Bioinformatics, 34(12), 2018, 2029–2036
doi: 10.1093/bioinformatics/bty039
Advance Access Publication Date: 6 February 2018
Original Paper

Sequence analysis

O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a *K*-means PCA oversampling technique

Cangzhi Jia[1,*], Yun Zuo[1] and Quan Zou[2,*]

[1]Department of Mathematics, Dalian Maritime University, Dalian 116026, China and [2]School of Computer Science and Technology, Tianjin University, Tianjin 300072, China

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on July 19, 2017; revised on November 23, 2017; editorial decision on January 22, 2018; accepted on February 5, 2018

Abstract

Motivation: Protein O-GlcNAcylation (O-GlcNAc) is an important post-translational modification of serine (S)/threonine (T) residues that involves multiple molecular and cellular processes. Recent studies have suggested that abnormal O-G1cNAcylation causes many diseases, such as cancer and various neurodegenerative diseases. With the available protein O-G1cNAcylation sites experimentally verified, it is highly desired to develop automated methods to rapidly and effectively identify O-GlcNAcylation sites. Although some computational methods have been proposed, their performance has been unsatisfactory, particularly in terms of prediction sensitivity.

Results: In this study, we developed an ensemble model O-GlcNAcPRED-II to identify potential O-GlcNAcylation sites. A K-means principal component analysis oversampling technique (KPCA) and fuzzy undersampling method (FUS) were first proposed and incorporated to reduce the proportion of the original positive and negative training samples. Then, rotation forest, a type of classifier-integrated system, was adopted to divide the eight types of feature space into several subsets using four sub-classifiers: random forest, k-nearest neighbour, naive Bayesian and support vector machine. We observed that O-GlcNAcPRED-II achieved a sensitivity of 81.05%, specificity of 95.91%, accuracy of 91.43% and Matthew's correlation coefficient of 0.7928 for five-fold cross-validation run 10 times. Additionally, the results obtained by O-GlcNAcPRED-II on two independent datasets also indicated that the proposed predictor outperformed five published prediction tools.

Availability and implementation: http://121.42.167.206/OGlcPred/

Contact: cangzhijia@dlmu.edu.cn or zouquan@nclab.net

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

GlcNAcylation is an O-linked β-N-acetylglucosamine (OGlcNAc) moiety linked to the side chain hydroxyl of a serine (S) or threonine (T) residue (Torres and Hart, 1984). GlcNAcylation was first discovered in 1984 by Hart (Torres and Hart, 1984) and was later found on numerous cytoplasmic and nuclear proteins. The addition of O-GlcNAc to proteins is catalyzed by O-GlcNActransferase (OGT) and its removal is catalyzed by O-GlcNAc–selectiveN-acetyl-β-D-glucosaminidase (O-GlcNAcase, OGA). This dynamic and reversible post-translational modification, which is analogous to phosphorylation, is emerging as a key regulator of protein function by regulating protein activity, protein-protein interaction, localization or protein degradation (Comer *et al.*, 1999). Recent studies have suggested that abnormal O-G1cNAcylation causes many

diseases, such as cancer and various neurodegenerative diseases. Accumulating evidence has shown that the expression of O-G1cNAcylation is significantly altered after Taxol treatment of breast cancer cells. Thus, it is crucial that O-G1cNAcylation sites are identified accurately.

Although the identification of O-G1cNAcylated using mass spectrometry technology has demonstrated great improvement, the precise identification of O-G1cNAcylation sites remains a challenge because of the following reasons: (i) it is highly dynamic in cells; (ii) the small molecular weight of O-G1cNAcylation; and (iii) the weak signal of the O-G1cNAc glycopeptide (Wang *et al.*, 2010). Because of an interest in better identifying O-G1cNAcylation sites, the computational prediction of O-G1cNAcylation sites has become an essential auxiliary tool.

During the last few years, many bioinformatics tools have been developed for identifying various PTM sites in proteins (Jia *et al.*, 2014, 2016; Liu *et al.*, 2017b; Qiu *et al.*, 2014, 2015, 2016a,b, 2017; Xu *et al.*, 2013a,b, 2014a,b, 2017; Zhang *et al.*, 2014). For O-linked glycosylation, several computational predictive techniques have been developed to identify protein O-G1cNAcylation sites in recent years. The first prediction tool, YinOYang (Gupta *et al.*, 2002), was built based on an artificial neural network system. The second prediction tool, OGlcNAcScan, based on a support vector machine (SVM) and trained using annotated O-GlcNAcylation sites from dbOGAP (Wang *et al.*, 2011). Chauhan *et al.* (2013) developed an SVM-based tool, GlycoEP, based on eukaryotic proteins of C-, N- and O-linked glycosylation collected from SWISS-PROT June 2011 release. Particularly, GlycoEP yielded a Sn of 35.75%, and Sp of 90.26% for O-linked glycosylation sites.The prediction model O-GlcNAcPRED was developed by Jia *et al.* (2013), and was also based on an SVM and the application of the adapted normal distribution bi-profile Bayes (ANBPB) feature encoding scheme. Zhao *et al.* (2015) built the predictor tool PGlcS based on using a k-means cluster to reduce the number of negative training samples, and two-step feature selection to reduce the dimension of the features. In 2014 and 2015, Lee's research group successively released two predictor tools (Kao *et al.*, 2015; Wu *et al.*, 2014), and the most current predictor tool adopted a two-layered machine learning method (i.e. the first layer is a profile hidden Markov model and the second layer uses an SVM), and the predictor tool significantly outperformed existing predictor tools. Li *et al.* (2015) constructed a model, GlycoMine, for predicting C-, N- and O-linked glycosylation sites. GlycoMine improve the prediction performance by coupling the RF algorithm with effective features selected by information gain (IG) and minimum redundancy maximum relevance (mRMR). More recently, Li *et al.* (2016) proposed GlycoMine^{struct} through considering protein-structural features. From another hand, Trost *et al.* (2016) described DAPPLE 2 for prediction of 20 different types of PTM sites by searching homology sequences, which is the update version of DAPPLE. Wang *et al.* (2017) applied the combination of multiple kernel support vector machines (SVM) for predicting PTM sites including phosphorylation, O-linked glycosylation, acetylation, sulfation and nitration. However, compared with approaches that predict other post-translation modification sites, the performance of these predictors remains unsatisfactory, and there remains considerable potential for improvement.

