Glycoconj J (2012) 29:551–564
DOI 10.1007/s10719-012-9434-x

A novel model to predict O-glycosylation sites using a highly unbalanced dataset

Kun Zhou · Chunzhi Ai · Peipei Dong · Xuran Fan · Ling Yang

Received: 26 March 2012 / Revised: 11 July 2012 / Accepted: 17 July 2012 / Published online: 3 August 2012

Abstract *In silico* approaches have become an alternative method to study O-glycosylation. In this paper, we developed a linear interpretable model for O-glycosylation prediction based on an unbalanced dataset, analyzing the underlying biological knowledge of glycosylation. A training set of 4446 sites involving 468 positive sites and 3978 negative sites was developed during this research. The sites were encoded using the amino acid index (AAindex), and the forward stepwise procedure utilized for feature selection. The linear discriminant analysis with an equal *a priori* probability (PP-LDA) was employed to develop the interpretable model. Performance of the model was verified using both the internal leave-one-out cross-validation and external validation methods. Two non-linear algorithms, the supervised support vector machine and the unsupervised self-organizing competitive neural network, were used as comparisons. The PP-LDA model exhibited improved classification results with accuracy of 82.1 % for cross-validations and 80.3 % for external prediction. Further analysis of this linear model indicated that the properties at position R_1 and the properties relative to hydrophobicity contributed more to the glycosylation prediction. However, the alpha and turn propensities at the C-terminal, together with physicochemical properties at the N-terminal, are also relative to the glycosylation activity. This model is not only capable of predicting the possibility of glycosylation using an unbalanced dataset, but is also helpful to understand the underlying biological mechanisms of glycosylation. Considering the publicly accessibility of our prediction model, a downloadable program is provided in our supply materials.

Keywords Protein glycosylation prediction · Amino acid index · Feature selection · PP-LDA

Electronic supplementary material The online version of this article (doi:10.1007/s10719-012-9434-x) contains supplementary material, which is available to authorized users.

K. Zhou · C. Ai · X. Fan · L. Yang (✉)
Laboratory of Pharmaceutical Resource Discovery,
Dalian Institute of Chemical Physics, Chinese Academy of Sciences,
457 Zhongshan Road,
Dalian 116023, China
e-mail: yling@dicp.ac.cn

K. Zhou
Graduate School of Chinese Academy of Sciences,
Beijing 100049, China

P. Dong
Research Institute of Integrated Traditional,
Western Medicine of Dalian Medical University,
Dalian 116044, China

Introduction

Glycosylation is one of the most common and therefore vitally important post-translational modifications (PTMs), which modulates a variety of biological processes, both at the cellular and protein level [1–3]. It is estimated that more than half of proteins in nature are glycosylated [4]. Glycoproteins are associated with diseases such as Alzheimer's disease and cancer [5–8]. Protein glycosylation can be divided into two main categories, N-linked glycosylation and O-linked glycosylation [9]. N-glycosylation is the modification of asparagine (N) residues in the sequence N-!P- S/T (where !P signifies any amino acid except proline), while O-glycosylation is only known to be Ser (S) or Thr (T) specific, and no consensus sequence has been identified [10].

Significant efforts have been made in understanding the rules or identifying a consensus sequence for O-linked glycosylation. The development of analytical methods, such as proteomics and mass spectrometry [11], has boosted the analysis of O-glycosylation. Previous investigations indicated that O-glycosylation is preferable to the site adjacent to

proline, serine, threonine and alanine [9, 12, 13]. Of these, based on studies on the effect of flanking residues on synthetic peptides, proline is considered crucial [14, 15]. However, due to the large variation in sequence surrounding the glycosylated residues it is difficult to manually inspect the protein sequences. Moreover, the analytical methods are still expensive and laborious. Therefore, it would be advantageous to simplify experimental steps by integrating computational approaches into validation procedures [16].

The *in silico* prediction, being an alternative method for glycosylation analysis, not only gives a preview of the glycosylated sites prior to the experiments, but also efficiently reduces the number of potential targets of glycosylation that require further *in vivo* and *in vitro* confirmation. According to previous studies, glycosylation is influenced by various factors, such as the primary sequences and the structure information of the protein or its location. However, scientists still attempted to explain and predict the complex glycosylation phenomena simply from the primary protein sequence [17, 18]. During this research, a series of prediction methods for O-glycosylation sites had been developed [11, 19]. Elhammer *et al.* applied a matrix statistics method to predict O-glycosylation sites [12], after which, a vector projection method was developed [17, 20]. Presently, machine learning methods such as neural network (NN) and support vector machine (SVM) are also employed to perform the prediction according to the peptide sequence [21, 22]. The NetOglyc 3.1 server was constructed based on a neural network using sequence contexts and surface accessibility. It can correctly predict 76 % of the glycosylated residues and 93 % of the non-glycosylated residues [23]. Also, there are other publically available O-glycosylation site prediction web-servers [11, 24] giving satisfactory performance. However, there are still some problems that need to be considered, including the identification of discriminatory features from the original feature pool. The more features we use, the larger the computational cost, and the dimensional curse occurs, which makes the predictions more difficult. Therefore, the extraction of useful descriptors from the original feature pool is essential for effective classification. Also, the interpretability of the model is of great importance to further exploit the potential bio-information of the glycosylation. Only a small amount of underlying information can be provided by the present *in silico* glycosylation models. This is due to the encoding methods, in which the codes themselves have little biological meanings; or due to the machine learning algorithms, which cannot formulize the relationship between variables or respond in clear numbers or coefficients. The highly unbalanced dataset was an inevitable challenge in computational studies for glycosylation. The unbalanced dataset made predictions more difficult because classifiers were trained to optimize the accuracy and performed rather poorly on the minority classes [25]. A commonly used solution for an unbalanced dataset in the prediction of glycosylation was to re-sample the original dataset, which was to randomly select a subset of non-glycosylation as a negative dataset. However, this method cannot utilize all the information available in the training set and increases the false positive rate [26]. To date, only Caragea *et al.* have used the ensemble SVM approach to predict the glycosylation sites with an unbalanced dataset [26]. However, it was still an integrated result of m individual SVM classifiers trained with a balanced subsample, where positive sites were repeatedly used. It is therefore necessary to develop an improved and interpretable model for accurate predictions of glycosylation with an unbalanced dataset.

The fundamental aim of this paper is to build an O-glycosylation site prediction model using an unbalanced dataset, and to interpret the model to explore the underlying bio-information for O-glycosylation. Combining with stepwise forward selection (SFS) methods, a linear discriminant analysis with an equal *a priori* probability (PP-LDA) was utilized to develop the interpretable model. Two non-linear algorithms, supervised support vector machine (SVM) and unsupervised self-organizing competitive neural network (SOCNN), were used as comparisons. In this manuscript, attention was focused on the statistical validity and model interpretability. We expect that such a prediction model will provide a helpful tool in identifying the O-glycosylation sites from proteins and also in understanding the biological process of glycosylation.

Material and methods

Dataset construction

The data used in this manuscript was from O-GlycBase v6.00 (www.cbs.dtu.dk/database/oglycbase), which contains 242 proteins from several different species. An entry in the database provided information about the glycan features involving, the species, experimentally verified glycosylation sites, literature references, protein sequence, and http-linked cross-references to other protein sequence databases (*e.g.* SWISS-PROT, PIR). Finally, 218 proteins with experimental verified O-glycosylation sites were included in our dataset.

