Bioinformatics, 31(9), 2015, 1411–1419
doi: 10.1093/bioinformatics/btu852
Advance Access Publication Date: 6 January 2015
Original Paper

Structural bioinformatics

GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome

Fuyi Li[1,†], Chen Li[2,†], Mingjun Wang[3], Geoffrey I. Webb[4], Yang Zhang[1,*], James C. Whisstock[2,5,*] and Jiangning Song[2,3,4,*]

[1]College of Information Engineering, Northwest A&F University, Yangling 712100, China, [2]Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia, [3]National Engineering Laboratory for Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, China, [4]Centre for Research in Intelligent Systems, Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia and [5]ARC Centre of Excellence in Advanced Molecular Imaging, Monash University, Melbourne, VIC 3800, Australia

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Alfonso Valencia

Received on July 11, 2014; revised on December 11, 2014; accepted on December 23, 2014

Abstract

Motivation: Glycosylation is a ubiquitous type of protein post-translational modification (PTM) in eukaryotic cells, which plays vital roles in various biological processes (BPs) such as cellular communication, ligand recognition and subcellular recognition. It is estimated that >50% of the entire human proteome is glycosylated. However, it is still a significant challenge to identify glycosylation sites, which requires expensive/laborious experimental research. Thus, bioinformatics approaches that can predict the glycan occupancy at specific sequons in protein sequences would be useful for understanding and utilizing this important PTM.

Results: In this study, we present a novel bioinformatics tool called *GlycoMine*, which is a comprehensive tool for the systematic *in silico* identification of C-linked, N-linked, and O-linked glycosylation sites in the human proteome. *GlycoMine* was developed using the random forest algorithm and evaluated based on a well-prepared up-to-date benchmark dataset that encompasses all three types of glycosylation sites, which was curated from multiple public resources. Heterogeneous sequences and functional features were derived from various sources, and subjected to further two-step feature selection to characterize a condensed subset of optimal features that contributed most to the type-specific prediction of glycosylation sites. Five-fold cross-validation and independent tests show that this approach significantly improved the prediction performance compared with four existing prediction tools: NetNGlyc, NetOGlyc, EnsembleGly and GPP. We demonstrated that this tool could identify candidate glycosylation sites in case study proteins and applied it to identify many high-confidence glycosylation target proteins by screening the entire human proteome.

Availability and implementation: The webserver, Java Applet, user instructions, datasets, and predicted glycosylation sites in the human proteome are freely available at http://www.structbioinfor.org/Lab/GlycoMine/.

Contact: Jiangning.Song@monash.edu or James.Whisstock@monash.edu or zhangyang@nwsuaf.edu.cn

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

1 Introduction

Glycosylation is a dynamic enzymatic process where glycans (or carbohydrates) are attached selectively to proteins or lipoproteins. Glycosylation is an important and ubiquitous post-translational modification (PTM) (Mazola *et al.*, 2011) and numerous studies have shown that glycosylation plays a crucial role in various BPs, ranging from protein folding, subcellular localization, and degradation (Dwek, 1998; von der Lieth *et al.*, 2004) to extracellular recognition and cell–cell interactions (Brennan *et al.*, 2011). It is commonly understood that glycosylation occurs in all cellular compartments and that >50% of all polypeptides are covalently modified (Hart and Copeland, 2010; Zaia, 2008). Serine (S), threonine (T), asparagine (N) and tryptophan (W) are the usual targets of protein glycosylation (Gavel and Vonheijne, 1990). Glycosylation can modulate the physicochemical and biological properties of proteins (Varki, 2007). For example, glycan occupancy in a protein can influence the ligand-binding mode and affect its biological activity. It can also influence protein stability and solubility, as well as modifying the susceptibility to proteolysis (Mazola *et al.*, 2011).

Glycosylation can be divided into four main categories according to the chemical linkage between the specific acceptor amino acid and the glycan, i.e. N-glycosylation, O-glycosylation, C-glycosylation (or C-mannosylation), and glycosyl-phosphatidylinositol anchor attachment (Ohtsubo and Marth, 2006). Each of these types has characteristic sequons or sequence motifs (Ohtsubo and Marth, 2006; Varki, 2007). The three most abundant types of glycosylation, i.e. N-, O- and C-glycosylation, are of particular interest. During N-glycosylation, the glycan is linked to the amide nitrogen in N residues. The sequon for N-glycosylation is N-X-T/S, or N-X-C in some rare cases, where X can be any residue except proline (Gavel and Vonheijne, 1990). During O-glycosylation, the glycan is attached to the hydroxy groups of S and T side-chains. No specific sequence motifs are defined for O-glycosylation, although experimental studies indicate that most O-linked glycosylation sites occur on S or T residues with a beta conformation that are in close proximity to proline residues (Christlet and Veluraja, 2001). In addition, O-glycosylation is frequent in secretory and membrane-binding proteins in a number of mammalian species. C-glycosylation involves the attachment of the glycan to the carbon of the first W residue in the following sequence motifs: W-X-X-W, W-X-X-C or W-X-X-F (Doucey *et al.*, 1998; Krieg *et al.*, 1998).

The key to understanding the mechanisms and other functional roles of glycosylation is identifying its substrates and the corresponding glycosylation sites. However, the experimental detection of glycosylation sites is still a challenging task, which often requires extensive laboratory work and considerable expense. Thus, computational approaches that accurately predict the glycan occupancy at specific sequons based on protein sequence information would be highly valuable and they may provide important insights into the functional roles of glycosylation. This is very important, because sequence-based analyses are currently the simplest and most readily deployed approach for predicting glycosylation targets and their respective glycosylation sites.

