IPSJ Transactions on Bioinformatics Vol. 2 25–35 (Mar. 2009)

Original Paper

Support Vector Machine Prediction of N- and O-glycosylation Sites Using Whole Sequence Information and Subcellular Localization

Kenta Sasaki,[†1] Nobuyoshi Nagamine[†1] and Yasubumi Sakakibara[†1]

Background: Glycans, or sugar chains, are one of the three types of chain (DNA, protein and glycan) that constitute living organisms; they are often called "the third chain of the living organism". About half of all proteins are estimated to be glycosylated based on the SWISS-PROT database. Glycosylation is one of the most important post-translational modifications, affecting many critical functions of proteins, including cellular communication, and their tertiary structure. In order to computationally predict N-glycosylation and O-glycosylation sites, we developed three kinds of support vector machine (SVM) model, which utilize local information, general protein information and/or subcellular localization in consideration of the binding specificity of glycosyltransferases and the characteristic subcellular localization of glycoproteins. **Results:** In our computational experiment, the model integrating three kinds of information achieved about 90% accuracy in predictions of both N-glycosylation and O-glycosylation sites. Moreover, our model was applied to a protein whose glycosylation sites had not been previously identified and we succeeded in showing that the glycosylation sites predicted by our model were structurally reasonable. **Conclusions:** In the present study, we developed a comprehensive and effective computational method that detects glycosylation sites. We conclude that our method is a comprehensive and effective computational prediction method that is applicable at a genome-wide level.

1. Introduction

Glycans, or sugar chains, are one of the three kinds of chain (DNA, protein and glycan) that constitute living organisms; they are often called "the third chain of the living organism". Within an organism, glycans mainly exist as glycolipid or glycoprotein. Efficient chemical synthesis of sugar chains has been well studied in combinatorial chemistry [1)–3)]. Recently, glycosyltransferases that catalyze the transfer of monosaccharides to specific residues in proteins have been well studied in biology and pathology [4)–6)]. In some glycoproteins, glycosylation or attachment of carbohydrate polymers to an amino acid residue has been studied in detail [7)–10)]. However there have been no general approaches that can comprehensively detect glycosylation sites and identify protein-bound glycan structures in living cells. Hence, though there exist several databases on glycans including KEGG GLYCAN [11)] and Glycan Database (http://www.functionalglycomics.org/glycomics/molecule/jsp/carbohydrate/carbMoleculeHome.jsp), there are currently no comprehensive and useful databases on glycosylation.

In the present study, we focused on glycosylation. Glycosylation is one of the most important post-translational modifications, affecting many critical functions of proteins, including cellular communication, and their tertiary structure [12)]. About half of all proteins are estimated to be glycosylated based on the Swiss-Prot database [13)]. There are four different types of glycosylation, namely, via N-glycosylation, O-glycosylation, C-mannosylation and glycophosphatidlyinositol (GPI) anchor attachments. In this study, we developed a method that predicts N-glycosylation, or glycosylation of Asn (N) residues, and O-glycosylation, or glycosylation of Ser (S) and Thr (T) residues, sites, in proteins.

Several computational approaches to predict O-glycosylation sites in proteins have been developed in recent years [14)–19)]. Statistical learning methods, such as artificial neural network (ANN) and support vector machine (SVM), have been widely utilized for this purpose. In these studies, each amino acid residue was represented by a feature vector in which only local information, or a window of fixed length surrounding the residue (**Fig. 1**), was considered. However, glycosyltransferases attach sugar chains to amino acid residues specifically by

Fig. 1 The sequence window used to encode local information of proteins. k upstream and downstream residues of the target residue (N in italic) were extracted (k=10, in this figure). To encode one residue in the sequence window, we utilized BLOSUM62 profile encoding (the corresponding row in the BLOSUM62 matrix).

†1 Department of Biosciences and Informatics, Keio University

recognizing the structure of the whole protein, rather than the individual residue only [20)–22)]. Thus, in predicting glycosylation sites, general protein information, or whole-sequence information should be considered. Moreover, the subcellular localization of glycoproteins is characteristic [15),23),24)]. For example, most membrane proteins have glycans outside the cell membranes and can be regarded as glycoproteins. Hence, we need to utilize not only local information, but also general information and subcellular localization, to predict glycosylation sites.

In this study, we constructed four kinds of SVM model to predict glycosylation sites. The *window* model was based on only local information. The *whole-sequence* and *localization* model utilized, in addition to local information, general information about the proteins and subcellular localization respectively. The *integral* model integrated local information, general protein information and subcellular localization. In our computational experiments, the *whole sequence*, *localization* and *integral* models showed better prediction performances than the *window* model. Moreover, we validated the effectiveness of our model by predicting glycosylation sites that were structurally reasonable in a protein whose glycosylation sites were unknown.

2. Results

2.1 Prediction Performance of the Proposed Method

Table 1 shows the prediction performances when our proposed four kinds of SVM model were applied to the N-glycosylation and O-glycosylation site datasets. Using only local information, the accuracy (described later in Methods) was 0.767 when the model was applied to the N-glycosylation site dataset and 0.784 when applied to the O-glycosylation site dataset. When utilizing all available information, the accuracy was 0.896 when the model was applied to the N-glycosylation site dataset and 0.897 when applied to the O-glycosylation site dataset. The prediction performances with several kernels were shown in Supplementary Material 1.