The O-glycosylation was a site specific process that mainly occurred on Ser or Thr residues [9]. The process involves enzymes (the transferase) that recognize a glycosylation site due to the surrounding residues [17]. Therefore, the experimentally verified glycosylation sites were extracted, and are represented by a subsequence fragment of 2n+1 amino acids as positive sites, where the glycosylated S or T site was in the central position, and n was the

number of amino acid neighbors on each side. Any of the S/T sites from the proteins in the dataset, which were not shown experimentally to be glycosylated, were extracted as negative. Considering that the number of upstream or downstream residues may be less than n for the sites located in N- or C-terminus, we assigned a non-existing amino acid O to fill in the corresponding positions, for the purpose of ensuring a sequence fragment with a unified length [24]. Finally, 21 different amino acids were used in the present study to reflect the sequence context of a glycosylation site, which were ordered as ACDEFGHIKLMNPQRSTVWYO. To remove redundant fragments within the dataset, the positive and negative datasets were further filtered respectively by a 60 % sequence identity cut-off. Negative sites sharing over 60 % identity with any of the positive sites were also discarded. The remaining sites were further randomly split into training and test sets in a ratio of 3:1. The remaining training set contained 468 positive sites and 3978 negative sites while the test set contained 160 positive and 1322 negative sites.

It is essential for a model to have a good classification on a known dataset. However, it is even more meaningful for the model to have predictability on unknown sites or sites that did not appear in the training set. Zsuzsanna *et al.* reported the unambiguous identification of 26 glycosylation sites using new methods, 21 of which were novel [27], and not included in our training dataset. Therefore, with the purpose to evaluate the predictability of our model for unknown or new glycosylation sites, the 21 novel glycosylation sites and their corresponding 511 non-glycosylation sites reported by Zsuzsanna *et al.* [27] were also used as the test set in our paper.

Feature construction

Protein structures and functions are defined by the combinations of physicochemical and biochemical properties of 20 naturally occurring amino acids, the building-blocks of proteins. Consequently, a wide variety of amino acid properties have been investigated through a large number of experiments and theoretical studies, including alpha and turn propensities, hydrophobicity and physicochemical properties. These amino acid properties can be represented by a set of 20 numerical values and are referred to as the amino acid index [28–30] (detailed information can be found on this website: http://www.genome.jp/aaindex/). In our work, the amino acid index (AAindex) was selected to represent the dataset. The AAindex is a number of descriptors representing various physicochemical and biochemical properties of amino acids and pairs of amino acids. AAindex1, including a total of 544 indices was used throughout this study. Excluding indices with missing values, 526 indices were left for encoding sample peptides. A peptide was denoted by R_n, R_{n-1}, R_{n-2},……R_1, R_0, R_1', R_2',……R_{n-1}', R_n' (where a number without or with ' refers to an amino acid that was located on either the N-terminal or the C-terminal side of the glycosylated amino acid, respectively) and can be encoded into 526*(2n+1) features. The 526 indices are listed in the Supplemental Materials I.

The feature construction detail is described as follows [18]:

$$F = \left(x_1, x_2, \cdots\cdots, x_i, \cdots\cdots, x_{526\times(2n+1)}\right)$$
$$(i = 1, 2, \cdots\cdots, 526 \times (2n+1))$$

'i' is the number of features, and can be calculated by the position and index of the residue:

$$\mathrm{i} = 526 \times \mathrm{Position}_{\mathrm{residue}} + \mathrm{Index}_{\mathrm{residue}} \quad (2)$$

The position and the index of the residue can be calculated reversely by the following equations:

$$Position_{residue} = \frac{i}{526} \quad (3)$$

$$\mathrm{Index}_{\mathrm{residue}} = i\%526 \quad (4)$$

Finally, in our paper, three datasets were used and abbreviated as the 21aa, 15aa and 9aa dataset where the n value was 10, 7 and 4, respectively. The minimum n value was set to 4 as the enzyme recognizes the neighboring four amino acids of the glycosylation sites [17], while the maximum n value was set to 10 to avoid overly optimistic estimates. Unless otherwise stated, the algorithm used the 21aa dataset.

Feature selection

In view of the fact that the final dataset contained as many as 11046 features, feature selection was necessary. The stepwise forward selection (SFS) method was conducted to extract informative features from the features pool in the training set. SFS was a standard procedure for variable selection, based on the procedure of sequentially introducing the predictors into the model one at a time. In this paper, the procedure begins by considering each of the features individually and selecting the one that provides higher performance (*e.g.* the feature that most reduces the prediction error). The next step was to calculate all the possible two variable models, and the variable is added to the model if when taken together with the selected variable produces the lowest prediction error. This process was iterated until the prediction error was not reduced further by including a new variable. The iteration also stopped when all variables have been added to the model or a stopping criterion was met [31–33].

Model building

Linear Discriminant Analysis (LDA)

The basic theory of LDA was to classify the dependent variable by dividing an n-dimensional feature space into two regions. The regions are separated by a hyperplane, which was defined by a linear discriminant function. LDA could be used to build a predictive model of the group membership based on the observed characteristics of each case. This procedure generated a discriminant function based on linear combinations of predictor variables that provide the best discrimination among the groups. In this paper, the LDA classification model (Eq. 5) was created to describe the glycosylation activity P as a linear combination of selected features:

$$\{x_1, x_2, \cdots\cdots, x_i, \cdots\cdots, x_N\}$$

weighted by coefficients

$$\begin{matrix}\{a_0, a_1, a_2, \cdots\cdots, a_i, \cdots\cdots, a_N\} \\ P = a_1x_1 + a_2x2 + \cdots\cdots + a_Nx_N + a_0\end{matrix} \quad (5)$$

Where P represents the discriminant score, a_0 was the intercept term, a_i (i=1, 2. . . N) represents the coefficient associated with the corresponding variable x_i (i=1, 2,. . . , N), N is the number of features selected by SFS. The P values of +1 and -1 were assigned to glycosylation and non-glycosylation, respectively. The model was further estimated by standard statistics such as the corresponding *p*-level (p).

Prior probabilities are the likelihood of data belonging to a particular group, which give no information about the data available. The P(wj) is defined as the prior probability of group j, the probability that a randomly selected object belongs to group j; f(x|wj) is the conditional probability density function for x being a member of group j. The posterior probability P(wj|x), which is the probability on object x belongs to group j, is obtained using the Bayes rule:

$$P(w_j|x) = \frac{f(x, w_j)}{f(x)}, \quad (6)$$

$$f(x, w_j) = f(x|w_j)P(w_j), \quad (7)$$

$$f(x) = \sum f(x,w_j). \quad (8)$$

Suppose x is observed, and thus assigned to a group. Let cij(x) be the cost of assigning x to group i when it actually belongs to group j. The expected cost of assigning x to group i is

$$C_i(x) = \sum c_{ij}(x)P(w_j|x) \quad (9)$$

Since x will be assigned to only one group, let C(x) be the resultant cost. The objective of a decision maker is to minimize the total expected cost,

$$C = \int C(x)f(x)dx \quad (10)$$

Function C is minimized when each term C(x) is minimized, and that is accomplished by

$$\text{Decide } w_k \text{ for x if } C_k(x) = \min_i C_i(x)$$

The above is known as the Bayesian decision rule in classification.