A number of computational approaches based on sequences or sequence-derived information have been developed to address this task. Gupta and Brunak developed NetNGlyc, which is a web tool based on neural networks that predicts N-glycosylation sites from protein sequences (Gupta and Brunak, 2002). Steentoft *et al.* (2013) developed NetOGlyc based on sequence context and surface accessibility. Cornelia *et al.* used ensembles of support vector machine (SVM) classifiers to predict amino acid residues that were likely to be glycosylated and implemented the EnsembleGly web server (Caragea *et al.*, 2007). The SVM classifiers were trained using a dataset of experimentally determined N-, O- and C-linked glycosylation sites in 242 proteins extracted from the O-GlycBase database (version 6.0) (Gupta *et al.*, 1999). Hamby and Hirst (2008) identified the pairwise patterns around glycosylation sites and combined these with the predicted secondary structure, predicted surface accessibility and hydrophobicity of the amino acids to train random forest (RF)-based predictors in an effort to improve the prediction accuracy. The developed GPP tool had an accuracy of 90.8% for S, 92.0% for T and 92.8% for N sites, respectively, based on an evaluation using a dataset of 242 proteins and 2415 glycosylation sites. Chen *et al.* (2008b) developed an SVM-based approach called CKSAAP_OGlySite, which employed a sequence-encoding scheme based on the composition of *k*-spaced amino acid pairs (CKSAAP) to predict O-linked glycosylation. Sasaki *et al.* (2009) trained SVM models using the local sequence features around glycosylation sites, and the characteristic subcellular localization of glycoproteins and achieved ~90% accuracy for both N-linked and O-linked glycosylation sites. Incorporation of structural features around the glycosylated sites, including the local contact order, surface accessibility, and secondary structure, have proved useful for improving the prediction accuracy of N-glycosylation (Chuang *et al.*, 2012). More recently, Chauhan *et al.* (2013) developed an SVM-based tool, GlycoEP, based on the datasets with eukaryotic proteins of C-, N- and O-linked glycosylation sites. GlycoEP yielded accuracy of 84.3, 86.9 and 91.4% for prediction of N-, O- and C-linked glycosylation sites, respectively.

There has been considerable success in the development of useful approaches for glycosylation site prediction, but several problems still exist in most of the currently available methods that need to be addressed to develop better methods, as follows. (i) The datasets used for model training are relatively small, which limits the predictive power of the method. In particular, many of the experimentally verified novel glycosylation sites that were discovered recently using high-throughput proteomic techniques were not included in the curated datasets. (ii) The majority of the available methods only predict specific types of glycosylation sites, such as N- or O-glycosylation. Few methods exist that predict all three of the major types of glycosylation sites, with the exceptions of NetNGlyc (Gupta and Brunak, 2002), NetOGlyc (Steentoft *et al.*, 2013) and SVMGlyc (Sasaki *et al.*, 2009), although they are outdated. (iii) The feature space used by existing methods to construct models is incomplete and not comprehensive. Other potentially useful features also remain that need to be characterized. (iv) Biological features are intrinsically heterogeneous, noisy, and multi-dimensional, but most existing methods do not employ feature selection techniques to quantify the importance and the contributions of the features used for the model performance, thereby leading to only a partial understanding of the sequence-glycosylation relationships. Given these deficiencies, it would be very useful to develop more accurate tools that facilitate the systematic prediction of all the major types of glycosylation.

In this study, we propose a novel bioinformatics approach called *GlycoMine* that addresses these problems and significantly improves the prediction performance for C-, N- and O-linked glycosylation by integrating various informative features derived from protein sequences. *GlycoMine* uses the RF algorithm coupled with a novel feature selection strategy to boost the performance. Specifically, an effective two-step feature selection method based on information gain (IG) (Kent, 1983) and minimum redundancy maximum relevance (mRMR) (Peng *et al.*, 2005) is used to determine the features that are important for glycosylation site specificity, as well as identifying condensed subsets of optimal features that contribute most to the prediction. A 5-fold cross-validation and independent tests using curated datasets demonstrated that *GlycoMine* outperformed other tools, including NetNGlyc (Gupta and Brunak, 2002), NetOGlyc (Steentoft *et al.*, 2013) EnsembleGly (Caragea *et al.*, 2007) and GPP (Hamby and Hirst, 2008). Two case studies demonstrated that *GlycoMine* can be applied rapidly to accurately identify potential novel glycosylation sites in a protein of interest.

2 Materials and methods

2.1 Overall framework

The overall framework of *GlycoMine* is illustrated in Figure 1, which shows that are four stages in the development of this tool, i.e. data collection and preprocessing, feature extraction, feature selection and model training and evaluation. The first stage involved data collection and preprocessing. In the second stage, a variety of features were extracted, including local sequence features, predicted structural features, protein functional features, and functional annotations. In the third stage, extensive feature selection was performed using a two-step procedure based on mRMR or IG in a cross-validation manner, where the optimal feature subsets were selected for each glycosylation type. In the final stage, three RF-based classifiers were respectively trained for C-, N- and O-linked glycosylation sites. The performance comparison with other existing methods was made via 5-fold cross-validation and independent tests.

2.2 Collection of datasets and preprocessing

The experimentally determined C-, N- and O-linked glycosylation sites were extracted from four public databases (Supplemental Methods). All of the experimentally verified glycosylation sites in human proteins were extracted from these databases. Sequence redundancy in the curated datasets was removed using the CD-HIT program (Huang *et al.*, 2010) to ensure that the sequence identity between any two proteins was no greater than 30%. This step was essential for eliminating sequence redundancy and avoiding overestimates of the performance of machine learning-based classifiers. As a result, we obtained 15, 168 and 208 glycosylated proteins with 68 C-, 416 N- and 649 O-linked glycosylation sites, respectively.

