Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Structure-based Comparative Analysis and Prediction of N-linked Glycosylation Sites in Evolutionarily Distant Eukaryotes

Phuc Vinh Nguyen Lam [1,3], Radoslav Goldman [2], Konstantinos Karagiannis [3], Tejas Narsule [3], Vahan Simonyan [4], Valerii Soika [4], Raja Mazumder [3,*]

[1] *Life Sciences Department, Paris Diderot University, Paris 75013, France*
[2] *Department of Oncology, Georgetown University, Washington, DC 20057, USA*
[3] *Department of Biochemistry and Molecular Biology, George Washington University Medical Center, Washington, DC 20037, USA*
[4] *Center for Biologics Evaluation and Research, Food and Drug Administration, Rockville, MD 20852, USA*

Received 17 September 2012; revised 2 November 2012; accepted 12 November 2012
Available online 28 February 2013

KEYWORDS

N-linked glycosylation;
Gain and loss of glycosylation;
nsSNP;
nsSNV;
Variation

Abstract The asparagine-X-serine/threonine (NXS/T) motif, where X is any amino acid except proline, is the consensus motif for N-linked glycosylation. Significant numbers of high-resolution crystal structures of glycosylated proteins allow us to carry out structural analysis of the N-linked glycosylation sites (NGS). Our analysis shows that there is enough structural information from diverse glycoproteins to allow the development of rules which can be used to predict NGS. A Python-based tool was developed to investigate asparagines implicated in N-glycosylation in five species: *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Arabidopsis thaliana* and *Saccharomyces cerevisiae*. Our analysis shows that 78% of all asparagines of NXS/T motif involved in N-glycosylation are localized in the loop/turn conformation in the human proteome. Similar distribution was revealed for all the other species examined. Comparative analysis of the occurrence of NXS/T motifs not known to be glycosylated and their reverse sequence (S/TXN) shows a similar distribution across the secondary structural elements, indicating that the NXS/T motif in itself is not biologically relevant. Based on our analysis, we have defined rules to determine NGS. Using machine learning methods based on these rules we can predict with 93% accuracy if a particular site will be glycosylated. If structural information is not available the tool uses structural prediction results resulting in 74% accuracy. The tool was used to identify glycosylation sites in 108 human proteins with structures and 2247 proteins without structures that have acquired NXS/T site/s due to

[*] Corresponding author.
E-mail: mazumder@gwu.edu (Mazumder R).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

ELSEVIER Production and hosting by Elsevier

1672-0229/$ - see front matter
http://dx.doi.org/10.1016/j.gpb.2012.11.003

non-synonymous variation. The tool, Structure Feature Analysis Tool (SFAT), is freely available to the public at http://hive.biochemistry.gwu.edu/tools/sfat.

Introduction

Co- and post-translational modifications (PTMs) modify the function of proteins by the addition of specific chemical groups that affect their thermodynamic, kinetic and structural properties. Glycosylation, one of the many types of PTMs, contributes to the diversification of proteins by the addition of structurally-diverse oligosaccharides. This modification is widespread and involved in a wide variety of biochemical and cellular processes including protein folding, maintenance of cell structure, receptor-ligand interaction, cell signaling and cell-cell recognition [1–3]. The function of glycosylation in health and disease attracts significant attention with recent reports on the effects of non-synonymous variations on glycosylation [4], study of glycosylation in cellular pathophysiology [5], pharmacological significance of glycosylation in therapeutic proteins [6], the significance of glycosylation in the development of biopharmaceuticals [7] and carbohydrate-based vaccines [8].

N-linked glycosylation (NGS) occurs as a post-translational modification and a co-translational process through which carbohydrates (glycans) are added to an asparagine (N) at the consensus motif asparagine-X-serine/threonine (NXS/T) in which X is any amino acid except proline [9]. There are reports of other NGS motifs such as asparagine-X-cysteine (NXC), but their frequency of occurrence is extremely low [10,11]. The attachment of the glycan is assisted by a hydrogen bond between the β-amide of asparagine as the hydrogen bond donor and the oxygen of threonine (serine) [12]. This process is catalyzed by the enzymatic action of N-glycosyltransferases which attach glycan to the unfolded protein during protein synthesis [1]. It has been suggested that NGS may contribute to the correct folding of proteins; experimental evidence shows that interactions between the sugars and the amino acids in the native state stabilizes the folding of glycoproteins [13]. It has been concluded that the primary structure of the NXS/T tri-peptide is necessary, but not sufficient, for glycosylation [10]. The most probable explanation for this observation is that in addition to other factors such as the localization of the protein, the adoption of an appropriate conformation and solvent accessibility of this tri-peptide is required for the glycosylation reaction [14,15].

