window size of this region using a 5-fold dataset, whereas a blind test using the GPI-AP and non-GPI-AP sequences, not including training datasets, detected 85.7% sensitivity at a 0.802 success rate, thus succeeding in detecting GPI-APs by combining hydropathy, residue volume screenings, and a PSSM around the omega-site. Thus, this method, high-throughput detection of novel GPI-APs in mammalian genomes can be achieved efficiently and accurately.
Certain amino acid j at position p. To avoid setting the denominator at zero in the PSSM calculation, a “constant mode” for the pseudo-count was introduced, as follows:

\[
f_p = \frac{n_p + \epsilon}{\sum_j n_{pj} + \epsilon}
\]

where \( n_p \) is the pseudo-count (=1). The position-specific score \( s_p \) was generally computed by dividing \( f_p \) by the background propensity, which was calculated from all the representative GPI-AP and non-GPI-AP sequences. The same background propensity was used for both GPI-APs and non-GPI-APs in the original calculation method employed in this study. Therefore, \( s_p \) was computed by the following equation:

\[
s_p = \ln \left( \frac{f_p^{\text{GPI-AP}}}{f_p^{\text{nonGPI-AP}}} \right)
\]

where \( f_p^{\text{GPI-AP}} \) expresses the amino acid propensity in the GPI-AP dataset and \( f_p^{\text{nonGPI-AP}} \) represents the propensity calculated for the non-GPI-AP dataset. The discrimination score (S) was estimated by taking the sum of the position-specific scores at the various alignment positions and averaging them with the number of added amino acids for positions \( M \) to \( N \), as shown in the following equation:

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

Evaluation of GPI-AP discrimination accuracy. The self-consistency and n-fold cross-validation test is often used for discriminant analysis of protein sequences to estimate prediction accuracy based on sensitivity, specificity, and success rate, as shown below:

\[
\text{Sensitivity} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \times 100
\]

\[
\text{Specificity} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \times 100
\]

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

The threshold of discrimination score (S) for discriminating GPI-APs from non-GPI-APs was fixed as follows: (i) If the frequency distributions of the discrimination scores for GPI-APs and non-GPI-APs showed no overlap, the threshold was the average of the score of the GPI-APs and the maximum score of the non-GPI-APs; (ii) if there was an overlap, the threshold was determined by scanning within the overlapping area to ensure, that the success rate (equation (9)) was maximized. If there were a number of candidates for the threshold, the average of the highest and lowest candidates was used in the discrimination. In the self-consistency test, the datasets (GPI-AP: 75 and non-GPI-AP: 88) were used for PSSM creation and accuracy estimation.

In adjusting for optimization of the window size of the calculation region using a 5-fold dataset, which is similar to the 5-fold cross-validation test, the average sensitivity, specificity, and success rate of an average of 1,000 random selections was calculated by randomly selecting four-fifths of the datasets to create the PSSM and using the remaining fifth to test the discrimination.

To evaluate discrimination performance objectively, GPI-AP and non-GPI-AP sequences that were not included in UniProt Knowledgebase/Swiss-Prot protein sequence database release 54.0 were extracted from release 2012.02 (Feb. 2012) by a procedure similar to that described previously (see “Dataset preparation” and “Hydropathy analysis and dataset screening”). Seven GPI-APs and 14 non-GPI-APs, which have lower than 30% similarity in full-length sequence identity to the previously mentioned training dataset, were used in objective evaluation as an independent dataset. GPI-AP entries with annotations about subcellular location, not only the cell membrane but other organelles, such as the Golgi apparatus, mitochondria, and nucleus, were removed.

### Results and Discussion

**Hydropathy profiles and screening**

Hydropathy profiles of full-length GPI-AP sequences were calculated for the purpose of finding high hydropathy regions. These calculations appear in Fig. 1, which shows the average hydropathy profiles for typical representatives of full-length GPI-APs. (a) UniProt ID: BY55_HUMAN/UniProt AC: O95979, and non-GPI-APs, (b) PMP22_HUMAN/Q01453 and (c) IL18_HUMAN/Q14116, calculated using a sliding window of 11 residues. The arrows marked “SP” and “GPI-AS” in GPI-AP show the N-terminus signal
peptide regions and the C-terminus GPI attachment signal regions annotated in UniProt Knowledgebase/Swiss-Prot, respectively. The arrows marked ‘TMR’ in nonGPI-APs show membrane spanning regions annotated in UniProt Knowledgebase/Swiss-Prot. In the GPI-APs, the most hydrophobic regions within 40 amino acid residues of the N-terminus and the C-terminus are included in the regions annotated as the signal peptide and the GPI attachment signal respectively. The signal peptide and the GPI attachment signal were found in the regions that included the most hydrophobic positions.