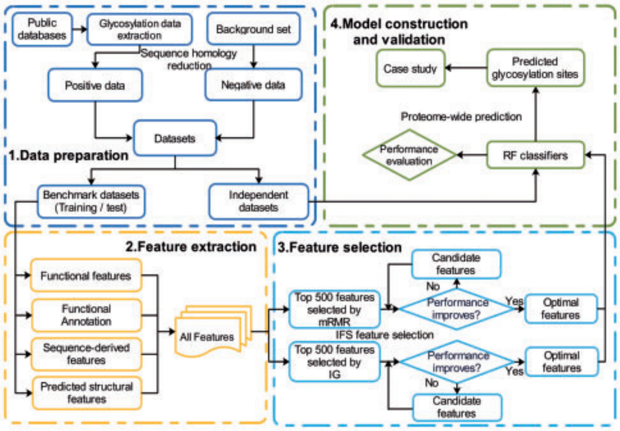

Fig. 1. Schematic framework of *GlycoMine*

The experimentally determined glycosylation sites were used as positive samples. However, it would be difficult to prove that a protein is not glycosylated in any specific conditions. Thus, it was difficult to collect proteins that could be considered as non-glycosylated. A background dataset that contained all human proteins was retrieved from UniProt. Negatives (no glycosylation sites) were selected from this background protein set, which excluded all experimentally verified glycosylated proteins. Amino acid residues (N, S, T and W) that were not experimentally verified as glycosylation sites were regarded as negatives. In addition, to obtain reliable positive and negative data, samples labeled as 'Probable', 'Potential' or 'By similarity' in UniProt were discarded.

The final datasets were divided into two random subsets, which are referred to as the benchmark dataset and the independent dataset (~20% of the size of the benchmark dataset). The performance of our method and other existing methods were compared using 5-fold cross-validation tests with the benchmark dataset and further validated with the independent dataset. In addition, feature selection and model training were performed with the benchmark dataset. To extract the local sequence or structural features around potential glycosylation sites, we used a local sliding window (Trost and Kusalik, 2011) that comprised 15 residues, where the potential glycosylation site was located at the center with seven neighboring residues upstream and downstream of the central site.

2.3 Feature extraction

Previous research into functional site prediction have demonstrated the value of features that capture sequence, structural and functional properties as well as information encoded in functional annotations (Chen *et al.*, 2008b; Chuang *et al.*, 2012; Wang *et al.*, 2014). To this end, we extracted major features of each of these types (see Supplementary Table S1 for a complete list of the extracted features).

2.3.1 Sequence-based features

(i) AAindex (Kawashima *et al.*, 2008), which is a database that contains 544 (release 9.1) amino acid properties collected from the literature; (ii) Physicochemical properties of proteins, which were generated by BioJava (Holland *et al.*, 2008); (iii) Position-specific scoring matrices (PSSMs), which were generated by performing PSI-BLAST searches (Altschul *et al.*, 1997) against the NCBI non-redundant database; (iv) Residue conversation score derived from the PSSM generated by PSI-BLAST defined as:

$$\text{Score}_i = -\sum_{j=1}^{20} p_{i,j}\log_2 p_{i,j},$$

where $p_{i,j}$ is the frequency of amino acid j at position i; (v) CKSAAP (Chen *et al.*, 2008b). This feature type was shown to be particularly useful for several PTM site predictions.

2.3.2 Predicted structural features

These include disordered regions predicted by DISOPRED2 (Ward *et al.*, 2004); secondary structures predicted by SABLE

(Wagner *et al.*, 2005) and backbone dihedral angles predicted by SpineX (Faraggi *et al.*, 2009).

2.3.3 Functional features

(i) BP from Gene Ontology (GO) (Ashburner *et al.*, 2000); (ii) cellular component from GO; (iii) molecular function from GO; (iv) functional domain from InterPro (Hunter *et al.*, 2012); (v) pathway information from KEGG (Kanehisa *et al.*, 2012); (vi) Functional domain from Pfam (Punta *et al.*, 2012) and (vii) Protein–protein interaction annotations from STRING (Franceschini *et al.*, 2013).

2.3.4 Functional annotations

The functional annotations were derived from UniProt: (i) Functional domain; (ii) nucleotide-binding site; (iii) disulfide bond; (iv) posttranslationally modified residues (glycosylation sites annotated by UniPort were removed); (v) active site; (vi) natural variants; (vii) metal ion-binding site and (viii) other binding sites for any chemical groups (such as co-enzymes and prosthetic groups). A detailed description of all the extracted features is provided in the Supplemental Methods.

2.4 Feature selection

It is likely that the initial feature sets contained certain redundant and noisy features, which were unwanted and which might have resulted in negative impacts on model training. Thus, it is imperative and often a common practice to employ feature selection techniques to reduce the dimensionality of the feature vectors in the initial sets (Saeys *et al.*, 2007; Song *et al.*, 2012) by eliminating features that do not contribute to the prediction. Thus, we developed a novel two-step feature selection strategy based on mRMR and IG, both of which are well-established feature selection methods, and they were applied initially to rank the importance and contributions of each feature type to glycosylation prediction.

2.4.1 Minimum redundancy maximum relevance

mRMR is a feature selection algorithm based on mutual information (MI). It has been used widely in the field of bioinformatics and computational biology (Li *et al.*, 2012, 2014). It can rank features based on their relevance to the target and the redundancy among the features. A ranked feature with a smaller index indicates that the feature has a better trade-off between the maximum relevance and the minimum redundancy. The mRMR-based feature selection strategy has been successfully applied in our previous studies of caspase-specific cleavage sites and has been shown to identify subsets of more important features relevant for the prediction (Wang *et al.*, 2014).

2.4.2 Information gain

IG is a feature selection method based on the information theoretic concept of entropy and it is a measure of the uncertainty of a random variable (Yu and Liu, 2003). IG evaluates each feature by measuring the IG with respect to the target class. IG has been used widely in bioinformatics as a fast and efficient feature selection method for high-dimensional data (Chen *et al.*, 2008a, 2009, 2013; Chen and Kurgan, 2007). See Supplemental Materials for more details. IG regards the features with the highest IG as important, whereas mRMR ranks features according to their relevance to the target class and the redundancy between the features.

2.4.3 Two-step feature selection based on mRMR or IG

Algorithm 1 describes the detailed procedure of our two-step feature selection method based on IG, mRMR and RF. In the first step, IG and mRMR were used to obtain the top 500 features as the optimal feature candidates (OFCs). In the second step, the most informative and contributory features were determined by the incremental feature selection (IFS). In step 1, the importance of each feature was evaluated by IG or mRMR.

