BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 13 2014, pages 1908–1916
doi:10.1093/bioinformatics/btu139

Data and text mining

Advance Access publication March 10, 2014

An adaptive workflow coupled with Random Forest algorithm to identify intact N-glycopeptides detected from mass spectrometry

Suh-Yuen Liang[1,*], Sz-Wei Wu[2], Tsung-Hsien Pu[1], Fang-Yu Chang[2] and Kay-Hooi Khoo[2,*]

[1]Core Facilities for Protein Structural Analysis at Institute of Biological Chemistry and [2]Institute of Biological Chemistry, Academia Sinica, Taipei 115, Taiwan

Associate Editor: Igor Jurisica

ABSTRACT

Motivation: Despite many attempts for algorithm development in recent years, automated identification of intact glycopeptides from LC-MS^2 spectral data is still a challenge in both sensitivity and precision.

Results: We implemented a supervised machine learning algorithm, Random Forest, in an automated workflow to identify N-glycopeptides using spectral features derived from ion trap-based LC-MS^2 data. The workflow streamlined high-confident N-glycopeptide spectral data and enabled adaptive model optimization with respect to different sampling strategies, training sample size and feature set. A critical evaluation of the features important for glycopeptide identification further facilitated effective feature selection for model improvement. Using split sample testing method from 577 high-confident N-glycopeptide spectral data, we demonstrated that an optimal true-positive rate, precision and false-positive rate of 73, 88 and 10%, respectively, can be attained for overall N-glycopeptide identification

Availability and implementation: The workflow developed in this work and the application suite, Sweet-Heart, that the workflow supports for N-glycopeptide identification are available for download at http://sweet-heart.glycoproteomics.proteome.bc.sinica.edu.tw/.

Contact: syliang@gate.sinica.edu.tw or kkhoo@gate.sinica.edu.tw

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

Received on November 7, 2013; revised on February 10, 2014; accepted on March 4, 2014

1 INTRODUCTION

Protein glycosylation is a prevalent post-translational modification, which involves intricate sets of glycan biosynthesis pathways to modify specific sites of the protein sequence with attached oligosaccharide chains (Apweiler *et al.*, 1999). Unlike other post-translational modification such as phosphorylation, ubiquitination and methylation, protein glycosylation is structurally and compositionally diverse (Ohtsubo and Marth, 2006). As glycan moiety on the protein sequence affects the physicochemical properties of proteins through the change of structural conformation and binding motif, glycan heterogeneity endows glycoproteins with graduated functional versatilities in many biological processes. For example, different glycoproteins on the membrane of immune cells are responsible for triggering immune responses to different antigens (Daniels *et al.*, 2002; Kolarich *et al.*, 2012; Rudd *et al.*, 2001; Tate *et al.*, 2011; Toscano *et al.*, 2007). Glycoproteins can also modulate cell adhesion and signaling by differentiating the receptors on cell surface or binding to specific lectins (Haines and Irvine, 2003; Haltiwanger, 2002; Zhao *et al.*, 2008). Moreover, the spatial and temporal dynamics of glycoproteins highly regulate the development of organisms (Flanagan-Steet and Steet, 2013; Haltiwanger and Lowe, 2004) and aberrant glycoproteins during growth or development could cause diseases (Durand and Seta, 2000; Lehle *et al.*, 2006; Vivekanandan-Giri *et al.*, 2011). Given the significant roles of glycosylation in mediating the biological functions of glycoproteins and thus their clinical implications, there is currently a growing awareness and increasing demand for highly efficient and precise protein glycosylation profiling.

Recent advances in mass spectrometry (MS) technology have enabled compositional, structural and quantitative profiling of the glycomes and glycoproteomes (Morelle, 2009; Wuhrer *et al.*, 2007). Owing to the relatively large glycan size and compositional diversity, high-throughput glycopeptide identification with intact glycan moiety attached is precluded from the conventional approach of bottom-up database search with tandem mass spectrometry (MS^2) data. Consequently, current MS-based glycoproteomics technologies relie mostly on separate analysis of released glycans and deglycosylated peptides (Sullivan *et al.*, 2004; Zhang *et al.*, 2003), instead of intact glycopeptides, as it should be. However, by doing so, the valuable information on site-specific glycosylation and its associated heterogeneity, which is important for delineating the function and activity of the glycoprotein (Dube *et al.*, 1988; Freeze and Aebi, 2005; Sumer-Bayraktar *et al.*, 2011), is lost in the process. Recognizing the need to automate the laborious and often manually impossible process of mining the intact glycopeptide MS and MS^2 data, computational tools including GlycoMiner (Ozohanics *et al.*, 2008), GlyPID (Mayampurath *et al.*, 2011; Wu *et al.*, 2010), GlycoPep Grader (Woodin *et al.*, 2012), Peptoonist (Goldberg *et al.*, 2007), GlyDB (Ren *et al.*, 2007), Medicel Integrator Protein N-glycosylation suite (Joenvaara *et al.*, 2008), GlycoPeptideSearch (Pompach *et al.*, 2012), Byonic (Saba *et al.*, 2012) and GlycoPeptide Finder (Strum *et al.*, 2013) have been developed in recent years to facilitate site-specific glycosylation analysis. Each of these tools differs by the computational strategies and algorithms and was often optimized for handling datasets from only one or two of the several available liquid chromatography-tandem mass spectrometry (LC-MS^2) platforms. In general, however, all share the same common

*To whom correspondence should be addressed.

approach for realizing the prospect of intact glycopeptide identification. The essential steps are (i) accurate mass measurement in precursor MS spectra; (ii) deducing possible glycan composition and peptide backbone sequence from the LC-MS^2 spectra; (iii) pattern matching the marker fragment ions for peptide backbone and glycan against theoretical spectra from protein and glycan database, or attempt to *de novo* sequencing the glycan moiety; and (iv) scoring for best matches. For example, using collision-induced dissociation (CID) on a quadrupole/time-of-flight (Q/TOF) mass spectrometry, the diagnostic sugar oxonium ions, sequential neutral losses of glycosyl residues and the Y1 ion, which is the peptide backbone plus one HexNAc attached to Asn, can usually be detected in the LC-MS^2 spectra of glycopeptides (Ritchie *et al*., 2002). These are particularly useful for N-glycopeptide analysis because all N-glycosylations comprise a common pentasaccharide core of $Hex_3HexNAc_2$. By matching for monosaccharide mass difference and peptide database search based on Y1 ion with precursor mass measured at high accuracy, possible glycopeptides can be deduced (Wuhrer *et al*., 2007). More recently, higher-energy collision dissociation and electron transfer dissociation modes of fragmentation have been introduced as alternatives or in combination with the ion trap-based CID mode on the LTQ-Orbitrap hybrid platform to improve the sensitivity and confidence of glycopeptide identification (Mayampurath *et al*., 2011; Saba *et al*., 2012).

