

Insight of the Signal Motif of GPI-(like)-anchored Proteins by using SVM

Wei Cao

Department of Biotechnology
The University of Tokyo
No.1-1-1 Yayoi, Bunkyo-ku,
Tokyo, Japan 113-8657

Tohru Terada

Programme of Agricultural Bioinformatics
The University of Tokyo
No.1-1-1 Yayoi, Bunkyo-ku,
Tokyo, Japan 113-8657

Kentaro Shimizu

Department of Biotechnology
The University of Tokyo
No.1-1-1 Yayoi, Bunkyo-ku,
Tokyo, Japan 113-8657

Kazuya Sumikoshi

Department of Biotechnology
The University of Tokyo
No.1-1-1 Yayoi, Bunkyo-ku,
Tokyo, Japan 113-8657

Shugo Nakamura

Department of Biotechnology
The University of Tokyo
No.1-1-1 Yayoi, Bunkyo-ku,
Tokyo, Japan 113-8657

Abstract - Many proteins contain a signal sequence at their COOH-terminus recognized by glycosylphosphatidylinositol (GPI) anchor and attached on the membrane. Experimental result suggests that the overall hydrophobicity of COOH-terminus is more important than precise sequence. We use a machine learning technique, support vector machine, to examine requirement for identifying this signal sequence measured at hydrophobicity scale by computational methodology. Effects of hydrophobicity of the signal sequence at multilevel and different segments of the signal motif proposed by previous works on their identification were investigated. For precisely identifying the signal sequence, we found that 15 residues (a hydrophobic domain) proximity to the COOH-terminus of the signal sequence is necessary. This result is consistent with observations from experimental methodology. Moreover, aside from 15 residues, the segment of 40 residues on COOH-terminus also contributes to its precise identification.

Keywords: GPI lipid modification, PTM, SVM.

1 Introduction

Glycosylphosphatidylinositol (GPI) lipid modification spotlighted as an important means for protein post-translational modification has been widely studied since the existence of GPI anchor was accepted in the mid 1980s [1]. Many proteins are anchored to the membrane via the GPI anchor or a similar functional anchor termed as

Glycosylsphingolipidinositol (GSI) anchor that is also called GPI-like-anchor. In GPI lipid modification, the COOH-terminal signal sequence of precursor proteins is cleaved followed by the addition of GPI moiety (the new COOH-terminus known as w -site). Distribution of GPI-(like)-anchored proteins among organisms is so wide ranging from bacteria to human beings and their functions also are diverse, such as Alkaline phosphatase[2]-hydrolytic enzymes, Neural cell adhesion molecule[3]-adhesion proteins, Folate receptor[4]-receptor, Scrapie prion protein[5], and Thy-1-antigens[6]. Previous experimental studies[7-9] with intact cells and cell-free systems have shown that the COOH-terminal signal sequence of GPI-anchored proteins has a feature of uncharged, predominant hydrophobicity and certain minimal length. Moreover, distinctive characteristics of GPI-anchored proteins are the location of a typical w -site is not far from the COOH-terminus aside from few exception reported[8] and only certain residues seem to be allowed at w , $w+1$ and $w+2$ positions[10]. Further, experimental result suggests that the overall hydrophobicity of COOH-terminus is more important rather than precise sequence [11]. Experimental identification[12] of GPI-(like)-anchored proteins is mainly accomplished by approaches: specific enzymatic, chemical cleavage in combination with detergent, antibody recognition and metabolic radioactive labeling. As currently practiced, all sorts of experimental techniques limit these approaches.

With the number of the identified GPI-(like)-anchored proteins uninterruptedly increasing in the existing protein sequence database, identification of the signal sequence has been paid much more attentions in the field of computational biology. Eisenhaber et al.[13] proposed a revised model of the GPI-modification motif on statistical analysis of known proteins. The signal sequence is composed of four components: unstructured linker region ($w-11\sim w-1$), the site of cleavage/GPI modification ($w-1\sim w+2$), a spacer region ($w+3\sim w+9$) and a hydrophobic tail ($w+9$ or $w+10\sim$ COOH-terminus). Recent years, machine learning techniques have been effective tools for solving tasks in many aspects of computational biology so far. These techniques could show a good balance between false positive and false negative errors for screening of unknown samples. Fankhauser et al.[14] presented an approach by using self-organizing map (SOM), an unsupervised learning method, for identification of the GPI-anchored signal sequence. For representing protein sequences in a numerical format for inputting into SOM, they evaluated several numerical formats generated by using hydrophobicity scale (Kyte-Doolittle hydrophobicity scale[15]), zentriole, virtual potential independently or combination of them. As they merely used hydrophobicity scale, prediction accuracy of $\sim 83\%$ is obtained. In their work, a protein sequence was represented by a vector of hydrophobic values of corresponding residues derived from 32 positions in protein COOH-terminus.