In the second step, the top 500 features were selected as OFCs ($n = 500$). *getTopfeatures* (F_X, *i*) is a function that can obtain the top *i* features from feature set F_X. To determine the optimal features in the OFC, an IFS strategy based on a RF classifier and a 5-fold cross-validation were applied to the benchmark dataset to assess all of the features included in the OFCs. Steps 3–15 in **Algorithm 1** describe the IFS strategy. In step 8, the variable ‘sum’ represents the sum of 20 calculation results. In step 10, ‘*average* (*sum*)’ is a function that calculates the average value of *sum*, which is defined as: *n*/20. In step 16, the variable ‘*num*’ represents the position of the best AUC score in the array AUC.

IFS can be described briefly as follows. It constructs *n* feature subsets by adding one feature at a time from OFC to the candidate feature subset F'_X. The *i*-th feature subset is defined as:

$$F'_X = \{f \mid f_1, f_2, \ldots, f_i\},$$

where f_i is the *i*-th feature from OFC. Each feature set F'_X was tested using the RF classifier and evaluated based on five-fold cross-validation tests to avoid overfitting. This process was repeated for 20 rounds. As a result, the feature set with the highest AUC value among the 500 AUC values was selected as the optimal feature set.

Algorithm 1 Framework of the two-step feature selection procedure.

```
Input:
    Initial feature set, $F$;
    Benchmark dataset, $S$;
    Number of features in the OFC set, $n$;
    Feature selection method, $X$
Output:
    Optimal feature set, $F'$.
 1: $F_X = RankByX\ (F)$;
 2: OFC = $getTopfeatures\ (F_X, n)$;
 3: $F'_X = \emptyset$;
 4: for each $i \in [1, n]$ do
 5:     $F'_X = F'_X \bigcup getFeature\ (F_X, i)$;
 6:     $M = \emptyset$;
 7:     temp = 0;
 8:     for each $k \in [1, 5]$ do
 9:         for each $t \in [1, 20]$ do
10:             $M = BuildRandomForest(S,\ F'_X)$;
11:             $sum = GetAUCFromCrossValidation(M)$
12:         end for;
13:       if $average$ (sum) > temp
14:         temp = $average$ (sum);
15:       end if;
16:     end for;
17:     AUC[i] = $average$ (sum);
18: end for;
19: $num = theBestAUC$ (AUC);
20: $F' = getTopfeatures\ (F_X, num)$;
21: return $F'$;
```

2.5 Model training and evaluation

2.5.1 RF classifier

RF (Breiman, 2001) is a popular machine learning algorithm, which has been used widely in bioinformatics and computational biology. RF is an ensemble of decision trees that are built by random bootstrapping with a training dataset and a feature space. RF grows many classification trees and selects the classification that receives the most votes from all the trees. RF has various advantages that make it suitable for our prediction task, including: (i) RF performs better with high-dimensional inputs; (ii) RF is a highly efficient machine learning model because the RF training process is usually faster than that of many other algorithms. Given its reliable performance and efficiency, the RF algorithm implemented using the R package was used to train the models in the present study.

2.5.2 Performance evaluation

We used five measures, i.e. sensitivity, specificity, precision, accuracy and Mathew's correlation coefficient (MCC) to evaluate the prediction performance in the present study. We also use AUC, the area under the receiver operating characteristic curve (ROC), which plots TPR against the false positive rate. We use the ROCR package (Sing *et al.*, 2005) to draw the ROC curves and to calculate the AUC values. To evaluate the performance and to facilitate comparisons with other methods, we performed a 5-fold cross-validation (based on the benchmark datasets) and independent tests (based on the independent test datasets). A detailed description of these measures can be found in the Supplemental Methods.

3 Results and discussion

3.1 Amino acid site specificity of glycosylation

Based on the curated datasets, we analyzed the site specificity of the three types of glycosylation sites. We calculated the frequencies of the amino acids at each position using a local window size of 14 residues (i.e. seven residues upstream and seven residues downstream) around the central glycosylated residue. The sequence logos shown in Supplementary Figure S1 allowed us to visualize and analyze the sequence-level site specificity of the three types of glycosylation sites, where each stack represents the corresponding amino acid position in the sequence. The degree of sequence conservation at the corresponding position is measured based on the total height of each stack, which indicates the relative occurrence of the amino acid at that position.

The distinct differences in the site motifs of the three types of glycosylation are reflected by the different requirements for amino acid types in the central position. For example, a hallmark of the C-glycosylation motif is the requirement for a W residue at the central position. A lesser requirement for W and cysteine (C) residues is also shown in the downstream (+3 position) region and W in the upstream (−3 position) region, thereby constituting the canonical C-linked glycosylation site motifs, including W-X-X-W and W-X-X-C (X can be any residue except proline) (Krieg *et al.*, 1998). For N-linked glycosylation, N residues are preferred in the central position, as well as a strong requirement for S or T at the +2 position, thereby forming the N-linked glycosylation sequence motifs N-X-T and N-X-S (Gavel and Vonheijne, 1990). For O-glycosylation, it can be seen that S and T residues are preferred in the central position. No specific sequence motifs can be clearly defined for this type of glycosylation. Thus, the sequence-derived features that describe the local site specificity might be useful for predicting the different types of glycosylation.

3.2 Enhancing the prediction performance using a two-step feature selection strategy

To the best of our knowledge, this study represents the first systematic effort to use intensive feature selection to select and characterize more relevant features for protein glycosylation prediction, which contrasts with most previous studies where few used feature selection methods to address this problem. In the first step of feature selection, we used IG or mRMR algorithms to rank all of the initial features and to generate the top 500 OFCs, which were ranked by the corresponding IG or mRMR scores. In the second step of feature selection, we performed IFS to characterize the final optimal features for each type of glycosylation. The performance changes in the RF-based classifiers during the course of this latter feature selection procedure were reflected by the respective FFS curves, as shown in Supplementary Figures S3 and S4. There was a general trend where the height (the AUC score) of the feature selection curve continued to increase until it reached its highest peak. As a result, we selected condensed subsets of the most relevant and informative features that contributed to the prediction of glycosylation sites.