While the hydropathy profile of one of the soluble nonGPI-APs (IL18_HUMAN) differed greatly from GPI-AP, the profile of PMP22_HUMAN, a transmembrane protein that has four membrane spanning regions, is similar to GPI-AP in the N-terminus and C-terminus regions in that it has high hydropathy regions as transmembrane helices.

After this, two high hydropathy regions of GPI-APs were observed, and the distributions of values and positions with maximum average hydropathy of the GPI-APs were calculated, as shown in Fig. 2. In Fig. 2a, the distribution of the maximum average hydropathy (window size 11) of GPI-AP sequences within the first 40 amino acids from the N-terminus is displayed. The average hydropathy of this region, which constitutes the signal peptide, was notably high, with a maximum value of more than 1.50. Figure 2b shows the distribution of the maximum average hydropathy (window size 11) and its distance from the C-terminus in GPI-APs of the 40 amino acid located nearest the C-terminus. The plots of GPI-APs were focused near the C-terminus, within 13 amino acids, at a maximum average hydropathy of more than 2.02, whereas those of the nonGPI-APs extend across the entirety of the second and third quadrant in this graph. This is consistent with the fact that the hydropathy of the GPI attachment signal in GPI-APs was critically high and occured close to the C-terminus.

Therefore three thresholds for hydropathy screening were set to detect nonGPI-APs that have hydropathy profiles similar to GPI-APs: (i) a maximum average hydropathy value (window size 11) of more than 1.50 within 40 amino acids of the N-terminus; (ii) a maximum average hydropathy value (window size 11) of more than 2.02; and (iii) a defined position 13 amino acids from the C-terminus of the sequence region. These three thresholds in the hydropathy screening were exceeded by 372 nonGPI-AP entries. While the non-GPI-APs, which had no high hydropathy regions in the N-terminus or C-terminus (including IL18_HUMAN, see Fig. 1), failed the screening, the nonGPI-APs that had hydropathy profiles similar to GPI-APs (including PMP22_HUMAN, see Fig. 1) passed. After more than 30% of the redundancy was removed, 130 nonGPI-APs remained.

Amino acid residue volume analysis and sequence alignment

The residue volume profiles of GPI-AP sequences were calculated to find the point with the lowest volume score. Figure 3a shows the average of the volume scores...
window sizes, a window size of 3 was judged to be most appropriate, since three or four low-volume residues were located successively around the \(\omega\)-sites in almost all the GPI-APs. Based on the above, a threshold was defined at a volume score of 3.83 for the selection of training data for discrimination, resulting in the selection of 88 nonGPI-AP entries by residue volume screening.

Because non-GPI-AP data are not able to discriminate simply by hydropathy profile, a method was constructed to distinguish accurately nonGPI-APs from GPI-APs. Therefore, the datasets of GPI-AP and nonGPI-AP sequences (75 GPI-APs and 88 nonGPI-APs, Table 1) were aligned at points of minimum residue score for 20–30 residues from the C-terminus. In this study, gaps were omitted in the residue volume alignment in order to circumvent the complex problem of inserting them at the appropriate positions in the multi-aligned sequences and that of choosing an appropriate gap penalty.