A particular case for the rule is when the cost is binary: cij(x)=0 if i=j and 1 otherwise. The cost function Ci(x) can be simplified to

$$C_i(x) = \sum_{i\neq j} P(w_j|x) = 1 - P(w_i|x) \quad (11)$$

and the Bayesian decision rule is reduced to

$$\text{Decide } w_k \text{ for x if } P(w_k|x) = \max_i P(w_i|x)$$

Based on this information, it could be concluded that prior probabilities can be transformed into the posterior probabilities of group membership, which could further affect the classification of x. In this paper, the overall classification rate of the discriminant model was optimized by adjusting *a priori* probabilities. The threshold for the *a priori* classification probability was estimated by means of the receiver operating characteristics (ROC) curve [34].

Support Vector Machine (SVM)

SVM, a machine learning technique based on statistical theory, has been widely applied in various pattern recognition problems. The basic idea of the SVM algorithm for classification is mapping input vectors into a higher dimension, and then constructing a hyper-plane to separate these vectors into different classes with the maximal margin or the least error. In our case, the training set consisted of N samples or input vectors

$$\{x_1, x_2, \cdots\cdots, x_i, \cdots\cdots, x_N\}$$

with known class labels

$$\{y_1, y_2, \cdots\cdots, y_i, \cdots\cdots, y_N\} y_i \in \{-1, +1\}$$

The x_i corresponded to amino acid properties of query peptides and y_i represented glycosylation (+1) or non-glycosylation (-1), and N was the number of selected features. The decision function can be written as follows:

$$f(x) = \text{sgn}\left(\sum_{i=1}^{N} \gamma_i\alpha_i k(x, x_i) + b\right) \quad (12)$$

where k was the kernel function that defines the feature space; b was the bias value, α_i was the number obtained by solving the quadratic programming (QP) problem that gave the maximum margin hyper plane. The aim was to maximize α_i

$$0 \leq a_i \leq C$$

where C was the regulatory parameter controlling the tradeoff between the margin and training error. More details on SVM can be found in Vapnik's publication [35]. In this paper, SVM-based classification was achieved by LibSVM, an integrated software that is freely available at http://www.csie.ntu.edu.tw/cjlin/libsvm [36]. Features selected by SFS were used as descriptors. For the purpose of obtaining SVM classifier with optimal performance, the penalty parameter C and the RBF kernel parameter γ are tuned based on the training set using the grid search strategy in LibSVM.

Self-Organizing Competitive Neural Network (SOCNN)

SOCNN was an effective unsupervised artificial neural network method. It can learn and organize data items without being given desired outputs for input vectors. The network was composed of two layers: input layer and competitive layer. The number of input elements defined the number of input layer neurons, and the number of classes decided the competitive layer neuron number.

In this study, the selected amino acid properties were used to form the input vector while the glycosylation and non-glycosylation were two competitive layer neurons. No final cluster information was required in the classification. The transfer function used in the competitive layer was a winner-takes-all rule. Competitive layer neurons competed with each other to determine a winner. The Kohonen rule [37] was used to adjust the weight vector of the neuron that won in the competition. The weight vectors of other neurons were not adjusted. The weight vector for the winning neuron was updated using the following rule:

$$w_{i+1} = w_i + \eta(x_i - w_i) \tag{13}$$

where x_i is the training example, w_i is the current weight vector, w_{i+1} is the new weight vector, and η is a learning rate. The weight matrix stored the whole standard vector of every class, and the winning neuron showed the classification result [38].

Performance validation

The generated model was validated by the internal leave-one-out cross-validation (LOOCV) and external validations. LOOCV was used to evaluate the performance of the classifiers, as it was regarded as the most objective evaluation method. The accuracy between the predicted and observed result was assessed.

The predictive performance of the trained models may be overestimated due to the over-fitting of a training set. Therefore, the predictabilities were also evaluated using an external test set including 160 positive sites and 1322 negative sites from O-glycbase 6.0. In this work, PP-LDA, SVM and SOCNN adopted the same training set and test set for external prediction. The predictability was also tested on several newly identified glycosylated sites [27]. To avoid confusion, the test set from O-glycbase 6.0 was named as test set I and the test set from Zsuzsanna *et al.* [27] was named as test set II.

Accuracy (Ac), sensitivity (Sn) and specificity (Sp) were used to evaluate prediction systems. Sn, Sp and Ac were expressed in terms of true positive (TP), false negative (FN), true negative (TN) and false positive (FP) predictions. Each measurement was given as follows:

$$S_n = \frac{TP}{(TP + FN)} \tag{14}$$

$$S_p = \frac{TN}{(TN + FP)} \tag{15}$$

$$A_c = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{16}$$

Accuracy was the overall classification accuracy of a prediction model; it corresponded with the ratio of correctly classified compounds to the total compounds. Sensitivity was the ratio of glycosylation sites that were correctly predicted, whereas specificity was the ratio of non-glycosylation sites that were correctly predicted.

Results

For predicting glycosylation sites accurately with an unbalanced dataset as well as to find the essential amino acid properties for glycosylation, PP-LDA in combination with feature selection has been used in model building. A set of informative features was extracted from the original 4446*11046 pool for further study, and discriminant coefficients given by LDA were used to evaluate the importance of the features. Some other computer-aided recognition and classification techniques, such as SVM and SOCNN, were also employed to build models. Meanwhile, three datasets with different lengths of amino acid residues were used to evaluate the affect of adjacent amino acids to glycosylation.

Characterization of glycosylation sites

In this paper, we first conducted the statistic analysis of the glycosylated S and T sites. After the redundant fragments were

removed, neighboring amino acids (R_{10}-R_{10}') of the glycosylated or non-glycosylated S or T residues (at position R_0) were graphically visualized (Table 1). WebLogo [39] was applied to give a view of the graphical sequence logo for the relative frequency of the corresponding amino acid at each position around the target sites. According to the sequence logo representation, amino acids in the flanking regions were not obviously conserved while the glycosylated S and T sites were relatively conserved at position R_1-R_3'. This was consistent with the experiment data [15, 40] and no motif sequence was identified at the O-glycosylation sites. However, this consideration of the neighboring amino acids might be varied with an increasing number of identified glycosylation sites [41].

Feature selection

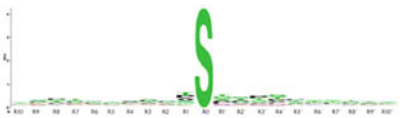
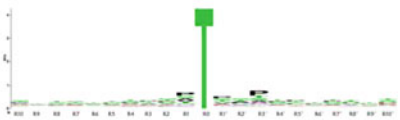
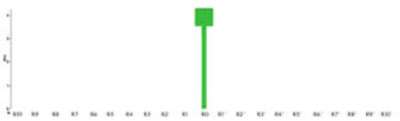
For the 21aa dataset (n=10), 58 features were selected from the training set using SFS, and they were presented in the form of a list with feature indices. According to the order they were selected, features were assigned as V_1, V_2, ……, V_{58} respectively. All selected indices were mapped back to the position and the index of the residue by Eqs. 3 and 4. Table 2 presented the detailed information about the selected 58 features, including the position and the index of features as well as their biological classes. The biological class of each feature can be downloaded from the website http://www.genome.jp/aaindex/AAindex/Appendix.