Studies by Beeley [16] and later by Bause et al. [17] demonstrated a statistical probability for glycosylated asparagine residues to be located within a turn/loop conformation. Availability of complete genomes, sensitive mass spectrometric tools, and bioinformatic methods has resulted in recent confirmation of these findings in many eukaryotes [10,11]. The authors show that eukaryotic N-glycoproteins have invariant sequence recognition patterns, structural constraints and subcellular localization. Their analysis suggests that a large number of N-glycoproteins evolved after the split between fungi, plants and animals to support organismal development, body growth and organ formation specific to the corresponding clade [11]. It has been shown by Park and Zhang [18] in a comparative genomic study involving higher eukaryotes that the glycosylated asparagines evolve more slowly than the non-glycosylated counterparts in the same set of proteins. The authors conclude that the solvent-accessible asparagines are most likely to be glycosylated and of biological importance [18]. A continued improvement of rule-based filters that predict occupancy of the large number of N-glycosylation sequons is therefore important.

In this study, we performed a comprehensive structural analysis of potential N-linked glycosylation sites in *Homo sapiens* (human), *Mus musculus* (mouse), *Saccharomyces cerevisiae* (yeast), *Drosophila melanogaster* (fly) and *Arabidopsis thaliana* (plant) to refine the structural constrains of N-glycosylation with the aim to formulate basic rules improving prediction accuracy. We then used these rules to predict N-glycosylation of NXS/T sequons created in the human genome by non-synonymous single nucleotide variation (nsSNV). These rules were incorporated into an N-linked glycosylation prediction tool: Sequence Structure Feature Analysis Tool (SFAT). Our analysis shows that current structural information is sufficient to develop such rules that are applicable to the entire proteome. Such analyses can be used to prioritize targets for further validation in the laboratory.

Results and discussion

Structural analysis of annotated and unannotated NXS/T motif

The occurrence of the N-linked glycosylation sequence motif is not sufficient to determine if a particular site will get glycosylated. To better understand and describe the sequence and structural parameters that allow a specific site to be glycosylated, and to see if these can be applied across evolutionarily distant organisms, we have performed a comprehensive analysis of the five following eukaryotic proteomes: human, mouse, fly, plant and yeast. **Table 1** provides details of the data sets used in this study.

Distribution of protein secondary structure elements in eukaryotes

To understand the distribution of NGSs (annotated in UniProtKB/Swiss/Prot) and unannotated NXS/T motifs, we have determined the distribution of α-helix, β-sheet and loop/turn elements in all the non-redundant protein structures. The percentage of amino acids in these structural elements was calculated for individual proteins and the percentage of α-helix, β-sheet and loop/turn conformations for the all the proteins with structures was then calculated. The results show that the distributions of the three structural elements in all five species are very similar with α-helix being the highest and β-Sheet the lowest secondary structure conformation (**Figure 1A**). More specifically, the frequency of α-helix, β-sheet and loop/turn conformation varies in the organisms studied, which is 38–47%, 21–28% and 31–33%, respectively. If asparagine is distributed evenly among all secondary structure elements, then one should expect to observe similar frequencies of occurrences of the amino acid in the three secondary structure elements. But this is not true as can be seen from the next analysis results.

Table 1 Structure datasets used in this study

Organism	Available structures[a]	No. of annotated NXS/T sites	No. of unannotated NXS/T sites	No. of N sites	Total length	Sheet total length	Helix total length	Loop/turn total length
Human	3094	2284	3779	30,762	1,627,531	377,793	713,587	536,151
Mouse	644	453	739	5984	91,718	31,568	24,182	35,968
Fly	103	42	103	1029	37,216	12,622	16,435	37,216
Plant	179	33	223	1834	136,158	30,062	62,978	43,118
Yeast	756	10	1163	16,745	191,581	41,428	87,412	62,741

Note: [a]Structures that have at least one asparagine in their sequence.

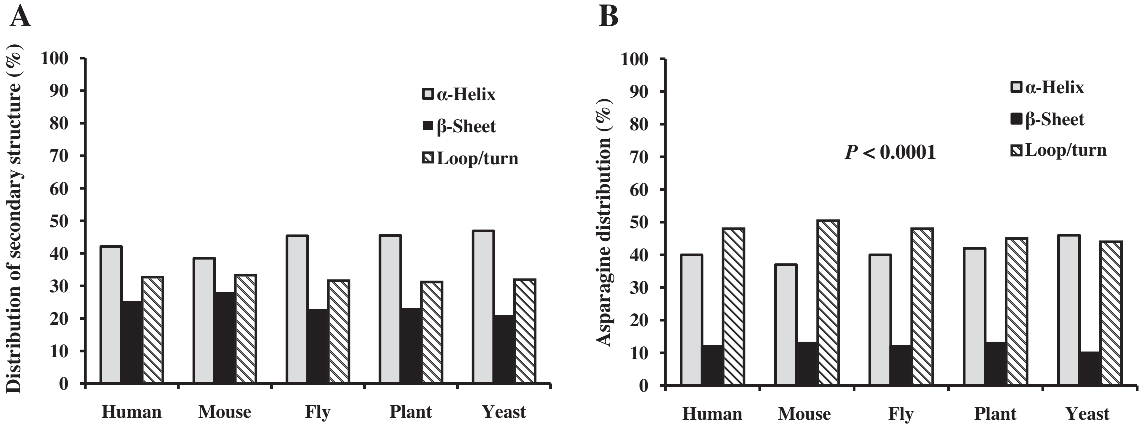


Figure 1 The distribution of secondary structure elements and asparagine

A. Distribution of secondary structural elements in proteins of human, mouse, fly, plant and yeast. **B**. Distribution of asparagine in secondary structural elements in proteins of human, mouse, fly, plant and yeast proteins. *P* values are calculated with χ^2 test by comparing the occurrence of asparagine in secondary structural elements to the overall distribution of α-helix, β-sheet and turns/loops in all available structures in the species of interest.