In all automated data analysis, scoring scheme is crucial to find the best glycopeptide from multiple candidates derived from MS spectra, especially for unknown protein with multiple glycosylation sites or complex sample. Most of the computational tools do provide this functionality, either using probability-based function, such as binominal distribution (Joenvaara *et al*., 2008; Mayampurath *et al*., 2011; Wu *et al*., 2010), or deterministic statistics, such as weighted average of important features associated with glycan composition and peptide backbones (Goldberg *et al*., 2007; Ozohanics *et al*., 2008; Strum *et al*., 2013; Woodin *et al*., 2012). Recently, we have developed a computational suite called Sweet-Heart (Wu *et al*., 2013), which implements a novel scoring scheme based on a supervised machine learning algorithm to first identify the glycosylation of intact N-glycopeptides, and then to further deduce the mass values of respective peptide backbones for further sequencing by either targeted multi-stage mass spectrometry (MS^3) or electron transfer dissociation experiments. Sweet-Heart requires no prior knowledge of glycan or peptide mass input and was shown to outperform currently available tools in its ability to process ion-trap–based CID data at higher sensitivity and specificity, especially in global proteomic applications. Supervised machine learning allows for systematic pattern identification from a large set of features in the training dataset and minimizes manual tuning for optimal model generalization (Barla *et al*., 2008), which is well-suited for an unbiased analysis of MS data and has been widely applied for identification of protein sequences, phosphorylation modification, metabolite fingerprinting and biomarkers (Anderson *et al*., 2003; Barla *et al*., 2008; Elias *et al*., 2004; Ge and Wong, 2008; Krambeck *et al*., 2009; Lahesmaa-Korpinen *et al*., 2010; Scott *et al*., 2010). Machine learning algorithms such as linear support vector machine and logistic regression have been used to facilitate glycosyl compositional annotation and scoring of released glycans at MS level based on known glycan compositions (Maxwell *et al*., 2012; Xu *et al*., 2012), but not yet adapted to distinguish the correct isomeric or isobaric entities from all possible peptide and glycan combinations, which requires tandem MS analysis. For instance, the mass of hexose+NeuAc is identical with fucose+NeuGc. Deamidation on peptide also confuses the assignment between 2 fucose and 1 NeuAc or between fucose+hexose and 1 NeuGc because of introduction of isobaric mass difference. The obstacle for intact glycopeptide identification using supervised machine learning on LC-MS^2 data is insufficient number of high-confident spectral data for model training. As ion fragmentation patterns of glycopeptides vary in different MS instrument, ionization and fragmentation techniques, it is not easy to obtain enough training data with comparable MS technologies from public domains for model building. Here, we describe a strategy to dynamically improve the supervised machine learning algorithm, Random Forest, implemented in Sweet-Heart, through the increment of high-confident N-glycopeptide spectral data collected from previous model predictions. With reciprocal model improvement, we aim for the flexibility of model optimization to enhance automated glycopeptide discovery with high sensitivity and precision.

2 METHODS

2.1 The source of the datasets

The glycopeptide MS^2 datasets were derived from LC-MS^2 analysis of samples prepared from human soluble epidermal growth factor receptor (sEGFR; amino acid sequence 25–650) and Human herpesvirus 2 (strain HG52) (HHV2H) envelope glycoprotein D expressed in Human Embryonic Kidney 293 (HEK 293) cells, mouse uterus fluid, mouse serum and digested membrane proteome of murine B-cell lymphoma (BCL1) cells. Glycopeptides were enriched from tryptic digests of each sample by using homemade amine-functionalized magnetic nanoparticles (Kuo *et al*., 2012) or by commercially available Oasis-Max cartridge (Waters). For the mouse serum, albumin was removed before trypsin digestion by following a previously reported trichloroacetic acid/acetone precipitation method (Chen *et al*., 2005). All samples were processed and analyzed by nanospray LC-MS^2 on an LTQ-Orbitrap Velos (Thermo Scientific) in CID mode as described in Wu *et al.* (2013).

2.2 Spectral feature extraction

Each raw spectral file was first processed using DeconMSn (Mayampurath *et al*., 2008) and then filtered using Mascot 2.3.02. All positively identified peptide spectra (with ion score greater than identity score) were removed from further N-glycopeptide analysis, as those were unlikely to be glycopeptides. The filtered LC-MS^2 spectra were analyzed for N-glycopeptides by Sweet-Heart. Detailed parameters used for database search and spectral feature extraction have previously been described (Wu *et al*., 2013). In brief, a semi-*de novo* module in Sweet-Heart finds partial glycan compositions based on sequential neutral loss of glycans and the oxonium ions in the LC-MS^2 spectrum. The N-GP combination module in Sweet-Heart generates all possible N-glycopeptides (glycan composition + peptide backbone) for each spectrum from species-specific protein database of the samples that contain the consensus sequence of Asn-X-Ser/Thr (X means any amino acid except proline) and are within ±10 ppm mass difference of the precursor from full scan mass spectrometry (MS^1) measurements. Knowledge-based rules are used as criteria to eliminate the unlikely glycan composition in mammals. All identified N-glycopeptide candidates are then in silico fragmented based on the rules schematically illustrated in Figure 1 and matched to the experimental