To better understand the relative importance and contribution of each group of the final optimal features to the performance of the *GlycoMine* models, we evaluated the performance decrease (based on the AUC score) by iteratively removing feature groups from the model one a time. The changes in the performance results are shown in Supplementary Table S8. Compared with the other groups of features, it is clear that functional features were the most important feature type for all three types of glycosylation because removing this group from the model caused the largest performance decrease (resulted in the lowest AUC scores in Supplementary Table S8). For example, for O-linked glycosylation prediction, removal of functional features resulted in 12.8 and 26.5% decreases in the AUC values for the IG+IFS and mRMR+IFS models, respectively. Moreover, we found that the mRMR+IFS model performed slightly better than the IG+IFS model in terms of the AUC for C- and N-linked glycosylation (Supplementary Table S8). For O-linked glycosylation, the IG+IFS model had a higher AUC. In addition, the number of final optimal features varied significantly, depending on the type of glycosylation (Supplementary Tables S9 and S10). Models with less final features did not necessarily deliver worse performance, but the opposite was also true, i.e. a model may requires more final features to achieve a better performance. In the latter case, many of the features complemented each other and contributed collectively to the performance. Supplementary Tables S9 and S10 show statistics for the final selected optimal features for each type of glycosylation based on the IG+IFS and mRMR+IFS models, respectively.

3.3 Comparison with other methods using a 5-fold cross-validation and independent tests

In this section, we compare the results of the prediction performance using *GlycoMine* and those with four other existing tools, i.e. NetNGlyc (Gupta and Brunak, 2002), NetOGlyc (Steentoft *et al.*, 2013), GPP (Hamby and Hirst, 2008) and EnsembleGly (Caragea *et al.*, 2007), which are currently publicly available. Note that NetNGlyc and EnsembleGLy were trained using human glycosylation data, thus it is reasonable to compare the performance of these two tools with *GlycoMine* based on our datasets. These different methods used different training datasets to train their prediction models, thus to objectively evaluate the performance of these methods, we performed both 5-fold cross-validation and independent tests using the same benchmark and independent datasets.

Table 1. The numbers of glycosylation sites

	Positive set		Negative set	
	Benchmark	Independent	Benchmark	Independent
C-linked	55	13	108	28
N-linked	333	83	667	166
O-linked	520	129	1018	258

In the 5-fold cross-validation tests, it should be noted that we only performed feature selection using the four subsets and we excluded the test subset, thereby facilitating objective evaluations of the performance of our method. The number of three kinds of glycosylation sites in both benchmark and independent test are shown in Table 1.

The sequences of these datasets were submitted to the servers of NetNGlyc, NetOGlyc, GPP and EnsembleGly to evaluate the prediction performance of these servers, where we used the default or recommended settings suggested by the authors. However, it is worth noting that *GlycoMine* was capable of predicting all three types of glycosylation, whereas the other tools could only be used to predict one or two types. The prediction performance comparison results for the three types of glycosylation sites are shown in Supplementary Table S12. We also used the ROC curves and AUC measures to evaluate the performance of the different methods. The ROC curves for the benchmark and independent datasets are shown in Supplementary Figure S5 and Figure 2, respectively.

The results on both 5-fold cross-validation (Supplementary Table S12) and independent tests (Fig. 2) indicate that for O-linked glycosylation sites, the IG+IFS models achieved slightly higher AUC scores, while the mRMR+IFS models achieved higher AUC scores for C- and N-linked glycosylation. Both types of *GlycoMine* models (trained using the optimal feature subsets selected by IG+IFS and mRMR+IFS) clearly outperformed the other four existing tools for all the three types of glycosylation sites. *GlycoMine* achieved significantly higher AUC scores than the other tools when tested using both the benchmark and independent datasets. For example, with the independent datasets, the *GlycoMine* models trained using the optimal feature sets (obtained by IG and IFS) obtained AUC values of 0.983 and 0.982 for N- and O-linked glycosylation, respectively, whereas the corresponding *GlycoMine* model trained using the optimal features selected by mRMR and IFS obtained AUC scores of 0.986 and 0.961, respectively. In contrast, the second best tool, EnsembleGly, obtained AUC scores of 0.939, 0.819 and 0.919, for the three types of glycosylation prediction. NetOGlyc obtained an AUC score of 0.850 for the O-linked glycosylation prediction. The other two tools, GPP and NetNGlyc, obtained much lower AUC values. For C-linked glycosylation, the *GlycoMine* models (IG+IFS and mRMR+IFS models) obtained an AUC score of 1.0. This was probably due to the limited size of the independent test dataset, where only 13 C-linked glycosylation sites were available.

Considering that certain negative samples (i.e. non-glycosylated sites) can potentially be mislabeled, we further investigated the performance of *GlycoMine* by randomly sampling negative samples. We randomly selected the same number of negative samples as positive samples from the benchmark datasets for C-, N- and O-linked glycosylation with different OFCs using mRMR+IFS and IG+IFS methods. This randomization process was repeated 10 times. For each randomization, we conducted five-fold cross-validation tests. We calculated the average performance for 10 times of 5-fold cross-validation tests and reported the results in Supplementary Table S11. It is clear that even with different selected negative samples, there is a slight performance difference across the 10 randomization trials. This indicates that the performance of *GlycoMine* is robust and insensitive to the potential impact posed by the randomly selected negative samples.