In this study, four strategies are used to improve the prediction accuracy of protein O-G1cNAcylation sites. First, we propose a K-means principal component analysis oversampling technique (KPCA) and fuzzy undersampling method (FUS) to reduce the proportion of the original positive and negative training samples. Second, eight types of feature are extracted and selected to encode each protein peptide. Third, four types of classifiers, random forest (RF), k-nearest neighbour (KNN), naive Bayesian (NB) and SVM, are used as the sub-classifiers of rotation forest. Fourth, majority voting is used to obtain a good ensemble classifier. By combining the aforementioned strategies, a novel tool called O-GlcNAcPRED-II is developed. Evaluated using five-fold cross-validation and on two independent test datasets, O-GlcNAcPRED-II proves to significantly outperform all the existing predictors for protein O-G1cNAcylation sites.

2 Materials and methods

2.1 Data collection and preprocessing

The datasets used to predict protein O-GlcNAcylation sites are generally constructed from the UniProtKB/Swiss-Prot database (Apweiler *et al.*, 2004), dbPTM (Lee *et al.*, 2006), dbOGAP (Wang *et al.*, 2011), O-GlycBase (Hansen *et al.*, 1999), PhosphoSitePlus (Hornbeck *et al.*, 2012) and PubMed literature. In this work, the datasets used to train and test the predictive model for identifying protein O-G1cNAcylation siteswere collected from the dbOGAPdatabase (Wang *et al.*, 2011) and the work of Jochmann *et al.* (2014). From dbOGAP, 392O-GlcNAcylated serine sites on 172 proteins were used to construct our positive training dataset. Additionally, from the work of Jochmann *et al.* (2014), 1,181O-GlcNAcylated sites on 520 proteins were collected to construct our positive training dataset. As indicated in (Wang *et al.*, 2016), a dataset that contained many redundant samples with high similarity would lack statistical representativeness. To eliminate redundancy, CD-HIT software (Fu *et al.*, 2012) was used to remove high-similarity protein sequences and protein peptide fragments. For a threshold of 40% similarity, 526 proteins with low sequence similarity were retained. Then, for a threshold of 30% similarity, we removed the protein fragments with high similarity. Finally, we obtained 945O-GlcNAcylated sites (547 serine and 398 threonine) that constituted our positive training dataset and 50 914 non-O-GlcNAcylated sites (29754 serine and 21160 threonine) that constituted our negative training dataset.

For no web-server is available by Jia *et al.* and Zhao *et al.*, the independent dataset simultaneously applied in their works, which consisted 67O-GlcNAcylation sites from 38 proteins, was firstly used to assess our predictor. Moreover, we collected the proteins available by Wu *et al.* (2014) and Li *et al.* (2015) to construct a new independent dataset.To avoid overestimate our model, we deleted those protein primary sequences that were also included in our training dataset. Furthermore, to minimize the similarity between the independent dataset and training dataset, the peptide sequences redundancy of length 23 were removed using the CD-HIT program (Huang *et al.*, 2010) with a threshold of 0.3. Finally, the independent test dataset consisted of 368 experimentally identified O-GlcNAcylation sites and 27 139 non-O-GlcNAcylation sites from 145 protein sequences. For convenience, the relevant training, test datasets are available both in Supplementary Material and web-server.

2.2 Feature extraction strategy

To build a superior predictor tool, we need to use a valid mathematical expression to formulate the protein peptide fragments, which can reflect more useful sequence information hidden in peptide fragments (Ahmad *et al.*, 2003a,b; Ward *et al.*, 2004). In this work, eight types of feature extraction strategies were used to formulate our protein peptide fragments. These features are BPB, adapted normal distribution BPB (ANBPB), di-amino acid BPB (DBPB), amino acid composition (AAC), di-AAC (DAAC), position-specific

amino acid propensity (PSAAP), position-specific di-amino acid propensity (PSDAAP) and position-specific tri-amino acid propensity (PSTAAP). The detailed feature extraction process is explained in the following subsections.

2.2.1 AAC and DAAC

AAC is the composition of 20 natural amino acids in one peptide. Similarly, DAAC is the composition of two adjacent amino acids in one peptide.

2.2.2 BPB

BPB is to encode one peptide by probability vector

$$P = (p_1, p_2, \ldots, p_n, p_{n+1}, \ldots, p_{2n}) \quad (1)$$

where $p_j (j = 1, 2, \ldots, n)$ denotes the posterior probability of each amino acid at the jth position in the positive training samples, and $p_j (j = n+1, n+2, \ldots, 2n)$ denotes the posterior probability of each amino acid at the jth position in the negative training samples. The two posterior probabilities were usually calculated using the frequency of each amino acid at each position in the positive and negative training datasets, respectively (Shao *et al.*, 2009; Song *et al.*, 2010).