Table 1 Prediction performances of four different models.

model	Accuracy	Sensitivity	Precision	MCC	AUC
N-glycosylation sites					
Window (N1)	0.767	0.494	0.658	0.412	0.814
Window + di-pep (N2)	0.884	0.766	0.840	0.721	0.942
Window + subcellular (N3)	0.822	0.640	0.743	0.568	0.891
Window + di-pep + subcellular (N4)	**0.896**	**0.808**	**0.844**	**0.752**	**0.952**
O-glycosylation sites					
Window (O1)	0.784	0.534	0.708	0.473	0.831
Window + di-pep (O2)	0.893	0.779	0.868	0.748	0.949
Window + subcellular (O3)	0.813	0.639	0.732	0.553	0.866
Window + di-pep + subcellular (O4)	**0.897**	**0.790**	**0.870**	**0.756**	**0.952**

Window means local information is used for prediction. Similarly, di-pep means the use of general information and subcellular means that of subcellular localization.

The *whole-sequence* model (N2 and O2 in Table1), using local information and general information, showed significantly better prediction performances than the *window* model (N1 and O1 in Table1). However, the *localization* model, integrating local information and subcellular localization, (N3 and O3 in Table1) showed smaller improvement in performance than the *whole-sequence* model (N2 and O2 in Table1). These results can be elucidated by biological properties represented by both whole-sequence information and subcellular localization information. Subcellular localization is determined partly by sorting signals, such as the secretory signal peptide "Ser-Lys-Leu" [25)]. In fact, the frequency of a particular peptide is used to predict subcellular localization by WoLF PSORT [26)]. Thus, counting the frequency of di-peptides in a protein sequence, which is used to represent general information about proteins, partly corresponds to counting signal peptides and considering subcellular localization information.

2.2 Comparison of Feature Representation of Local Information with the Previous Studies

Several approaches to encode local information have been proposed [14)–17)]. We compared these approaches using several lengths of the sequence window (**Fig. 2**). As shown in Fig. 2, among the BLOSUM62 profile encoding, 0/1 encoding and physico-chemical property encoding, the BLOSUM62 profile encoding system, which was used in our method, was, except when using the window of length 4, better than the other two encodings in the N-glycosylation prediction. On the other hand, in the O-glycosylation prediction (Fig. 2), the 0/1 encoding system was better than the other two encodings except when using the window of length 10. However, the difference between the performances of the 0/1 encoding system and the BLOSUM62 profile encoding system was very small. As for the window

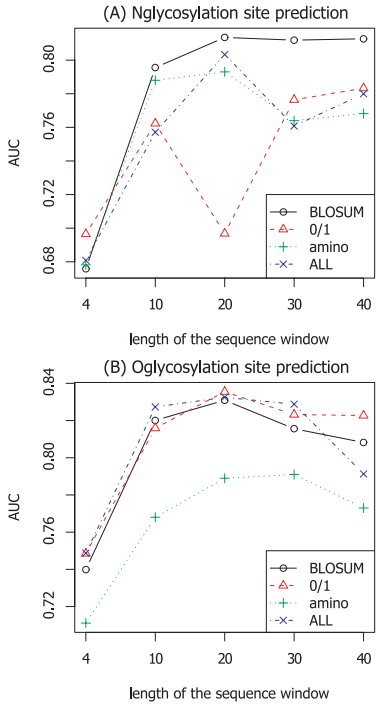

Fig. 2 Comparison of encoding systems. The transition of prediction performances in N-glycosylation sites (A) and O-glycosylation sites (B) were shown. The lengths of sequence window were 4, 10, 20, 30 and 40. Four encoding systems, BLOSUM62 profile encoding system, 0/1 encoding system, amino acid physico-chemical properties encoding system and integrated encoding system which was combined by three encoding systems, were applied.

length, the prediction performances almost generally peaked when using the sequence window of length 20. Thus we adopted the BLOSUM 62 profile encoding system, using the window of length 20. Here, we confirm the superiority of our feature representation method to those used in previous studies [14)–17)]. These studies considered only local information; hence their method performance was estimated to be nearly the same as "Window" in Table 1. Therefore, as our method, utilizing the whole sequence information and subcellular localization, improved the prediction accuracy of "Window" by more than 10 percent in both

Table 2 O-glycosylation site prediction in three protein sequences.

method	Sensitivity	Balanced accuracy
NetOGlyc [15)]	0.563	0.728
EnsembleGly [16)]	0.375	0.679
Our method	**1.000**	**0.766**

The BSP30, Kallikrein-1 and Ig delta chain C region have sixteen experimentally validated O-glycosylation sites. Previous methods (NetOGlyc and EnsembleGly) and our method, which used almost the same positive data to train the prediction model, were applied to these sites. Our method achieved 1.000 (16/16) sensitivity, while the previous methods showed 0.563 (9/16) and 0.375 (6/16) sensitivity respectively. Balanced accuracy was calculated as follows;

$$\text{Balanced accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right)$$

N-glycosylation and O-glycosylation site prediction (Table 1), our method is competitive with and sometimes surpass the previous methods.

2.3 Comparison of Prediction of Known O-glycosylation Sites

Our method and several previous methods [15),16)] were applied to the sixteen experimentally validated O-glycosylation sites of three protein sequences, which are BSP-30, Kallikrein-1 and Ig delta chain C region (**Table 2**). Our method and the previous methods used almost the same positive data, which didn't contain BSP30, Kallikrein-1 and Ig delta chain C region, to train the prediction model. As shown in Table 2, our method achieved 1.000 (16/16) sensitivity, while the previous methods showed 0.375 (6/16) and 0.563 (9/16) sensitivity respectively. Moreover, our method showed better balanced accuracy than the previous methods. Hence we can conclude that in predicting O-glycosylation sites our method is competitive with and sometimes superior to the previous methods.

2.4 Validation of Biological Application of the Proposed Model to the N-glycosylation Site Prediction

Although the previous studies [14)–17)] focused on the O-glycosylation, our study also produced the prediction model for the N-glycosylation. To confirm the biological applicability of our prediction model, we predicted the N-glycosylation sites of a protein, envelope glycoprotein gp120 precursor, whose glycosylation sites have been identified. Gp120 was not included in the dataset.