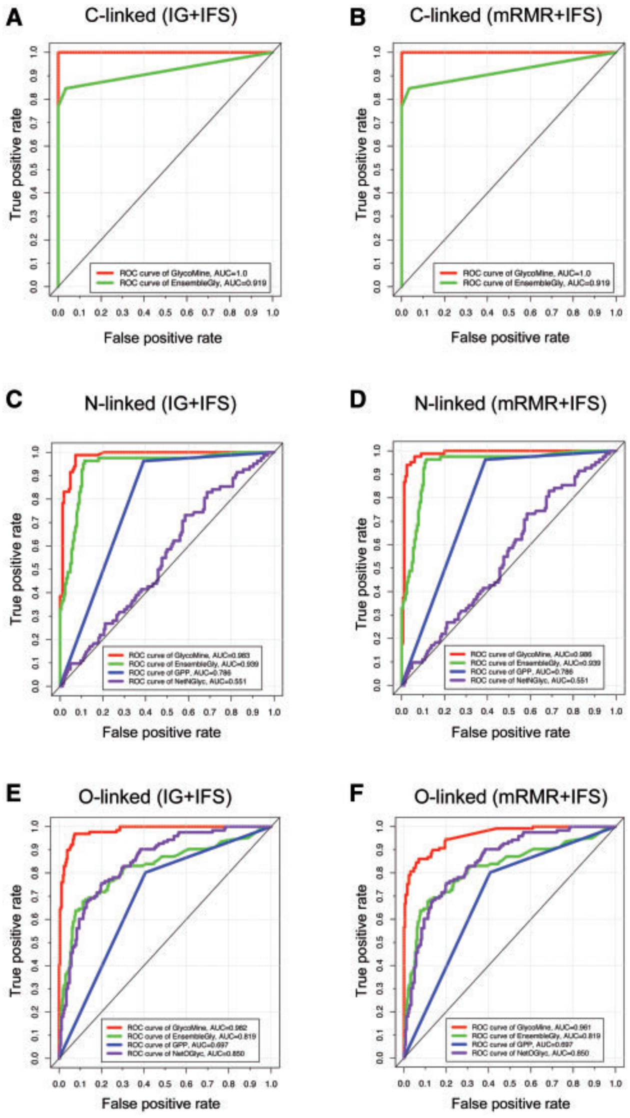

Fig. 2. ROC curves for *GlycoMine*, NetNGlyc, NetOGlyc, EnsembleGly and GPP glycosylation site prediction based on the independent datasets. (**A**) and (**B**): C-linked glycosylation; (**C**) and (**D**) N-linked glycosylation; (**E**) and (**F**) O-linked glycosylation

In addition to the AUC scores, we calculated the MCC, ACC, sensitivity, specificity and precision to compare the performance of the tools (Supplementary Table S12). These results demonstrated that *GlycoMine* obtained better performance than the other four tools, NetNGlyc, NetOGlyc, GPP and EnsembleGly, for the prediction of N-, C- and O-linked glycosylation sites, with the sole exception of the sensitivity score for N-linked glycosylation where *GlycoMine* obtained a slighter lower sensitivity than GPP (Supplementary Table S12). The improved prediction performance of our method compared with GPP and EnsembleGly might be attributable to three important factors: (i) the extraction of more useful and complementary features from multiple sources yielded better descriptions of protein glycosylation sites; (ii) the use of RF as the algorithm to learn the underlying rules of glycosylation and to train the models; and (iii) the use of an efficient two-step feature selection strategy to remove noisy and irrelevant features, and to select most informative features that contributed to the predictions.

3.4 Web server and local Java tool implementation

We developed a web server and a local Java Applet for *GlycoMine* to allow users to perform high-throughput bioinformatics analyses of novel glycosylation sites, which are freely available at http://www.structbioinfor.org/Lab/GlycoMine. The *GlycoMine* server was implemented using Java Server Pages running Tomcat7 and configured in the Linux environment on a 16-core server machine with 50 GB memory and a 4 TB hard disk. The server uses the best-performing models trained using the final optimal features selected by the two-step feature selection procedure. For O-linked glycosylation, the optimal features were selected by IG+IFS, while for C- and N-linked glycosylation, the optimal features were determined by mRMR+IFS. The server requires protein amino acid sequences in the FASTA format as the input. In cases where the user has multiple sequences to predict, they are encouraged to download the Java Applet of *GlycoMine* and submit a multiple FASTA formatted input to this Java program to predict the potential glycosylation sites of multiple proteins. Supplementary Figure S6 shows an example of the interfaces of web server and the Java program together with its prediction output. The computational time required for a prediction task depends on the length of the submitted sequence. For a protein sequence with 500 amino acid residues, the prediction task requires 3 min and 50 s to generate and return the prediction results.

3.5 Proteome-wide prediction of glycosylation sites

The development of *GlycoMine* allows users to perform the proteome-wide prediction of glycosylation sites in an automated manner. Thus, we used *GlycoMine* to screen potential glycosylated proteins and their respective glycosylation sites in the complete human proteome (a total of 84 843 proteins). The prediction models were trained using the final optimal features based on the complete training datasets. We adjusted the prediction thresholds to set the specificity level at 99.0% to generate high-confidence prediction results. The statistics for the predicted N-, C- and O-linked glycosylated proteins and glycosylation sites are shown in Supplementary Table S13. In total, 5846 and 6358 proteins were predicted to be N- and O-glycosylated, which contained 24 174 and 97 042 predicted N- and O-glycosylation sites, respectively. In contrast, the number of predicted C-linked glycosylation sites was lowest (only 18 926 sites), possibly due to the limited availability of experimentally verified glycosylation data for model training. A complete list of the predicted glycosylated proteins and their glycosylation sites is available from the *GlycoMine* website.