We introduce a strategy of identifying GPI-(like)-anchored signal sequence by using support vector machine and a single protein descriptor derived from Kyte-Doolittle hydrophobicity scale, and achieve the prediction accuracy of 95.5%. The single protein descriptor based on hydrophobicity scale makes it easy to examine requirement for identifying this signal sequence measured at hydrophobicity scale by computational methodology. Effects of hydrophobicity of the signal sequence at multilevel scales and different segments of the signal motif proposed by previous works[13] on their identification were investigated. For precisely identifying the signal sequence, we found that 15 residues (a hydrophobic domain) proximity to the COOH-terminus of the signal sequence is necessary and indispensable. This result is consistent with observations from experimental methodology. In addition, the segment of 40 residues on the COOH-terminus also contributes to its precise identification.

2 Results

2.1 Prediction accuracy and sliding window size

To evaluate the performance of our trained classifier, prediction accuracy and the area under the Receiver Operating Characteristic (ROC) curve so called AUC value

under 5-fold cross-validation (CV) test were adopted. A dataset consisting of 520 positive and 429 negative entries whose lengths of protein sequences are larger than or equal to 100 residues was employed. Last 60 residues at COOH-terminus for each protein sequence were selected as the length of input sequence for generating the protein descriptor (see also methods), i.e. a feature vector for representing the protein sequence and training a SVM classifier. Prediction accuracies and AUC values are summarized and shown in Table 1. Observing the Table 1, prediction accuracies under 5-fold CV test are all over 90% except for the situation as window size was set to one residue. Furthermore, AUC values of corresponding accuracy of above 90% are also close to the ideal value (ideal value is 1.0). The closer AUC gets to 1.00, the better classification rule is while AUC of less than or equal to 0.5 indicates that the classification rule is no better than random selection in respect to a two-class classifier. A series of sliding window size (of an odd number) ranging from 1 to 21 was employed to investigate effect of sliding window size on prediction accuracy and AUC. Table 1 shows that prediction accuracy and the highest AUC value of the first rank corresponds to sliding window size of 9 residues. Besides it, if we observe Table 1 vertically, another remarkable result is shown that a fluctuation of prediction accuracy corresponding to window size of 1 and 3 respectively is more distinguished than that of any subsequently consecutive two pairs of window size.

Table 1 Window size vs. Prediction accuracy

Window size	Accuracy	AUC
1	70.60	0.7663
3	94.63	0.9570
5	95.36	0.9606
7	95.47	0.9568
9	95.50	0.9631
11	94.94	0.9563
13	94.73	0.9498
15	94.63	0.9561
17	94.63	0.9568
19	94.63	0.9533
21	94.42	0.9579

2.2 Effect of different regions

In order to investigate effects of hydrophobicity of different segments of the signal motif proposed by previous works[13] on their identification, we conducted elongation

and deletion simulation (see details in Materials and Methods). In the former simulation, tendency of the accuracy along with addition of residues uninterruptedly appears to go up but decreases again after total amount of residues is over 64 residues as input. When the length of input sequence reaches to the range of 60-64 residues, the SVM classifier shows the best performance as shown in Figure 1. In the elongation simulation, it is observed that prediction accuracy of about 3.5% (92.0~95.5%) is increased. In the latter simulation, accuracy of 95.5% is obtained at the initial step, prediction accuracy under 5-fold CV test decreases sharply along with removal of residues from the end of the COOH-terminus in order and ~20% reduced when 15 residues were removed as can be seen in Figure 2. Not only experimental results[16] that GPI-anchored protein can be converted to a secreted protein after deletion of the hydrophobic domain (15-20 residues) but also the result obtained here suggests the hydrophobic tail for GPI attachment is very important. As another demonstration, mean hydrophobic values of positions of 100 consecutive residues counted starting from COOH-terminus derived from 520 positive and 429 negative entries were calculated. As shown in Figure 3, mean hydrophobic values of each position occupied by the last 15-20 residues for positive entries are apparently higher than that of corresponding position for negative entries. In addition, hydrophobicity of this region is also higher than other parts inside 100-residue at COOH-terminus with respect to positive dataset.