Model building and validation

PP-LDA classification

A model generated by the LDA algorithm was beneficial and simple to be interpreted with target descriptors, as it allowed interpreting individual descriptor contributions by the magnitude and sign of its discriminant coefficient. Therefore, the LDA classifier was performed for accurate glycosylation site prediction, as well as to identify which

Table 1 The statistics and sequence logos of nonhomologous glycosylated sites. The color table can be viewed in the online issue, which is available at wileyonlinelibrary.com

Residues	No.[3] of nonhomologous sites	No.[3] of proteins	Window lengths	Sequence logos
S(G.)[1]	239	108	R10-R10'	
S(N.G.)[2]	2641	113	R10-R10'	
T(G.)[1]	389	143	R10-R10'	
T(N.G.)[2]	2659	145	R10-R10'	

[1] *G* Abbreviation of glycosylation, [2] *NG* Abbreviation of nonglycosylation, [3] *No* Abbreviation of number

Table 2 The list for the features selected by SFS. The features were assigned as V_1, V_2……, V_{58} according to the order they selected and were represented as the corresponding position, index, standard coefficient and bioinformation class. The position and index of residues were calculated by the Eqs. 2 and 3 from the output list of SFS; the standard coefficient was the corresponding standard coefficient a_i to V_i given by PP-LDA. The bioinformation class of the features was obtained according to the AAindex website

Features	Position	Index	Standard coefficient	Classes
V_1	R_3'	369	-0.43	Alpha and turn propensities
V_2	R_1	488	0.23	Not defined
V_3	R_1'	445	0.24	Not defined
V_4	R_{10}'	91	0.41	Alpha and turn propensities
V_5	R_1	27	0.23	Physicochemical properties
V_6	R_8	72	0.00	Physicochemical properties
V_7	R_2'	27	0.15	Not defined
V_8	R_2	177	0.23	Physicochemical properties
V_9	R_1	194	-0.12	Composition
V_{10}	R_7'	350	-0.07	Alpha and turn propensities
V_{11}	R_{10}	112	0.18	Physicochemical properties
V_{12}	R_7	272	-0.15	Hydrophobicity
V_{13}	R_0	242	-0.17	Hydrophobicity
V_{14}	R_5'	319	0.19	Physicochemical properties
V_{15}	R_4'	350	-0.14	Alpha and turn propensities
V_{16}	R_3'	257	-0.15	Beta propensity
V_{17}	R_8'	273	-0.24	Hydrophobicity
V_{18}	R_3	95	0.12	Hydrophobicity
V_{19}	R_4	460	-0.08	Not defined
V_{20}	R_2'	23	0.11	Alpha and turn propensities
V_{21}	R_2	372	0.13	Other properties
V_{22}	R_1'	294	0.13	Alpha and turn propensities
V_{23}	R_9'	219	-0.10	Physicochemical properties
V_{24}	R_2	202	0.15	Hydrophobicity
V_{25}	R_{10}'	413	-0.17	not defined
V_{26}	R_5'	231	-0.21	Alpha and turn propensities
V_{27}	R_4'	188	0.24	Alpha and turn propensities
V_{28}	R_5	463	-0.08	Not defined
V_{29}	R_9'	333	0.08	Alpha and turn propensities
V_{30}	R_4'	272	-0.17	Hydrophobicity
V_{31}	R_5'	329	0.13	Hydrophobicity
V_{32}	R_{10}	201	0.14	Composition
V_{33}	R_8'	455	0.12	Not defined
V_{34}	R_8'	464	-0.09	Not defined
V_{35}	R_3'	386	-0.09	Hydrophobicity
V_{36}	R_6	41	-0.18	Hydrophobicity
V_{37}	R_6'	367	-0.07	Alpha and turn propensities
V_{38}	R_1	251	0.00	Beta propensity
V_{39}	R_{10}'	435	0.11	Not defined
V_{40}	R_1	438	0.11	Not defined
V_{41}	R_4	65	-0.07	Physicochemical properties
V_{42}	R_6	357	0.10	Hydrophobicity
V_{43}	R_2'	442	-0.07	Not defined
V_{44}	R_8	471	-0.15	Not defined
V_{45}	R_{10}	17	0.24	Physicochemical properties
V_{46}	R_9	405	-0.09	Not defined

Table 2 (continued)

Features	Position	Index	Standard coefficient	Classes
V_{47}	R_7'	21	0.09	Other properties
V_{48}	R_6	219	0.08	Physicochemical properties
V_{49}	R_1'	283	-0.11	Hydrophobicity
V_{50}	R_1'	371	0.08	Hydrophobicity
V_{51}	R_1	283	-0.14	Hydrophobicity
V_{52}	R_4'	313	-0.09	Hydrophobicity
V_{53}	R_8'	272	0.12	Hydrophobicity
V_{54}	R_7	334	-0.09	Alpha and turn propensities
V_{55}	R_2	493	-0.08	Not defined
V_{56}	R_2	17	0.07	Physicochemical properties
V_{57}	R_7	213	0.10	Hydrophobicity
V_{58}	R_5	155	0.07	Other properties

features characterized the glycosylation by using the selected 58 features as discriminating variables. However, in the case of O-glycosylation, a true distribution of glycosylation and non-glycosylation remained unknown. Moreover, due to the unbalanced dataset, the likelihood that a case belonged to the higher dispersion group can be increased. In this sense, *a priori* probabilities were adjusted to improve the overall classification rate from the discriminant model. In order to identify the proper value of *a priori* probability, the receiver operating characteristics (ROC) curve was adopted, which is a useful technique for obtaining the best thresholds for the *a priori* classification probability [34]. According to the ROC curve, the optimal threshold for predicting the glycosylation sites in our prediction model was 0.5 (Fig. 1). Moreover, this model was not random, but a truly statistically significant classifier, because the area under the ROC curve was 0.89, which was significantly larger than that of the random classifier curve (diagonal line).

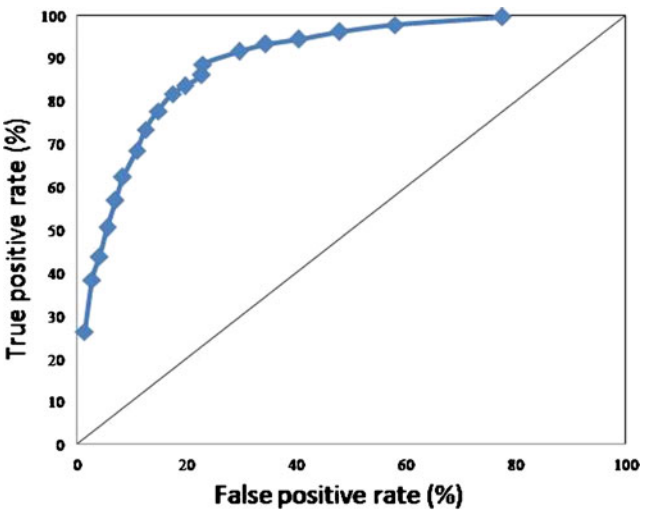


Fig. 1 Receiver operating characteristic (ROC) curve for the classification model Eq. 5

Based on the optimal *a priori* probability, the classification model derived from the training set was created by combining the LDA and SFS techniques. The p value was <0.001, indicating the model was statistically significance. From the classification results in Table 3, it was observed that, 382 out of 468 were correctly classified as glycosylation sites and 3297 out of 3978 non-glycosylation sites. The total accuracy, Sn and Sp for the training set were 82.7 %, 81.6 % and 82.9 %, respectively. This result suggested that the PP-LDA model could accurately classify the glycosylation sites from the original dataset, and created a linear relationship between the effective parameters and glycosylation.