Envelope glycoprotein gp120 precursor is a part of envelope glycoprotein from AIDA virus [27)] and has 17 consensus N-glycosylation motifs (Asn-Xaa-Ser/Thr). Among them, 14 sites are validated to be glycosylated in the PDB database

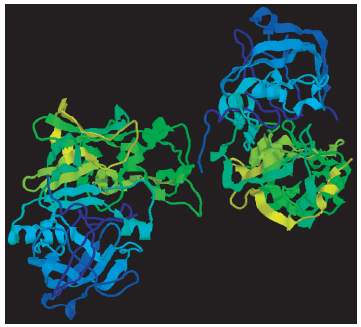

Fig. 3 3D structure of beta-secretase 1 (PDB ID:1FKN) BACE1 is an enzyme that breaks down proteins and regulates functions of membrane proteins. Furthermore, it is known to be associated with Alzheimer's disease. BACE1 forms a homo-dimer.

(PDB ID: 1G9M).

10 out of these 14 sites were correctly predicted as glycosylation sites. Moreover, 2 of 3 non-glycosylation sites were successfully identified. Thus we conclude that our model can be applied to glycoproteins with sufficient reliability.

2.5 Predictions for Unknown Glycosylation Sites

To validate the applicability of our prediction model at a genome-wide level, we predicted the N-glycosylation sites of beta-secretase 1 (BACE1) whose glycosylation sites have not been identified. BACE1 (**Fig. 3**) is an enzyme that breaks down proteins, and which regulates the function of membrane proteins [28]. Moreover, it is known to be associated with Alzheimer's disease [29].

The BACE1 protein sequence has four consensus N-glycosylation motifs (Asn-Xaa-Ser/Thr). We predicted whether these four sites would be glycosylated or not using our method (**Table 3**). Three sites were predicted to be glycosylated and the other one was predicted to be non-glycosylated. The prediction for these four sites was finished within 0.3 seconds on a 2-CPU cluster (Opteron 275 2.2 GHz processors). This fast computation suggests our method can be applied at a genome-wide level.

To confirm the validity of our predictions, the local structure around the predicted N-glycosylation sites in BACE1 as well as known N-glycosylation sites in the training dataset were shown in **Fig. 4**. The molecular mechanism of N-glycosylation is that a glycan moiety is attached to an asparagine residue by binding to the amido group in the target residue. As glycan moieties are larger than amino acids with several monosaccharides that have a ring structure, some space around the amido group of the asparagine is necessary for glycosylation to occur. In particular, the amido group of the asparagine residue shown in Fig. 4 (B), a known glycosylation site, has plenty of space around it and sticks out. Similarly, the amido group of the 153rd asparagine residue predicted to be a glycosylation site, shown in Fig. 4 (A), is likely to bind to a glycan moiety since there is a lot of space around it and the amido group is very exposed. On the other hand, the amido group of the 223rd asparagine residue predicted to be non-glycosylated, shown in Fig. 4 (C), is less likely to be glycosylated, because the space surrounding it is as small as a known non-glycosylation site, shown in Fig. 4 (D).

Table 3 N-glycosylation site prediction in BACE1.

Residue number	Sequence window	Prediction result	SES (Å^2)
153	TDLVSIPHGP*N*VTVRANIAAI	Glycosylation site	26.88
172	AITESDKFFI*N*GSNWEGILGL	Glycosylation site	29.02
223	ISLYMGENVT*N*QSFRITILPQ	Non-glycosylation site	5.99
354	AITESDKFFI*N*GSNWEGILGL	Glycosylation site	24.20

The BACE1 has four consensus N-glycosylation motifs (Asn-Xaa-Ser/Thr). Among these, three sites (153rd, 172nd and 354th residue) were predicted to be glycosylated and the other (223rd residue) was predicted to be non-glycosylated. SES areas of amido group of these 3 positive sites are clearly larger than that of the negative site.

To assess our prediction quantitatively, we calculated the solvent-excluded surface (SES) area by MSMS [30]. MSMS is a software which has been shown to be fast and reliable in computing molecular surfaces. The SES is the topological boundary of the union of all possible probes that do not overlap with the molecule (**Fig. 5**) and is used to visualize and study molecular properties [30]. The SES area of each amido group of the asparagine which we predicted as glycosylation sites, 153th, 172th and 354th residues, is obviously larger than that of the amido group of the asparagine which we predicted as a non-glycosylation site, 223rd residue (Table 3). Here, even if the molecular dynamics simulations were performed, the SES area of the amido group of the asparagine residue didn't fluctuate significantly (See Supplementary Material 3). The SES area of the glycosylated amido