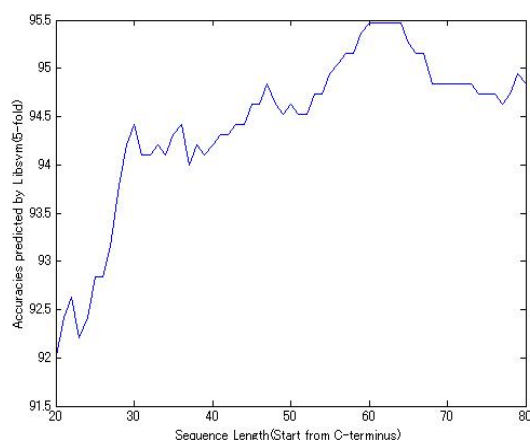


Figure 1 Elongation simulation of GPI-(like)-anchoring signals

3 Discussion

3.1 Hydrophobicity of GPI-(like)-anchoring signals

Experimental observations suggest us that the overall hydrophobicity of the signal sequence is more important than precise sequence. From the viewpoint of prediction

accuracy, window size of larger than one residue is necessary as using Kyte-Doolittle hydrophobicity scale to depict hydrophobic property of the signal sequence.

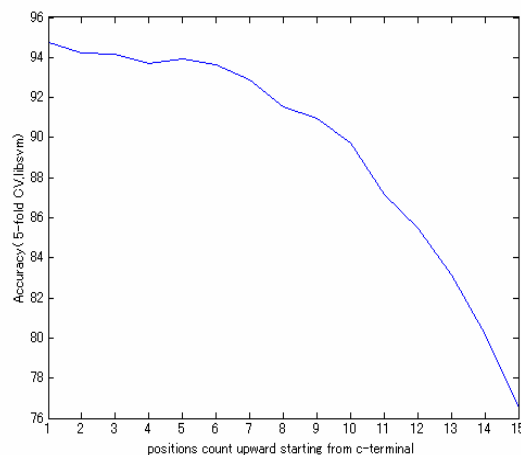


Figure 2 Deletion simulation of GPI-(like)-anchoring signals

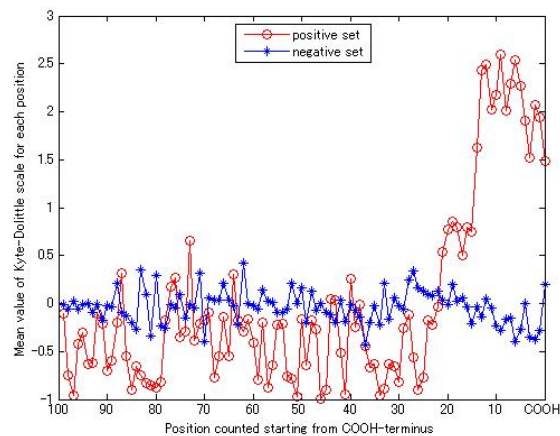


Figure 3 Plot of mean hydrophobic values of positions occupied by 100 COOH-terminal residues for the positive dataset (open red circles) and the negative dataset (blue stars).