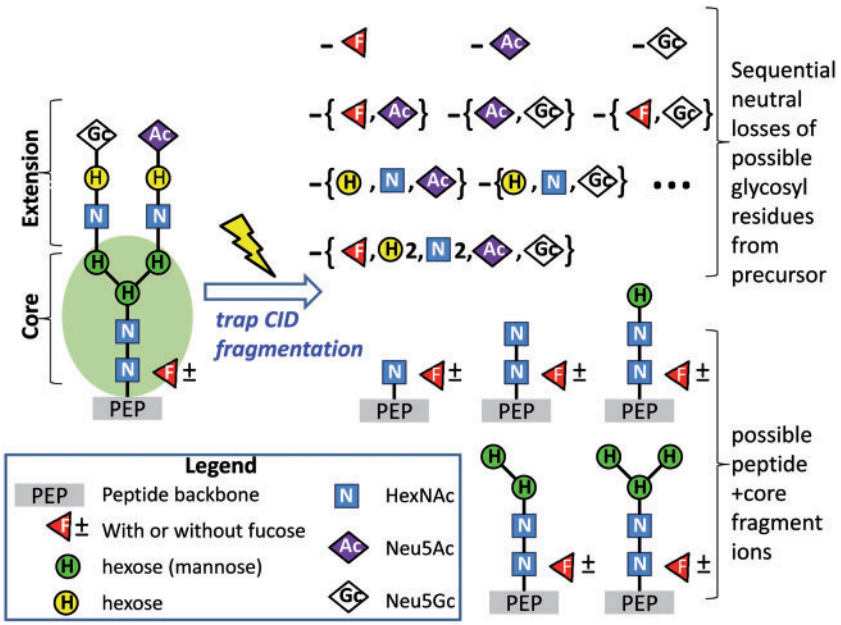

Fig. 1. A schematic illustration of the in silico fragmentation used as a basis for spectral features extracted from ion trap CID-based LC-MS^2 for prediction of N-glycopeptide with supervised machine learning. Sequential neutral losses of possible glycosyl residues from precursor and possible peptide+core fragment ions are spectral-specific and restricted by the minimal number of glycosyl residues as determined by the semi-*de novo* algorithms and the largest precursor found in the database. Only the common fragment ions associated with mammalian type of glycosylation as afforded by ion trap-based MS^2 are considered

MS^2 data for spectral feature extraction by considering only the b/y-ions of the same charge state z as the precursor, or z − 1. The spectral features typically observed in ion trap-based CID spectra, namely the sequential losses of common glycosyl residues found in mammals (fucose, hexose, HexNAc, Neu5Ac and Neu5Gc) from the trimannosyl core ($Man_3GlcNAc_2$) of an N-glycan attached to the peptide backbone and its probable fragmentation were included for machine learning in this study (Fig. 1). Features based on intensity or counts were normalized to the base peak of the spectrum (the most intense peak of the MS^2 spectrum) or the total matched peaks to ensure a uniform scale across all variables. All other features associated with the N-glycopeptides but not used for machine learning were descriptive features for dataset or glycopeptide, the original raw count features, or the oxonium ions used as glycan filter during semi *de novo*. A sample of feature data for all N-glycopeptide candidates and column descriptions can be found in Supplementary Tables S1.1 and S1.2, respectively.

2.3 Data for initial machine learning

The initial training dataset for machine learning was based on 106 manually validated spectra from human sEGFR and mouse uterus fluid. Sweet-Heart typically generates many possible N-glycopeptide candidates per spectrum, and true positive was defined here as the one best supported by manually verifiable MS^2 spectral features. All other candidate matches for the same spectrum were treated as true negatives. A pair of true-positive N-glycopeptide and one randomly selected true-negative counterpart from each spectrum were included in the training dataset, except for one spectrum that had only one true-positive candidate without any true-negative counterpart. Random sampling was applied to minimize any bias of the negative candidates, and paired sampling was taken to keep the training dataset as balanced as possible for better model performance (data not shown). To minimize the memory loading and computing time, five randomly selected instead of all true negatives for each of the 131 manually validated N-glycopeptide true positives from the sEGFR sample were included in the initial testing dataset. For all subsequent studies after the initial model building, all candidates computed by Sweet-Heart were included using the automated workflow developed.

Several well-known supervised machine learning algorithms including support vector machine [C-SVM function from LIBSVM (Chang and Lin, 2011)], Logistic Regression, Multi-Layer Perceptron, Naïve Bayes, C4.5 Decision Tree and Random Forest were evaluated on Weka 3.6.3 platform (Hall *et al.*, 2009) using 36 spectral features (Supplementary Table S1.2). Random Forest algorithm outperformed all others when the performances on the testing dataset were compared (Supplementary Table S2). Random Forest is an ensemble of decision trees generated by the bootstrap sampling method in which each tree is grown without pruning by a subset of variables randomly selected at each node. The data are classified according to the majority assignment of the class from the trees in the forest, and the internal error of misclassification is estimated by out-of-bag data not included in the bootstrap sampling for each tree (Breiman, 2001). Given that Random Forest performed well in the testing dataset and the ensemble algorithm is often suggested for pursuing better classification accuracy (Kotsiantis, 2007), Random forest was therefore selected for N-glycopeptide scoring based on the assigned probability value ranging from 0 to 1.

2.4 Workflow for adaptive machine learning for N-glycopeptide identification

The initial Random Forest model built on the 106 spectra using 36 normalized spectral features was incorporated in Sweet-Heart to facilitate data collection for future model improvement. We implemented a separate automated workflow to integrate all spectral data with N-glycopeptide confirmation and spectral features derived from Sweet-Heart in an embedded relational database for model improvement and optimization (Fig. 2). The workflow can dynamically evaluate the performance of Random Forest with respect to feature and sample selection (including

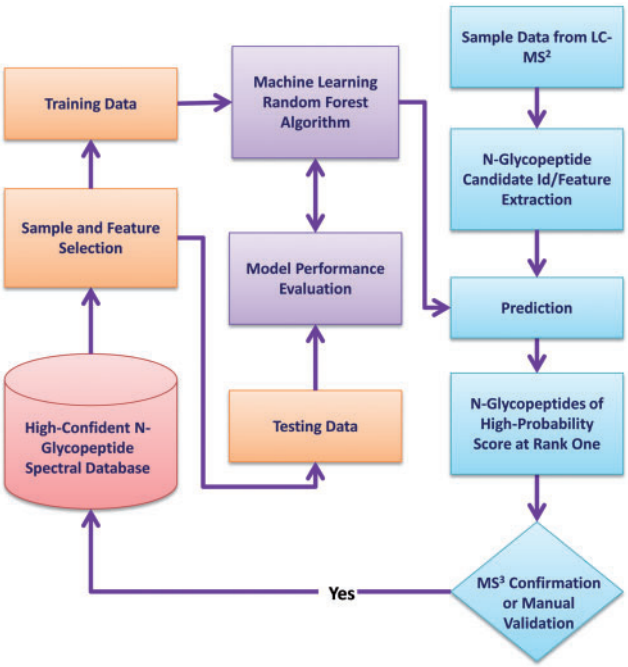

Fig. 2. An automated workflow, which dynamically optimizes the performance of a machine learning model, Random Forest, for N-glycopeptide prediction by incorporating the high-confident N-glycopeptide spectral data from previous model predictions confirmed by MS^3 experiment or manual validation

sampling methods and training data size). The number of trees in Random Forest and k-fold cross-validation can also be optimized or specified. These variables can be evaluated simultaneously or stepwise, depending on the purpose of the study. Demonstration of feature and sample selection is described in Section 2.5.

