Med Biol Eng Comput (2015) 53:535–544
DOI 10.1007/s11517-015-1268-9

ORIGINAL ARTICLE

Prediction of *O*-glycosylation sites based on multi-scale composition of amino acids and feature selection

Yuan Chen[1] · Wei Zhou[2] · Haiyan Wang[3] · Zheming Yuan[1]

Received: 22 April 2014 / Accepted: 2 March 2015 / Published online: 10 March 2015

Abstract Protein glycosylation is one of the most important and complex post-translational modification that provides greater proteomic diversity than any other post-translational modification. Fast and reliable computational methods to identify glycosylation sites are in great demand. Two key issues, feature encoding and feature selection, can critically affect the accuracy of a computational method. We present a new *O*-glycosylation sites prediction method using only amino acid sequence information. The method includes the following components: (1) on the basis of multi-scale theory, features based on multi-scale composition of amino acids were extracted from the training sequences with identified glycosylation sites; (2) perform a two-stage feature selection to remove features that had adverse effects on the prediction, including a stage one preliminary filtering with Student's *t* test, and a second stage screening through iterative elimination using novel pairwise comparisons conducted in random subspace using support vector machine. Important features retained are used to build prediction model. The method is evaluated with sequence-based tenfold cross-validation tests on balanced datasets. The results of our experiments show that our method significantly outperforms those reported in the literature in terms of sensitivity, specificity, accuracy, Matthew's correlation coefficient. The prediction accuracy of serine and threonine residues sites reached 95.7 and 92.7 %. The Matthew correlation coefficient of our method for *S* and *T* sites is 0.914 and 0.873, respectively. This method can evaluate each feature with the interactions of the rest of the features, which are still included in the model and have the advantage of high efficiency.

Keywords *O*-glycosylation · Multi-scale composition of amino acids · Paired comparison through random subspace screening · Support vector machine

1 Introduction

Protein glycosylation is the process of attaching a branched oligosaccharide to a protein through a glycopeptide linkage and is controlled by a large genes family [1]. As one of the most important and complicated protein post-translational modifications, protein glycosylation occurs frequently at the cell surface, on intercellular substances, in the Golgi apparatus, in plasma and in mucus [22, 32, 36, 39]. It directly affects the function of proteins and is involved in numerous biological processes [14, 15, 33]. *N*-linked glycosylation (*N*-glycosylation) and *O*-linked glycosylation (*O*-glycosylation) are two major types of glycosylation. The consensus sequence motif Asn-X-Ser/Thr (X can be any amino acid except Pro) is essential in *N*-glycosylation [3]. *O*-glycosylated proteins are also called mucin-type glycoproteins, and they are usually modified at the site of Ser (*S*)

Electronic supplementary material The online version of this article (doi:10.1007/s11517-015-1268-9) contains supplementary material, which is available to authorized users.

✉ Zheming Yuan
zhmyuan@sina.com

[1] Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization, Hunan Agricultural University, Changsha 410128, China

[2] Hunan Provincial Key Laboratory for Biology and Control of Plant Diseases and Insect Pests, Changsha 410128, China

[3] Department of Statistics, Kansas State University, Manhattan, KS, USA

or Thr (*T*) without a consensus motif [23]. The identification of glycosylation sites in this study is only for *O*-glycosylation.

Aberrations in *O*-glycosylation are responsible for certain human diseases and are associated with disease risk factors. Recent studies have demonstrated essential roles for mucin-type *O*-glycosylation in protein secretion, stability, processing and function [37]. The identification of *O*-glycosylation sites is very useful for us to understand its biological structure and function. The development of genomic and proteomic methods has facilitated the rapid experimental identification of glycosylation sites. Therefore, high-accuracy computational prediction methods are in great demand to analyse these massive data. A sequence-based vector-projection model was developed to predict *O*-glycosylation sites during the early stages of model development [11]. Furthermore, neural networks and support vector machines (SVM) were also popular strategies to predict protein modification sites [6, 7, 10, 27], and there were a series of predictive tools for predicting *O*-glycosylation sites, such as NetOGlyc [23], Oglyc [27] and CKSAAP-OGlySite [10]. The prediction accuracy of these methods is generally below 0.86 and is unbalanced for positive and negative sequences.

The input feature vector is a key component for obtaining a prediction model based on a machine learning algorithm. A 2n + 1-residue-long sequence with S/T in the centre is used as the input for an *O*-glycosylation site predictor. The encoding scheme for the input feature includes binary encoding [6, 23, 27], second structure information encoding [16], evolutionary information encoding [19] and so on. Due to a deficiency in annotation of information for secondary structures, its practical application is limited. Subsequently, encoding schemes based only on the characteristics of sequences have been developed, such as the composition of k-spaced amino acid pairs [10], as well as the physical and chemical properties of amino acids [27]. When the feature vector has a high dimensionality, it usually contains redundant information. Feature selection is another key issue for computational identification of *O*-glycosylation sites. Information entropy (IE) and correlation coefficient (CC) were used for feature selection by Chen et al. [10]; however, the IE and CC feature selections did not significantly improve prediction accuracy. A new method is proposed in this article to detect *O*-glycosylation sites with the aim of jointly tackling the challenge of feature encoding and feature selection. The proposed method performs feature encoding by multi-scale composition of amino acids (MSCAA) and performs feature selection through two-stage filtering, one using the Student's *t* test and the other through a paired comparison via random subspace screening (PCRSS). This new feature selection method has significantly improved the accuracy.