It is supported by observations from effects of different sliding window sizes on prediction accuracy. i.e. prediction accuracy of 94.6% corresponding to window size of 3 residues is obtained while as the window size is set to one residue, we could only achieve prediction accuracy of 70.6% (see also Table 1). Further, Table 1 also shows that fluctuation of prediction accuracy is very small after applying the window size of larger than one residue. As a result, depicting the characteristic of GPI-(like)-anchored proteins by the protein sequence descriptor derived from the Kyte-Doolittle hydrophobic scale is one of key factors in our case. We have shown the impact of the sliding window size on prediction accuracy using the

method of hydrophobicity plot in Table 1. The result shows that the window size of 9 residues is the best choice for our task. Relative hydrophobic values of residues measured by the Kyte-Doolittle scale used for transformation of GPI-(like)-anchored proteins (i.e. the sliding window size=1) was also reported in the paper of Fankhauser et al[14]. However, prediction accuracy is about 83%. We still must notice the discrepancy between the existing work of Fankhauser et al. and ours that may result from selection of the length, position of input sequences and training method. The merit of single protein sequence descriptor generated by only using the Kyte-Doolittle scale presented in this work, which is based on physical-chemical characteristic of amino acids, is that it is simpler and more competitive than that generated by combination of different scales for its ease understanding.

3.2 Optimal length for GPI attachment

In respect of effects of different segments on prediction accuracy, the results from deletion simulation and observation of statistical mean value for each position emphasize its importance and indispensability and are in accord with experimental observation[16]. Moreover, prediction accuracy of ~3.5% improved as addition of the segment of 40 residues on COOH-terminus in the elongation simulation shows that hydrophobicity of that segment strengthens identification of the signal sequence as an effective complement. Regarding to optimal length of the signal sequence, an unexpected discrepancy occurs that the optimal length deduced from experimental observations and our work could not be consistent with each other. Experimental results of fusion proteins[17] suggest that the signal sequence (29-37 residues located at COOH terminus) for optimal GPI attachment includes two elements at the minimum, a hydrophobic domain(15-20 residues) and a pair of small residues (positioned 10-12 residues preceding the hydrophobic domain). However, the length of 29-37 residues suggested by experimental results is much less than optimal length of 60-64 residues deduced from the elongation simulation in this work. Unfortunately, little information obtained from experimental methodology is available.

In summary, to examine requirement for identifying GPI-(like)-anchored signal sequence measured at hydrophobicity scale, we first found that protein sequence descriptor generated by the sliding window algorithm is useful for the identification of the signal sequence. Second, a hydrophobic tail of GPI-(like)-anchored signal sequence is very significant for the identification in accordance with results of experimental observations. Third, the segment of 40 residues counted starting from twentieth residue of COOH-terminus unconcerned antecedently can strengthen its identification. Further works are still needed to be carried out to investigate the region encapsulated by these 40 residues of GPI-(like)-anchoring signals in more detail,

especially using experimental methodology and effect of NH₂-terminal signal of GPI-(like)-anchored proteins on identification of GPI anchoring signal sequence.

4 Materials and Methods

4.1 Materials

4.1.1 Positive dataset

We collected 587 entries from the database of UniProtKB/Swiss-Prot. Those are labeled by 'GPI-ANCHOR' or 'GPI-LIKE-ANCHOR' in the field of KW by searching the database online using a keyword "GPI ANCHOR" (Note that the number of GPI-anchor entries could be increasing). After getting rid of 56 entries, 531 entries annotated with definite comments of position of lipid modification (*w*-site) in feature table (FT) field were left and finally taken as the positive dataset (i.e. GPI-(like)-anchored proteins). Among the dataset of 531 entries, lengths of sequences of 520 entries are ≥ 100 residues, the others are between 50 and 100 residues.

4.1.2 Negative dataset

A 441-entry dataset as the negative dataset selected from GenBank by text-based searching was supplied by Professor Pascal Mäser, Institute of Cell Biology, University of Bern, who used it as a test dataset in his published work[14]. There are 429 entries of sequence length ≥ 100 residues among the dataset. Components of the 429-entry negative dataset are composed of four resources from the raw dataset, 105 of 107 cytosolic proteins, 63 of 68 secreted proteins, 104 of 107 N-TM-C (transmembrane proteins with an NH₂-terminal export signal predicted by SignalP as well as a hydrophobic COOH-terminus), and 157 of 159 transmembrane proteins. Homology similarity of any two protein sequences in the negative set was less than 50% reduced by the employment of Smith/Waterman algorithm. Numbers of positive and negative samples are in proportions about 1:1.