2.5 Implementation of dynamic feature set, sampling methods and training sample size for model optimization

To enable the adaptability of the N-glycopeptide identification tool to new samples or sampling strategies, we built the workflow with flexible specifications in feature sets, sampling methods for training dataset and training sample sizes. For demonstration, we included 577 high-confident N-glycopeptide spectra from samples described in method Section 2.1 for the following analyses. Among them, 131 spectra from human sEGFR sample and 74 spectra from mouse serum were confirmed both by manual validation on MS^2 evidence and MS^3 analyses for peptide sequence. All others were manually validated based solely on MS^2 data (Supplementary Table S4.1).

2.5.1 Effect of feature set on model performance and feature importance evaluation Two sets of features, each including 36 and 42 features (the same 36 features plus 6 new ones), were compared to assess how the model performed with different feature sets (Supplementary Table S1.2). Both feature sets differed by six features extracted from the error match of peptide + HexNAc and its fragment series in the core, including +HexNAcFuc, +HexNAc2Fuc, +HexHexNAc2Fuc, +Hex2HexNAc2Fuc and +Hex3HexNAc2Fuc. An error match of peptide + HexNAc is defined when there is a mass match of peptide + HexNAc plus any one of mannose, Neu5Ac and Neu5Gc residue or a mass match of peptide + HexNAc minus any one of fucose, mannose, Neu5Ac and Neu5Gc residue. All other five error-matched features are identified when the N-glycopeptide candidate does not have any fucose residue in the glycan composition, but a match of peptide core series with fucose is found. These six new features represent negatively matched features inconsistent with existing patterns in the MS^2 spectrum or glycan composition. A pair of one true-positive N-glycopeptide and one randomly selected true-negative N-glycopeptide candidate was selected from each of the 577 spectra for training dataset. For an unbiased evaluation of model performance between the two different feature sets, model performance was assessed based on 5-fold cross-validation method and was averaged over 30 training datasets in which the data only varied in the randomly selected negative N-glycopeptide candidates. Three parameters were used for model performance, including true-positive rate [true-positive rate (TPR) = total number of top-ranked true-positive candidates with probability ≥ 0.8/total number of true-positive candidates], precision (total number of top-ranked true-positive candidates with probability ≥ 0.8/total number of top-ranked candidates with probability ≥ 0.8) and false-positive rate [false-positive rate (FPR) = total number of top-ranked true-negative candidates with probability ≥ 0.8/total number of true-negative candidates). To optimize the number of trees in Random Forest algorithm, different numbers of trees, $n_1 + v(k-1)$, were simultaneously evaluated by iterating through k times defined by user along with the initial value n_1 and the incremental interval v. The optimal value was determined based on the internal error of misclassification estimated by out-of-bag data not included in the bootstrap sampling (Supplementary Table S3). Moreover, variable importance score generated by Random Forest was used to evaluate the relative importance of variable for each feature set. The variable importance score for the m^{th} variable is calculated by averaging the difference in the sum of corrected predictions from all the trees in the forest with and without random permutation of the value in the m^{th} variable using the out-of-bag data (Breiman, 2001). The higher the importance score, the greater influence of the variable on the model prediction.

2.5.2 Effect of sampling methods and training sample sizes on model performance Two sampling methods were compared to evaluate the difference in random and balanced glycoform sampling strategy for training dataset effect on the model performance using the optimized feature set following the result from Section 2.5.1. The glycoforms used here were based on the glycan composition, considering only fucose, hexose, HexNAc, Neu5Ac and Neu5Gc. The random sampling method chose a portion of 557 high-confident true-positive N-glycopeptide spectral data for training dataset. The corresponding true-negative candidate of the spectrum was randomly chosen to make both positive and negative candidates balanced in the training dataset. Balanced glycoform sampling method randomly chose a portion of glycoforms from high-confident true-positive N-glycopeptide data for training dataset. In each glycoform, equal numbers of true-positive N-glycopeptides were randomly chosen for the training dataset according to the selected portion. When the number of N-glycopeptides for a particular glycoform was not enough, it included only what was available and would not resample the same data twice. Similarly, the same number of true-negative candidates from the corresponding spectrum was randomly chosen to make both classes balanced in the training dataset. A total of 30 training datasets were generated to estimate the variability of the model because of different datasets. The portion unused for training dataset was included for testing data to ensure independence between the training and testing dataset.

Different training sample sizes were evaluated simultaneously with the sampling methods to identify the effect on model performance. To make the random and balanced glycoform sampling method comparable, the training sample size for both methods was kept the same. Five different training sample sizes were used for comparison. For balanced glycoform sampling method, 90% of 103 glycoforms was included for training dataset. The five different sizes of training datasets ranging from one to five true-positive N-glycopeptides per glycoform were chosen. The same

Table 1. The mean (±95% confidence interval) of TPR, FPR and precision from 5-fold cross-validation of Random Forest based on 30 randomly selected training datasets (577 true-positive and 544 false-positive N-glycopeptide data) with 36 and 42 features

Model performance	Number of features			Mean difference	*P* value
	36	42	N		
TPR	0.9650 ± 0.0015	0.9675 ± 0.0016	30	0.0025	0.0219*
FPR	0.03495 ± 0.00072	0.03247 ± 0.00077	30	−0.00248	0.0219*
Precision	0.96640 ± 0.00069	0.96879 ± 0.00074	30	0.00239	0.0219*

Student's *t*-test was used to test the difference in means. * denotes significant mean difference at $P < 0.05$.

numbers of true-negative N-glycopeptides in the corresponding spectrum were randomly selected to make the training dataset balanced. For random sampling method, 16, 28, 36, 43 and 48% of 577 true-positive N-glycopeptides were randomly selected for the five different training datasets. One true-negative N-glycopeptide from the corresponding spectrum was randomly selected to make the training dataset balanced. All spectra not included in the training dataset were used as testing dataset for evaluation of model performance.

2.6 Statistical analysis

R Statistics was used for statistical analyses (R Development Core Team, 2012). To test the difference of means, Shapiro and Bartlett test were conducted first to determine the normality of data distribution and homogeneity of variance. If the distribution is normal and variance is homogenous, ANOVA (for three groups or more) or Student's *t* test (for two groups) is used, otherwise Kruskal–Wallis test. For multiple comparisons, the Tukey procedure was chosen after ANOVA or the kruskalmc procedure (Siegel and Castellan, 1988) after Kruskal–Wallis test.