2 Methods

2.1 Positive datasets

The *O*-glycosylation sites of proteins were collected from the Swiss-Prot database [10], including 116 *S* and 212 *T* sites from 103 mammalian protein sequences. Each *O*-glycosylation site is represented by a sequence fragment of 41 amino acids with S/T in the centre. We used a non-existing amino acid named O to fill in the N- or C-terminus where the number of upstream or downstream residues may be <20.

2.2 Negative datasets

We selected all *S* and *T* residues in these 103 protein sequences without annotation as negative sites and got 1506 *S* residues and 2529 *T* residues. Samples containing 116 *S* sites and 212 *T* sites were selected randomly each time to form datasets in which the ratio of positive to negative sequences was 1:1.

2.3 The support vector machine

SVM is a machine learning algorithm based on statistical learning theory [38]. It includes support vector classification (SVC) and support vector regression (SVR). It is based on structural risk minimisation and has solved the defects in conventional machine learning algorithms, such as the over-fit, local minimum, the curse of dimensionality. It also allows nonlinear modelling and strong generalisation ability [42]. SVM has been widely applied to the field of bioinformatics [12, 13, 18–21, 28, 34, 43]. In this study, we adopt the SVM in a toolkit named LIBSVM whose parameters can be automatically set to the best value [9]. The radial basis function (RBF) is selected as the kernel function. The instructions for LIBSVM can be found at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

2.4 Tenfold cross-validations

Cross-validation can assess the utility of the prediction method. There are different kinds of cross-validation according to the size of the dataset (fivefold cross-validation, tenfold cross-validation, leave-one-out, etc.). In this study, the SVM prediction method was tested using tenfold cross-validation. Firstly, a dataset is partitioned into ten subsets of equal sizes. Nine subsets were used to construct the training set and predict the remaining one until all the ten subsets had been predicted.

2.5 Evaluating indicator

Accuracy (Ac), sensitivity (SN), specificity (SP), Matthew's correlation coefficient (MCC) and the ROC were used to evaluate the prediction performance. They are given as follows:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\ \% \tag{1}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \times 100\ \% \tag{2}$$

$$\mathrm{Ac} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \times 100\ \% \tag{3}$$

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FN} \times \mathrm{FP})}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}} \tag{4}$$

Here, TP, TN, FP, FN denote true positives, true negatives, false positives and false negatives, respectively. Sn and Sp can be understood as the accuracy of positive and negative sequences. When there is a big difference between Sn and Sp, then the accuracy cannot truly reflect the predictive ability of a method and MCC can be used as an alternative. Greater Ac and MCC represent a better prediction ability of a method.

As a widely used evaluation method [2, 8], the ROC curve plots true-positive rate (Sn) as a function of false-positive rate (1 − Sp) for all possible thresholds. The area under the ROC curve (AUC) is analysed to give a comprehensive metric for evaluating prediction methods. The closer value of AUC is to 1, the better the prediction method is.

2.6 Feature coding: multi-scale composition of amino acids (MSCAA)

In the coding scheme of 0/1 system, each amino acid is coded by a binary vector. For common 20 amino acids and the filler of 'O', a 21-dimensional vector only composed of 0 and 1 is constructed to represent these residues. For example, amino acid A is represented with 000000000000000000001 and C is represented with 000000000000000000010, and similarly for other amino acids. Under this coding scheme, each sequence fragment of 40 residues (exclude the S/T in the centre) is coded by 40 × 21 dimension vectors. This coding scheme is used commonly in the literature [6, 25, 28]. However, it only accounts for single amino acid and completely ignores the relationship of amino acids at different locations in a sequence. To consider pairwise relationship, the coding scheme of k-spaced amino acid pairs CKSAAP is proposed by Chen et al. [10]. For 21 amino acids, there are 441 possible amino acid pairs (AA, AC, AD,…, OO) in each sequence fragment. The k-spaced amino acid pairs are the pairs that are separated by k other amino acids. If the pairs for $k = 0, 1, \ldots, k\mathrm{max}$ are jointly considered, the total dimension of the CKSAAP feature vector is $441 \times (k\mathrm{max} + 1)$.

To consider contribution of further relationship between more than two amino acids, we propose a new sequence coding scheme named multi-scale composition of amino acids in this paper. Let k be the scale of the composition. The composition of k amino acids $\alpha_1\,\alpha_2 \ldots \alpha_k$ is a fragment, where α_i is one of the 21 amino acid residues (including the O residue used to complete the terminus of a sequence). For integer k between 1 and 3, there are total 9723 multi-scale composition fragments, among which 21 are single amino acid, 441 double amino acids fragments and 9261 fragments each with three amino acids. Multi-scale theory is more reasonable than single-scale theory on the study of the object because different scales of amino acid composition can reveal interactions between different positions of the sequence. For a sequence of 41 amino acids with S/T in the centre, let $f(\alpha_1\,\alpha_2 \ldots \alpha_k)$ denote the frequency of the composition of k amino acid fragments of $\alpha_1\,\alpha_2 \ldots \alpha_k$ appearing in the sequence. We use $f(\alpha_1\,\alpha_2 \ldots \alpha_k)$ as a possible feature for predicting O-glycosylation sites. For the value of k ranging from 1 to 3, there are 9723 multi-scale composition features.

The above-mentioned three encodings are sparse, and another non-sparse encoding named AA531 was used. In the coding scheme of AA531, 531 physicochemical and biochemical properties of amino acids [24] were used to code each amino acid. Each sequence fragment is coded by 41 × 531 dimension vectors.

