BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 17 2012, pages 2249–2255
doi:10.1093/bioinformatics/bts426

Structural bioinformatics

Advance Access publication July 10, 2012

Computational prediction of N-linked glycosylation incorporating structural properties and patterns

Gwo-Yu Chuang, Jeffrey C. Boyington, M. Gordon Joyce, Jiang Zhu, Gary J. Nabel, Peter D. Kwong and Ivelin Georgiev*

Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MA 20892, USA

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: N-linked glycosylation occurs predominantly at the N-X-T/S motif, where X is any amino acid except proline. Not all N-X-T/S sequons are glycosylated, and a number of web servers for predicting N-linked glycan occupancy using sequence and/or residue pattern information have been developed. None of the currently available servers, however, utilizes protein structural information for the prediction of *N*-glycan occupancy.

Results: Here, we describe a novel classifier algorithm, NGlycPred, for the prediction of glycan occupancy at the N-X-T/S sequons. The algorithm utilizes both structural as well as residue pattern information and was trained on a set of glycosylated protein structures using the Random Forest algorithm. The best predictor achieved a balanced accuracy of 0.687 under 10-fold cross-validation on a curated dataset of 479 N-X-T/S sequons and outperformed sequence-based predictors when evaluated on the same dataset. The incorporation of structural information, including local contact order, surface accessibility/composition and secondary structure thus improves the prediction accuracy of glycan occupancy at the N-X-T/S consensus sequon.

Availability and Implementation: NGlycPred is freely available to non-commercial users as a web-based server at http://exon.niaid.nih.gov/nglycpred/.

Contact: ivelin.georgiev@nih.gov

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

Received on April 3, 2012; revised on June 15, 2012; accepted on July 2, 2012

1 INTRODUCTION

N-linked glycosylation, the attachment of glycan to the amide nitrogen of asparagine, is an ubiquitous co-/post-translational modification in eukaryotic cells that occurs as the nascent protein is extruded into the endoplasmic reticulum. Glycosylation plays an important role in a number of biological processes ranging from protein folding, sorting and degradation (Helenius and Aebi, 2004) to immune response (Rudd *et al.*, 2001). N-linked glycosylation occurs predominantly at the sequons N-X-T and N-X-S, with rare occurrence at N-X-C, where X is any standard amino acid except proline (Gavel and von Heijne, 1990). Studies suggest that about half of N-X-T/S sequons might be glycosylated (Petrescu *et al.*, 2004; Zielinska *et al.*, 2010), and an accurate algorithm that could predict the glycan occupancy (the presence or absence of a glycan) at these sequons would be useful for understanding and utilizing this ubiquitous co-/post-translational protein modification.

A number of algorithms trained with sequence or sequence-based information have been developed to improve the prediction of N-linked glycosylation (Caragea *et al.*, 2007; Gupta *et al.*, 2004; Hamby and Hirst, 2008; Sasaki *et al.*, 2009). To that end, Gupta *et al.* (2004) trained artificial neural networks on the surrounding sequence of the N-X-T/S sequons; Caragea *et al.* (2007) trained ensembles of support vector machine classifiers based on sequence neighbors; Hamby and Hirst (2008) used the random forest (RF) algorithm (Breiman, 2001) with pairwise patterns; and Sasaki *et al.* (2009) trained support vector machine classifiers based on surrounding sequence, whole protein sequence and sub-cellular localization information. Statistical analyses, however, have demonstrated that in addition to sequence differences, there are also structural differences between protein residues in spatial proximity to glycosylated versus non-glycosylated sites (Petrescu *et al.*, 2004). Thus, the incorporation of protein structural information might be expected to improve the prediction of the glycan occupancy. The incorporation of structural features in the prediction of N-linked glycan occupancy was previously reported in a conference proceeding (Karnik *et al.*, 2009); however, the statistical differences between structure-based and sequence-based predictors were not analyzed, and no publicly available software or web server was provided.

In this article, we report the development of, the theoretical basis for, and the properties of a novel predictor, NGlycPred, which predicts the glycan occupancy of N-X-T/S sequons of eukaryotic proteins. NGlycPred was trained on N-linked glycosylated proteins with structures available in the RCSB PDB database (Berman *et al.*, 2000) using the RF algorithm (Breiman, 2001). In essence, the RF algorithm uses multiple classification trees constructed with different boot strap samples from the original data, with each tree participating in the final classification. The algorithm is efficient and can handle numeric and nominal values simultaneously. The ability to predict *N*-glycan occupancy should allow for a better understanding of overall protein structure as well as the aforementioned biological processes and may also assist in the design of hyperglycosylated immunogens (Pantophlet *et al.*, 2003). Statistical analyses showed that

*To whom correspondence should be addressed.