3.6 Case study

To further illustrate the capacity of *GlycoMine*, we performed a case study of two proteins with respect to N- and O-linked glycosylation. The first protein was perforin-1 (PRF1; UniProt ID: P14222), a cytotoxic pore-forming glycoprotein that plays an important role in killing virus-infected and transformed cells by cytotoxic T lymphocytes and natural killer cells (Brennan *et al.*, 2011). One N-linked glycosylation site has been experimentally confirmed by a previous study (Chen *et al.*, 2009). Previous study also indicates that transport from the Golgi to secretory granules requires N-linked glycosylation of perforin-1 (Brennan *et al.*, 2011). The second protein was glycophorin-A (GYPA, UniProt ID: P02724), a major intrinsic membrane protein with a high proportion of O-glycosylated residues in erythrocytes (Wilson *et al.*, 1993). The O-linked glycosylation sites (labeled in blue) in Figure 3B were experimentally verified (Pisano *et al.*, 1993). Glycophorin-A is expressed exclusively in erythroid cells and their precursors and is a marker for identifying erythroid differentiation in hematopoietic malignancies. The sequence scanning results and the predicted glycosylation sites for the domains of each protein are shown in Figure 3. By assigning a higher prediction cutoff threshold of 0.5, *GlycoMine* predicted six and 34 high-confidence N- and O-linked glycosylation sites, respectively, in these two proteins. One of these

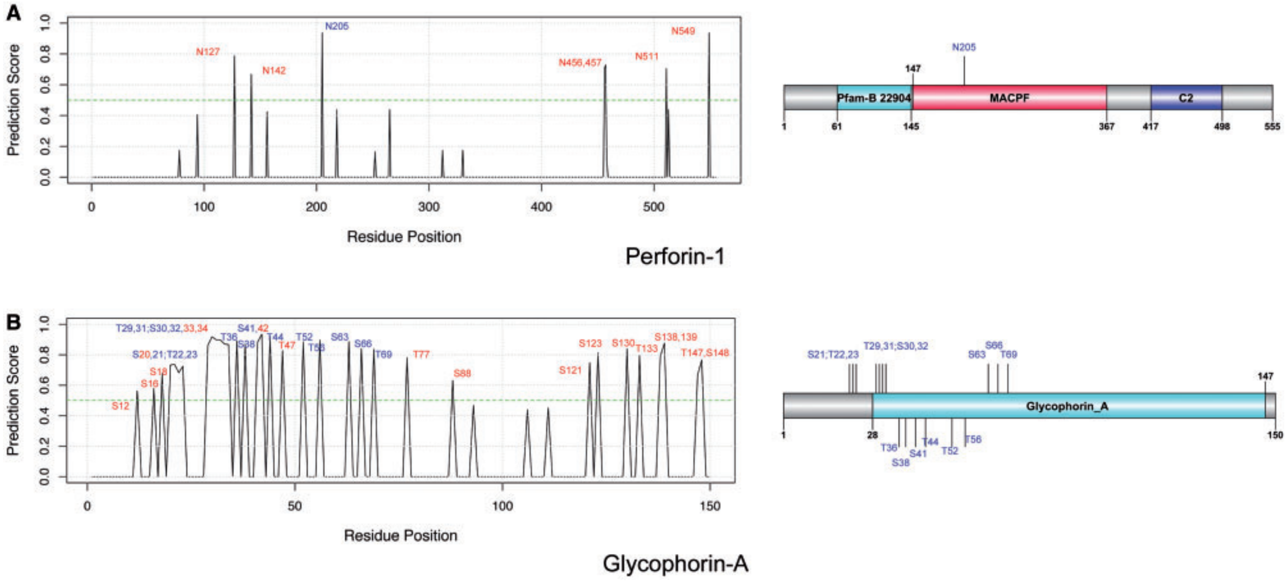

Fig. 3. The predicted glycosylation site probability using *GlycoMine* for two case study proteins (experimentally verified glycosylation sites are denoted as blue, whereas predicted glycosylation sites are shown in red). **(A)** The sequence scanning results and predicted N-glycosylation site locations with respect to the protein domains of perforin-1 (UniProt ID: P14222). **(B)** The sequence scanning results and predicted O-glycosylation site locations with respect to the protein domains of glycophorin-A (UniProt ID: P02724). The Pfam domains of these two proteins were drawn using DOG 2.0 (Ren *et al.*, 2009) (Color version of this figure is available at Bioinformatics online.)

sites in perforin-1 and 16 sites in glycophorin-A have been confirmed experimentally as glycosylation sites. These results suggest that *GlycoMine* can be a useful tool for *in silico* glycosylation site prediction.

4 Conclusion

In this study, we developed a novel machine-learning approach called *GlycoMine* for predicting C-, N- and O-linked glycosylation sites in the human proteome. *GlycoMine* combines various informative features from multiple sources. A two-step feature selection method based on mRMR or IG with IFS is used by *GlycoMine* to select the most contributory and useful features. Extensive cross-validation and independent tests using benchmark and independent datasets suggested that *GlycoMine* outperformed existing tools: GPP, EnsembleGly and NetNGlyc. This study represents the first systematic effort to select and characterize the most important features that are relevant for predicting the three major types of glycosylation using intensive feature selection techniques. A user-friendly web interface and a Java Applet are available as an implementation of this approach. Compared with experimental approaches, bioinformatics tools such as *GlycoMine* can provide a powerful and cost-effective approach that allows the proteome-wide prediction of potential glycosylation sites. These predictions are likely to have a number of important applications, such as identifying novel glycosylated substrate proteins, examining the effect of point mutations on the gain or loss of glycosylation sites, and the rational design of glycoproteins by inserting neo-glycosylation sites. We consider that the web server and Java program will be useful tools for large-scale glycosylation site prediction and they may facilitate hypothesis-driven experimental design.

Funding

This work was supported by grants from a Major Inter-disciplinary Project Grant Awarded by Monash University, National Natural Science Foundation of China [61202167, 61303169, 11250110508, 31350110507], Australian National Health and Medical Research Council (NHMRC) [490989] and Chinese Academy of Sciences (CAS). J.S. is an NHMRC Peter Doherty Fellow and a recipient of the Hundred Talents Program of CAS. J.C.W. is an Australia Research Council (ARC) Federation Fellow and an NHMRC Senior Principal Research Fellow.