One of the challenging problems in the classification of the unbalanced dataset was the aspect of over-fitting. When models become too powerful on the training set, they might not be useful for the classification of unseen data. Thus, the internal and external validations were applied for evaluating the PP-LDA models. We strictly divided the sites collected from the O-GlycBase 6.0 into a training set and test set I where the latter values are not shown to the classifier during the learning process. The 21 novel glycosylation sites, combined with their non-glycosylation sites from Zsuzsanna *et al.*, also evaluated the predictability of the model with unknown sites (test set II) [27]. Results are shown in Table 3. LOOCV using the training group showed that 369 out of 468 glycosylation sites (Sn=78.85 %) and 3281 out of 3978 non-glycosylation sites (Sp=82.5 %) were correctly classified. The LOOCV correct classification rate was 82.1 %. In the test set I, 118 out of 160 glycosylation sites (Sn= 73.75 %) and 1052 out of 1322 non-glycosylation sites (Sp=79.5 %) were correctly classified. The accuracy for newly reported sites (test set II) was 80.3 %, with the Sn and Sp being 80.95 % and 80.3 %, respectively. These internal and external validation results indicated that the classifier was not only robust, but also has good predictability for unknown glycosylation sites.

Table 3 The classification and performance results of the model built with PP-LDA for the dataset with 21 amino acid residues using 58 features

21aa			Predicted group membership		Total accuracy (%)
			N.G.[1]	G.[2]	
Original	Count	N.G.	3297	681	82.7
		G.	86	382	
	%	N.G.	82.9	17.1	
		G.	18.4	81.6	
Cross-validated	Count	N.G.	3281	697	82.1
		G.	99	369	
	%	N.G.	82.5	17.5	
		G.	21.15	78.85	
Test set I	Count	N.G.	1051	271	
		G.	42	118	
	%	N.G.	79.5	20.5	78.9
		G.	26.25	73.75	
Test set II	Count	N.G.	452	111	
		G.	4	17	
	%	N.G.	80.3	19.7	80.3
		G.	19.05	80.95	

[1]*NG* nonglycosylation, [2]*G* glycosylation

SVM classification

In this paper, the SVM was utilized to predict glycosylation. The performance of the classifier is shown in Fig. 2. For the 21aa dataset, with the penalty parameter C and the RBF kernel parameter γ were 32768 and $1.22*10^{-4}$, the SVM model showed high accuracy values of 89.5 % and 89.7 % for the training and test set I, respectively. Further analysis of the Sp and Sn for test set I indicated that the Sn was only 15 % (25 out of 160 sites correctly classified) though the Sp was as high as 98.7 %. The relatively high prediction of accuracy and Sp, comparing with low Sn, indicated that after being trained, the hyperplane outputs of SVMs had grasped the complicated relationship between the effective parameters and the major non-glycosylation group, but not the minor glycosylation group. Considering the unsatisfied Sn was probably caused by the unbalanced dataset, we further randomly omitted some negative sites in the training set and the test set I. This resulted in 468 glycosylation sites/531 non-glycosylation sites for the training dataset and 140 glycosylation sites/374 non-glycosylation sites for test set I. The best total accuracy for the balanced training datasets was 80.98 %. The Sn and Sp for the test set were 77.14 % and 77.29 %, when the penalty parameter C and the RBF kernel parameter γ were 32 and $7.81*10^{-3}$. According to the comparison of results of the balanced dataset and the unbalanced dataset, the SVM method was good for predicting glycosylation with the balanced dataset but not the unbalanced dataset.

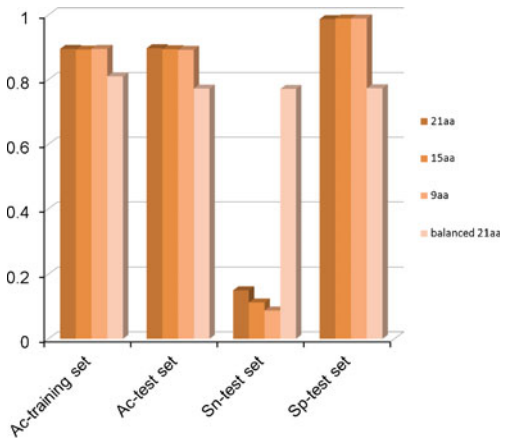


Fig. 2 The SVM classification result for glycosylation with unbalanced dataset and the balanced dataset . 21aa: Abbreviation of dataset with 21 amino acid residues; 15aa: Abbreviation of dataset with 15 amino acid residues; 9aa: Abbreviation of dataset with 9 amino acid residues

SOCNN classification

SOCNN, another machine learning algorithm, was also used for model building. Table 4 illustrates the number of cases that were mapped to each cluster as well as the performance evaluation of this classifier. The total accuracy was 66.06 % and 66.3 % for the training set and test set I, respectively. The Sn and Sp for test set I was 45.6 % and 68.8 %, respectively. Though the Sn was higher than the SVM classification, the total accuracy and Sp were low compared with other reported algorithms. As SOCNN was unsupervised it did not provide information on the final glycosylation, and the classification result might only be a reflection of differences among the amino acid properties between two clusters, and not a reflection on the glycosylation information.

Table 4 The SOCCN classification result for glycosylation with unbalanced dataset

21aa			Predicted group membership		Total accuracy (%)
			N.G.[1]	G.[2]	
Original	Count	N.G.	2742	1236	66.06
		G.	273	195	
	%	N.G.	68.9	31.1	
		G.	58.3	41.7	
Test set I	Count	N.G.	910	412	
		G.	87	73	
	%	N.G.	68.8	31.2	66.3
		G.	54.4	45.6	

[1]*NG* nonglycosylation, [2]*G* glycosylation

PP-LDA model built with different datasets

Comparing the classification results with SVM and SOCNN (Fig. 3), the LDA model with an equal *a priori* probability exhibited robust predictability for preliminary classification of glycosylation sites according to their neighboring amino acid properties even with an unbalanced dataset. Therefore, further models were built with 15aa and 9aa datasets to investigate the effect on the length of the neighboring amino acids. Both the 15aa and the 9aa datasets were analyzed using the same process as the 21aa dataset. From the 15aa and the 9aa, 41 and 30, feature indices were respectively selected, and the corresponding LDA classification results are shown in Fig. 4. All of the selected indices were mapped back to the position and the index of the residue by Eqs. 3 and 4. Comparing the classification results with different amino acid residue datasets, the accuracy increased as more amino acid residues were taken into account. Thus, the four neighboring amino acids played a pivotal role for glycosylation prediction, while the comprehensive understanding of glycosylation was also supported by the knowledge of the role of the amino acid in the position R_{10} (')-R_5 (').

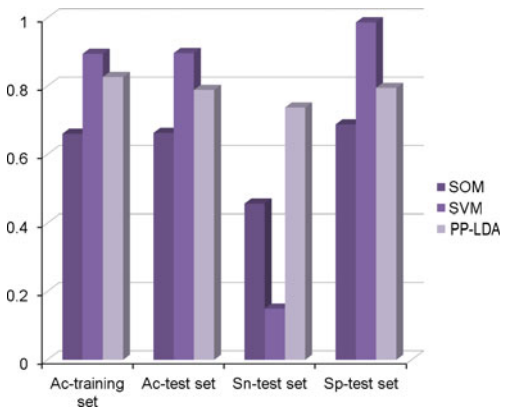


Fig. 3 Comparison of the classification results of SOM, SVM and PP-LDA for the imbalanced dataset with 21 amino acid residues