2.7 Feature selections

In addition to the high dimensionality of the feature space, the signal of MSCAA and CKSAAP features is sparse. That is, most features are 0 since the fragment of a sequence is very short. Including a large number of near-zero features hinders accurate prediction. Therefore, it is necessary to conduct a feature selection. In this study, we propose using a two-stage filtering in which the composition features are filtered initially by a Student's t test [40] and then by a new method of feature selection named paired comparison through random subspace screening (PCRSS). The t test briefly rules out those features that are irrelevant to the status of a sequence being positive or negative. After the initial screening, there are still a large number of composition features that require additional screening. We provide an iterative method PCRSS with the assistance of SVM to conduct the additional screening. PCRSS applies SVM to randomly selected feature subspaces and creates two sets of contrasting conditions so that paired comparisons can be applied later to assess the contribution of one feature in the presence of a possible interaction with other composition features. Only the significant features will be retained to construct the model for prediction.

2.7.1 Feature selection using a Student's t test

Consider the initial training dataset with R features from K O-glycosylation or negative O-glycosylation sequences. For each feature $z = 1,\ldots, R$, the test is performed as follows.

Step 1: according to the classification labels, O-glycosylation or negative O-glycosylation, the entries of zth feature are divided into two groups.

Step 2: apply a two-sample t test to the feature values in the two groups and obtain the p value from the test. If the test is significant at a 0.05 level (i.e. p value < 0.05), there is some evidence that this feature is correlated with the classification labels. Otherwise, this feature does not offer a good distinction between glycosylation site and negative O-glycosylation site and is permanently eliminated.

Repeat steps 1–2 for all features by letting $z = 1, 2,\ldots, R$ and eliminate all the features with p values > 0.05.

The test statistic from a two-sample t test is closely related to the point-biserial correlation coefficient between a dichotomous variable and a continuous random variable. However, it is difficult to assign a cut-off threshold to a point-biserial correlation coefficient for elimination. On the other hand, a typical significance level can easily be used as a reasonable threshold for the p value.

2.7.2 Feature selection through PCRSS

Suppose there are M features $\{f_1, f_2, \ldots, f_M\}$ retained after the initial filtering by the t test. These features are subjected to additional screening. We will consider the importance of a feature by examining the contribution of this feature in explaining the variation between the two sequence classes when the feature is included in a large number of different feature subspaces. In particular, we will generate random combinations of feature sets; each of them defines a feature subspace to be used together with others for feature screening. The following subsections describe the details.

Step 1: Initial standard of comparison

Before the PCRSS filtering, a fivefold cross-validation is used to predict the class labels with all features. The obtained AC is recorded. It represents the ability of the classifier with all features. Denote this initial AC value as AC_{all}.

Step 2: Generate random combinations of feature sets

We will generate random combinations of feature sets and use them to evaluate the effects of features. For convenience of operation, we generate an $N \times M$ dimensional random matrix X with entries X_{ij} being an indicator function of whether the ith random combination of feature set contains feature f_j. That is, the entries of X are either 1 or 0. A value of 1 at the ith row and jth column of X indicates that the ith random combination of the feature set contains feature f_j, and a value of 0 indicates that the ith random combination of the feature set does not contain feature f_j. Each row of X is an M-dimensional vector that defines a feature set to be considered for model evaluation in the next step. The row dimension N specifies that the number of feature subsets to be used to evaluate the training model is N instead of all feature combinations. The value of N can be set to a large even number (such as 500 and 1000) according to the computing resources capacity. The columns of X are filled randomly with $N/2$ ones and $N/2$ zeros such that there are equal numbers of ones and zeros in each column of X and they are randomly located. This basically specifies that each feature is included in half of the random combination of feature sets and excluded from the remaining half of the feature sets. Different rows of X give different feature sets by different combinations of the random numbers.

Step 3: Evaluation of the prediction ability of a feature subset

Each row of the binary random matrix X generated in the previous subsection specifies a feature subset, which can be obtained by selecting the features that correspond to value one in the row vector. We extract a new dataset from the initial dataset. The new dataset contains the class labels for all the training sequences and all features in the feature subset. We then apply the SVM to this dataset using fivefold CV to get the accuracy AC value that can represent the prediction ability of this feature subset. This operation is repeated for all the rows of X to get an N-dimensional vector of accuracy values, which represent the performance of each of the randomly formed feature sets. Denote the vector of AC values as *Y-new*. It will be used as the dependent variable in the next step.

Step 4: Assessing the individual feature's contribution

Taking into account the fact that a feature may play a role through its interaction with other features, it is important to consider the impact of other features while assessing the contribution of each feature. Based on the feature subset indicator matrix X, the contribution of each feature is assessed by a paired comparison of two cross-validated prediction performances. For $j = 1,\ldots, M$, this assessment of the jth feature is performed as follows.

(1) Create two indicator matrices $X_0^{(j)}$ and $X_1^{(j)}$ that are identical to X in all columns except for the jth column. The entries in the jth column of $X_0^{(j)}$ are set to be zero, and all entries in the jth column of $X_1^{(j)}$ are set to be one. Each row of matrix $X_0^{(j)}$ specifies a feature subset similar to that specified by X except that the jth feature is excluded in the modelling. Similarly, each row of $X_1^{(j)}$ defines a feature set that is the same as that given in X except that the jth feature is included in the modelling.
(2) Train a SVM regression model M^{train} using *Y-new* as the dependent variable and the columns of the binary random matrix X as the independent variables.