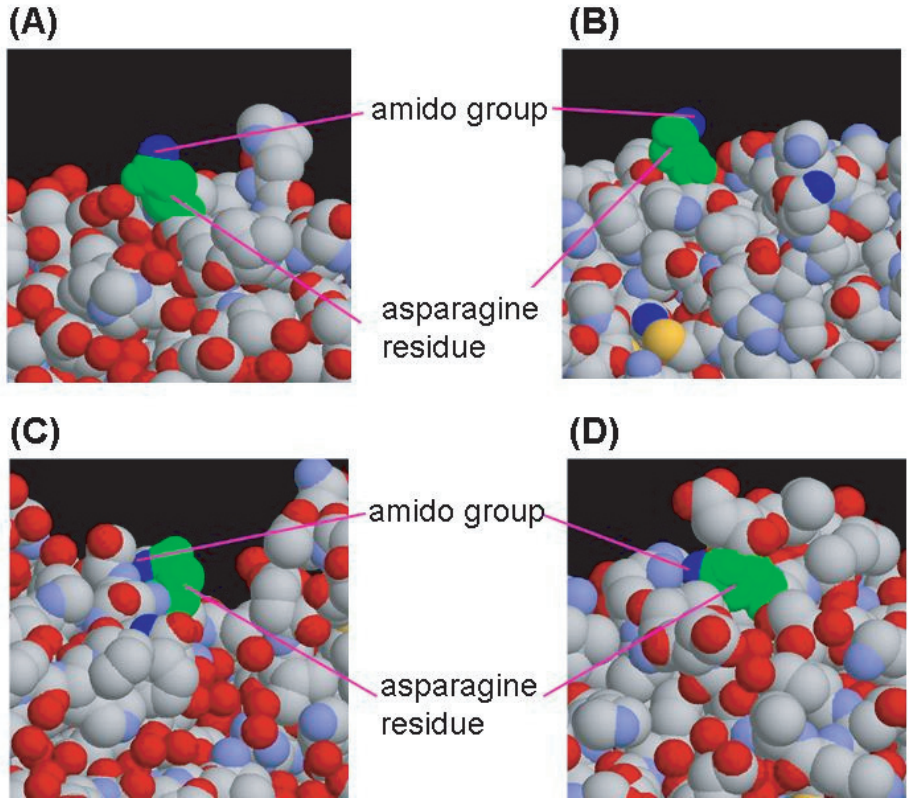

Fig. 4 Local structure around N-glycosylation sites and non-glycosylation sites. The atoms shown in green correspond to asparagine residues and atoms shown in blue illustrate an amido group in the asparagine residue. (A) The local structure around the 153rd residue in BACE1, which is predicted to be a glycosylation site. (B) The local structure around a known glycosylation site in the training dataset. (C) The local structure around the 223rd residue in BACE1, which is predicted to be a non-glycosylation site. (D) The local structure around a known non-glycosylation site in the training dataset.

group was constantly larger than that of the non-glycosylated amido group.

We also applied the same evaluation approach to the O-glycosylation site prediction. O-glycosylation sites of leptin precursor which is the causal factor of adipositas were predicted [31]. The molecular mechanism of O-glycosylation is

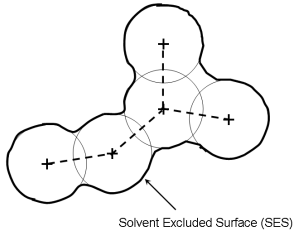

Fig. 5 The solvent-excluded surface (SES). SES is the topological boundary of the union of all possible probes having no intersection with a set of overlapping spheres M. This surface is used to not only describe hydration effects, but also to visualize protein surfaces and to study molecular properties.

that a glycan moiety is attached to a serine or threonine residue by binding to the hydroxyl group in the target residue.

Leptin precursor has twenty-two candidate sites of O-glycosylation. Among these candidates, seven sites were predicted to be glycosylated (Supplementary Material 1).

We analyzed the local structure around the predicted O-glycosylation sites in leptin precursor (**Fig. 6**). The hydroxyl group of the 138th serine residue, predicted as a glycosylation site, was shown in Fig. 6 (A). On the other hand, the hydroxyl group of the 73rd serine residue, predicted as a non-glycosylation site, was shown in Fig. 6 (B). As shown in Fig. 6, the hydroxyl group of the 138th serine was spatially more suitable for an approach of glycosyltransferases than that of the 73rd serine. SES areas of the hydroxyl group in the 7 predicted glycosylation residues are significantly larger than those in the non-glycosylation residues (P-value ¡ 0.02 by t test) (See Supplementary Material 2).

Therefore, we conclude that our model can predict structurally reasonable both N- and O- glycosylation sites in proteins.

3. Discussion

Our model, which predicts glycosylation sites using not only local information, but also general information and subcellular localization of proteins, showed better prediction performances than previous models [14)–17)], which only considered local information (Table 1). These findings suggest that it is important to con-

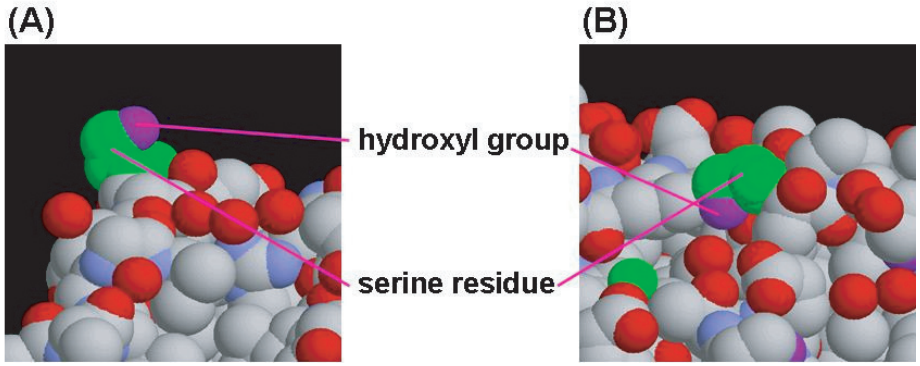

Fig. 6 Local structure around O-glycosylation sites and non-glycosylation sites. The atoms shown in green correspond to serine residues and atoms shown in purple illustrate a hydroxyl group in the serine residue. (A) The local structure around the 117$^{\text{th}}$ residue in leptin precursor, which is predicted to be a glycosylation site. (B) The local structure around the 52$^{\text{nd}}$ residue in leptin precursor, which is predicted to be a non-glycosylation site.

sider whole-protein-sequence information and subcellular localization when predicting glycosylation sites. Furthermore, in our computational experiment, in which our model was applied to a protein whose glycosylation sites had not been identified, glycosylation sites predicted by our model were shown to be structurally reasonable (Fig. 4 and Fig. 6). Therefore, we conclude that our method is a comprehensive and effective computational method that is applicable at a genome-wide level.