2.7 Software implementation

A fully automated workflow was developed in Java 1.6 with embedded relational database using JavaDB. Source codes for Random Forest and other machine learning modules were adopted from Livingston (2005) and Weka 3.6.3 (Hall *et al.*, 2009), respectively. The software had the following specifications, including (i) allowing the multiple file inputs of N-glycopeptide candidate data with spectral features in text format previously generated from the modules of Sweet-Heart (Supplementary Table S1) and storing them in an embedded relational database, (ii) accepting high-confident N-glycopeptide data in text format with MS^3 confirmation or manual validation (when MS^3 spectra are not available), as the true-positive candidates for model construction (Supplementary Table S1.3), (iii) providing two sampling strategies (random or balanced method in current version) with user-defined variables for construction of training dataset, (iv allowing dynamic evaluation of different number of trees grown in Random Forest model and feature set for model optimization and (v) generating a summary report of variable importance score and model performance based on k-fold cross-validation or split sample test.

3 RESULTS AND DISCUSSIONS

3.1 The optimization of the spectral feature set for identification of N-glycopeptides

To optimize the spectral feature set used in the model, we compared two sets of spectral features that differed by six features extracted from the error match of peptide + HexNAc and its fragment series in the core. The training dataset for Random Forest model was built using all 577 true-positive N-glycopeptides, and one true-negative N-glycopeptide was randomly selected from each respective spectrum to make the training dataset balanced. However, there were 23 spectra, which afforded only true-positive without any true-negative N-glycopeptide candidate. The training dataset was therefore slightly unbalanced. For these spectra with single N-glycopeptide candidate, only one glycan composition is possible for their respective precursor masses to find a matched peptide mass in the database after considering all possible combinations of glycosyl residues allowed by predefined biosynthesis rule. Based on the result of 5-fold cross-validation averaging 30 training datasets, the model built with 42 features consistently performed better ($P < 0.05$) in TPR, precision and FPR than that of 36 features (Table 1). The results indicate that the performance of the model could benefit from incorporating six additional features from the error match of peptide+HexNAc and its fragment ion series, albeit amounting to only a relatively small improvement.

3.2 Feature importance using variable importance score

We examined the key features for identification of the N-glycopeptides using the variable importance score of Random Forest algorithm from the same 30 training datasets described above. We found that the key features were similar between the models based on the 36 and 42 feature set (Fig. 3 and Supplementary Table S5). The percent intensity of matched peptide +HexNAc (PPEPN) was a predominant feature with an importance score more than twice of the second ranked feature, which was the percent intensity of matched peptide $+HexNAc_2$ (PPEPNN). PPEPN is commonly known as Y1 ion, which is the key ion used in most informatics solutions to identify a particular N-glycopeptide by CID MS^2 (Joenvaara *et al.*, 2008; Mayampurath *et al.*, 2011; Ozohanics *et al.*, 2008; Woodin *et al.*, 2012). The top five important features for the 36 and 42 feature set were the same and accounted for 55 and 49% of the total importance score, respectively. Four of the top five important features were extracted from the glycan core fragments with the peptide backbone. Among the features, which include ions from both the peptide + glycan core fragments and the glycan extension, the percent total number of matched peaks >5% of base peak intensity (PMTACH_F) was ranked the fifth. Proportionally, features from peptide + glycan core fragments had higher rank than those features with all ion types considered,

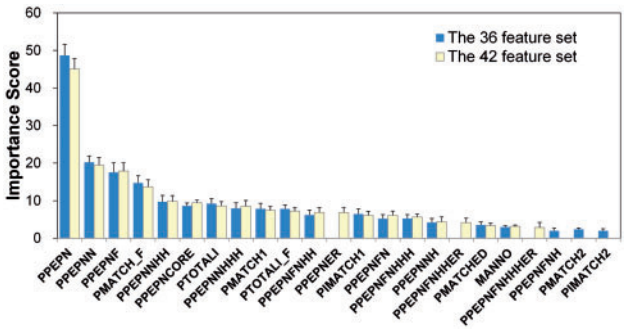

Fig. 3. Averaged variable importance score of the top 20 features from the Random Forest model based on the 36 and 42 feature sets. Error bar denotes one standard deviation

suggesting a stronger influence of peptide + glycan core fragments on model performance. It was also noted that the percent intensity of error-matched peptide + HexNAc (PPEPNER), percent intensity of error-matched peptide + $Hex_2HexNAc_2Fuc$ (PEPNFNHHER) and percent intensity of error-matched peptide + $Hex_3HexNAc_2Fuc$ (PEPNFNHHHER) were all ranked in the top 20 among the 42 feature set. The relatively higher importance scores of these error-matched features also signify their contribution to data classification and the improvement of model performance in the dataset consisting of the 42 feature set.

3.3 Effect of sampling method and training set size on the model performance

We also evaluated the model performance by using the portion of the spectra not included in the model so that the testing data were independent to the data used in the model. Unlike data for k-fold cross-validation, the testing datasets from split sample reflect the actual data distribution in a spectrum where there are usually large numbers of true-negative N-glycopeptides but one true-positive N-glycopeptide per spectrum. TPR, precision and FPR were again used to evaluate the effect of sampling method and training sample size on model performance based on the optimized feature set of 42 described in Section 3.1. Because N-glycopeptides of high mannose type ($Hex_{5-9}HexNAc_2$) often yielded peptide+HexNAc fragment ions at distinctively higher intensity, we compared the model performance between N-glycopeptides of high mannose and non-high mannose (with spectral evidence of additional fucose, HexNAc, Neu5Ac or Neu5Gc extending from the glycan core moiety) types. Figure 4 and Supplementary Table S6 show the averaged TPR, precision and FPR with ± standard deviation for total tested spectra and two subgroups (high mannose and non-high mannose type) with respect to five different training sample sizes and two sampling methods. The letters on top of each bar indicate multiple comparison results from Kruskal–Wallis test. Any two groups are considered significantly different ($P < 0.05$) if they do not share any letters between each other.

TPR was generally improved as the training sample size increased. However, the improvement was not as significant once it reached the size of 162 spectra. Random sampling and balanced glycoform sampling methods had similar effect on the TPR of total tested spectra but significantly differed in their effect on high mannose type. The TPR for high mannose type based on the random sampling method was consistently higher across the five training sets than that of the balanced glycoform sampling method. Although the effect of both sampling methods on non-high mannose type was not statistically different, there was a trend in which the balanced glycoform sampling method had slightly better TPR than the random sampling method. Regardless of the sampling method and training sample size, the TPR of the model was higher for high mannose type than non-high mannose type. The best TPR for predicting N-glycopeptides was all found at the largest training dataset (277 spectra). The best TPR ±1 SD for total tested spectra, high mannose type and non-high mannose type based on random sampling method were 0.732 ± 0.053, 0.920 ± 0.037 and 0.649 ± 0.076, respectively. The best TPR for total tested spectra, high mannose type and non-high mannose type based on balanced glycoform sampling method were 0.739 ± 0.046, 0.798 ± 0.065 and 0.677 ± 0.065, respectively.