2.2.3 ANBPB

In this method, the frequency of each amino acid at each position was encoded as random variables $X_{ij} (i = 1, 2, \ldots, 21; j = 1, 2, \ldots, 22)$, which were independent and obeyed binominal distribution $b(m, p)$, where m is the number of positive/negative training samples, and $p = \frac{1}{21}$. From the de Moivre-Laplace theorem, when m is sufficiently large, $\frac{X_{ij} - mp}{\sqrt{mp(1-p)}}$ is approximated to obey the standard normal distribution $N(0, 1)$. If we make V_j express the standard variance of $X_{ij} (i = 1, 2, \ldots, 21)$ (i.e. the deviation of the fraction of each amino acid at the same jth position), $X'_{ij} = \frac{X_{ij} - mp}{\sqrt{V_j}}$ is used as the new normalization of X_{ij}. Thus, $p_j (j = 1, 2, \ldots, n, \ldots, 2n)$ is encoded by the following expression: $p_j = P(X \leq X_{ij}) = \varphi\left(X'_{ij}\right)$, where $\varphi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$; see Jia *et al.* (2013).

2.2.4 DBPB

Let the probability vector

$$P = \left(p_1, p_2, \ldots, p_{n-1}, p_n, \ldots, p_{2\times(n-1)}\right)$$

encode the sequence S in the training dataset, where $(p_1, p_2, \ldots, p_{n-1})$ denotes the posterior probability of two adjacent amino acids at each position in positive training samples and $\left(p_n, \cdots, p_{2\times(n-1)}\right)$ denotes the posterior probability of two adjacent amino acids at each position in negative training samples. The posterior probability ofboth the O-GlcNAcylation and non-O-GlcNAcylationpeptides in the training dataset was calculated usingthe frequency of two adjacent amino acids at each position in thepositive and negative training datasets, respectively.

2.2.5 PSAAP

PSAAP matrix $Z = \left(z_{i,j}\right)_{21*22}$ constructed by

$$z_{i,j} = \frac{a_{P,i,j} - \bar{a}_{N,i,j}}{\sigma_{N,i,j}}, i = 1, 2, \cdots, 21; j = 1, 2, \cdots, 22. \quad (2)$$

where $a_{P,i,j}$ denotes the frequency of the ith amino acid at the jth position in the positive training samples, $\bar{a}_{N,i,j}$ and $\sigma N, i, j$ denote the average value and variance of the ith AACat the jth position in 10 negative training sample subsets (the size of each subset is same as that of the positive training samples), respectively. Applying PSAAP matrix $Z = \left(z_{i,j}\right)_{21*22}$, an n-dimensional feature vector for a sequence peptide was constructed, where n is the length of the sequence peptide, omitting the residue 'S' or 'T' of the central position (i.e. $n = 22$).

2.2.6 PSDAAP and PSTAAP

PSDAAP was defined by the probability difference of ith di-amino acid appeared in the jth position between positive and negative samples, while PSTAAP was defined by the probability difference of ith tri-amino acid appeared in the jth position between positive and negative samples (Supplementary Material).

It should be mentioned that the extracted features AAC, DAAC and BPB are just different models of general pseudo amino acid composition (PseAAC), first proposed by Chou (2001). With the wide application of PseAAC, Liu *et al.* developed a powerful web server called 'Pse-in-One' (Liu *et al.*, 2015), which can generate PseAAC. For detailed information on Pse-in-One and its updated version, please refer to references (Liu *et al.*, 2015, 2017a).

2.3 KPCA and FUS

The ratio of O-GlcNAcylated sites and non-O-GlcNAcylated sites (945: 50914) is extremely imbalanced; thus, it is essential to research new approaches to build balanced datasets. In this work, we propose KPCA and FUS methods.

2.3.1 KPCA

KPCA is proposed in this study based on applying the K-means clustering algorithm and principal components analysis (PCA) to add some synthetic positive samples into the positive training dataset. The detailed description is as follows:

1. First, randomly choose K positive training samples as initial cluster centres based on

$$P(InitialCenter(j)), \ j = 1, 2, \ldots, K, \quad (3)$$

where the initial K centres are defined as

$$InitialCenter(j) = [rand(0, 1) \times m] \quad (4)$$

where $[X]$ indicates that the maximum integer is not greater than X, m is the number of original positive training samples and $P(t), t = 1, 2, \ldots, m$ denotes the t original positive training samples.

2. According to the following computational formula, the positive training samples can be divided into K clusters:

$$I_t = \min_{1 \leq j \leq K} \|P(t) - P(InitialCenter(j))\|^2, \quad (5)$$

where I_t is the minimum squared Euclidean distance of the t original positive training samples and K initial cluster centres. $P(t)$ is divided at the kth cluster if

$$I_t = \|P(t) - P(initialCenter(k))\|^2. \quad (6)$$

3. For each cluster of the original positive training samples, calculate the average value of all samples to create the new cluster centre.
4. Repeat the second and third steps 10 times, and the original positive training samples can be divided into K new clusters.