NGlycPred performed significantly better than the sequence-based predictors generated using the same dataset. NGlycPred is available in the form of a web server and—to our knowledge—is the first publicly available web server to predict *N*-glycan occupancy at N-X-T/S sequons that is trained on information derived from experimental structures.

2 METHODS

2.1 Dataset

2.1.1 Structure Selection The pipeline for selecting the proteins and the N-X-T/S sequons to be used for training and testing is shown in Figure 1. Potential N-linked glycosylated PDB entries were downloaded from RSCB PDB on 29 July 2011 using the advanced search functionality with the following criteria: (1) Macromolecule Type: Contains Protein: Yes; (2) Chemical ID: NAG; (3) X-Ray Resolution: ≤2.5Å; and (4) Remove similar sequence at 70% identity, resulting in 525 PDB files. Structures with missing atoms were removed, resulting in 336 PDB files. Only the PDB files with N-X-T/S sequons and at least one direct asparagine-*N*-acetylglucosamine (ASN–NAG) linkage were kept, resulting in 262 PDB files. Finally, only eukaryotic expressed proteins (as annotated in the PDB header) were considered, resulting in 154 final PDB files.

2.1.2 Sequon Selection For each PDB file, N-X-T/S sequons were extracted from only one chain (with the highest number of NAGs) per unique protein molecule, resulting in 522 N-X-T/S sequons. Sequons were kept for further consideration if there were at least 10 residues flanking upstream and downstream of the sequon ASN residue, resulting in 479 N-X-T/S sequons. The sequons with ASN–NAG linkage were considered glycosylated. The sequons without ASN–NAG linkage were considered potentially non-glycosylated, except for those listed as N-linked glycosylated (experimentally verified sites only; potential and predicted sites were not taken into account) under 'Sequence annotation (Features)' section in UniProt (Apweiler *et al.*, 2004)—in these cases, the sequons were also considered glycosylated. The structure factor files corresponding to the PDB files with potential non-glycosylated residues, if available, were downloaded from the Electron Density Server (Kleywegt *et al.*, 2004) and $2F_O - F_C$ ($s = 1$) and $F_O - F_C$ ($s = 3$) maps were visualized using COOT (Emsley *et al.*, 2010) (F_O: structure factor observed; F_C: structure factor calculated). Electron density adjacent to the ASN ND2 was examined to assess whether a NAG group could be modeled. In a number of cases significant electron density was visible and real-space refinement of a NAG group into the density resulted in at least 10 out of the 14 non-H atoms of the NAG group inside either the $2F_O - F_C$ or $F_O - F_C$ electron density. Refinement using the deposited structure factors and PDB file with the newly introduced NAG groups using PHENIX (Adams *et al.*, 2010) indicated that the introduction of a NAG group to these residues was correct. These potential non-glycosylated ASN residues were then considered as glycosylated. The final dataset consisted of 97 non-glycosylated and 382 glycosylated N-X-T/S sequons (Supplementary Tables S1 and S2). To avoid positive bias due to the unbalanced dataset, four identical copies of the non-glycosylated sequons were presented in the dataset to oversample the minority class.

2.2 Learning properties

A number of structural, sequence and pattern properties were determined for each sequon to be used as variables for the learning process. All non-amino acid molecules, including glycans, were stripped off from the PDB file for structural property calculations.

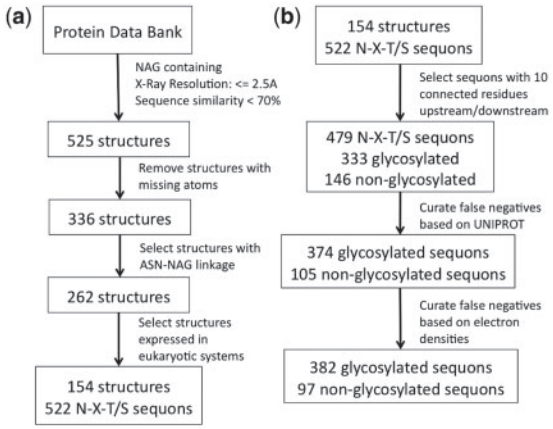

Fig. 1. Workflow for obtaining the input dataset. (**a**) Pipeline for selecting the PDB files. (**b**) Pipeline for selecting glycosylated and non-glycosylated sequons.