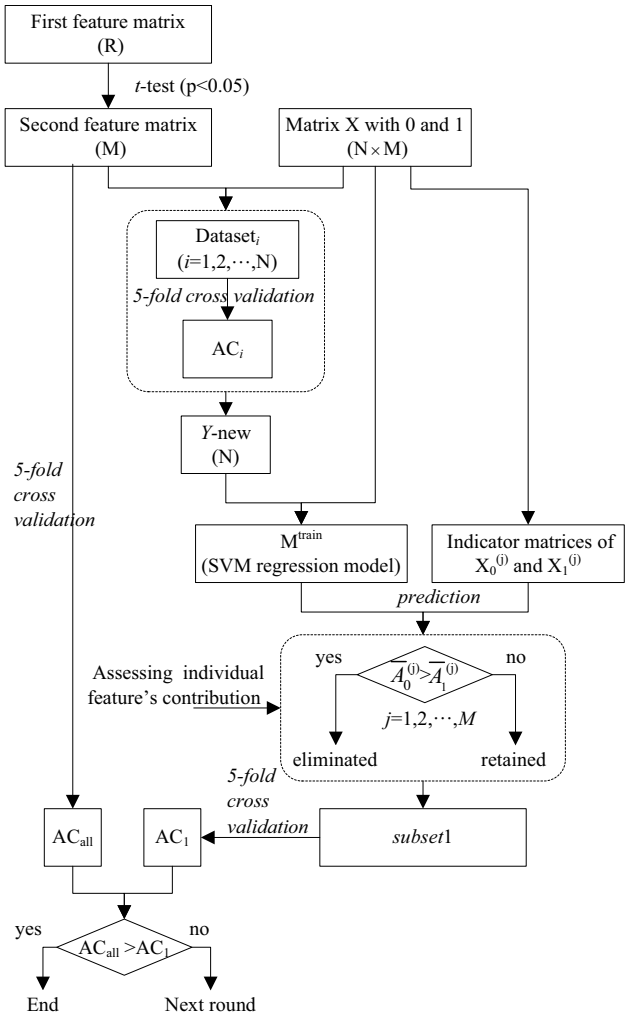

Fig. 1 Flowchart of PCRSS

(3) Predict the value of the dependent variable for each row of $X_0^{(j)}$ and $X_1^{(j)}$. We denote the two N-dimensional vectors of predicted dependent variable as $A_0^{(j)}$ and $A_1^{(j)}$, respectively. Entries in $A_0^{(j)}$ and $A_1^{(j)}$ are paired in the sense that the conditions of all other features are identical except for the difference of inclusion or exclusion of the *j*th feature.

(4) Calculate the average value of entries in $A_0^{(j)}$ and $A_1^{(j)}$ and denote them as $\overline{A}_0^{(j)}$ and $\overline{A}_1^{(j)}$, respectively. The result of $\overline{A}_0^{(j)} > \overline{A}_1^{(j)}$ indicates that excluding the *j*th feature tends to give better class prediction performance measured by accuracy AC. In this case, the *j*th feature is permanently eliminated at this stage. If $\overline{A}_0^{(j)} \leq \overline{A}_1^{(j)}$, the *j*th feature is retained.

Perform (1)–(4) for all features by letting $j = 1, 2, \ldots, M$ to determine which features will be eliminated and which features will be retained.

Step 5: Filtering termination condition

After the first round of filtering, the resultant set of retained features is denoted as subset 1. Apply the SVM with fivefold cross-validation as in Step 1 to the training dataset that includes the features in subset 1. Denote the fivefold CV accuracy as AC_1.

If $AC_1 \geq AC_{all}$, it suggests that the retained features in subset 1 have a better performance than all of the features in the prediction of *O*-glycosylation sites, and a second round of filtering will be performed by repeating the procedures in Steps 2–4 to features in subset 1. If $AC_1 \geq AC_{all}$, it suggests that the reduction in feature space reduces the predictive performance. Therefore, the filtering should stop.

At the end of each round of filtering, we can determine whether the filtering should be terminated or not by comparing the AC performance for SVM model with the AC performance of the previous round of filtering. If the new AC value is better than the previous AC value, the filtering continues; otherwise, the filtering stops. See Fig. 1 for the procedure of PCRSS.

3 Results

3.1 Prediction performance

In this section, we present the comparative results of different encoding schemes as well as different feature selection methods. There is no feature selection in the binary encoding, which is included for benchmark purposes. To be consistent with the literature, the parameter *k*max of CKSAAP is set to be four as in [10] such that the k-spaced amino acid pairs were considered for $k = 0$, 1, 2, 3 and 4. The parameter k of MSCAA is set to be three. The CKSAAP and MSCAA encoding schemes give 2205 and 9723 features, respectively. The accuracy of prediction using MSCAA encoding is compared to that using CKSAAP and binary encoding. For the comparison of feature selection methods, the t test alone and t test followed by PCRSS (t test + PCRSS) screening are compared to the no feature selection, CC, IE, minimum redundancy maximum relevance (mRMR) [30] and SVM recursive feature elimination (SVM-RFE)-based [29] methods on CKSAAP-encoded features. The t test alone and t test + PCRSS screening are also compared to the no feature selection, mRMR and SVM-RFE cases on the MSCAA-encoded features. Applying each feature encoding scheme and feature selection method combination corresponds to one selected model. Therefore, we also use the term model to refer to a feature encoding scheme and feature selection method combination. For all models, the radial basis function (RBF) is used as the kernel function of SVM and the parameters in each model are separately optimised. The feature selection processes of mRMR and SVM-RFE involve the following

Table 1 Prediction accuracy of *O*-glycosylation sites based on different encoding schemes and feature selection methods