5. For the samples of each cluster gained from the fourth step, suppose $P = [p_1, p_2, \ldots, p_{m1}]^T$ represents all the samples of the kth cluster, where $p_i = (p_{i1}, p_{i2}, \ldots, p_{iDim}), i = 1, 2, \ldots, m1$ denotes the ith sample of the kth cluster. Then, $p_i = (p_{i1}, p_{i2}, \ldots, p_{iDim})$, $i = 1, 2, \ldots, m1$ is standardized using $p'_i = \frac{p_i - u}{\sigma}$, where u and σ are the mean and standard deviation of each column of P, respectively.
6. Calculate the covariance matrix of $P' = [p'_1, p'_2, \ldots, p'_{m1}]^T$ and generate synthetic positive samples using above equation and

$$Y = [y_1, y_2, \ldots, y_{m1}]^T, y_i = (y_{i1}, y_{i2}, \ldots, y_{iDim}), i = 1, 2, \ldots, m1 \quad (7)$$

$$y_{ij} = p_i'^* A(:, j), i = 1, 2, \ldots, m1, j = 1, 2, \ldots, Dim, \quad (8)$$

where A is the corresponding eigenvector matrix, with the eigenvalues of the covariance matrix sorted in descending order.

2.3.2 FUS

FUS (Hosseinzadeh *et al.*, 2016) uses a fuzzy membership function to take the hidden information of training samples into account, and we applied it to delete a certain number of negative training samples to balance the positive and negative datasets.

The first step is to calculate the average value of the positive and negative training samples for each feature using

$$C^j_{Pos} = \frac{\sum_{i=1}^{PosNum} Pos(i, j)}{PosNum}, j = 1, 2, \ldots, Dim, \quad (9)$$

$$C^j_{Neg} = \frac{\sum_{i=1}^{NegNum} Neg(i, j)}{NegNum}, j = 1, 2, \ldots, Dim, \quad (10)$$

respectively, where $PosNum/NegNum$ is the number of positive/negative training samples, $Pos(i, j)/Neg(i, j)$ represents the value of the ith positive/negative training sample for the jth feature and Dim is the number of extracted features.

The second step is to calculate the standard deviation of the positive and negative training samples for each feature using

$$\sigma^j_{Pos} = \sqrt{\frac{1}{PosNum} \sum_{i=1}^{PosNum} \left(Pos(i, j) - C^j_{Pos}\right)^2}, j = 1, \ldots, Dim; \quad (11)$$

$$\sigma^j_{Neg} = \sqrt{\frac{1}{NegNum} \sum_{i=1}^{NegNum} \left(Neg(i, j) - C^j_{Neg}\right)^2}, j = 1, \ldots, Dim. \quad (12)$$

Then, the average value and standard deviation of the positive/negative training samples were used to construct the fuzzy membership functions on the training dataset:

$$u^j_{Pos}(i) = GaussMF\left(Data(i, j); C^j_{Pos}, \sigma^j_{Pos}\right), \quad (13)$$

$$u^j_{Neg}(i) = GaussMF\left(Data(i, j); C^j_{Neg}, \sigma^j_{Neg}\right), \quad (14)$$

$$GaussMF(x; c, \sigma) = \exp\left(-\frac{1}{2}\left(\frac{x - c}{\sigma}\right)^2\right), \quad (15)$$

where $Data(i, j)$ is the value of the ith sample for the jth feature on the training samples and $u^j_{Pos}(i)/u^j_{Neg}(i)$ represents the fuzzy membership degree of $Data(i, j)$ that belongs to the positive/negative sample $(i = 1, 2, \ldots, PosNum, PosNum + 1, \ldots, PosNum + NegNum, j = 1, 2 \cdots, Dim)$. Hereafter, the fuzzy value of the ith sample for the jth feature on the training samples is expressed as

$$Fval(i, j) = u^j_{Pos}(i) + \left(1 - u^j_{Neg}(i)\right). \quad (16)$$

Then, we deleted unnecessary and uninformative negative training samples. To achieve this, all the positive training samples were retained. For each negative training sample, we defined a score function as follows:

$$Score(i) = \sum_{j=1}^{Dim} Fval(i, j) \quad (17)$$

$$i = PosNum + 1, \ldots, PosNum + NegNum.$$

Finally, we removed some negative training samples that gained the highest score, and the following was used to determine the number of removed negative training samples:

$$Neg\text{R}emove = NegNum - \frac{1 - \alpha}{\alpha} PosNum \quad (18)$$

For a specific description of FUS, see Hosseinzadeh *et al.* (2016).

2.4 Rotation Forest algorithm

Rotation forest is a type of classifier integrated system proposed by Rodriguez *et al.* (2006), and its basic design concept is based on the RF algorithm (Breiman *et al.*, 2001; Rodriguez *et al.*, 2006). Rotation forest divides the original feature space into several subsets, and then performs a linear transformation, such as PCA, for each subset. The resulting transform components are merged according to the original order of the subsets and preserve all the principal components so that the data obtained after each random segmentation is projected into different coordinate spaces, which results in a large difference quantum set. These components are used to train the classifier, and it is possible to obtain a classifier with a large difference and high classification performance to improve the classification performance of the integrated system.

Four types of classifier, RF, KNN, NB and SVM, are used as the sub-classifiers of Rotation Forest. For a specific description of the algorithm for rotation forest, see (Rodriguez *et al.*, 2006).