4. Conclusions

In the present study, we developed a comprehensive and effective computational method that detects glycosylation sites. Identification of the structure of glycans attached to glycosylation sites is a challenge that follows the identification of glycosylation sites. To resolve this problem, it is necessary to construct a comprehensive database, which contains information about glycosylation sites and glycan structures at each glycosylation site. Identification of glycosylation sites and protein-bound glycan structures will contribute to further understanding of the functions of glycosylation and glycans that have not been fully elucidated.

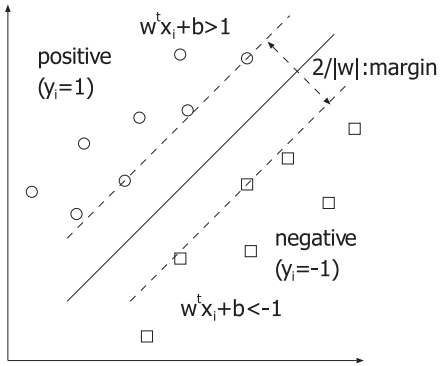

Fig. 7 Schematic diagram of SVM separating positives (circles) and negatives (squares) in a higher dimensional feature space. Hyperplanes (dotted lines) are determined so that $|w|$, the Euclidean norm of weights for each dimension or feature, is minimized, or the margin $(2/|w|)$ is maximized.

Moreover, if we can overcome these problems, the field of glycoinformatics will be established next to bioinformatics and chemoinformatics.

5. Methods

5.1 Support Vector Machine

SVM is a new technique for data classification that has better performance than ANN [32]. SVM has been used to solve a variety of biological classification problems [33)–37)].

The concept of SVM is based on the structural risk minimization principle to minimize both training and generalization errors [38)]. When used for classification, SVM separates positive (for example, glycosylation sites) and negative (for example, non-glycosylation sites) training samples in a multidimensional space by constructing a hyperplane optimally positioned between the positive and negative samples (**Fig. 7**). A testing sample is then projected onto this multidimensional space to determine its class affiliation based on its relative position to the hyperplane.

SVM produces the classifier shown in Equation (1). In SVM, each feature vector x_i is projected into a higher dimensional feature space using a kernel function such as the RBF kernel, or $K(x_i, x_j)$ in Eq. (1).

$$f(x) = \mathrm{sign}\left(\sum_{i=1}^{n} y_i \lambda_i^* K(x_i, x) + b^*\right), \qquad (1)$$
$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

where λ_i^* is a Lagrange multiplier, b^* is a parameter which is determined by the hyperplane and σ is a parameter of RBF Kernel.

In this paper, we used the SVM software named *LIBSVM*[39] to perform the prediction task. RBF kernel was selected as it showed the best performances (See Supplementary Material 1). Kernel functions used were as follows,

Linear kernel : $K(x_i, x_j) = x_i^T x_j$

Polynomial kernel : $K(x_i, x_j) = \left(\gamma x_i^T x_j\right)^3$

Sigmoid kernel : $K(x_i, x_j) = \tanh\left(\gamma x_i^T x_j\right)$

5.2 Extraction of a Sequence Descriptor

5.2.1 Local Information

We encoded local information of glycosylation sites by extracting a subsequence within a window of fixed size (Fig. 1). We extracted k upstream and downstream residues of Asn (N), Ser (S) or Thr (T) residues that were predicted to be glycosylated. In this paper, we set $k = 10$, constituting the sequence window of 20 residues (Fig. 1). In case the full sequence window cannot be extracted, we define 'Z' as the 21st amino acid to fill blanks (**Fig. 8**). To encode one residue in the sequence window, we utilized the BLOSUM62 profile encoding (the corresponding row in the BLOSUM62 matrix). For example, the BLOSUM62 profile for alanine is equal to the vector (0,1,-1,-1,4,0,-1,-2,-1,-1,-2,-1,-1,-1,-1,-1,-2,-2,-2,-3) and that for 'Z', the 21st amino acid, is equal to the vector (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0). Therefore, a 20×20 dimension vector was calculated for each sequence window. In the previous study[40], BLOSUM encoding, where each row in BLOSUM matrix was utilized to encode each amino acid, was used to predict T-cell class 1 epitopes by neural network. In this study, the prediction performance with this encoding was better than the other method.

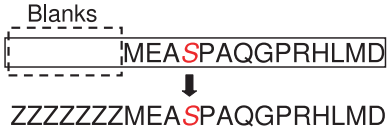

Fig. 8 'Z' as the 21st amino acid. When the glycosylation site is near the ends of protein sequence, the full sequence window cannot be extracted. In this situation, we define 'Z' as the 21st amino acid to fill blanks.

5.2.2 General Information about Proteins

We counted the frequency of di-peptides in a whole protein sequence to encode general protein information. Glycans are attached to proteins by glycosyltransferases, which interact with the target proteins. The interaction with the objective protein depends not only on the local site but also on the whole protein structure. In order to consider the effects of glycosyltransferases, the structures of proteins should be taken into account. In a previous study, it was shown that protein structural classes can be predicted by counting the frequency of di-peptides[41]. Thus, we assume that counting the frequency of di-peptides enables consideration of protein structures. As there are 20 amino acids and 20×20 kinds of di-peptides, a 400-dimension vector was calculated for each protein.