**Position-specific scoring matrix around the \(\omega\)-site and discrimination score estimation**

PSSM has been applied effectively to various problems, including the identification of motifs and domains in nucleic acid and protein sequences. In the present study, a PSSM was created to identify the characteristics of individual amino acids and to discriminate GPI-APs from nonGPI-APs using those characteristics. The position-specific score for each amino acid at each position was calculated using 20 position-specific amino acid propensities, taking the pseudo-count into consideration using equation (4). In Table 2, a positive score is given when the frequency of GPI-APs exceeds that of nonGPI-APs, whereas a score with an absolute value near zero indicates little difference in propensity between GPI-APs and nonGPI-APs. Although it is known that a small amino acid residue is present at the \(\omega\)-site, shown at alignment position −1, the Cys (C), Asn (N), and Ser (S) propensities in GPI-APs were higher than those in the nonGPI-APs. On the contrary, the Ala (A) and Gly (G) propensities at alignment position −1 in nonGPI-APs were higher than those in the GPI-APs. In the case of the other amino acid residues at the \(\omega\)-site, the propensity of His (H) was significantly higher in GPI-APs than in nonGPI-APs. Excluding the \(\omega\)-site, the propensity of Asp (D), a negatively charged amino acid, was lower in GPI-APs than nonGPI-APs throughout the region. This difference was also witnessed in the propensities of Trp (W) and Tyr (Y), aromatic residues, which tended to be much lower in the GPI-APs than in the nonGPI-APs due to interaction with the ER membrane, as can be seen at the +10 amino acid alignment position, where the hydrophathy residues were consecutive. Trp propensity in GPI-APs was lower in the sequence region from −12 to −1 as well as in Pro (P), a peculiar amino acid with an imino group, but the propensity of Pro was much higher at +1, a position next the residue separated from the \(\omega\)-site by transamidase.

Thus the ability of the PSSM to reflect the difference in amino acid propensity at each alignment position shows its advantages distinctly. Based on the PSSM, the discrimination scores for the GPI-AP and nonGPI-AP datasets were calculated. The discrimination score distributions (Fig. 4) for the GPI-APs and nonGPI-APs

---

**Fig. 3.** Average of Volume Scores (window size 3) of 75 GPI-APs Aligned at the \(\omega\)-Site Using the DAWD720101 Index (a), Schematic Diagram of the Relation between Volume Score and Sequence (b), and Distribution of the Volume Score (window size 3) and Its Distance from the C-Terminus in GPI-APs (hollow circle) for the 40 Amino Acids Located Nearest the C-Terminus and nonGPI-APs (solid circle) for the 20 to 30 Amino Acids from C-Terminus (c).
Table 2. PSSM Calculated on the Basis of Residue Volume Alignment (window size 3) and Position-Specific Amino Acid Propensities for GPI-APs and nonGPI-APs

<table>
<thead>
<tr>
<th>Residue</th>
<th>Alignment position (amino acid)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-12</td>
</tr>
<tr>
<td>A</td>
<td>0.51</td>
</tr>
<tr>
<td>C</td>
<td>-4.17</td>
</tr>
<tr>
<td>D</td>
<td>-0.35</td>
</tr>
<tr>
<td>E</td>
<td>-0.93</td>
</tr>
<tr>
<td>F</td>
<td>0.16</td>
</tr>
<tr>
<td>G</td>
<td>-0.45</td>
</tr>
<tr>
<td>H</td>
<td>5.34</td>
</tr>
<tr>
<td>I</td>
<td>-5.26</td>
</tr>
<tr>
<td>K</td>
<td>0.32</td>
</tr>
<tr>
<td>L</td>
<td>-1.91</td>
</tr>
<tr>
<td>M</td>
<td>0.32</td>
</tr>
<tr>
<td>N</td>
<td>-5.26</td>
</tr>
<tr>
<td>P</td>
<td>-0.93</td>
</tr>
<tr>
<td>Q</td>
<td>-4.17</td>
</tr>
<tr>
<td>R</td>
<td>-5.26</td>
</tr>
<tr>
<td>S</td>
<td>1.26</td>
</tr>
<tr>
<td>T</td>
<td>0.16</td>
</tr>
<tr>
<td>V</td>
<td>-4.17</td>
</tr>
<tr>
<td>W</td>
<td>0.32</td>
</tr>
<tr>
<td>Y</td>
<td>5.34</td>
</tr>
</tbody>
</table>

indicate that the GPI-APs can be clearly distinguished from the nonGPI-APs. The threshold for discrimination was set based on the discrimination score distribution (see “Materials and Methods” above). In Fig. 4 with the threshold at -0.003, the entries, including false negatives, were SACA4_MOUSE (Q80ZQ0), XPP2_HUMAN (O43895), and CBPM_MOUSE (Q80V42), all including “by similarity” annotations in the FT LIPID lines, which means that their annotations were not given by experimental methods but by a similarity search. Among the false positives, CA054_HUMAN (Q8WWF1) and ASM3B_HUMAN (Q92485) are annotated as secreted proteins. Additionally, ASM3B_HUMAN is known to have isoforms. Both Q300_MOUSE (Q02722) and UPK2_BOVIN (Q08537), also false positives, are described as single-pass membrane proteins with “potential” annotations according to prediction, utilizing a prediction algorithm for trans-
membrane regions of membrane proteins. Experimental corroborations are not included in these entries at present, leaving the possibility that they belong to a different group that can be discovered by further experimental analysis. All entries that have GPI annotations in the FT LIPID lines by the experimental methods were predicted correctly by the method above described.