2.5 Model construction and evaluation

To improve the predictive performance of O-GlcNAcylation sites, an ensemble learning predictor was used in this study, which used majority voting to integrate the output of the four individual models: RF, KNN, NB and SVM.The performance of O-GlcNAcPRED-II was evaluated using four measurements derived by Chen *et al.* (2013) and Lin *et al.* (2014) based on the symbols introduced by Chou in predicting signal peptides. Particularly, its advantages have been analyzed and endorsed by a series of studies published very recently (Jia *et al.*, 2014, 2016; Liu *et al.*, 2017b; Qiu *et al.*, 2014, 2015, 2016a,b, 2017; Xu *et al.*, 2013a,b, 2014a,b, 2017; Zhang *et al.*, 2014). The four measurements are given as follows:

$$Sn = 1 - \frac{N^+_-}{N^+} \quad (19)$$

$$Sp = 1 - \frac{N^-_+}{N^-} \quad (20)$$

$$Acc = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \quad (21)$$

$$MCC = \frac{1 - \left(\frac{N_+^+ + N_+^-}{N^+ + N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N_+^+}{N^+}\right)\left(1 + \frac{N_+^+ - N_+^-}{N^-}\right)}} \quad (22)$$

where N^+ is the total number of O-GlcNAcylation sequences and N_-^+ is the number of O-GlcNAcylation sequences incorrectly predicted as non-O-GlcNAcylation sequences; and N^- is the total number of non-O-GlcNAcylation sequences and N_+^- is the number of non-O-GlcNAcylation sequences incorrectly predicted as O-GlcNAcylation sequences. Note that the set of metrics is valid only for single-label systems. For multi-label systems whose existence has become more frequent in system biology (Lin *et al.*, 2013; Wu *et al.*, 2011) and system medicine (Cheng *et al.*, 2017a,b; Qiu *et al.*, 2016b; Xiao *et al.*, 2013), a completely different set of metrics defined by (Chou, 2013) is required.

3 Results and discussion

3.1 Processing of the positive and negative training datasets

Applying the KPCA, there were 945 synthetic positive training samples generated and combined with the original 945 positive samples to create a new positive training dataset. Additionally, by setting the parameter $\alpha = 0.3$ in equation of *NegRemove*, there were 2204 negative training samples retained after removing 48 710 negative training samples. It should be noted that only the original 945 positive training samples were experimentally validated, and the non-redundant negative training samples were adopted as the testing dataset for five-fold cross-validation.

3.2 Combining various features to optimize the prediction model

For the four classifiers, RF, KNN, NB and SVM, the optimal feature combination was different; therefore, we ran the selection of features four times. In the following, we only consider an example of constructing the optimal prediction model for the RF classifier to describe the specific process.

To determine the optimal prediction model, we not only individually used eight types of feature coding strategy, described in Section 2.2, but also two or more combinations of different types of features to encode each peptide. Five-fold cross-validation was used to train our prediction model. Sn was used as the evaluation criterion to select features because the higher the sensitivity, the fewer wrongly predicted O-GlcNAcylated sites.

Initially, each of eight types of feature coding strategy was individually used to encode our samples, and seven types of feature coding strategy with Sn values $\geq 63\%$ for five-fold cross-validation were retained to further optimize prediction performance. This is because the Sn was just 31.98% using DAAC coding, and so DAAC was not considered in the following steps. The specific results are listed in Supplementary Table S1a and b. Then, any combination of two types of features among seven feature coding strategies was further evaluated. Among the 21 ($C_7^2 = 21$) models, the top four models with Sn values $\geq 80\%$ for five-fold cross-validation were retained to further optimize the prediction model (Supplementary Table S1c). By increasing the types of features combined, only the combination AAC + ANBPB + BPB + PSAAP + PSTAAP achieved the best prediction performance, with a Sn of 82.26%, Sp of 90.99%, Acc of 88.41% and MCC of 0.7248 for five-fold cross-validation. The detailed evaluation results are listed in Supplementary Table S1d.

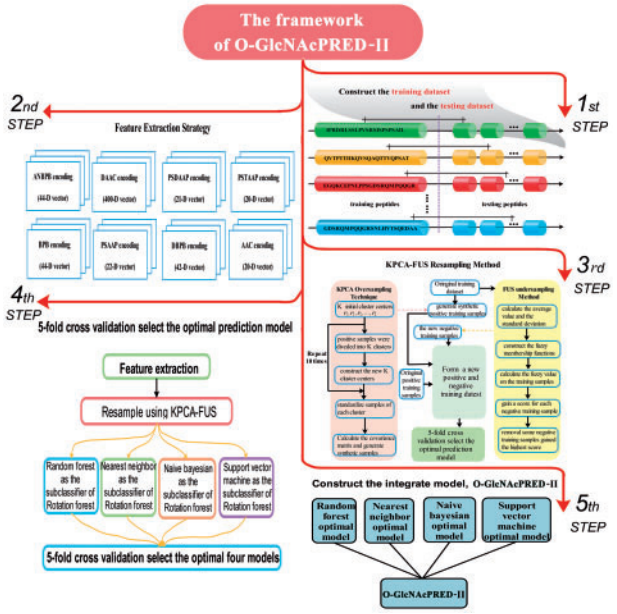

Fig. 1. Conceptual framework of O-GlcNAcPRED-II

Thus, an RF-based prediction model was built using the feature combination AAC + ANBPB + BPB + PSAAP + PSTAAP. Additionally, we used the same procedure to construct the other three optimal prediction models: the nearest neighbour optimal model built on the feature combination AAC + DAAC + BPB + DBPB + ANBPB + PSTAAP + PSAAP, NB optimal model built on the feature combination PSAAP + AAC, and SVM optimal model built on the feature combination ANBPB + DAAC + PSTAAP + BPB + PSAAP + DBPB. The detailed results for feature selection for the three classifiers are listedin Supplementary Tables S2–S4, respectively. Finally, the ensemble learning predictor, O-GlcNAcPRED-II, was constructed using a majority voting strategy for the prediction values of the four optimal models. Given a potential O-GlcNAcylat site, each of the classification models derives a prediction on the class (O-GlcNAcylated or non-O-GlcNAcylated). If N1 models derived an O-GlcNAcylated site, whereas N2 models derived a non-O-GlcNAcylated site among four classifiers, then we compared N1 and N2. If $N1 \geq N2$, the ensemble model assessed this site to be an O-GlcNAcylated site; otherwise, the ensemble model assessed it to be a non-O-GlcNAcylatedsite. Figure 1 shows the flow diagram for constructing the four optimal models and ensemble predictor, O-GlcNAcPRED-II. The parameters selected in O-GlcNAcPRED-II are given in Supplementary Table S5.