Feature analysis

Further analyses of selected features were performed to investigate the effect of neighboring amino acids. The standardized discriminant function coefficients for each feature are listed in Table 2, showing how much each individual predictor adds to the LDA. The larger the standardized discriminant function coefficients, the more related the feature is to glycosylation. According to the coefficients, it was suggested that the V_1 (the normalized frequency of chain reversal R at position R_3'), V_4 (helix-coil equilibrium constant at R_{10}'), V_3 (propensity of amino acids within pi-helices at position R_1'), V_{27} (normalized frequency of coil) and V_{45} (alpha-CH chemical shifts) had the biggest contributions. However, there are still 20 features whose absolute values of coefficients were less than 0.1, and were considered to have no classification ability. Excluding these 20 features, we repeated the LDA classification based on the remaining 38 features, and the results are presented in Table 5. Approximately 80.1 % of the glycosylation sites and 81.6 % of the non-glycosylation sites were correctly classified. The LOOCV

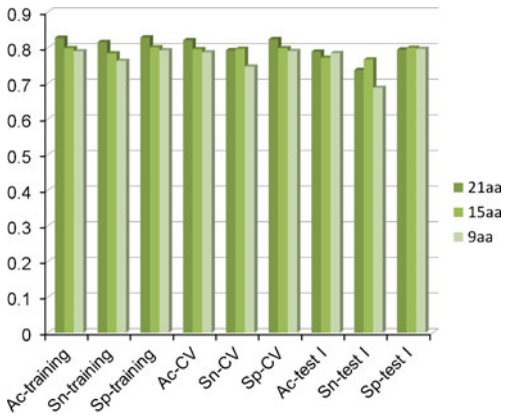


Fig. 4 The classification and performance results of the PP-LDA model built with different datasets. CV: Abbreviation of cross validation; 21aa: Abbreviation of dataset with 21 amino acid residues; 15aa: Abbreviation of dataset with 15 amino acid residues; 9aa: Abbreviation of dataset with 9 amino acid residues

Table 5 The classification and performance results of the model built with PP-LDA for the dataset with 21 amino acid residues using 38 features

21aa			Predicted group membership		Total accuracy (%)
			N.G.[1]	G.[2]	
Original	Count	N.G.	3245	733	81.4
		G.	93	375	
	%	N.G.	81.6	18.4	
		G.	19.9	80.1	
Cross-validated	Count	N.G.	3239	739	81.1
		G.	102	366	
	%	N.G.	81.4	18.6	
		G.	21.8	78.2	
Test set I	Count	N.G.	1052	270	
		G.	45	115	
	%	N.G.	79.6	20.4	78.7
		G.	28.1	71.9	

[1]*NG* nonglycosylation, [2]*G* glycosylation

correct classification rate was 81.1 %. In the test set I, 115 out of 160 glycosylation sites (Sn=71.9 %) and 1052 of 1322 non-glycosylation sites (Sp=79.6 %) were correctly classified, with the total accuracy reaching 78.7 %. The classification results were comparable with the results from the model with 58 features, which further indicated that the discarded features had little contribution to the classification. Thus, further feature analysis was conducted based on the 38 features.

The analysis of the 38 features was conducted based on the distribution of their position combination with the coefficient. According to the position distribution and coefficients of the 38 features (Fig. 5), we can observe that the properties at position R_1 contributed more to the glycosylation, following with R_{10}', R_3', R_1', R_5', R_2, R_{10} and R_8', while R_9, R_5, R_4 and R_6' positions contributed the least to glycosylation. The biological class analysis of the selected features was studied based on the classification in the AAindex. In total, 402 out of the 526 features were clustered into six groups: the alpha and turn propensities, the beta propensity, the composition, the physicochemical properties, the hydrophobicity and others. The remaining 124 features were not defined. According to our analysis, 34.2 %, 18.4 % and 15.8 % of the selected features were hydrophobicity related, alpha and turn propensity related and physicochemical related properties, respectively. The hydrophobicity related properties at site R_7-R_8' are essential for glycosylation, especially at position R_6, R_0, R_1' and R_8' (Fig. 6). Meanwhile, the roles of the alpha and turn propensities and physicochemical properties cannot be ignored in glycosylation analysis. All alpha and turn propensity related properties were distributed at the C-terminal while 66.7 % of the physicochemically related features were at the N-terminal. Therefore, we speculated that the secondary structure at the C-terminal as well as physicochemical properties at the N-terminal was also playing a role in glycosylation. Besides, 21.0 % of the selected features were thought to affect glycosylation, while the AAindex website did not give a classification. Due to the lack of information on these properties, we did not conduct further analysis. Also, as shown in Figs. 5 and 6, there are many other factors influencing the glycosylation besides the hydrophobicity and secondary structure, which maybe caused by the diversity of the enzyme participating in the process. Recently, about 20 different GalNAc-transferases were reported to be involved in the mediation of O-glycosylation [16]. Therefore, the more precise prediction of the glycosylation site was dependent on the detailed study of enzymes.

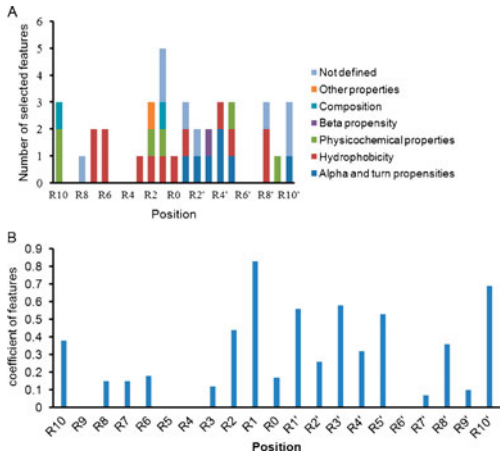


Fig. 5 The position distribution and coefficients of the selected features and the frequencies of different classes of the selected features according to their biological meanings. **a** The position distribution of the selected features and the frequencies of different classes of the selected features according to their biological meanings. **b** The coefficients of the selected features at different positions. The data are given in the form of the absolute value of coefficients

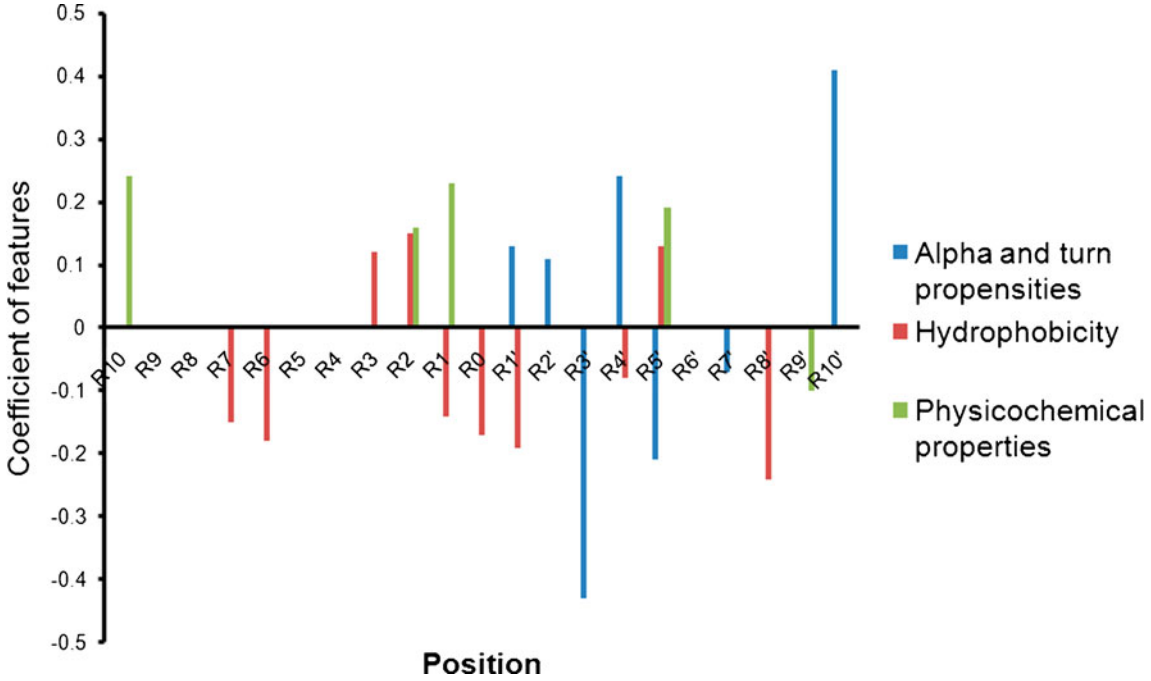


Fig. 6 The position distribution and relative contribution of the top 3 classes of features according to their biological meanings. If more than 1 features of a certain class were selected at a certain position, the coefficient with the largest absolute value was used here to describe the contribution of this class at this position