2.2.1 Structural properties

2.2.1.1 Surface accessibility The surface accessible area of the sequon ASN side chain was calculated by NACCESS (Hubbard and Thorton, 1993) using a probe of radius = 3Å (roughly the size of a monosaccharide). Surface accessibility (SA) was encoded as a binary variable with a cutoff of 3.7Å^2, as the entries with SA under this cutoff were highly biased toward non-glycosylated.

2.2.1.2 Surface composition Surface composition (SC) was defined as the amino acid composition of the residues surrounding the sequon ASN side chain amide nitrogen. Residues were considered if any of their side chain atoms were within a certain distance *r* to the the ASN side chain amide nitrogen. Six different types of encoding for the surface composition variable were evaluated, given by the combinations of three different distance thresholds *r* (4, 4.5 and 5Å) and two different length vectors: (a) a 20-bit integer vector with each bit representing the number of surrounding residues of each of the 20 standard amino acid types and (b) an 8-bit integer vector where amino acids with similar properties were grouped together (PRO, small hydrophobic (GLY, ALA), large hydrophobic (LEU, VAL, MET, ILE, CYS), aromatic (TYR, PHE, TRP), positively charged (ARG, LYS, HIS), negatively charged (ASP, GLU), polar O (SER, THR) and polar N (ASN, GLN)).

2.2.1.3 Secondary structure The secondary structures (SS) for the sequon ASN and five residues prior to/after the ASN were calculated using DSSP (Kabsch and Sander, 1983). 66 different types of encoding for the SS variables were evaluated: one bit (11) or two bits (55) encoded with the DSSP output of either one or two residues from the 11 residues. Each bit could be eight different values (alpha helix, isolated beta-bridge, extended strand, 3_{10} helix, pi helix, hydrogen bonded turn, bend and irregular).

2.2.1.4 Local contact order The local contact order (CO), used as an estimate of the degree of local protein folding, was defined as the following:

$$\mathrm{CO} = \frac{1}{l \cdot N_L} \cdot \sum^{N_L} S_{i,j}, \quad (1)$$

where *l* is the full length of the protein, L is the number of local residues under consideration (see below), N_L is the total number of residue contacts involving at least one of the local residues and $S_{i,j}$ is the sequence separation in terms of residue number between contacting residues *i* and *j*. Two residues *i* and *j* were defined as contacting residues

if any atom of one residue i from the L local residues was within 6 Å of any atom of the other residue j. Contact order was encoded as a binary variable (gT for local contact order equal to or greater than a threshold T and lT for local contact order less than T). Different combinations of L ($2w+1$, where $w=0$ to 10, centered at the sequon ASN) and T (0.1, 0.15, 0.2, 0.25 and 0.3) were used to calculate the weighted average of the Gini impurities (I_G) of the two subsets from the dataset split by CO:

$$I_G = \frac{N_{gT}}{N_{gT}+N_{lT}} \cdot \sum_{i=0}^{1} f_{gT,i}(1-f_{gT,i}) + \frac{N_{lT}}{N_{gT}+N_{lT}} \cdot \sum_{i=0}^{1} f_{lT,i}(1-f_{lT,i}), \quad (2)$$

where N_{gT} (N_{lT}) is the number of entries where local contact order is greater (less) than the threshold T and $f_{gT,i}$ ($f_{lT,i}$) is the frequency of entries in the gT (lT) set of $i=1$ (glycosylated) or 0 (non-glycosylated). To reduce the computational complexity, only the 10 combinations of L and T that gave the lowest Gini impurity when splitting the dataset were evaluated during the 10-fold cross-validation step (Supplementary Table S3).

2.2.2 Sequence (SEQ) w residues immediately prior to and after the sequon ASN residue. Nine different windows ($w=2$–10) were evaluated. Three different encoding methods were evaluated for the SEQ variable, resulting in a total of 27 encoding schemes:

- a K (=2w)-bit vector with each bit encoding the amino acid type at each position (standard scheme).
- a 20 K-bit binary-valued vector with each bit encoding the presence or absence of a specific amino acid at each position (0/1-based scheme).
- a 20 K-bit binary-valued vector with each 20 bits encoding the Blosum62 (Henikoff and Henikoff, 1992) string of a specific amino acid at each position (Blosum62-based scheme).