3.3 Effectiveness of the resampling approach

The average result of five-fold cross-validation run 20 times was used to demonstrate the effectiveness of our combination of KPCA and FUS. The comparison results are listed in Table 1 for original extremely imbalanced positive and negative samples, random undersampling, conducting KPCA oversampling, FUS undersampling and conducting KPCA oversampling s and FUS undersampling. For the random undersampling, we randomly selected the same number of negative samples as positive samples. We repeated the randomization for 10 times and compared the results in Supplementary Table S7 and Table 1, respectively. As can be seen, the best Sn,

Table 1. Comparison of different resampling methods on 5-fold cross validation

Resample method	Sn (%)	Sp (%)	Acc (%)	MCC
Without resampling	27.75	99.72	98.41	0.4160
Random undersampling	73.73	90.35	81.61	0.6459
Fuzzy undersampling	76.68	94.28	89.01	0.7322
KPCA oversampling	16.70	99.95	98.43	0.3717
O-GlcNAcPRED-II	77.89	98.38	91.90	0.8112

Table 2. Comparison of O-GlcNAcPRED-II with other four classifiers on our independent test set

Classifier	Sn (%)	Sp (%)	Acc (%)	MCC
Random forest	61.41	76.25	76.02	0.1010
K nearest neighbor	64.13	71.19	71.09	0.0892
Naive Bayes	65.76	67.83	67.80	0.0823
SVM	55.16	78.07	77.76	0.0916
O-GlcNAcPRED-II	67.12	72.46	72.39	0.1012

the best Sp and the average result among 10 times randomization are all lower than the result of combination of KPCA oversampling and Fuzzy undersampling. We found that O-GlcNAcPRED-II for a combination of KPCA and FUS achieved the best prediction performance, with aSn of 77.89%, Sp of 98.38%, Acc of 91.90% and Mcc of 0.8112. Additionally, the second best prediction performance was achieved by FUS for negative samples with a Sn of 76.68%, Sp of 94.28%, Acc of 89.01% and Mcc of 0.7322. The Sn values achieved by the without resampling and KPCA oversampling were less than 30%; therefore, we shall not consider them further.

3.4 Improving predictive performance using ensemble learning

The ensemble of the sub-classifier of Rotation Forest was first used to resolve the prediction of protein post-translational modification. To intuitively reflect the effectiveness of ensemble learning, we listed the prediction results of the four single classifiers and ensemble O-GlcNAcPRED-II for an independent test dataset that contained 368 experimentally identified O-GlcNAcylation sites and 27 139 non-O-GlcNAcylation sites. As listed in Table 2, the ensemble model O-GlcNAcPRED-II achieved the best Sn of 67.12% and the best MCC of 0.1012. The two best Sp were achieved by SVM and RF; however, the Sn achieved by SVM and RF were only 55.16% and 61.41%, respectively. Therefore, the ensemble model using majority voting was selected to construct our prediction model.

3.5 Comparison with existing tools

Regarding the performance of O-GlcNAcylation prediction, we compared O-GlcNAcPRED-II with other ten predictors as we known from different aspects.

Since the original training datasets were friendly offered by GlycoEP (Chauhan *et al.*, 2013), GlycoMine (Li *et al.*, 2015) and kernel SVM (Wang *et al.*, 2017), O-GlcNAcPRED-II was compared with these methods using 5/10 fold cross-validation according to the results listed in their works. As shown in Table 3, O-GlcNAcPRED-II generally results in an improvement of 2.76-29% with respect to GlycoMine, kernel SVM, MDDLogo-clustered SVM models and GlycoEP. From the results can also be concluded that the predictor was greatly influence by the selected sample. However, O-GlcNAcPRED-II showed better robustness, and obtained the accuracy ranged from 89.43 to 95.58%. Additionally, the best performance also indicated that ensemble classifier is more suitable than single classifier for O-GlcNAcylation sites prediction.

Table 3. Comparison of O-GlcNAcPRED-II with other four popular predictors for different methods on k-fold cross-validation

Classifier	Sn (%)	Sp (%)	Acc (%)	MCC
GlycoMine (IG+IFS)	90.00	84.97	86.67	0.7228
GlycoMine (mRMR+IFS)	84.23	87.82	86.61	0.7078
O-GlcNAcPRED-II	86.07	90.93	89.43	0.7573
GlycoEP	63.40	62.13	62.77	0.2700
O-GlcNAcPRED-II	84.43	96.22	92.60	0.8233
MDDLogo-clusterd SVM model	76.00	80.00	78.00	0.3700
O-GlcNAcPRED-II	78.48	98.70	92.55	0.8222
kernel SVM	62.85	95.00	92.55	–
O-GlcNAcPRED-II	76.15	97.21	95.58	0.7789

Table 4. Comparison of O-GlcNAcPRED-II with other popular predictors on the JIA independent test set

Classifier	Sn (%)	Sp (%)	Acc (%)	MCC
YinOYang	34.33	89.36	88.85	0.0725
O-GlcNAcscan	31.34	92.45	91.89	0.0847
O-GlcNAcPRED	56.72	64.77	64.70	0.0428
PGlcS	64.62	68.40	68.37	0.0697
Two-layered model of Lee	49.25	87.92	87.57	0.1074
O-GlcNAcPRED-II	71.64	70.18	70.20	0.0868