Discussion

To date, the experimental analysis of glycosylation sites is still a challenging area due to the diverse structural modifications ranging from a few monosaccharide residues to heavily branched oligosaccharides. In organisms, the glycosylation of the protein was affected by various factors, including the neighboring O-glycans and their carbohydrate structure, the enzyme involved, and the tissue where the protein in located. Computational methods are efficient alternatives for the study of glycosylation. However, these methods were unable to consider all possible factors. In our paper, we were only concerned with whether a site could be glycosylated according to its sequence, while the status of its neighboring environment was not considered. When predicting glycosylation from the sequence, a number of obstacles were encountered. The unbalanced datasets created a well-known machine learning challenge where classifiers tended to become biased towards the majority class [25], while the non-interpretability of the present models makes it difficult to understand the underlying information for the study of glycosylation. Moreover, though groups have attempted to predict the glycosylation sites simply according to the amino acid sequences and their properties, the number of amino acid properties is almost unlimited, creating a computational challenge for the hardware. A hypothesis that we proposed here was that, the prediction ability was only related with a small number of amino acid properties. Based on our hypothesis, the SFS method was initially used to select the features of critical importance for glycosylation, and then the LDA was applied for model building and to develop a primary view of the contribution of each feature. However, the LDA model for glycosylation does not prevent the bias for the majority class. In this sense, adjusting the *a priori* probabilities can greatly improve the overall classification rate of the discriminant model (Fig. 1).

Large datasets presented a computational challenge for the hardware (such as the demand for memory) as well as the classification algorithm (most algorithms could not handle enormous amounts of data). Previous studies employed different methods to solve this problem. For example, Cai *et al.* used the mRMR (maximum relevance, minimum redundancy) method combined with feature selection methods to reduce the dimensionality of the dataset. This method has been successfully applied for the prediction of PTMs, such as mucin O-glycosylation prediction [42] and the protein palmitoylation site prediction [43]. However, the mRMR only gave pre-evaluated features in the original feature pool and it does not have the ability to select the most appropriate features. Therefore, other selection methods are required. The SFS method can directly provide information about how many and which features should be selected, reducing the time for computation. Through this step, the dimensionality was largely reduced as only 58 features were selected. This result made a satisfied classifier performance possible, and focused our attention to limited features with useful biological knowledge for further study [44].

Based on the SFS selected features, the LDA model was successfully created using an unbalanced dataset, which compared with the SVM and the SOCNN models. The LDA has been used widely in many applications including cancer research [34], face recognition [45] and microarray data classification [46]. In this paper, it was applied for the glycosylation classification and an improved performance was achieved by adjusting the *a priori* probabilities. We focused the glycosylation prediction methods on the ensemble SVM approach presented by Caragea *et al.* [26], which is the only glycosylation model created using unbalanced datasets. The data used in the ensemble SVM model and ours were both from O-GlycBase v6.00. Differences between the two models can be outlined as the encoding method, the modeling method and the performance evaluation. In the ensemble SVM model, the classification result was still an integrated result of m individual SVM classifiers trained with a balanced subsample. The positive sites were repeatedly used while the negatives were not, and the model was only validated internally. In our model, the classification model was trained with an unbalanced dataset and validated externally. The predicted accuracy for ensemble SVM was 89 % while the sensitivity was only 68 % [26]. Comparing with the ensemble SVM, the accuracy in our method can reach 82 % by adjusting the *a priori* probabilities. Though the accuracy in our method was lower than the ensemble SVM, the Sn was much higher (78.85 % *vs.* 68 %). Moreover, the Sn for external validation (test set II) reached as high as 80.95 %, which further validated the reliability of our model. Considering that Sn is the percentage of observed positives that are correctly predicted, these results indicated the PP-LDA model can extract positive sites from unknown proteins.

Our model, not only gives significant predictability, but can also be useful in understanding some of the underlying biological knowledge for glycosylation as shown in Figs. 5 and 6. The amino acid residue at R_1 position contributed more to whether the site was glycosylated, while the other positions contributed to glycosylation with varying degrees (Fig. 5). Previous papers reported the position basis to glycosylation, for example, Aruto Yoshida *et al.* observed that five amino acids from the position R_1 to R_3' were regarded as compulsory for glycosylation [15]; O'Connel *et al.* found that positions -1 (R_1), and +3 (R_3') were of particular significance [40]. The results we obtained were consistent with other experiment results, which further verified the rationality of our model. Further frequency analysis of different classes of features indicated that the hydrophobicity was essential for glycosylation activity. A possible reason for this is that hydrophilic residues are more likely to be located on the surface of the proteins. Therefore, the hydrophilic residues such as Arg and Lys in such positions will contribute to glycosylation. The alpha and turn propensities at the C-terminal and physicochemical properties at N-terminal also play a role in the glycosylation activity. Therefore, a greater tendency to glycosylation might be shared in an alpha or turn conformation in the C-terminal, to some extent, by Met, Gla, Phe, Arg and Leu residues [47]. Besides, 21.0 % of the selected features were thought to play the important role for glycosylation, while the AAindex website did not give a classification. Though we did not analyze these properties in this paper, further understanding about these features might be of great importance for glycosylation studies.

In conclusion, a linear interpretable prediction model (PP-LDA) was created to aid glycosylation prediction. Based on the selected features, the PP-LDA classification model exhibited acceptable predictability for an unbalanced dataset that accuracies were 82.7 % and 78.9 % for the training and the external test set I, respectively. Moreover, this model exhibited 80.3 % accuracy for unknown glycosylation sites (test set II). Further analysis of selected features was conducted, which indicated that properties at position R_1 and properties relating to hydrophobicity contributed more to the prediction. The alpha and turn propensity properties and the physicochemical properties displayed an obvious preference in the C-terminal and N-terminal, respectively. This newly developed interpretable model, not only provided an effective method to solve issues for the prediction of glycosylation sites with an unbalanced dataset, but also helped to exploit the potential biological information for further understanding the mechanism of glycosylation. Considering the publicly accessibility of our prediction model, a downloadable program is provided in our supply materials.

Acknowledgments Thanks to Mr. Chuanliang Li for technical assistance in programing. This work was supported by the National High Technology Research and Development Program of China (863 Program) (No. 2009AA02Z205) , the National Nature Science Foundations of China (No. 81072698 and 81173124) and the Key Direction Project of the Chinese Academy of Sciences (approved No. KSCX2-YW-G-050).