4.2 Methods

4.2.1 Support vector machine

As a supervised learning algorithm, SVM introduced by Vapnik and his coworkers[18, 19] is known for its outstanding performance and successfully applied in many issues of computational biology so far. In present work, we use 1-norm soft margin SVM. With respect to a dichotomic classification problem, the basic idea behind SVM is to map feature vectors by which each sample in a training dataset is represented into a high dimensional feature space and then construct an optimal separating plane so called hyperplane in this space. Subsequently, a boundary of the margin between positive and negative samples is

maximized for giving good generalization properties. The decision boundary is used for classification of unknown samples. In order to overcome the dimension disaster in computation caused by mapping, kernel functions are proposed for implicit mapping of input data. In our work, radial basic function (RBF) is taken as the kernel function for implicitly mapping input vectors into the high dimensional feature space. A regularization parameter C of SVM and a parameter γ of RBF kernel function, $\exp(-\gamma \|x_i - x_j\|^2)$, were respectively set to 1 and $1/n$, where n is the number of elements of a feature vector (i.e. a protein sequence descriptor in present work). C and γ were not optimized for tasks in present work. To implement the algorithm of 1-norm soft margin SVM, a Matlab interface v2.81 of the software package named Libsvm[20] was employed in present work. K -fold cross validation test, where K was fixed to 5, was used for evaluating performances of SVM classifiers.

4.2.2 Protein sequence descriptor

A numerical sequence (called protein sequence descriptor here) readable for support vector machine was generated by using a sliding window-based method called hydrophobicity plot for transformation of a protein primary structure. Hydrophobicity plot is that a window of a given size slides along the protein sequence from NH_2 -terminus to COOH-terminus (one residue at a time in present work) and mean value within the window is placed in the numerical sequence at each time. For example, 60 residues taken from COOH-terminus of a protein sequence and window size of 9 residues adopted, the protein sequence descriptor for representing this protein sequence generated by using hydrophobicity plot consists of 52 elements, i.e. a 52-Dimensional vector. In present work, a widely applied scale called Kyte-Doolittle scale used for detecting hydrophobic regions in proteins was adopted for delineating hydrophobic character of 20 standard amino acids. Hydrophobic regions will be represented by a positive value. Sliding window sizes of 5-7 residues will work well for identifying surface-exposed regions whereas window sizes of 19-21 is well suited for finding transmembrane domains. Applying different window sizes make us explore hydrophobicity of protein sequence at multilevel scales.

4.2.3 Elongation and deletion simulations

A systematic study that the effect of varied lengths of the COOH-terminal signal sequence with shortening or lengthening on correct processing of the GPI-anchored protein was carried out by Berger. Here we performed a similar consideration *in silico* to investigate different segments of the signal motif proposed by previous works on their identification. Two datasets are composed of the whole entries and 949 entries (520 positive and 429 negative entries, ≥ 100 residues for each protein) selected

for deletion and elongation simulations (see also Materials), respectively. Window Size was set to 9 residues as transforming protein sequences by using the method of hydrophobicity plot described above into readable data for SVM and prediction accuracy was calculated under 5-fold CV test. Deletion simulation of COOH-terminal residues was carried out through the following steps: 1) 50 residues counted starting from the end of COOH-terminus for each entry to construct the input data for SVM; 2) one residue was deleted from the end of COOH-terminus at each time; 3) prediction accuracy of the trained SVM classifier was obtained (parameters were set by default not optimized, see also the above section); 4) repeat 2-3 steps and terminated when 15 residues from the COOH-terminus for each entry were removed. On the contrary, in elongation simulation the initial dataset was established with 20 residues counted starting from each COOH-terminus of the entire 949 protein sequences. Then prediction accuracy was calculated by adding one residue on the anterior end of each input sequence used in last calculation at each time. The calculation was terminated when the number of residues counted starting from COOH-terminus for each protein sequence reached to 80.