Site	Encoding scheme	Feature selection	Sn (%)	Sp (%)	Ac (%)	MCC
S	Binary[b]	None[a]	74.2	81.9	78.0	0.567
	CKSAAP[b]	None[a]	77.9	86.5	82.2	0.655
	CKSAAP[b]	CC	80.7	85.6	83.1	0.671
	CKSAAP[b]	IE	82.1	83.9	83.0	0.665
	CKSAAP	SVM-RFE	91.4	89.7	90.5	0.818
	CKSAAP	mRMR	94.0	92.2	93.1	0.870
	CKSAAP	*t* test	89.7	87.9	88.8	0.783
	CKSAAP	*t* test + PCRSS	**95.7**	**95.7**	**95.7**	**0.914**
	MSCAA	None[a]	76.7	90.5	83.6	0.636
	MSCAA	SVM-RFE	88.8	92.2	90.5	0.792
	MSCAA	mRMR	87.1	92.2	89.7	0.774
	MSCAA	*t* test	82.8	90.5	86.6	0.708
	MSCAA	*t* test + PCRSS	**93.1**	**94.9**	**94.0**	**0.872**
T	Binary[b]	None[a]	74.8	78.3	76.6	0.536
	CKSAAP[b]	None[a]	80.4	82.3	81.3	0.631
	CKSAAP[b]	CC	80.3	82.5	81.4	0.632
	CKSAAP[b]	IE	80.8	81.9	81.3	0.631
	CKSAAP	SVM-RFE	88.2	84.4	86.3	0.741
	CKSAAP	mRMR	92.5	85.9	89.2	0.812
	CKSAAP	*t* test	87.2	89.6	88.4	0.760
	CKSAAP	*t* test + PCRSS	**93.0**	**89.6**	**91.3**	**0.839**
	MSCAA	None[a]	72.2	90.1	81.1	0.583
	MSCAA	SVM-RFE	89.2	88.7	88.9	0.780
	MSCAA	mRMR	90.6	86.8	88.7	0.789
	MSCAA	*t* test	86.3	88.7	87.5	0.742
	MSCAA	*t* test + PCRSS	**94.8**	**90.6**	**92.7**	**0.873**

Bold values indicate the best prediction model on the basis of MCC

[a] No feature selection is used

[b] Results as in [10]

steps: firstly, features are ranked in order of importance by SVM-RFE or mRMR. Secondly, features are introduced one by one and evaluated by tenfold cross-validation with the assistance of SVM. Lastly, the feature set with highest accuracy is considered to be the target feature set. The evaluation criterion of mRMR is mutual information quotient (MIQ) in this paper [31].

The tenfold cross-validation results from SVM based on different feature encoding and feature selection methods are reported in Table 1. The results of binary, CKSAAP with no feature selection, CKSAAP with CC feature selection and CKSAAP with IE feature selection are quoted from Ref. [10].

The binary encoding method has an accuracy of <80 % for both S and T site prediction, which is the lowest among all the methods tested. This suggests that the simple binary encoding method alone has limited potential in automatic information mining of *O*-glycosylation sites. When no feature selection was applied, the prediction accuracy (Ac) for S sites using all 9723 MSCAA features (83.6 %) was slightly higher than that using all 2205 CKSAAP features (82.2 %). Notice that the dimension of the CKSAAP feature space is much less than that of the MSCAA feature space. It is well known that irrelevant features add more noise and difficulty to all learning methods. Therefore, the relatively better performance using MSCAA features suggests that MSCAA encoding introduced some important features that are missing from the CKSAAP encoding. These two encoding methods have comparable accuracy for T site prediction (81.3 % for CKSAAP and 81.0 % for MSCAA) when no feature selection is performed.

From Table 1, we can see that the accuracy of the SVM model based on all MSCAA-encoded features without feature selection is obviously less than that with PCRSS and the *t* test feature selection-based SVM model. When we use multi-scale composition to encode amino acid sequence, the number of input features grows exponentially as the scale increases. When the length of sequence is short (the site of *O*-glycosylation is represented by a sequence fragment of 41 amino acids), some amino acid composition

may not exist in the sequence and the frequency is low. In such a case, the information in the corresponding feature is insufficient to distinguish fragments of *O*-glycosylation from non-*O*-glycosylation, and this feature serves as noise. Including noise features can reduce the accuracy of SVM. We use the *t* test to preliminarily remove some noise features. As the accuracy of the SVM model for S/T sites improves, the MCC also increases to 0.7 and the accuracy of prediction for positive and negative samples becomes more even. We comment that screening using the *t* test on the 9723 MSCAA-encoded features drastically reduced the dimension of the feature space to 362 and 578 for *S* and *T* site prediction, respectively. Compared to the number of sequences in our balanced dataset, the dimensions of the feature space remain high. As a result, there is a lot of noise that prevents SVM from achieving high accuracy (86.6 and 87.5 % for *S* and *T* sites, respectively). We then performed further feature selection using the PCRSS method on these features to yield a final set of 38 features to predict *S* sites and 53 features to predict *T* sites. The accuracy has been greatly improved to 94.0 % [sensitivity (SN) = 93.1 %, specificity (SP) = 94.9 %, Matthew's correlation coefficient (MCC) = 0.872] for *S* site prediction and 92.7 % (Sn = 94.8 %, Sp = 90.6 %, MCC = 0.873) for *T* site prediction. Compared with the initial results without feature selection, the prediction accuracy for *S* and *T* sites is gained by 12.4 and 14.3 %, respectively.

The CKSAAP encoding with CC or IE feature selection barely showed any improvement on accuracy (81–83 %) compared to the same coding without feature selection. On the other hand, when the feature selection using a *t* test is used, the accuracy for the CKSAAP-encoding method improved significantly (>86 %). With the filtering of the *t* test applied to the original 2205 CKSAAP features, there were 539 and 691 features retained for prediction of *S* and *T* sites, respectively. With further feature selection using our PCRSS filtering, the dimensions of features to predict *S* and *T* sites were reduced to 76 and 108, respectively. Now, the prediction accuracy of *S* sites is 95.7 % (Sn = 95.7 %, Sp = 95.7 %, MCC = 0.914), which is even slightly better than that based on MSCAA features applying the same feature selection methods. The prediction accuracy for *T* sites with CKSAAP features plus *t* test and PCRSS selection (Ac = 91.3 %, Sn = 93.0 %, Sp = 89.6 %, MCC = 0.839) is less than that based on MSCAA features using the same feature selection methods.