Considering the unavailability of the two previously developed webservers O-GlcNAcPRED and PGlcS, we further compared our O-GlcNAcPRED-II with the existing tools on the same independent dataset simultaneously applied in O-GlcNAcPRED and PGlcS. As can be observed in Table 4, O-GlcNAcPRED-II achieved the best Sn of 71.64%, which generally resulted in an improvement of 7.02 and 14.92% with respect to the second best PGlcS and third best O-GlcNAcPRED, respectively. Regarding another assessment, Sp, O-GlcNAcPRED-II achieved the fourth best Sp of 70.18%, O-GlcNAcscan achieved the best Sp of 92.45%, YinOYang achieved the second best Sp of 89.36% and the two-layered model of Lee *et al.* (Kao *et al.*, 2015) achieved the third best Sp of 89.83%. However, it is instructive that we found that the sensitivities attained by the O-GlcNAcscan, YinOYang and two-layered model of Lee *et al.* (Kao *et al.*, 2015) were far less than 50%, which is the probability of a random guess.

To further show the effectiveness of our O-GlcNAcPRED-II predictor, we also compared O-GlcNAcPRED-II with the accessible web-servers GlycoEP and DAPPLE2 on our new independent test set. May be affected by the number of known sites for O-GlcNAcylation sites, DAPPLE2 captured 45 proteins among 145 test proteins. On the other hand, GlycoEP and O-GlcNAcPRED-II achieved pretty close value of Sn, 64.13 and 67.12%, respectively. However, GlycoEP got the Sp of 37.17%, which is far below the Sp of 72.46% reached by O-GlcNAcPRED-II. It should be pointed out that we used the default threshold suggested by GlycoEP. These results indicate that our method represents a significant improvement in sensitivity over the existing prediction algorithm apparently. Due to the dynamics of O-GlcNAcylation, how to increase sensitivity at the same time as enhancing specificity is future work that we should research.

Considering the similarity between the independent test dataset and training dataset can influence the prediction results, we further investigated the performance of O-GlcNAcPRED-II by setting similarity thresholds from 0.9 to 0.3 with CD-HIT software. We reported the Sn, Sp, Acc, MCC and AUC obtained on different thresholds in Supplementary Table S6. It is noted that the prediction performance was not sensitive to the similarity between the independent test samples and training samples.

4 Conclusion

This study proposed a novel resampling method, KPCA-FUS, to reduce the imbalanced ratio of positive and negative training samples. Based on the training dataset processed by KPCA-FUS, different optimal features of eight types of sequence information were selected according to four sub-classifiers (KNN, RF, NB and SVM) of rotation forest. Then, using voting methods, we built the ensemble predictor O-GlcNAcPRED-II based on the rotation forest algorithm. The prediction results demonstrated that our approach was more accurate than the other five methods, which demonstrates the usefulness of the KPCA-FUS resampling technique and ensemble prediction algorithm. It is anticipated that O-GlcNAcPRED-II will be a helpful tool for predicting O-GlcNAcylation sites, and the KPCA-FUS resampling technique and ensemble prediction algorithm can be used in other protein post-translational modification predictions.

Acknowledgements

We thank Prof. Jiangning Song (Monash University) and Prof. Ao Li (University of Science and Technology of China) for providing datasets.

Funding

This work was supported by the Fundamental Research Funds for the Central Universities (number 3132016306, 3132017048 and 3132017085).The National Social Science Foundation of China (Grant No. 15CGL031) and the Program for Dalian High Level Talent Innovation Support (Grant No. 2015R063).

Conflict of Interest: none declared.

References

Ahmad,S. *et al*. (2003a) RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. *Bioinformatics*, **19**, 1849–1851.

Ahmad,S. *et al*., (2003b) Real value prediction of solvent accessibility from amino acid sequence. *Proteins*, **50**, 629–635.

Apweiler,R. *et al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*., **32**, 115–119.

Breiman,L. *et al*. (2001) Rotation forest. *Mach. Learn*., **45**, 5–32.

Chauhan,J.S. *et al*. (2013) Insilico platform for prediction of N-, O- and C-glycosites in eukaryotic protein sequences. *PLoS One*, 28, **8**, e67008.

Chen,W. *et al*. (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res*., **41**, e68.

Cheng,X. *et al*. (2017a) iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals. *Oncotarget*, **8**, 58494.

Cheng,X. *et al*. (2017b) iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. *Bioinformatics*, **33**, 341–346.

Chou,K.C. (2001) Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins Struct. Funct. Bioinf*., **44**, 246.

Chou,K.C. (2013) Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst*., **9**, 1092–1100.

Comer,F.I. *et al*. (1999) O-GlcNAc and the control of gene expression. *Biochim. Biophys. Acta*, **1473**, 161–171.

Fu,L. *et al*. (2012) CD-HIT. *Bioinformatics*, **28**, 3150–3152.

Gupta,R. *et al*. (2002) Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac. Symp. Biocomput. Pac. Symp. Biocomput*., **7**, 310–322.

Hansen,J.E. *et al*. (1999) O-GLYCBASE: a revised database of O-glycosylated proteins. *Nucleic Acids Res*., **27**, 370–372.

Hornbeck,P.V. *et al*. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res*., **40**, D261.