2.2.3 Pattern (PA) The ability for all single-position patterns within three residues from the sequon ASN to classify glycosylated from non-glycosylated sites was first evaluated in terms of Gini impurity as described above. The 10 patterns with the lowest Gini impurity are shown in Supplementary Table S4. The PA variable was implemented as a binary-valued vector of one to five bits where each bit represents the presence or absence of 1 of the top 10 patterns. The total number of possible encoding schemes for the PA variable was 637.

2.3 Algorithm training

RF (Breiman, 2001) models were generated and tested on our dataset using the 10-fold cross-validation scheme. We used the RF package implemented in Weka 3.6.5 (Hall *et al.*, 2009), with each model containing 1000 trees (−I 1000). The number of input variables to be used to determine the decision at a node was equal to $\log_2 M+1$, where M is the total number of input variables (the default setting). The maximum depth of each tree was set to 'unlimited'. The models were ranked in terms of balanced accuracy (BACC), defined as follows:

$$\mathrm{BACC} = \frac{\mathrm{TP}}{2\cdot(\mathrm{TP}+\mathrm{FN})} + \frac{\mathrm{TN}}{2\cdot(\mathrm{TN}+\mathrm{FP})}, \quad (3)$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. AUC (area under the ROC [receiver operating characteristic] curve) for each of the 10 tests in 10-fold cross-validation was also calculated by Weka, and the mean and 95% confidence interval (95% CI, estimated using mean $\pm$ ($2.262\cdot \mathrm{SE}/\sqrt{10}$), assuming a t-distribution with nine degrees of freedom) of AUC of the 10 tests were calculated (SE: standard error). Five different classes of RF predictors were generated and evaluated:

(1) Structural properties only: combinations of 3 or all 4 of SA, SC, SS and CO: a total of 9036 encoding combinations.
(2) Sequence only: a total of 27 combinations.
(3) Pattern only: a total of 637 combinations.
(4) Structural properties and sequence: top 200 encoding schemes from (1) (in terms of BACC) combined with all sequence combinations: a total of 5400 encoding combinations.
(5) Structural properties and pattern: top 200 encoding schemes from (1) combined with top 100 encoding schemes from (3) a total of 20 000 encoding combinations.

2.4 Statistical analysis

Welch's t-test (Welch, 1947) was used to evaluate the significance of the performance difference between all pairs of the best RF predictors (ranked by overall balanced accuracy) from each of the five classes. The balanced accuracy values from each of the 10-fold experiments were used for comparison. The calculated P-values were further adjusted using the false discovery rate approach (Benjamini and Hochberg, 1995). To analyze the stability of the performances of the best RF predictors from each of the five classes, a total of 200 cross-validation repeats were performed for each predictor with a shuffled dataset and a different random number seed for RF generation. The mean, standard deviation and 95% CI of the balanced accuracy values from the 200 runs were determined. The calculations were performed with the software R (R Development Core Team, 2005).

2.5 Comparison with available methods

The sequences of our dataset were submitted to the NetNGlyc (Gupta *et al.*, 2004), EnsembleGly (Caragea *et al.*, 2007) and GPP (Hamby and Hirst, 2008) servers to evaluate the prediction accuracies of these servers on the dataset described here. For each server, the default settings were used, except for EnsembleGly where the Blosum62 string kernel option was used as suggested in the original publication.

3 RESULTS

Five classes of RF predictors were defined: predictors trained on structural properties only, predictors trained on sequence only, predictors trained on pattern only, predictors trained on structural properties and sequence, and predictors trained on structural properties and pattern (see 'Methods' section). The performances of these predictors were evaluated based on 10-fold cross-validation of our dataset. The performances of the best RF predictors, ranked by balanced accuracy (the average value of true-positive rate and false-positive rate), from each of the five classes are shown in Table 1. The adjusted P-values from Welch's t-test on the pairwise comparison of these five predictors are shown in Supplementary Table S5. The 10 best models from each class are shown in Tables 2 and 3, Supplementary Tables S6 and S7. The performance of the best predictors from each of the five classes was generally stable, as suggested by the results from a total of 200 cross-validation calculations performed on different shuffled datasets (Supplementary Table S8).