5.2.3 Subcellular Localization Information about Proteins

We used the output of WoLF PSORT[26] to encode subcellular localization information. Proteins are synthesized in the ribosome and modified with glycans in the endoplasmic reticulum or Golgi. The resultant glycoproteins are distributed throughout cells. In particular, most membrane proteins are glycoproteins[24]. For example, the subcellular localization of glycoproteins and non-glycoproteins in our datasets is shown in **Fig. 9**. As shown in Fig. 9, the subcellular localization of glycoproteins is specific, as about half of all glycoproteins localize extracellularly, while only 15% non-glycoproteins localize extracellularly. In WoLF PSORT, localization of the target sequence is determined based on the localization of training proteins that have sequence similarity with the target. To encode subcellular localization information, we utilized the frequency of each subcellular localization in the output of Wolf PSORT. The value for the subcellular localization x is calculated as the number of proteins localizing in x divided by the total number of proteins similar to the target. As there are 23 subcellular localizations

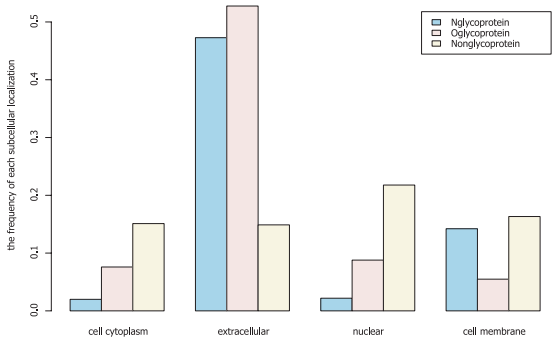

Fig. 9 The frequency of each subcellular localization. Distribution of subcellular localization prediction outputs of Wolf PSORT for glycoproteins and non-glycoproteins in our datasets is illustrated. It should be noted that the prediction output of Wolf PSORT is based on localization of proteins similar to a query and thus several localizations where N-linked glycoproteins don't exist, for example, are observed.

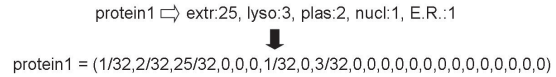

Fig. 10 Encoding the output of WoLF PSORT. In this example, WoLF PSORT exhibits that, among 32 sequences similar to protein1, there are 25 proteins that localize extracellularly (*extra*). Therefore, the 3rd element of the feature vector, which corresponds to an *extra* localization, is 25/32 for protein1. The value for the subcellular localization x is calculated as the number of proteins localizing in x divided by the total number of proteins similar to the target. Here, *extr* stands for extracellular, *lyso* for lysosome, *plas* for plasmalemma, *nucl* for nuclear, *E.R.* for endoplasmic reticulum.

in the output of WoLF PSORT, a 23-dimension vector was calculated for each target protein (**Fig. 10**).

5.2.4 The Structure of the Feature Vector

To utilize all information (local information, general information and subcellular localization information), the each vector was combined respectively (**Fig. 11**). If a protein has more than one glycosylation sites, vectors derived from protein whole sequence and subcellular localization information are identical for these sites. We use the combined vector as an input for *LIBSVM*.

5.3 Prediction Performance Assessment

The performance of SVM has often been assessed using the five-fold cross validation method [42]. The dataset was randomly divided into five subsets of approximately equal size. One of the five subsets was used as a test set, and the remaining four subsets were used as training sets. This process was repeated five times so that every subset was used as a test set once. The performance of SVM can be assessed on the basis of accuracy, sensitivity, precision, MCC (Matthew's correlation coefficient) and AUC defined as follows:

Fig. 11 The structure of the feature vector. The each vector (vectors derived from local information, general information or subcellular information) was combined respectively. We use the combined vector as an input for *LIBSVM*.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}$$

Here, TP, TN, FP, FN stand for true positive, true negative, false positive and false negative, respectively. MCC ranges between -1 and 1. If there is no relationship between the predicted values and the real values, MCC should be around 0. In contrast, there is strong relationship between the predicted values and the real values, MCC should be close to 1. AUC represents the Area Under the (ROC) Curve which draws the evolution of the true positive rate versus the false positive rate. The AUC of an ideal classifier would be 1, while for a random classifier it would be 0.5.

5.4 Dataset Construction

5.4.1 N-glycosylation Site Dataset

From the glycosciences.de database [43], we collected N-glycosylation sites in human proteins that were validated in PDB database [44] as positives. As putative negative data, we randomly extracted 700 Asn residues attached no N-

glycosylation annotation and with a consensus motif in the form of "Asn-Xaa-Ser/Thr" (Xaa represents all kinds of amino acid except for proline) from the glycoproteins which have some annotations about glycosylation (such as "Potential", "Probable" and "By similarity") in UniProtKB/Swiss-Prot. This extraction significantly reduces the possibility to unexpectedly pick up false negatives, because in the glycoproteins with glycosylation annotation, every Asn residue site must have been examined and therefore Asn residue site with no glycosylation annotation is quite certainly true non-glycosylation site. The N-glycosylation dataset consisted of 308 positives from 125 human proteins and 700 negatives from 648 human proteins.

5.4.2 O-glycosylation Site Dataset

From the O-GLCBASE database [45], we collected O-glycosylation sites in mammalian proteins that were evidenced experimentally as positives. As putative negative data, we picked up the glycoproteins by choosing the proteins which have some annotations about glycosylation (such as "Potential", "Probable" and "By similarity") in UniProtKB/Swiss-Prot as mentioned above. From these limited proteins, we randomly extracted 1200 Ser/Thr residues in mammalian sequences with no annotation (such as "Potential", "Probable" and "By similarity") related to O-glycosylation in UniProtKB/Swiss-Prot. Since the mucin protein sequence has repeat sequences, several identical subsequences were generated within the window. These identical subsequences were counted as one positive or one negative in the dataset. The O-glycosylation dataset was composed of 551 positives from 242 mammalian proteins and 1200 negatives from 1160 mammalian proteins.

These N-glycosylation and O-glycosylation site dataset are available in our web site (http://www.dna.bio.keio.ac.jp/glycan/).