The precision of the model for predicting total tested spectra and high mannose type was relatively independent of the size of training dataset and the sampling methods with an average value >0.8 for total tested data and 0.9 for high mannose type (Fig. 4). One exception is the significantly low precision from 92 training dataset based on the balanced glycoform sampling method. For non-high mannose type, the precision based on random sampling method was significantly ($P < 0.05$) higher than balanced glycoform sampling method across five different training datasets. Similar to TPR, the best precision was afforded by the largest training dataset, and high mannose type had better precision than non-high mannose type. The best precision ±1 SD for total tested spectra, high mannose type and non-high mannose type based on random sampling method were 0.878 ± 0.043, 0.9933 ± 0.0084 and 0.816 ± 0.066, respectively. The best precision for total tested spectra, high mannose type and non-high mannose type based on balanced glycoform sampling method was 0.842 ± 0.048, 0.9770 ± 0.0070 and 0.723 ± 0.077, respectively.

The FPR of the model for predicting total spectra and subgroups based on the random sampling method consistently outperformed the balanced glycoform sampling method across all five training datasets (Fig. 4). Increase in the size of training dataset generally improved the FPR of the model in most of the data groups, although the trend was less evident than that was found in TPR. The best FPR ±1 SD based on the random sampling method for total tested spectra and those of high mannose and non-high mannose types were obtained by using the largest training dataset at a value of 0.103 ± 0.036, 0.0067 ± 0.0084 and 0.150 ± 0.051, respectively. In contrast, the best FPR was not always attained by the largest training dataset using the balanced glycoform sampling method, although the corresponding values were not statistically different. For the purpose of data comparison, the FPR based on balanced glycoform sampling method for total tested spectra and those of high mannose and non-high mannose type were 0.138 ± 0.045, 0.0267 ± 0.0088 and 0.234 ± 0.083, respectively, by the largest training dataset.

Our studies indicate that all factors including feature set, type of glycoforms, training sample size and sampling method

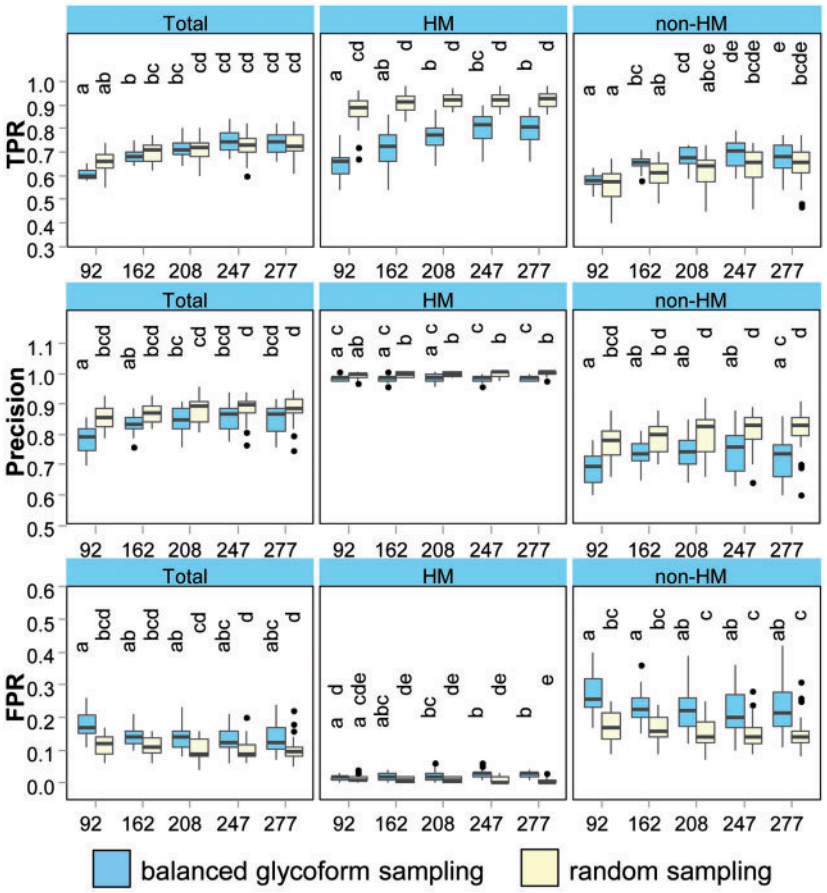

Fig. 4. Comparison of model performance with respect to five different training datasets (92, 162, 208, 247 and 277 spectra) and two sampling methods (balanced glycoform and random sampling). TPR, precision and FPR of total tested spectra (Total) and subgroups based on high mannose (HM) and non-high mannose type (non-HM) were calculated. The same letter above each bar in the graphs denotes no significant difference between the means at $P<0.05$ by Kruskal–Wallis test. The box corresponds to the 25th and 75th percentiles of the data. The whiskers of the box are the highest and lowest data points that are within 1.5 fold of the box length (inter-quartile range, IQR). Outliers that are outside 1.5×IQR are labeled in black dot

contributed to model performance. Random sampling method seems to have better overall model performance than the balanced glycoform sampling method, but the difference can be compensated by increasing the training sample size. Random sampling method allows a similar data distribution between training and testing dataset, whereas balanced glycoform sampling method ensures equal representation of glycoforms in the sample data to the training dataset. The glycoform distribution in our sample data is highly diverse and unbalanced. High mannose type consisted of only six different glycoforms ($Hex_{5-9}HexNAc_2$) but accounted for 31% of the data. Non-high mannose type consisted of the total 97 glycoforms, but one-fourth of them were unique (i.e. 26 non-high mannose glycoforms only occur once) (Supplementary Table S4.1). The high proportion of unique glycoforms in the sample data results in a higher dissimilarity in glycoform distribution between training and testing dataset with the balanced glycoform sampling method. As supervised machine learning performs better when training and testing data are similar, it is not surprising that the random sampling method is found better than the balanced glycoform sampling method in our study. However, the results from the balanced glycoform sampling method help to evaluate how our model would perform in the real scenario when the training dataset is not a sufficient representative of the real testing dataset. Under sub-optimal sampling strategy by the balanced glycoform sampling method, the overall Random Forest model performance was still robust with 74% TPR, 84% precision and 14% FPR.