Optimization of the calculation region and accuracy evaluation of discrimination method

To optimize the calculation region around the ω-site, a 5-fold dataset was used, one similar to a general 5-fold cross-validation test. The sensitivity, specificity, and success rates for the discrimination of GPI-APs from nonGPI-APs were calculated using the original datasets (75 GPI-APs and 88 nonGPI-APs), and the lower and upper boundary positions used to calculate the discrimination score were set by a best-subset selection procedure from −12 to +12. Maximum accuracy was obtained by adding the scores of the nine residues located within a range of 0 to +8. In view of this result, it can be seen that the position-specific amino acid compositions are essential to discriminate GPI-APs from nonGPI-APs among the nine amino acids starting at the position of the amino acid with the minimum volume score, that is, the transamidase cutting position adjacent to the ω-site.

The precision of the method is displayed in Table 3, as evaluated in a self-consistency test and by adjusting for optimization of the window size of the calculation region using a 5-fold dataset. In this study, GPI-APs were distinguished from nonGPI-APs with high accuracy using an algorithm that combines hydropathy screening, residue volume alignment, and a PSSM calculated from the ratios of position-specific amino acid propensities for GPI-APs and nonGPI-APs.

To compare accuracy with other prediction methods, a blind test was also performed using independent datasets (GPI-AP: 7 and nonGPI-AP: 14) that were not included in UniProt Knowledgebase/Swiss-Prot protein sequence database release 54.0. These datasets were extracted from release 2012_02, and they reduced the redundancy of sequence identity of the original datasets below 30% by applying the PSSM and specified parameters (Table 4). That is, discrimination threshold = −0.003, M = 0, and N = +8 obtained the highest success rate on a self-consistency test compared to the original dataset. TDFG1_MOUSE (P51865), a false negative, has "potential" single-pass membrane proteins. The proposed method discriminated GPI-APs with 85.7% sensitivity, 75.1% specificity, and a 0.802 success rate on a blind test.

Table 4 also shows the discrimination results for the independent dataset by big-PI (sensitivity: 28.6%/specificity: 16.8%/success rate: 0.219), FragAnchor (28.6%/22.2%/0.252), and PredGPI (57.1%/66.7%/19).
the only systems available as internet tools at the present time. In the case of big-PI and FragAnchor, false positives and negatives were noticed. Furthermore, all three systems, including PredGPI, had lower accuracy than the original blind test.\textsuperscript{34} The approach of big-PI is similar to the method proposed here, but after performing the blind test, a lack of sensitivity was found in big-PI. This might be a result of insufficient annotations of GPI-APs when big-PI was developed. Credibility increases when similar results are produced by two or more tools of differing methodologies for GPI-AP prediction; it is crucial that many methods are utilized to ensure high accuracy. Furthermore, because prediction accuracy rises as the number of training datasets increases, it is important to add the latest sequence data to the training datasets and to update the system periodically.

In fact, high-accuracy discrimination of GPI-APs, which was independent of signal peptide prediction and could not be achieved by previous methods,\textsuperscript{17–20,26} was accomplished. This method can effectively identify GPI-APs in undiscovered protein sequences. Present study is expected to contribute to genome-wide screening for GPI-APs, not only in mammalian genomes but also in all other eukaryotic genomes. Comprehensive identification of GPI-APs is expected to yield contributions to the treatment of serious human diseases and to reveal the mechanisms of biochemical activities.

Acknowledgments

We are grateful to Dr. Adam Seth Magdalena for comments and proofreading. This work was financially supported in part by the Research for Promoting Technological Seeds from the Japan Science and Technology Agency (JST) and by a Grant-in-Aid for Young Scientists, B (JSPS).

References