Acknowledgments This work was supported in part by a Grant program for bioinformatics research and development from the Japan Science and Technology Agency, and a Grant-in-Aid for Scientific Research on Priority Area No.17018029.

Supplementary Materials

(1) Comparison with the performances by using other kernels.
Integral model was utilized and different kernels (RBF, linear, polynomial and sigmoid) were applied in SVM computation.

(2) O-glycosylation site prediction in leptin precursor.
Leptin precursor has twenty two candidate sites of O-glycosylation. Sequence windows around the candidate sites and the SES area of hydroxyl group are shown as well as the prediction result.

(3) Effect of conformational change on SES area.
The average SES area of both the glycosylated and non-glycosylated amido group in several conformation of the endothelial protein C receptor precursor (PDB ID: 1L8J) is shown. One nanosecond molecular dynamics simulation was performed with AMBER 9 [46], and the SES area was calculated every 200 picoseconds.

References

1) Salas, J. and Mendez, C.: Engineering the glycosylation of natural products in actinomycetes, *Trends Microbiol.*, Vol.15, pp.219–232 (2007).
2) Saxon, E. and Bertozzi, C.: Chemical and biological strategies for engineering cell surface glycosylation, *Annu. Rev. Cell Dev. Biol.*, Vol.17, pp.1–23 (2001).
3) Plante, O.: Combinatorial chemistry in glycobiology, *Comb. Chem. High Throughput Screen.*, Vol.8, pp.153–159 (2005).
4) Breton, C., Snajdrova, L., Jeanneau, C., Koca, J. and Imberty, A.: Structures and mechanisms of glycosyltransferases, *Glycobiology*, Vol.16, pp.29R–37R (2006).
5) Breton, C. and Imberty, A.: Structure/function studies of glycosyltransferases, *Curr. Opin. Struct. Biol.*, Vol.9, pp.563–571 (1999).
6) Imberty, A., Wimmerova, M., Koca, J. and Breton, C.: Molecular modeling of glycosyltransferases, *Methods Mol. Biol.*, Vol.347, pp.145–156 (2006).
7) Goletz, S., Thiede, B., Hanisch, F., Schultz, M., Peter-Katalinic, J., Muller, S., Seitz, O. and Karsten, U.: A sequencing strategy for the localization of O-glycosylation sites of MUC1 tandem repeats by PSD-MALDI mass spectrometry, *Glycobiology*, Vol.7, pp.881–896 (1997).
8) Sadeghi, H. and Birnbaumer, M.: O-Glycosylation of the V2 vasopressin receptor, *Glycobiology*, Vol.9, pp.731–737 (1999).
9) Skropeta, D., Settasatian, C., McMahon, M., Shearston, K., Caiazza, D., McGrath, K., Jin, W., Rader, D., Barter, P. and Rye, K.: N-Glycosylation regulates endothelial lipase-mediated phospholipid hydrolysis in apoE- and apoA-I-containing high density lipoproteins, *J. Lipid Res.*, Vol.48, pp.2047–2057 (2007).
10) Wojczyk, B., Takahashi, N., Levy, M., Andrews, D., Abrams, W., Wunner, W. and Spitalnik, S.: N-glycosylation at one rabies virus glycoprotein sequon influences N-glycan processing at a distant sequon on the same molecule, *Glycobiology*, Vol.15, pp.655–666 (2005).

11) Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K., Ueda, N., Hamajima, M., Kawasaki, T. and Kanehisa, M.: KEGG as a glycome informatics resource, *Glycobiology*, Vol.16, pp.63R–70R (2006).
12) Jenkins, N., Parekh, R. and James, D.: Getting the glycosylation right: implications for the biotechnology industry, *Nat. Biotechnol.*, Vol.14, pp.975–981 (1996).
13) Apweiler, R., Hermjakob, H. and Sharon, N.: On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database, *Biochim. Biophys. Acta*, Vol.1473, pp.4–8 (1999).
14) Li, S., Liu, B., Zeng, R., Cai, Y. and Li, Y.: Predicting O-glycosylation sites in mammalian proteins by using SVMs, *Comput Biol Chem*, Vol.30, pp.203–208 (2006).
15) Julenius, K., Molgaard, A., Gupta, R. and Brunak, S.: Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites, *Glycobiology*, Vol.15, pp.153–164 (2005).
16) Caragea, C., Sinapov, J., Silvescu, A., Dobbs, D. and Honavar, V.: Glycosylation site prediction using ensembles of Support Vector Machine classifiers, *BMC Bioinformatics*, Vol.8, p.438 (2007).
17) Chen, Y., Tang, Y., Sheng, Z. and Zhang, Z.: Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs, *BMC Bioinformatics*, Vol.9, p.101 (2008).
18) Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S. and Brunak, S.: Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence, *Proteomics*, Vol.4, pp.1633–1649 (2004).
19) Gupta, R. and Brunak, S.: Prediction of glycosylation across the human proteome and the correlation to protein function, *Pac Symp Biocomput*, pp.310–322 (2002).
20) Petrescu, A., Milac, A., Petrescu, S., Dwek, R. and Wormald, M.: Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding, *Glycobiology*, Vol.14, pp.103–114 (2004).
21) Baenziger, J.: Protein-specific glycosyltransferases: how and why they do it!, *FASEB J.*, Vol.8, pp.1019–1025 (1994).
22) Opdenakker, G., Rudd, P., Ponting, C. and Dwek, R.: Concepts and principles of glycobiology, *FASEB J.*, Vol.7, pp.1330–1337 (1993).
23) vonder Lieth, C., Bohne-Lang, A., Lohmann, K. and Frank, M.: Bioinformatics for glycomics: status, methods, requirements and perspectives, *Brief. Bioinformatics*, Vol.5, pp.164–178 (2004).
24) Spiro, R.: Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds, *Glycobiology*, Vol.12, pp.43R–56R (2002).
25) Nielsen, H., Brunak, S. and von Heijne, G.: Machine learning approaches for the prediction of signal peptides and other protein sorting signals, *Protein Eng.*, Vol.12, pp.3–9 (1999).
26) Horton, P., Park, K., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. and Nakai, K.: WoLF PSORT: protein localization predictor, *Nucleic Acids Res.*, Vol.35, pp.W585–587 (2007).
27) Ohgimoto, S., Shioda, T., Mori, K., Nakayama, E., Hu, H. and Nagai, Y.: Location-specific, unequal contribution of the N glycans in simian immunodeficiency virus gp120 to viral infectivity and removal of multiple glycans without disturbing infectivity, *J. Virol.*, Vol.72, pp.8365–8370 (1998).
28) Zacchetti, D., Chieregatti, E., Bettegazzi, B., Mihailovich, M., Sousa, V., Grohovaz, F. and Meldolesi, J.: BACE1 expression and activity: relevance in Alzheimer's disease, *Neurodegener Dis*, Vol.4, pp.117–126 (2007).
29) Heneka, M. and O'Banion, M.: Inflammatory processes in Alzheimer's disease, *J. Neuroimmunol.*, Vol.184, pp.69–91 (2007).
30) Sanner, M., Olson, A. and Spehner, J.: Reduced surface: an efficient way to compute molecular surfaces, *Biopolymers*, Vol.38, pp.305–320 (1996).
31) Sone, M. and Osamura, R.: Leptin and the pituitary, *Pituitary*, Vol.4, pp.15–23 (2001).
32) Byvatov, E. and Schneider, G.: Support vector machine applications in bioinformatics, *Appl. Bioinformatics*, Vol.2, pp.67–77 (2003).
33) Bhasin, M. and Raghava, G.: GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors, *Nucleic Acids Res.*, Vol.32, pp.W383–389 (2004).
34) Yabuki, Y., Muramatsu, T., Hirokawa, T., Mukai, H. and Suwa, M.: GRIFFIN: a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model, *Nucleic Acids Res.*, Vol.33, pp.W148–153 (2005).
35) Gubbi, J., Shilton, A., Parker, M. and Palaniswami, M.: Protein topology classification using two-stage support vector machines, *Genome Inform*, Vol.17, pp.259–269 (2006).
36) Han, L., Zheng, C., Xie, B., Jia, J., Ma, X., Zhu, F., Lin, H., Chen, X. and Chen, Y.: Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness, *Drug Discov. Today*, Vol.12, pp.304–313 (2007).
37) Burbidge, R., Trotter, M., Buxton, B. and Holden, S.: Drug design by machine learning: support vector machines for pharmaceutical data analysis, *Comput. Chem.*, Vol.26, pp.5–14 (2001).
38) Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. and Nielsen, H.: Assessing the accuracy of prediction algorithms for classification: an overview, *Bioinformatics*, Vol.16, pp.412–424 (2000).
39) Chang, C.-C. and Lin, C.-J.: *LIBSVM: A library for support vector machines* (2001). Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm.
40) Nielsen, M., Lundegaard, C., Worning, P., Lauemoller, S., Lamberth, K., Buus,

S., Brunak, S. and Lund, O.: Reliable prediction of T-cell epitopes using neural networks with novel sequence representations, *Protein Sci.*, Vol.12, pp.1007–1017 (2003).
41) Chen, C., Zhou, X., Tian, Y., Zou, X. and Cai, P.: Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network, *Anal. Biochem.*, Vol.357, pp.116–121 (2006).
42) Han, L., Cui, J., Lin, H., Ji, Z., Cao, Z., Li, Y. and Chen, Y.: Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity, *Proteomics*, Vol.6, pp.4023–4037 (2006).
43) Lutteke, T., Bohne-Lang, A., Loss, A., Goetz, T., Frank, M. and vonder Lieth, C.: GLYCOSCIENCES.de: An Internet portal to support glycomics and glycobiology research, *Glycobiology*, Vol.16, pp.71R–81R (2006).
44) Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. and Bourne, P.: The Protein Data Bank, *Nucleic Acids Res.*, Vol.28, pp.235–242 (2000).
45) Gupta, R., Birch, H., Rapacki, K., Brunak, S. and Hansen, J.: O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins, *Nucleic Acids Res.*, Vol.27, pp.370–372 (1999).
46) Case, D., Cheatham, T., Darden, T., Gohlke, H., Luo, R., Merz, K., Onufriev, A., Simmerling, C., Wang, B. and Woods, R.: The Amber biomolecular simulation programs, *J Comput Chem*, Vol.26, pp.1668–1688 (2005).

(Received September 16, 2008)
(Accepted November 25, 2008)
(Released March 24, 2009)

(Communicated by *Takenao Ookawa*)

Kenta Sasaki was born in 1984. He received his B.E. from Keio University in 2007 and will receive his M.E. from Keio University in 2009. His research interests include machine learning to bioinformatics and chemoinformatics.

Nobuyoshi Nagamine is currently a Ph.D. candidate at Graduate School of Science and Technology, Keio University. He received his Bachelor of Science and Master of Science from Keio University in 2005 and 2006 respectively. His research interests include Bioinformatics, Chemoinformatics and particularly application of machine learning methods to the field of virtual screening in drug discovery.

Yasubumi Sakakibara is a full professor at the department of Biosciences and Informatics at Keio University. He received his degree of Doctor of Science from Tokyo Institute of Technology in 1991. He spent one year as postdoc at UC Santa Cruz and collaborated with Prof. David Haussler on the project of stochastic context-free grammars for modeling RNAs. He also worked for Fujitsu Laboratory. His research interests include bio-informatics, DNA computers, and computer science. He is a member of IPSJ, JSBi, and ISCB.