Our studies further revealed that high mannose type was consistently better predicted by the model than non-high mannose type regardless of the sampling method. Both TPR and precision were above 90%, whereas FPR was <1% for high mannose type. According to our observation from the 577 N-glycopeptide spectra, the percent intensity of peptide+HexNAc and peptide+$HexNAc_2$ in high mannose type was usually at least five times higher than that of non-high mannose type on average (Supplementary Table S4.2). Because peptide+HexNAc and peptide+$HexNAc_2$ were the two most important features found in the model for N-glycopeptide identification, it may explain why high mannose type was better predicted by the model than non-high mannose type. At optimal performance, the latter afforded TPR, precision rate and FPR of 65, 82 and 15%,

respectively. Non-high mannose type as defined here includes the hybrid and complex forms, which result from further processing of the high mannose types in Golgi involving a series of glycosyltransferases (Neelamegham and Liu, 2011). The accessibility of the high mannose type glycans to these downstream diversifications is apparently dictated by the inherent physicochemical properties of the underlying peptide backbone (Thaysen-Andersen and Packer, 2012), in particular, the local primary structure around the glycosylation site (Petrescu *et al*., 2004). In fact, peptide sequence and structure features have been used for identification of glycosylation sites with Random Forest algorithm with accuracy ranging from 72.8 to 92.8% (Hamby and Hirst, 2008; Karnik *et al*., 2009). Inclusion of these peptide-specific features in our model should be beneficial, especially for prediction of non-high mannose type glycopeptides. The overall performance, as demonstrated by the relatively high precision rates and low FPR found for both types of glycoform, suggests the usability of the workflow for high-throughput N-glycopeptide discovery. Our results were also consistent with other findings, which demonstrated the robustness of Random Forest algorithm in MS-based proteomic applications, such as biomarker identification (Barrett and Cairns, 2008; Fusaro *et al*., 2009; Ge and Wong, 2008; Izmirlian, 2004; Wu *et al*., 2003). Future efforts should focus on the expansion of sample size and peptide-specific features to better represent the wide array of glycoforms in the training sample data.

ACKNOWLEDGEMENTS

The author gratefully acknowledge Drs Ying-Chih Liu (sEGFR), Cheng-Chung Lee (HHV2H envelope glycoprotein D), Yi-Yun Chen (mouse serum), Chu-Wei Kuo (mouse uterus fluid) and Chih-Wei Chien (digested membrane proteome of BCL1 cells) for providing the biological samples. The LC-MS^2 data were acquired at the Core Facilities for Protein Structural Analysis at Academia Sinica. The authors thank Dr Chien-Yu Chen (Department of Bio-industrial Mechatronics Engineering, National Taiwan University) for her critical comments on this work.

Funding: Taiwan National Core Facility Program for Biotechnology (NSC grant 100-2325-B-001-029, 101-2319-B-001-003) and Taiwan National Research Program for Genomic Medicine (NSC grant 99-3112-B-001-025).

Conflict of Interest: none declared.

REFERENCES

Anderson,D.C. *et al*. (2003) A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: Support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J. Proteome Res*., **2**, 137–146.

Apweiler,R. *et al*. (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta*, **1473**, 4–8.

Barla,A. *et al*. (2008) Machine learning methods for predictive proteomics. *Brief. Bioinform*., **9**, 119–128.

Barrett,J.H. and Cairns,D.A. (2008) Application of the random forest classification method to peaks detected from mass spectrometric proteomic profiles of cancer patients and controls. *Stat. Appl. Genet. Mol. Biol*., **7**, 1–20.

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

Chang,C.C. and Lin,C.J. (2011) LIBSVM: a Library for support vector machines, *ACM Trans. Intel. Syst. Technol*., **2**, 21–27.

Chen,Y.Y. *et al*. (2005) A modified protein precipitation procedure for efficient removal of albumin from serum. *Electrophoresis*, **26**, 2117–2127.

Daniels,M.A. *et al*. (2002) Sweet 'n' sour: the impact of differential glycosylation on T cell responses. *Nat. Immunol*., **3**, 903–910.

Dube,S. *et al*. (1988) Glycosylation at specific sites of erythropoietin is essential for biosynthesis, secretion, and biological function. *J. Biol. Chem*., **263**, 17516–17521.

Durand,G. and Seta,N. (2000) Protein glycosylation and diseases: blood and urinary oligosaccharides as markers for diagnosis and therapeutic monitoring. *Clin. Chem*., **46**, 795–805.

Elias,J.E. *et al*. (2004) Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol*., **22**, 214–219.

Flanagan-Steet,H.R. and Steet,R. (2013) "Casting" light on the role of glycosylation during embryonic development: insights from zebrafish. *Glycoconj. J*., **30**, 33–40.

Freeze,H.H. and Aebi,M. (2005) Altered glycan structures: the molecular basis of congenital disorders of glycosylation. *Curr. Opin. Struct. Biol*., **15**, 490–498.

Fusaro,V.A. *et al*. (2009) Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat. Biotechnol*., **27**, 190–198.

Ge,G.T. and Wong,G.W. (2008) Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. *BMC Bioinformatics*, **9**, 275–286.

Goldberg,D. *et al*. (2007) Automated N-glycopeptide identification using a combination of single- and tandem-MS. *J. Proteome Res*., **6**, 3995–4005.

Haines,N. and Irvine,K.D. (2003) Glycosylation regulates Notch signalling. *Nat. Rev. Mol. Cell Biol*., **4**, 786–797.

Hall,M. *et al*. (2009) The WEKA data mining software: an update. *SIGKDD Explor*., **11**, 10–18.

Haltiwanger,R.S. (2002) Regulation of signal transduction pathways in development by glycosylation. *Curr. Opin. Struct. Biol*., **12**, 593–598.

Haltiwanger,R.S. and Lowe,J.B. (2004) Role of glycosylation in development. *Ann. Rev. Biochem*., **73**, 491–537.

Hamby,S.E. and Hirst,J.D. (2008) Prediction of glycosylation sites using random forests. *BMC Bioinformatics*, **9**, 500–512.

Izmirlian,G. (2004) Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial. *Ann. N. Y. Acad. Sci*., **1020**, 154–174.

Joenvaara,S. *et al*. (2008) N-glycoproteomics–an automated workflow approach. *Glycobiology*, **18**, 339–349.

Karnik,S. *et al*. (2009) Identification of N-glycosylation sites with sequence and structural features employing random forests. In: Chaudhury,S. *et al*. (ed.) *Pattern Recognition and Machine Intelligence*. Springer, Berlin/Heidelberg, pp. 146–151.