3.1 Predictors trained on structural properties only

The best predictor from structural properties only, trained on all four structural properties, achieved a balanced accuracy of 0.663 in the 10-fold cross-validation of our dataset, with true-positive rate of 0.770 and false-positive rate of 0.443. Six of the top 10 predictors trained on all four structural properties (Table 2);

three trained with surface accessibility, secondary structure and surface composition and one trained with surface accessibility, secondary structure and local contact order. For all top 10 predictors, the secondary structure property was represented by DSSP values at the sequon ASN position and the fifth residue position preceding the ASN. In the top 10 list, eight of the nine predictors that were trained with the surface composition property used the eight-bit vector implementation, suggesting that grouping of residue types for the surface composition attribute can improve the accuracy.

3.2 Predictors trained on sequence only

The best predictor trained on sequence achieved a balanced accuracy of 0.547 in the 10-fold cross-validation of our dataset, with true-positive rate of 0.950 and false-positive rate of 0.856. The balanced accuracy of this predictor was significantly lower than that of the best predictor trained on structural properties ($P < 0.005$). From the list of the top 10 predictors of the class (Table 3) it can be seen that 0/1- or Blosum62-based schemes generated more accurate predictions. Compared with the predictors trained on structural properties, predictors trained on sequence had a much higher false-positive rate. The same was observed in the evaluation of other sequence-based web servers (NetNGlyc, EnsembleGly, GPP) on the dataset described here (Supplementary Table S10). For all three methods, the false-positive rates were >0.7.

3.3 Predictors trained on patterns only

The best predictor trained on sequence-based patterns, defined as the presence of a specific amino acid type at a specific position relative to sequon ASN, achieved a balanced accuracy of 0.623 in the 10-fold cross-validation of our dataset, with true-positive rate of 0.607 and false-positive rate of 0.361. The balanced accuracy of this predictor was lower than that of the best predictor trained on structural properties but higher than that of the best predictor trained on sequence. Compared with predictors trained on sequence, the predictors trained on pattern had a much lower false-positive rate. Notably, 9 of the top 10 models (Table 4) used patterns +2 THR and −1 GLY, emphasizing the importance of these two patterns.

Table 1. The performances of the best RF predictors, ranked by balanced accuracy, from each of the five classes

Properties	True-positive rate	False-positive rate	AUC[a]	Balanced accuracy
Structure + pattern	0.694	0.320	0.703	0.687
Structure	0.770	0.443	0.665	0.663
Pattern	0.607	0.361	0.636	0.623
Structure + sequence	0.932	0.784	0.646	0.574
Sequence	0.950	0.856	0.518	0.547

[a]Average AUC values of the 10 separate predictions in 10-fold cross-validation.

Table 3. The performances of the top 10 predictors trained on sequence, ranked by balanced accuracy

Encoding	Window (w)	True-positive rate	False-positive rate	AUC[a]	Balanced accuracy
0/1	3	0.950	0.856	0.518	0.547
Blosum62	3	0.961	0.876	0.548	0.542
0/1	5	0.974	0.907	0.510	0.533
Blosum62	2	0.877	0.814	0.561	0.531
0/1	4	0.955	0.897	0.525	0.529
0/1	7	0.976	0.918	0.500	0.529
0/1	6	0.976	0.918	0.481	0.529
standard	2	0.741	0.691	0.570	0.525
0/1	8	0.976	0.928	0.487	0.524
Blosum62	7	0.984	0.938	0.564	0.523

[a]Average AUC values of the 10 separate predictions in 10-fold cross-validation.

Table 2. The performances of the top 10 predictors trained on structural properties, ranked by balanced accuracy

Surface accessibility	Secondary structure	Surface composition		Local contact order		True-positive rate	False-positive rate	AUC[a]	Balanced accuracy
+	pos = −5,0	grouped	$r = 5.0$	$L = 13$	$T = 0.3$	0.770	0.443	0.665	0.663
+	pos = −5,0	grouped	$r = 5.0$	—	—	0.764	0.443	0.668	0.661
+	pos = −5,0	—	—	$L = 13$	$T = 0.3$	0.712	0.402	0.639	0.655
+	pos = −5,0	grouped	$r = 5.0$	$L = 13$	$T = 0.3$	0.772	0.464	0.666	0.654
+	pos = −5,0	ungrouped	$r = 5.0$	—	—	0.751	0.443	0.661	0.654
+	pos = −5,0	grouped	$r = 5.0$	$L = 13$	$T = 0.3$	0.759	0.454	0.651	0.653
+	pos = −5,0	grouped	$r = 5.0$	$L = 13$	$T = 0.3$	0.767	0.464	0.664	0.652
+	pos = −5,0	grouped	$r = 4.5$	—	—	0.707	0.402	0.631	0.652
+	pos = −5,0	grouped	$r = 5.0$	$L = 13$	$T = 0.3$	0.746	0.443	0.659	0.651
+	pos = −5,0	grouped	$r = 5.0$	$L = 13$	$T = 0.3$	0.764	0.464	0.660	0.650