Hosseinzadeh,M. *et al*. (2016) Using fuzzy undersampling and fuzzy PCA to improve imbalanced classification through rotation forest algorithm.In: *International Symposium on Computer Science and Software Engineering*, pp. 1–7.

Huang,Y. *et al*. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **5**, 680–682.

Jia,C.Z. *et al*. (2013) O-GlcNAcPRED: a sensitive predictor to capture protein O-GlcNAcylation sites. *Mol. Biosyst*., **9**, 2909–2913.

Jia,C.Z. *et al*. (2014) Prediction of protein S-nitrosylation sites based on adapted normal distribution bi-profile Bayes and Chou's pseudo amino acid composition. *Int. J. Mol. Sci*., **15**, 10410–10423.

Jia,C.Z. *et al*. (2016) pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. *Bioinformatics*, **32**, 3133–3141.

Jochmann,R. *et al*. (2014) Validation of the reliability of computational O-GlcNAc prediction. *BBA Proteins Proteomics*, **1844**, 416–421.

Kao,H.-J. *et al*. (2015) A two-layered machine learning method to identify protein O-GlcNAcylation sites with O-GlcNActransferase substrate motifs. *BMC Bioinformatics*, **16**, S10.

Lee,T.Y. *et al*. (2006) dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res*., **34**, 622–627.

Li,F. *et al*. (2015) GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome. *Bioinformatics*, **31**, 1411–1419.

Li,F. *et al*. (2016) GlycoMinestruct: a new bioinformatics tool for highly accurate mapping of the human N-linked and O-linked glycoproteomes by incorporating structural features. *Sci. Rep*., **6**, 34595.

Lin,W.Z. *et al*. (2013) iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. *Mol. BioSyst*., **9**, 634–644.

Lin,H. *et al*. (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res*., **42**, 12961–12972.

Liu,B. *et al*. (2015) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res*., **43**, W65–W71.

Liu,B. *et al*. (2017a) Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein Sequences. *Nat. Sci*., **9**, 67–91.

Liu,L.M. *et al*. (2017b) iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. *Med. Chem*., **13**, 552–559.

Qiu,W.-R. *et al*. (2014) iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. *Biomed. Res. Int*., **2014**, 947416.

Qiu,W.-R. *et al*. (2015) iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model. *J. Biomol. Struct. Dyn*., **33**, 1731–1742.

Qiu,W.-R. *et al*. (2016a) iHyd-PseCp: identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. *Oncotarget*, **7**, 44310–44321.

Qiu,W.-R. *et al*. (2016b) iPTM-mLys: identifying multiple lysine PTM sites and their different types. *Bioinformatics*, **32**, 3116–3123.

Qiu,W.-R. *et al*. (2017) iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory. *Mol. Inf*., **36**.

Rodriguez,J.J. *et al*. (2006) Rotation forest: a new classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell*., **28**, 1619–1630.

Shao,J. *et al*. (2009) Computational identification of protein methylation sites through bi-Profile bayes feature extraction. *PLoS One*, **4**, e4920.

Song,J. *et al*. (2010) Cascleave: towards more accurate prediction of caspase substrate cleavage sites. *Bioinformatics*, **26**, 752–760.

Torres,C.R. and Hart,G.W. (1984) Topography and polypeptide distribution of terminal N- acetylglucosamine residues on the surfaces of intact lymphocytes. *J. Biol. Chem*., **259**, 3308–3317.

Trost,B. *et al*. (2016) DAPPLE 2: a tool for the homology-based prediction of post-translational modification sites. *J. Proteome Res*., **15**, 2760–2767.

Wang,Z. *et al*. (2010) Enrichment and site mapping of O-linked N-acetylglucosamine by a combination of chemical/enzymatic tagging, photochemical cleavage, and electron transfer dissociation mass spectrometry. *Mol. Cell. Proteomics MCP*, **9**, 153–160.

Wang,J. *et al*. (2011) dbOGAP-an integrated bioinformatics resource for protein O-GlcNAcylation. *BMC Bioinformatics*, **2**, 91.

Wang,J. *et al*. (2016) SOHPRED: a new bioinformatics tool for the characterization and prediction of human S-sulfenylation sites. *Mol. Biosyst*., **12**, 2849.

Wang,B. *et al*. (2017) Prediction of post-translational modification sites using multiple kernel support vector machine. *PeerJ*, **5**, e3261.

Ward,J.J. *et al*. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol*., **337**, 635.

Wu,Z.-C. *et al*. (2011) iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Mol. BioSyst*., **7**, 3287–3297.

Wu,Z.-C. *et al*. (2014) Characterization and identification of protein O-GlcNAcylation sites with substrate specificity. *BMC Bioinformatics*, **15**, S1.

Xiao,X. *et al*. (2013) iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem*., **436**, 168–177.

Xu,Y. *et al*. (2013a) iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS One*, **8**, e55844.

Xu,Y. *et al*. (2013b) iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *Peerj*, **1**, e171.

Xu,Y. *et al*. (2014a) iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. *Int. J. Mol. Sci*., **15**, 7594–7610.

Xu,Y. *et al*. (2014b) iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *PLoS One*, **9**, e105018.

Xu,Y. *et al*. (2017) iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. *Med. Chem*., **13**, 544.

Zhang,J. *et al*. (2014) PSNO: predicting cysteine S-nitrosylation sites by incorporating various sequence-derived features into the general form of Chou's PseAAC. *Int. J. Mol. Sci*., **15**, 11204–11219.

Zhao,X. *et al*. (2015) PGlcS: prediction of protein O-GlcNAcylation sites with multiple features and analysis. *J. Theor. Biol*., **380**, 524.