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Brennan,A.J. *et al.* (2011) Protection from endogenous perforin: glycans and the C terminus regulate exocytic trafficking in cytotoxic lymphocytes. *Immunity*, **34**, 879–892.

Caragea,C. *et al.* (2007) Glycosylation site prediction using ensembles of Support Vector Machine classifiers. *BMC Bioinformatics*, **8**, 438.

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

Chen,K. *et al.* (2009) Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. *J. Comput. Chem.*, **30**, 163–172.

Chen,K. and Kurgan,L. (2007) PFRES: protein fold classification by using evolutionary information and predicted secondary structure. *Bioinformatics*, **23**, 2843–2850.

Chen,K. *et al.* (2008a) Prediction of protein structural class using novel evolutionary collocation-based sequence representation. *J. Comput. Chem.*, **29**, 1596–1604.

Chen,R. *et al.* (2009) Glycoproteomics analysis of human liver tissue by combination of multiple enzyme digestion and hydrazide chemistry. *J. Proteome Res.*, **8**, 651–661.

Chen,X. *et al.* (2013) Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. *Bioinformatics*, **29**, 1614–1622.

Chen,Y.Z. *et al.* (2008b) Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. *BMC Bioinformatics*, **9**, 101.

Christlet,T.H.T. and Veluraja,K. (2001) Database analysis of O-glycosylation sites in proteins. *Biophys. J.*, **80**, 952–960.

Chuang,G.Y. *et al.* (2012) Computational prediction of N-linked glycosylation incorporating structural properties and patterns. *Bioinformatics*, **28**, 2249–2255.

Doucey,M.A. *et al.* (1998) Protein C-mannosylation is enzyme-catalysed and uses dolichyl-phosahate-mannose as a precursor. *Mol. Biol. Cell*, **9**, 291–300.

Dwek,R.A. (1998) Biological importance of glycosylation. *Dev. Biol. Stand.*, **96**, 43–47.

Faraggi,E. *et al.* (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins*, **74**, 847–856.

Franceschini,A. *et al.* (2013) STRING v9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.

Gavel,Y. and Vonheijne,G. (1990) Sequence differences between glycosylated and nonglycosylated Asn-X-Thr Ser acceptor sites—implications for protein engineerin. *Protein Eng.*, **3**, 433–442.

Gupta,R. *et al.* (1999) O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res.*, **27**, 370–372.

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

Hamby,S.E. and Hirst,J.D. (2008) Prediction of glycosylation sites using random forests. *BMC Bioinformatics*, **9**, 500.

Hart,G.W. and Copeland,R.J. (2010) Glycomics hits the big time. *Cell*, **143**, 672–676.

Holland,R.C.G. *et al.* (2008) BioJava: an open-source framework for bioinformatics. *Bioinformatics*, **24**, 2096–2097.

Huang,Y. *et al.* (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.

Hunter,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database (vol 40, pg D306, 2011). *Nucleic Acids Res.*, **40**, 4725–4725.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Kawashima,S. *et al.* (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.*, **36**, D202–D205.

Kent,J.T. (1983) Information gain and a general measure of correlation. *Biometrika*, **70**, 163–173.

Krieg,J. *et al.* (1998) Recognition signal for C-mannosylation of Trp-7 in RNase 2 consists of sequence Trp-x-x-Trp. *Mol. Biol. Cell.*, **9**, 301–309.

Li,B.Q. *et al.* (2012) Prediction of protein cleavage site with feature selection by random forest. *PLoS One*, **7**, e45854.

Li,B.Q. *et al.* (2014) Classification of non-small cell lung cancer based on copy number alterations. *PLoS One*, **9**, e88300.

Mazola,Y. *et al.* (2011) Integrating bioinformatics tools to handle glycosylation. *PLoS Comput. Biol.*, **7**, e1002285.

Ohtsubo,K. and Marth,J.D. (2006) Glycosylation in cellular mechanisms of health and disease. *Cell*, **126**, 855–867.

Peng,H. *et al.* (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.

Pisano,A. *et al.* (1993) Glycosylation sites identified by solid-phase Edman degradation: O-linked glycosylation motifs on human glycophorin A. *Glycobiology*, **3**, 429–435.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

Ren,J. *et al.* (2009) DOG 1.0: illustrator of protein domain structures. *Cell Res.*, **19**, 271–273.

Saeys,Y. *et al.* (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.

Sasaki,K. *et al.* (2009) Support vector machine prediction of N- and O-glycosylation sites using whole sequence information and subcellular localization. *IPSJ Trans. Bioinformatics*, **2**, 11.

Sing,T. *et al.* (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

Song,J. *et al.* (2012) PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites. *PLoS One*, **7**, e50300.

Steentoft,C. *et al.* (2013) Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.*, **32**, 1478–1488.

Trost,B. and Kusalik,A. (2011) Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics*, **27**, 2927–2935.

Varki,A. (2007) Glycan-based interactions involving vertebrate sialic-acid-recognizing proteins. *Nature*, **446**, 1023–1029.

von der Lieth, C.W. *et al.* (2004) Bioinformatics for glycomics: status, methods, requirements and perspectives. *Brief. Bioinform.*, **5**, 164–178.

Wagner,M. *et al.* (2005) Linear regression models for solvent accessibility prediction in proteins. *J. Comput. Biol.*, **12**, 355–369.

Wang,M. *et al.* (2014) Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets. *Bioinformatics*, **30**, 71–80.

Ward,J.J. *et al.* (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, **20**, 2138–2139.

Wilson,B.S. *et al.* (1993) Endoplasmic reticulum-through-Golgi transport assay based on O-glycosylation of native glycophorin in permeabilized erythroleukemia cells: role for Gi3. *Proc. Natl. Acad. Sci. U S A.*, **90**, 1681–1685.

Yu,L. and Liu,H. (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In *Proceedings of the 10th International Conference on Machine Learning*. pp. 856–863.

Zaia,J. (2008) Mass spectrometry and the emerging field of glycomics. *Chem. Biol.*, **15**, 881–892.