Presence of a property is indicated by '+' or specification of the parameters. Absence of a property is indicated by '−'. 'Secondary structure' is specified by the residue positions for which the DSSP results were used, relative to sequon ASN. 'Surface composition' is specified by two parameters: grouped (eight-bit) or ungrouped (20-bit), and distance threshold r. 'Local contact order' is specified by length L and threshold T.
[a]Average AUC values of the 10 separate predictions in 10-fold cross-validation.

Table 4. The performances of the top 10 predictors trained on patterns, ranked by balanced accuracy

Patterns	True-positive rate	False-positive rate	AUC[a]	Balanced accuracy
+2T/−1G/−1V/−3W/−2E	0.607	0.361	0.636	0.623
+3P/−1V/−1N/−3Y/−1P	0.846	0.608	0.614	0.619
+2T/−1G/−1V/−2E	0.599	0.361	0.627	0.619
+2T/−1G/+1L/−1V/−2E	0.599	0.361	0.631	0.619
+2T/−1G/+1L/−3W/−2E	0.628	0.392	0.633	0.618
+2T/−1G/+1L/−1V/−3W	0.626	0.392	0.642	0.617
+2T/−1G/−1V/−3W	0.626	0.392	0.635	0.617
+2T/−1G/−3W/−2E	0.626	0.392	0.631	0.617
+2T/−1G/−3W/−2E/−1P	0.613	0.381	0.637	0.616
+2T/−1G/−1V/−2E/−1P	0.581	0.351	0.637	0.615

Each pattern is shown as the residue number relative to sequon ASN, followed by amino acid type.
[a]Average AUC values of the 10 separate predictions in 10-fold cross-validation.

3.4 Predictors trained on structural properties and sequence

The best predictor trained on structural properties and sequence achieved a balanced accuracy of 0.574 in the 10-fold cross-validation of our dataset, with true-positive rate of 0.932 and false-positive rate of 0.784 (Supplementary Table S6). The balanced accuracy of this predictor was higher than that of the best predictor trained on sequence but was significantly lower than that of the best predictor trained on structural properties ($P < 0.05$).

3.5 Predictors trained on structural properties and patterns

The best predictor trained on structural properties and sequenced-based patterns achieved a balanced accuracy of 0.687 and an average ROC AUC of 0.703 in the 10-fold cross-validation of our dataset, with true-positive rate of 0.694 and false-positive rate of 0.320. This predictor was the best predictor among all five classes. The balanced accuracy of this predictor was significantly higher than that of the best predictor trained on sequence ($P < 0.001$) and that of the best predictor trained on structure and sequence ($P < 0.05$). Of note, this predictor used only three out of the four structural properties (SA, SS and CO; Supplementary Table S7).

3.6 NGlycPred server

The best predictor trained on structural properties and sequenced-based patterns was implemented as the NGlycPred Server, available to the general public. The server identifies the N-X-T/S sequons from the given input PDB file and predicts the glycan occupancy of each sequon. The computational time for each execution depends on the size of the input PDB file. For an input PDB file of <1000 amino acid residues, the predictions could be generated within <30 s. For an input PDB file of >2000 amino acid residues, run-times could take up to a few minutes.

4 DISCUSSION

Our results indicate that the incorporation of structural information can improve the prediction of glycan occupancy of N-X-T/S sequons. In our study, the RF predictors generated using structural information, with or without additional pattern information, outperformed the predictors trained on sequence information with statistical significance, as well as a number of other sequence-based servers that were evaluated on our dataset (Supplementary Table S10). Overall, a comparison between different predictors and their underlying algorithms is complicated by the fact that the predictors (and their respective servers) were developed and optimized on different datasets. Nevertheless, analyses based on our dataset clearly demonstrated that predictors generated using structural properties could give better predictions than those generated solely by sequence information. Differences in local structural features between glycosylated and non-glycosylated sequons were not only observed in our dataset but also in previous studies (Petrescu *et al.*, 2004). Moreover, recent evidence has shown that specific amino acid side chains could directly stabilize the first *N*-acetylglucosamine of the glycan (Culyba *et al.*, 2011), suggesting that in addition to sequence, structural features could directly affect glycan occupancy.