With the filtering of *t* test applied to the original 21,771 AA531 features, there were 1777 and 2676 features retained for prediction of *S* and *T* sites, with accuracies of 86.2 and 84.2 %, respectively. With further feature selection using our PCRSS filtering, the dimensions of features to predict *S* and *T* sites were reduced to 218 and 110, with accuracies of 93.1 and 87.7 %, respectively. The AA531 encoding is less accurate than that of MSCAA and CKSAAP; however, it establishes that the proposed feature selection method can be used in non-sparse encoding methods.

The CKSAAP-encoding features as well as MSCAA-encoded features were also screened with mRMR and SVM-RFE. The two feature selection methods have been significantly applied in many fields [5, 17, 25, 26, 35, 41], and they perform better in the accuracy of tenfold cross-validation compared to the CC, IE and *t* test-based methods used in this study. However, the two feature selection methods are not as accurate as the *t* test + PCRSS-based method.

The CKSAAP encoding with the new feature selection resulted in the highest accuracy in predicting *S* sites, and the MSCAA encoding with the new feature selection resulted in the highest accuracy in predicting T sites. This suggests that the two encoding methods have different advantages. We further considered combining all the features from the MSCAA and CKSAAP encoding and applied the *t* test to the entire set of 11,928 features to perform preliminary screening. After discarding duplicate features (MSCAA features of $k = 2$ and CKSAAP features of $k = 0$ are the same), there were 792 and 1136 features retained for *S* and *T* sites, respectively. Additional feature selection was then applied using PCRSS. In the end, 83 and 97 features for predicting *S* and *T* sites were retained.

Table 2 Prediction accuracy of *O*-glycosylation sites based on combined features

Site	Encoding scheme	Feature selection	Sn (%)	Sp (%)	Ac (%)	MCC
S	CKSAAP + MSCAA	*t* test	85.3	87.9	86.6	0.724
	CKSAAP + MSCAA	*t* test + PCRSS	**95.7**	**94.8**	**95.3**	**0.909**
T	CKSAAP + MSCAA	*t* test	88.2	87.3	87.7	0.758
	CKSAAP + MSCAA	*t* test + PCRSS	**92.9**	**91.5**	**92.2**	**0.851**
Average	CKSAAP	*t* test + PCRSS	94.35	92.65	93.50	0.876
	MSCAA		93.95	92.75	93.35	0.872
	CKSAAP + MSCAA		**94.30**	**93.15**	**93.75**	**0.880**

Bold values indicate the best prediction model on the basis of MCC

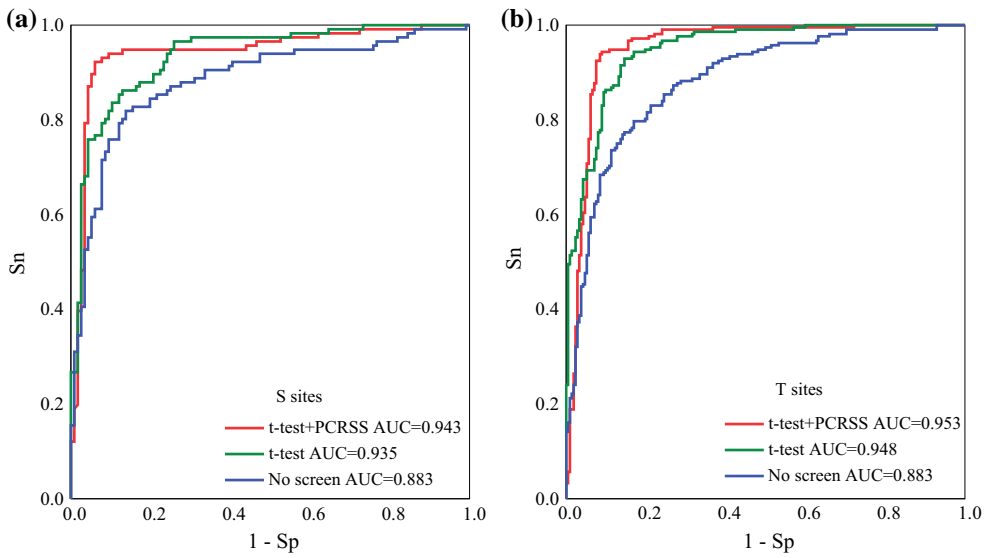

Fig. 2 ROC *curves* of *O*-glycosylation S site (**a**) and T site (**b**)

The combined features plus *t* test and PCRSS selection produced an accuracy of 95 % for predicting *S* sites and an accuracy of 92.2 % for predicting *T* sites (Table 2). That is, they produced the average performance of the best two models.

3.2 Comparison of different feature selection methods based on ROC curves

To obtain a better understanding of the relationship between the true-positive and false-positive rates of our MSCAA encoding followed by *t* tests and PCRSS feature selection, we plotted the receiver operating characteristics (ROC) curve for these three cases: the MSCAA encoding with no feature selection, the MSCAA encoding with *t* test screening and the MSCAA encoding with *t* test and PCRSS feature selection (Fig. 2).