References

1. Sola, R.J., Rodriguez-Martinez, J.A., Griebenow, K.: Modulation of protein biophysical properties by chemical glycosylation: biochemical insights and biomedical implications. Cell. Mol. Life Sci. **64**(16), 2133–2152 (2007)
2. Geyer, H., Geyer, R.: Strategies for analysis of glycoprotein glycosylation. BBA-Proteins Proteom **1764**(12), 1853–1869 (2006)
3. Gupta, R., S. Brunak: Prediction of glycosylation across the human proteome and the correlation to protein function. Pac. Symp. Biocomput. 310–322 (2002)
4. Hart, G.W.: Glycosylation. Curr. Opin. Cell Biol. **4**(6), 1017–1023 (1992)
5. Ohtsubo, K., Marth, J.D.: Glycosylation in cellular mechanisms of health and disease. Cell **126**(5), 855–867 (2006)
6. Li, M., Song, L.J., Qin, X.Y.: Glycan changes: cancer metastasis and anti-cancer vaccines. J. Biosciences. **35**(4), 665–673 (2010)
7. Gong, C.X., *et al.*: Post-translational modifications of tau protein in Alzheimer's disease. J. Neural Transm. **112**(6), 813–838 (2005)

8. Saeland, E., van Kooyk, Y.: Highly glycosylated tumour antigens: interactions with the immune system. Biochem. Soc. Trans. **39**, 388–392 (2011)
9. Christlet, T., Veluraja, K.: Database analysis of O-glycosylation sites in proteins. Biophys. J. **80**(2), 952–960 (2001)
10. Walsh, G., Jefferis, R.: Post-translational modifications in the context of therapeutic proteins. Nat. Biotechnol. **24**, 1241–1252 (2006)
11. Blom, N., *et al.*: Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics **4**(6), 1633–1649 (2004)
12. Elhammer, A., *et al.*: The specificity of UDP-GalNAc:polypeptide N- acetylgalactosaminyltransferase as inferred from a database of *in vivo* substrates and from the *in vitro* glycosylation of proteins and peptides. J. Biol. Chem. **268**, 10029–10038 (1993)
13. Wilson, B., Gavel, Y., von Heijne, G.: Amino acid distributions around O-linked glycosylation sites. Biochem. J. **275**, 529–534 (1991)
14. Oconnell, B.C., Hagen, F.K., Tabak, L.A.: The influence of flanking sequence on the O-glycosylation of threonine *invitro*. J. Biol. Chem. **267**(35), 25010–25018 (1992)
15. Yoshida, A., *et al.*: Discovery of the shortest sequence motif for high level mucin-type O-glycosylation. J. Biol. Chem. **272**(27), 16884–16888 (1997)
16. Jensen, O.N.: Interpreting the protein language using proteomics. Nat. Rev. Mol. Cell Biol. **7**(6), 391–403 (2006)
17. Chou, K.: A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. Protein Sci. **4**, 1365–1383 (1995)
18. Lu, L., *et al.*: GalNAc-transferase specificity prediction based on feature selection method. Peptides **30**(2), 359–364 (2009)
19. Eisenhaber, B., Eisenhaber, F.: Prediction of posttranslational modification of proteins from their amino acid sequence. Methods Mol. Biol. **609**, 365–384 (2010)
20. Chou, K., *et al.*: A vector projection method for predicting the specificity of GalNAc-transferase. Proteins **21**, 118–126 (1995)
21. Hansen, J., *et al.*: Prediction of O-glycosylation of mammalian proteins: specificity patterns of UDP-GalNac:polypeptide N-acetylgalactosaminyltransferase. Biochem. J. **308**, 801–813 (1995)
22. Li, S., *et al.*: Predicting O-glycosylation sites in mammalian proteins by using SVMs. Comput. Biol. Chem. **30**(3), 203–208 (2006)
23. Julenius, K., *et al.*: Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. Glycobiology **15**, 153–164 (2005)
24. Chen, Y.-Z., *et al.*: Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. BMC Bioinformatics **9**(1), 101 (2008)
25. Weiss, G.M., Provost F.: The effect of class distribution on classifier learning Technical Report ML-TR-44, (2001)
26. Caragea, C., *et al.*: Glycosylation site prediction using ensembles of Support Vector Machine classifiers. BMC Bioinformatics **8**(1), 438 (2007)
27. Darula, Z., Medzihradszky, K.F.: Affinity enrichment and characterization of mucin core-1 type glycopeptides from bovine serum. Mol. Cell. Proteomics **8**(11), 2515–2526 (2009)
28. Kawashima, S., Kanehisa, M.: AAindex: amino acid index database. Nucl. Acids. Res. **28**(1), 374 (2000)
29. Kawashima, S., *et al.*: AAindex: amino acid index database, progress report 2008. Nucl. Acids. Res **36**(suppl_1), D202–D205 (2008)
30. Tomii, K., Kanehisa, M.: Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng. **9**(1), 27–36 (1996)
31. Eklöv, T., Mårtensson, P., Lundström, I.: Selection of variables for interpreting multivariate gas sensor data. Anal. Chim. Acta **381**(2–3), 221–232 (1999)
32. Xu, L., Zhang, W.-J.: Comparison of different methods for variable selection. Anal. Chim. Acta **446**(1–2), 475–481 (2001)
33. Gualdrón, O., *et al.*: Coupling fast variable selection methods to neural network-based classifiers: application to multisensor systems. Sens. Actuator. B-Chem. **114**(1), 522–529 (2006)
34. Morales Helguera, A., *et al.*: Probing the anticancer activity of nucleoside analogues: a QSAR model approach using an internally consistent training set. J. Med. Chem. **50**(7), 1537–1545 (2007)
35. Vapnik, V.N.: An overview of statistical learning theory. IEEE Trans. Neural Netw. **10**(5), 988–999 (1999)
36. Chih-Chung Chang, C.-J.L.: LIBSVM: a library for support vector machines. (2001)
37. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biol. Cybern. **43**(1), 59–69 (1982)
38. Xu, P., Xu, S.J., Yin, H.W.: Application of self-organizing competitive neural network in fault diagnosis of suck rod pumping system. J. Pet. Sci. Eng. **58**(1–2), 43–48 (2007)
39. Crooks, G.E., *et al.*: WebLogo: a sequence logo generator. Genome Res. **14**(6), 1188–1190 (2004)
40. Oconnell, B., Tabak, L.A., Ramasubbu, N.: The influence of flanking sequences on O-glycosylation. Biochem. Biophys. Res. Commun. **180**(2), 1024–1030 (1991)
41. Liu, B., *et al.*: Predicting the protein SUMO modification sites based on Properties Sequential Forward Selection (PSFS). Biochem. Biophys. Res. Commun. **358**(1), 136–139 (2007)
42. Cai, Y., He, J., Lu, L.: Prediction of mucin-type O-glycosylation sites by a two-staged strategy. Mol. Divers. **15**(2), 427–433 (2011)
43. Hu, L.L., *et al.*: Prediction and analysis of protein palmitoylation sites. Biochimie **93**(3), 489–496 (2011)
44. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. **3**, 1157–1182 (2003)
45. Chen, L.F., *et al.*: A new LDA-based face recognition system which can solve the small sample size problem. Pattern. Recogn. **33**, 1713–1726 (2000)
46. Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc. **97**(457), 77–87 (2002)
47. Jiang, B., *et al.*: Folding type-specific secondary structure propensities of amino acids, derived from alpha-helical, beta-sheet, alpha/beta, and alpha+beta proteins of known structures. Biopolymers **45**(1), 35–49 (1998)