A comparison of the best predictors built on all four and three of the four structural properties (Table 5 and Supplementary Table S9) suggested that the SS property might be the most important factor in the improvement of the prediction accuracy. Surface accessibility played a lesser role: although the sequons with SA less than the 3.7Å^2 threshold had a much higher tendency to be non-glycosylated (13 of the 97 non-glycosylated sequons compared with 2 of the 382 glycosylated sequons), SA of >96% of the entries (464 out of 479) were above the threshold, and thus the property was less effective for improving prediction accuracy. The reason for the lesser effect of local contact order on prediction accuracy could be similar: for example, in the case of L = 13, non-glycosylated entries were enriched in the set with CO >0.3 (4 of the 97 non-glycosylated sequons compared with 4 of the 382 glycosylated sequons); CO of >98% (471 out of 479), however, were below the 0.3 threshold.

Table 5. The performances of the top predictors encoded by three or four structural properties, ranked by balanced accuracy

Properties	True-positive rate	False-positive rate	AUC[a]	AUC 95% CI	Balanced accuracy
SA + SC + SS + CO	0.770	0.443	0.665	0.591–0.739	0.663
SA + SC + SS	0.764	0.443	0.668	0.595–0.741	0.661
SA + SS + CO	0.712	0.402	0.639	0.584–0.694	0.655
SC + SS + CO	0.762	0.464	0.638	0.595–0.681	0.649
SA + SC + CO	0.736	0.526	0.597	0.551–0.643	0.605

[a]Average AUC values of the 10 separate predictions in 10-fold cross-validation.

Several limitations are inherent to our current approach. The 154 protein structures used to generate the dataset only encompass a very small subset of the eukaryotic proteome. Although we only selected crystal structures with resolution better than 2.5Å to generate the dataset, the presence or the conformation of specific amino acid or sugar residues could still be ambiguous in some cases. To increase the reliability of the computational predictions, the dataset used in our study was extensively curated. Since the eukaryotic and prokaryotic N-linked glycosylation schemes are different (Kowarik *et al*., 2006), sequons were only selected from PDB files where the proteins were expressed in eukaryotic systems. To reduce the incidences of false negatives in the dataset, the sequons without ASN-NAG linkage from the PDB files were considered glycosylated if they were annotated as such in UniProt (Apweiler *et al*., 2004). The incidences of false negatives were further reduced by considering as glycosylated sequons for which the ASN-NAG linkage could be modeled in the electron densities from the PDB Structure Factor file. Nevertheless, false negative sequons could still be present in the dataset, in cases where the glycosylation site is occupied but both the glycan electron densities were absent and the site was not annotated as glycosylated in UniProt. Furthermore, the ratio of glycosylated to non-glycosylated entries in the dataset was adjusted to roughly 1:1, while the actual ratio of glycosylated to non-glycosylated sites in reality is unknown. Additional curation of the dataset could thus further improve the accuracy of the predictions. Tuning of some of the RF input parameters, such as the maximum depth of each tree and the number of input variables to be randomly selected at each node, could further optimize the performance of the different predictors.

The structural features chosen for the NGlycPred algorithm are less sensitive to the exact coordinates of the protein, and therefore should be suitable for use with homology models. In our analysis, we noticed that the knowledge of side chain torsion angles improved the prediction of N-linked glycan occupancy (data not shown). However, since side chain torsions are more difficult to predict for homology models and might differ dramatically before/after glycosylation, we chose not to include this feature in our models so that the NGlycPred algorithm would be applicable to both crystal structures and homology models. Nonetheless, it should be noted that the accuracy of the predictions may be affected by the quality of the homology models. Also, as NGlycPred uses structural properties of the protein as input, different predictions would be generated for sequons on sequence-identical domains if the tertiary/quaternary context is different (for example, sequons on the outer-domain of HIV-1 gp120 monomer versus sequons on an outer-domain-only construct).