The area under an ROC curve (AUC) suggests the probability of correct discrimination between the positive and negative sequences. A method with the AUC close to one is a good classifier. Correspondingly, the nearer the curve is to the upper left corner, the better its corresponding method is. The results established that the *t* test plus PCRSS-based feature selection was optimal for both *S* and *T* site prediction in terms of AUC (Fig. 1). Specifically, the AUC of *t* test plus PCRSS-based feature selection for *S* site prediction was 0.943, which is higher than the AUC of the *t* test-based feature selection (0.935) and non-selection model (0.883). The AUC of *t* test plus PCRSS-based feature selection for *T* site prediction was 0.953, which is also higher than the AUC of the *t* test-based feature selection (0.948) and non-selection model (0.883).

4 Discussion

Feature encoding and feature selection are two key issues in computational prediction of *O*-glycosylation sites based only on sequence information. Binary encoding has been widely used in the literature [6, 10, 23, 27]. Even though it is capable of identifying the type of amino acid in different locations of the fragment, such coding cannot reflect the relevancy of different locations and therefore has low accuracy [10]. Subsequently, for a primary sequence of protein with a determined structure and function, feature encoding based only on sequence has more practical applications [34]. A new MSCAA feature encoding based on sequence was proposed in this study. Each *O*-glycosylation site was represented by a sequence fragment of 41 amino acids. The related information on the relative amino acid locations is revealed by features obtained with multi-scale composition. Predictive results in this article have validated its reliability. As the dimensionality of the feature space grows exponentially when the scale increases, scale step selection is limited and more efficient information encoding methods based on sequence still have room to improve.

Feature selection has a critical contribution to the accuracy of the prediction model [4, 12]. The IE- and CC-based feature selection methods adopted by Chen et al. are all based on individual feature evaluation [9]. However, K individually best features are not necessarily the best combination of K features since individual evaluation tends to ignore the interaction between features. We proposed a SVM-based PCRSS method to perform feature selection. This method has two advantages: (1) the contribution of each feature to the prediction model is assessed conditionally on the interaction of

other features. This process is built into the feature selection procedure. The random indicator matrix represents various combinations of features. When evaluating each feature, only one column of the matrix is changed to include or exclude the feature. The interactions of the rest of the features are still included in the model. (2) Computational efficiency is one key issue of high-dimensional feature selection. A great deal of time is saved with the PCRSS method proposed in this study because only one cross-validated training is performed, and the assessments are predicted by the same model. This procedure takes full advantage of the fact that SVM is faster in predicting than in model training.

In the actual protein sequences, there are many more non-glycosylated S/T residues than *O*-glycosylated S/T residues. Schemes based on different ratios of positive/negative sites (unbalanced dataset) extract the information implicit in a large number of negative sequences. However, methods evaluated on severely unbalanced datasets tend to report increased total accuracy at the expense of low accuracy of the true positives. Therefore, those methods are no longer effective in predicting true *O*-glycosylation. The ability to fully exploit the hidden information in a large number of non-glycosylation sequences to improve prediction performance of *O*-glycosylation sites is one of the major tasks of the future.

5 Conclusions

Protein glycosylation is one of the most important and complex post-translational modifications that provides greater proteomic diversity than any other post-translational modification. It is critically involved in many important biological processes. The detection of *O*-glycosylation sites in a query protein is very helpful to understand its biological function. With the rapid increase in the amount of omics data, fast and reliable computational methods to identify *O*-glycosylation sites in protein sequences are in great demand. Two key issues, feature encoding and feature selection, can critically affect the accuracy of a computational method. In this article, we proposed a SVM-based paired comparison through random subspace screening method to perform feature selection and established that it significantly increases model prediction accuracy.

Acknowledgments This work was supported in part by the Research Foundation for the Doctoral Program of Higher Education of China (No. 20124320110002) and Haiyan Wang's work is partly supported by a grant from the Simons Foundation (#246077).