Kolarich,D. *et al*. (2012) Glycomics, glycoproteomics and the immune system. *Curr. Opin. Chem. Biol*, **16**, 214–220.

Kotsiantis,S.B. (2007) Supervised machine learning: a review of classification techniques. *Front. Artif. Intel. Appl*., **160**, 3–24.

Krambeck,F.J. *et al*. (2009) A mathematical model to derive N-glycan structures and cellular enzyme activities from mass spectrometric data. *Glycobiology*, **19**, 1163–1175.

Kuo,C.W. *et al*. (2012) Rapid glycopeptide enrichment and N-glycosylation site mapping strategies based on amine-functionalized magnetic nanoparticles. *Anal. Bioanal. Chem*., **402**, 2765–2776.

Lahesmaa-Korpinen,A.M. *et al*. (2010) Integrated data management and validation platform for phosphorylated tandem mass spectrometry data. *Proteomics*, **10**, 3515–3524.

Lehle,L. *et al*. (2006) Protein glycosylation, conserved from yeast to man: a model organism helps elucidate congenital human diseases. *Angew. Chem. Int. Ed. Engl*., **45**, 6802–6818.

Livingston,F. (2005) Implementation of Breiman's random forest machine learning algorithm. ECE591Q. *Mach. Learn. Conf. Pap*, .

Maxwell,E. *et al*. (2012) GlycReSoft: A Software Package for Automated Recognition of Glycans from LC/MS Data. *PLoS ONE*, **7**, e45474.

Mayampurath,A.M. *et al*. (2008) DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. *Bioinformatics*, **24**, 1021–1023.

Mayampurath,A.M. *et al*. (2011) Improving confidence in detection and characterization of protein N-glycosylation sites and microheterogeneity. *Rapid Commun. Mass Spectrom*., **25**, 2007–2019.

Morelle,W. (2009) Analysis of glycosylation and other post-translational modifications by mass spectrometry. *Curr. Anal. Chem.*, **5**, 144–165.
Neelamegham,S. and Liu,G. (2011) Systems glycobiology: biochemical reaction networks regulating glycan structure and function. *Glycobiology*, **21**, 1541–1553.
Ohtsubo,K. and Marth,J.D. (2006) Glycosylation in cellular mechanisms of health and disease. *Cell*, **126**, 855–867.
Ozohanics,O. *et al.* (2008) GlycoMiner: a new software tool to elucidate glycopeptide composition. *Rapid Commun. Mass Spectrom.*, **22**, 3245–3254.
Petrescu,A.J. *et al.* (2004) Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. *Glycobiology*, **14**, 103–114.
Pompach,P. *et al.* (2012) Semi-automated identification of N-Glycopeptides by hydrophilic interaction chromatography, nano-reverse-phase LC-MS/MS, and glycan database search. *J. Proteome Res.*, **11**, 1728–1740.
R Development Core Team. (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Ren,J.M. *et al.* (2007) N-glycan structure annotation of glycopeptides using a linearized glycan Structure Database (GlyDB). *J. Proteome Res.*, **6**, 3162–3173.
Ritchie,M.A. *et al.* (2002) Precursor ion scanning for detection and structural characterization of heterogeneous glycopeptide mixtures. *J. Am. Soc. Mass Spectrom.*, **13**, 1065–1077.
Rudd,P.M. *et al.* (2001) Glycosylation and the immune system. *Science*, **291**, 2370–2376.
Saba,J. *et al.* (2012) Increasing the productivity of glycopeptides analysis by using higher-energy collision dissociation-accurate mass-product-dependent electron transfer dissociation. *Int. J. Proteomics*, **2012**, 560391.
Scott,I.M. *et al.* (2010) Enhancement of plant metabolite fingerprinting by machine learning. *Plant Physiol.*, **153**, 1506–1520.
Siegel,S. and Castellan,N.J.J. (1988) *Nonparametric Statistics for the Behavioral Sciences*. MacGraw Hill. Int. New York.
Strum,J.S. *et al.* (2013) Automated assignments of N- and o-site specific glycosylation with extensive glycan heterogeneity of glycoprotein mixtures. *Anal. Chem.*, **85**, 5666–5675.
Sullivan,B. *et al.* (2004) Selective detection of glycopeptides on ion trap mass spectrometers. *Anal. Chem.*, **76**, 3112–3118.
Sumer-Bayraktar,Z. *et al.* (2011) N-glycans modulate the function of human corticosteroid-binding globulin. *Mol. Cell. Proteomics*, **10,** M111 009100.
Tate,M.D. *et al.* (2011) Specific sites of N-linked glycosylation on the hemagglutinin of H1N1 subtype influenza A virus determine sensitivity to inhibitors of the innate immune system and virulence in mice. *J. Immunol.*, **187**, 1884–1894.
Thaysen-Andersen,M. and Packer,N.H. (2012) Site-specific glycoproteomics confirms that protein structure dictates formation of N-glycan type, core fucosylation and branching. *Glycobiology*, **22**, 1440–1452.
Toscano,M.A. *et al.* (2007) Differential glycosylation of TH1, TH2 and TH-17 effector cells selectively regulates susceptibility to cell death. *Nat. Immunol.*, **8**, 825–834.
Vivekanandan-Giri,A. *et al.* (2011) Urine glycoprotein profile reveals novel markers for chronic kidney disease. *Int. J. Proteomics*, **2011**, 214715.
Woodin,C.L. *et al.* (2012) GlycoPep grader: a web-based utility for assigning the composition of N-linked glycopeptides. *Anal. Chem.*, **84**, 4821–4829.
Wu,B.L. *et al.* (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, **19**, 1636–1643.
Wu,S.W. *et al.* (2013) Sweet-Heart—an integrated suite of enabling computational tools for automated MS2/MS3 sequencing and identification of glycopeptides. *J. Proteomics*, **84**, 1–16.
Wu,Y. *et al.* (2010) Mapping site-specific protein N-glycosylations through liquid chromatography/mass spectrometry and targeted tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, **24**, 965–972.
Wuhrer,M. *et al.* (2007) Glycoproteomics based on tandem mass spectrometry of glycopeptides. *J. Chromatogr. B*, **849**, 115–128.
Xu,G. *et al.* (2012) Improve accuracy and sensibility in glycan structure prediction by matching glycan isotope abundance. *Analytica Chimica Acta*, **743**, 80–89.
Zhang,H. *et al.* (2003) Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat. Biotechnol.*, **21**, 660–666.
Zhao,Y.Y. *et al.* (2008) Functional roles of N-glycans in cell signaling and cell adhesion in cancer. *Cancer Sci.*, **99**, 1304–1310.