5 References

- [1] R. C. Ajit Varki, Jeffrey Esko, Hudson Freeze, Jamey Marth, *Essentials of Glycobiology*, 1st ed: Cold Spring Harbor Laboratory Pr 1999.
- [2] H. Ikezawa, M. Yamanegi, R. Taguchi, T. Miyashita, and T. Ohyabu, "Studies on phosphatidylinositol phosphodiesterase (phospholipase C type) of *Bacillus cereus*. I. purification, properties and phosphatase-releasing activity," *Biochim Biophys Acta*, vol. 450, pp. 154-64, 1976.
- [3] S. K. Powell, B. A. Cunningham, G. M. Edelman, and E. Rodriguez-Boulant, "Targeting of transmembrane and GPI-anchored forms of N-CAM to opposite domains of a polarized epithelial cell," *Nature*, vol. 353, pp. 76-7, 1991.
- [4] S. Rijnboutt, G. Jansen, G. Posthuma, J. B. Hynes, J. H. Schornagel, and G. J. Strous, "Endocytosis of GPI-linked membrane folate receptor- α ," *J Cell Biol*, vol. 132, pp. 35-47, 1996.
- [5] N. Stahl, M. A. Baldwin, R. Hecker, K. M. Pan, A. L. Burlingame, and S. B. Prusiner, "Glycosylated phospholipid anchors of the scrapie and cellular prion proteins contain sialic acid," *Biochemistry*, vol. 31, pp. 5043-53, 1992.
- [6] B. J. Dowsing, A. A. Gooley, P. Gunning, A. Cunningham, and P. L. Jeffrey, "Molecular cloning and

- primary structure of the avian Thy-1 glycoprotein," *Brain Res Mol Brain Res*, vol. 14, pp. 250-60, 1992.
- [7] J. Berger, A. D. Howard, L. Brink, L. Gerber, J. Hauber, B. R. Cullen, and S. Udenfriend, "COOH-terminal requirements for the correct processing of a phosphatidylinositol-glycan anchored membrane protein," *J Biol Chem*, vol. 263, pp. 10016-21, 1988.
- [8] I. W. Caras, "An Internally Positioned Signal Can Direct Attachment of a Glycophospholipid Membrane Anchor," *Journal of Cell Biology*, vol. 113, pp. 77-85, 1991.
- [9] S. Udenfriend and K. Kodukula, "How glycosylphosphatidylinositol-anchored membrane proteins are made," *Annu Rev Biochem*, vol. 64, pp. 563-91, 1995.
- [10] K. Kodukula, L. D. Gerber, R. Amthauer, L. Brink, and S. Udenfriend, "Biosynthesis of glycosylphosphatidylinositol (GPI)-anchored membrane proteins in intact cells: specific amino acid requirements adjacent to the site of cleavage and GPI attachment," *J Cell Biol*, vol. 120, pp. 657-64, 1993.
- [11] I. W. Caras and G. N. Weddell, "Signal peptide for protein secretion directing glycopospholipid membrane anchor attachment," *Science*, vol. 243, pp. 1196-8, 1989.
- [12] M. G. Low, "Biochemistry of the glycosylphosphatidylinositol membrane protein anchors," *Biochem J*, vol. 244, pp. 1-13, 1987.
- [13] B. Eisenhaber, P. Bork, and F. Eisenhaber, "Sequence properties of GPI-anchored proteins near the omega-site: constraints for the polypeptide binding site of the putative transamidase," *Protein Engineering*, vol. 11, pp. 1155-1161, 1998.
- [14] N. Fankhauser and P. Maser, "Identification of GPI anchor attachment signals by a Kohonen self-organizing map," *Bioinformatics*, vol. 21, pp. 1846-1852, 2005.
- [15] J. Kyte and R. F. Doolittle, "A Simple Method for Displaying the Hydrophobic Character of a Protein," *Journal of Molecular Biology*, vol. 157, pp. 105-132, 1982.
- [16] I. W. Caras, G. N. Weddell, and S. R. Williams, "Analysis of the Signal for Attachment of a Glycophospholipid Membrane Anchor," *Journal of Cell Biology*, vol. 108, pp. 1387-1396, 1989.
- [17] P. Moran and I. W. Caras, "A nonfunctional sequence converted to a signal for glycoposphatidylinositol membrane anchor attachment," *J Cell Biol*, vol. 115, pp. 329-36, 1991.
- [18] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
- [19] V. N. Vapnik, *Statistical Learning Theory*: John Wiley & Sons, Inc, 1998.
- [20] C. C. Chang and C. J. Lin, "LIBSVM : a library for support vector machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>," 2001.