References

1. Bennett EP, Mandel U, Clausen H, Gerken TA, Fritz TA, Tabak LA (2012) Control of mucin-type *O*-glycosylation: a classification of the polypeptide GalNAc-transferase gene family. Glycobiology 22:736–756
2. Bewick V, Cheek L, Ball J (2004) Statistics review 13: receiver operating characteristic curves. Crit Care 8:508–512
3. Blom N (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics 4:1633–1649
4. Cabrera AF, Farina D, Dremstrup K (2010) Comparison of feature selection and classification methods for a brain–computer interface driven by non-motor imagery. Med Biol Eng Comput 48:123–132
5. Cai Y, Huang T, Hu L, Shi X, Xie L, Li Y (2012) Prediction of lysine ubiquitination with mRMR feature selection and analysis. Amino Acids 42:1387–1395
6. Cai YD, Chou KC (1996) Artificial neural network model for predicting the specificity of GalNAc-transferase. Anal Biochem 243:284–285
7. Cai YD, Liu XJ, Xu XB, Chou KC (2002) Support vector machines for predicting the specificity of GalNAc-transferase. Peptides 23:205–208
8. Centor RM (1991) Signal detectability: the use of ROC curves and their analyses. Med Decis Mak 11:102–106
9. Chang CC, Lin CJ (2011) LIBSVM : a library for support vector machines. ACM T Intell Syst Techn 2:1–27. http://www.csie.ntu.edu.tw/~cjlin/libsvm
10. Chen YZ, Tang YR, Sheng ZY, Zhang Z (2008) Prediction of mucin-type *O*-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. BMC Bioinform 9:101
11. Chou KC (1995) A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. Protein Sci 4:1365–1383
12. Dias NS, Kamrunnahar M, Mendes PM, Schiff SJ, Correia JH (2010) Feature selection on movement imagery discrimination and attention detection. Med Biol Eng Comput 48:331–341
13. Ding JD, Zhou SG, Guan JH (2011) miRFam: an effective automatic miRNA classification method based on n-grams and a multiclass SVM. BMC Bioinform 12:216
14. Geoghegan KF, Song X, Hoth LR, Fenga X, Shankera S, Quazib A, Luxenbergb DP, Wrightb JF, Griffora MC (2013) Unexpected mucin-type *O*-glycosylation and host-specific *N*-glycosylation of human recombinant interleukin-17A expressed in a human kidney cell line. Protein Expr Purif 87:27–34
15. Gill DJ, Chia J, Senewiratne J, Bard F (2010) Regulation of *O*-glycosylation through Golgi-to-ER relocation of initiation enzymes. J Cell Biol 189:843–858
16. Hansen JE, Lund O, Tolstrup N, Gooley AA, Williams KL, Brunak S (1998) NetOglyc: prediction of mucin type *O*-glycosylation sites based on sequence context and surface accessibility. Glycoconjugate J 15:115–130
17. Hidalgo-Muñoz AR, López MM, Galvao-Carmona A, Pereira AT, Santos IM, Vázquez-Marrufo M, Tomé AM (2014) EEG study on affective valence elicited by novel and familiar pictures using ERD/ERS and SVM-RFE. Med Biol Eng Comput 52:149–158
18. Hou TJ, Li N, Li YY, Wang W (2012) Characterization of domain-peptide interaction interface: prediction of SH3 domain-mediated protein-protein interaction network in yeast by generic structure-based models. J Proteome Res 11:2982–2995
19. Hou TJ, Xu Z, Zhang W, McLaughlin WA, David CA, Xu Y, Wang W (2009) Characterization of domain-peptide interaction interface: a generic structure-based model to decipher the binding specificity of SH3 domains. Mol Cell Proteomics 8:639–649
20. Hou TJ, Zhang W, David CA, Wang W (2008) Characterization of domain-peptide interaction interface: a case study on the amphiphysin-1 SH3 domain. J Mol Biol 376:1201–1214
21. Hou TJ, Zhang W, Wang J, Wang W (2009) The prediction of HIV-1 protease drug resistance by analyzing the protease/drug

decomposed interaction energy components. Proteins Struct Funct Bioinform 74:837–846
22. Jenkins NP, James DC (1996) Getting the glycosylation right: implications for the biotechnology industry. Nat Biotechnol 14:975–981
23. Julenius K, Molgaard A, Gupta R, Brunak S (2005) Prediction, conservation analysis, and structural characterization of mammalian mucin-type *O*-glycosylation sites. Glycobiology 15:153–164
24. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36:D202
25. Li BQ, Hu LL, Chen L, Feng KY, Cai YD, Chou KC (2012) Prediction of protein domain with mRMR feature selection and analysis. PLoS ONE 7:e39308
26. Li BQ, Huang T, Liu L, Cai YD, Chou KC (2012) Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. PLoS ONE 7:e33393
27. Li S, Liu B, Zeng R, Cai Y, Li Y (2006) Predicting *O*-glycosylation sites in mammalian proteins by using SVMs. Comput Biol Chem 30:203–208
28. Li XB, Peng SH, Chen J, Lü B, Zhang H, Lai M (2012) SVM-T-RFE: a novel gene selection algorithm for identifying metastasis-related genes in colorectal cancer using gene expression profiles. Biochem Biophys Res Commun 419:148–153
29. Liang Y, Zhang F, Wang J, Joshi T, Wang Y, Xu D (2011) Prediction of drought-resistant genes in Arabidopsis thaliana using SVM-RFE. PLoS ONE 6:e21750
30. Ma C, Dong X, Li R, Liu L (2013) a computational study identifies HIV progression-related genes using mRMR and shortest path tracing. PLoS ONE 8:e78057
31. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238
32. Reynders E, Foulquier F, Annaert W, Matthijs G (2011) How Golgi glycosylation meets and needs trafficking: the case of the COG complex. Glycobiology 21:853–863
33. Schjoldager KTBG, Clausen H (2012) Site-specific protein *O*-glycosylation modulates proprotein processing deciphering specific functions of the large polypeptide GalNAc-transferase gene family. BBA Gen Subj 1820:2079–2094
34. Shen JW, Zhang J, Luo XM, Zhu W, Yu K, Chen K, Jiang H (2007) Predicting protein-protein interactions based only on sequences information. PNAS 104:4337–4341
35. Shieh MD, Yang CC (2008) Multiclass SVM-RFE for product form feature selection. Expert Syst Appl 35:531–541
36. Sparrow LG, Gorman JJ, Strike PM, Robinson CP, McKern NM, Epa VC, Ward CW (2007) The location and characterisation of the *O*-linked glycans of the human insulin receptor. Proteins 66:261–265
37. Tran DT, Ten Hagen KG (2013) Mucin-type *O*-glycosylation during development. J Biol Chem 288:6921–6929
38. Vapnik V (1998) Statistical learning theory. Wiley, New York
39. Walsh G, Jefferis R (2006) Post-translational modifications in the context of therapeutic proteins. Nat Biotechnol 24:1241–1252
40. Yang ZH, Fang KT, Kotzc S (2007) On the Student's t-distribution and the t-statistic. J Multivariate Anal 98:1293–1307
41. Yoon S, Kim S (2009) Mutual information-based SVM-RFE for diagnostic classification of digitized mammograms. Pattern Recogn Lett 30:1489–1495
42. Yuan ZM, Zhang YS, Xiong JY (2008) Multidimensional time series analysis based on support vector machine regression and its application in agriculture. Sci Agric Sin 41:2485–2492
43. Zaki N, Wolfsheimer S, Nuel G, Khuri S (2011) Conotoxin protein classification using free scores of words and support vector machines. BMC Bioinform 12:217