Finally, we note that differences in glycosylation do occur between different eukaryotic species (e.g. mammalian versus insect) as well as in different tissues of the same organism; further improvements in *N*-glycan prediction may be needed to incorporate these variables. Furthermore, as the addition of N-linked glycosylation typically occurs during protein translation, with the N-X-T/S sequon recognized by the glycosylation machinery as the nascent polypeptide is synthesized and extruded into the endoplasmic reticulum, the theoretical link between structure-based information and *N*-glycan prediction is unclear: the protein is not yet folded when *N*-glycans are incorporated. Differences in *N*-glycan prediction accuracy between artificially incorporated sites and naturally evolved sites may provide insight into this conundrum.

The NGlycPred algorithm described here should provide better prediction of glycan occupancy, and such a prediction is likely to have a number of applications. For example, the ability to silence immunodominant epitopes reliably, through targeted addition of *N*-glycans, should contribute to immunogen design. *N*-glycan can also affect half-life and trafficking of protein therapeutics and correct prediction of glycan occupancy would be of utility. Thus, despite recent advancement in experimental technology to detect N-linked glycosylation (Kaji *et al*., 2007; Zielinska *et al*., 2010), a computational algorithm that can quickly identify glycosylation sequons with higher probabilities of glycan occupancy should be of use.

ACKNOWLEDGEMENTS

We thank members of the Structural Biology Section and Structural Bioinformatics Core at the NIH Vaccine Research Center for comments on this article, Jeff Skinner from Office of Cyber Infrastructure and Computational Biology (OCICB/NIAID/NIH) for assistance with statistical analysis, and David Liou and Yong-Jian Guo from OCICB for assistance with server implementation.

Funding: Intramural Research Program (National Institute of Allergy and Infectious Diseases, NIH, USA).

Conflict of Interest: none declared.

REFERENCES

Adams,P.D. *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 213–221.

Apweiler,R. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B. Methodol.*, **57**, 289–300.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

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

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

Culyba,E.K. *et al.* (2011) Protein native-state stabilization by placing aromatic side chains in N-glycosylated reverse turns. *Science*, **331**, 571–575.

Emsley,P. *et al.* (2010) Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 486–501.

Gavel,Y. and von Heijne,G. (1990) Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. *Protein Eng.*, **3**, 433–442.

Gupta,R. *et al.* (2004) Prediction of N-glycosylation sites in human proteins. http://www.cbs.dtu.dk/services/NetNGlyc/.

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

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

Helenius,A. and Aebi,M. (2004) Roles of N-linked glycans in the endoplasmic reticulum. *Annu. Rev. Biochem.*, **73**, 1019–1049.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA.*, **89**, 10915–10919.

Hubbard,S. and Thorton,J. (1993) NACCESS. Software.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kaji,H. *et al.* (2007) Proteomics reveals N-linked glycoprotein diversity in *Caenorhabditis elegans* and suggests an atypical translocation mechanism for integral membrane proteins. *Mol. Cell Proteomics*, **6**, 2100–2109.

Karnik,S. *et al.* (2009) Identification of n-glycosylation sites with sequence and structural features employing random forests. In Santanu,C., Sushmita,M., Murthy,C.A., Sastry,P.S. and Pal,S.K. (eds.) *Pattern Recognition and Machine Intelligence, Third International Conference, PReMI 2009*. Springer, New Delhi, India, pp. 146–151.

Kleywegt,G.J. *et al.* (2004) The Uppsala electron-density server. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 2240–2249.

Kowarik,M. *et al.* (2006) N-linked glycosylation of folded proteins by the bacterial oligosaccharyltransferase. *Science*, **314**, 1148–1150.

Pantophlet,R. *et al.* (2003) Hyperglycosylated mutants of human immunodeficiency virus (HIV) type 1 monomeric gp120 as novel antigens for HIV vaccine design. *J. Virol.*, **77**, 5889–5901.

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.

R Development Core Team. (2005) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rudd,P.M. *et al.* (2001) Glycosylation and the immune system. *Science*, **291**, 2370–2376.

Sasaki,K. *et al.* (2009) Support vector machine prediction of n- and o-glycosylation sites using whole sequence information and subcellular localization. *IPSJ Trans. Bioinform.*, **2**, 25–35.

Welch,B.L. (1947) The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, **34**, 28–35.

Zielinska,D.F. *et al.* (2010) Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell*, **141**, 897–907.