Distribution of asparagines in protein secondary structural elements

It has been shown that NGS are more prevalent in turns [14,16]. This observation would not have functional implications if the abundance of asparagines (N) in turns is similar to the abundance of NGS sites in turns. There are 30,762 N-containing sites in 3094 proteins with crystallographic PDB structures for the human proteome; 5984 sites in 644 proteins for mouse, 1029 sites in 103 proteins for fly; 1834 sites in 179 proteins for plant and 16,745 sites in 756 proteins for yeast. **Figure 1B** shows that asparagines are located preferentially in the loop/turn conformation with a frequency of 44−50%. Compared to the results shown in Figure 1A, the percentage of asparagines found in α-helix appears to be close to the expected (40.20% vs. 42.10%). However, the percentage of asparagines is higher in turns/loops and lower in β-sheets than expected. These results prompted us to examine whether the distribution of potential NGS (annotated in UniProtKB/Swiss/Prot) and unannotated NXS/T sites follows the same pattern.

Distribution of unannotated NXS/T motif in protein secondary structure elements

All UniProtKB/Swiss-Prot records are manually curated. Even the prediction results of every protein are manually checked before they are entered into the database. Therefore, annotations available from UniProtKB records are considered ‘gold standard’ in terms of functional annotation. It is expected that unannotated NXS/T motifs not known to carry a glycan (and therefore functionally comparable to any N), should reflect the overall distribution of asparagines. If we consider proteins with crystallographic structures, there are 3779 unannotated NXS/T motifs in human proteins, 739 sites for mouse, 1163 sites for yeast, 103 sites for fly and 223 sites for plant. According to **Figure 2A**, the percentage of asparagines in the unannotated NXS/T motif is slightly higher than that of all asparagines (Figure 1B) in a turn conformation. It is possible that this is a function of the tripeptide property. We therefore wanted to see if the distribution of annotated NXS/T motifs is significantly different than that of the unannotated NXS/T motifs.

Distribution of annotated NXS/T motif in protein secondary structure elements

There are 2284 annotated NXS/T sites in 592 proteins which have PDB structures in the human dataset. Among these, 1779 sites are in turn (78%), 222 sites in the α-helix (9.7%) and 283 sites in the β-sheet (12.3%) (**Figure 2B**). This distribution, based on analysis of the entire set of available structures, is consistent with recent results in the mouse showing 75% of NGSs in turns and 15% in β-sheets [10]. The same tendency is observed in our analysis for mouse, fly, plant and yeast. In all

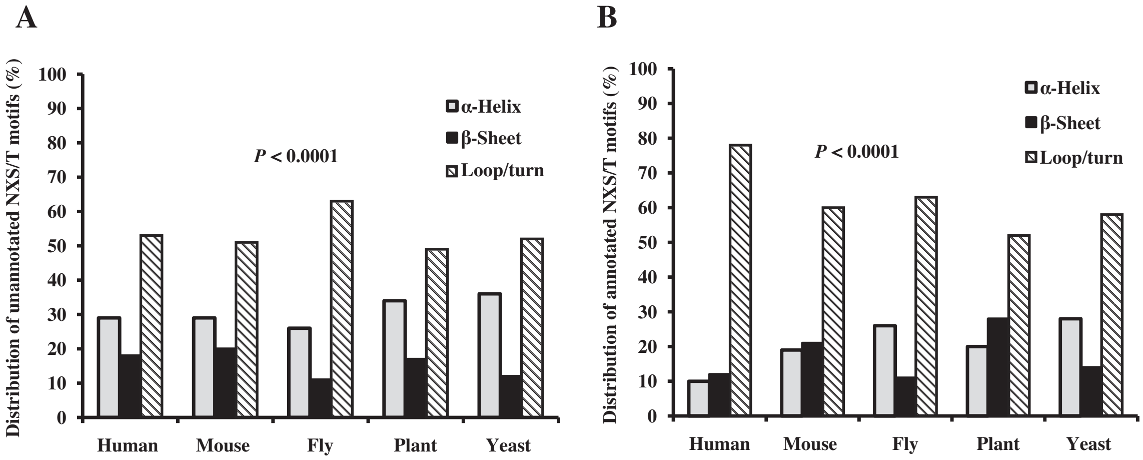


Figure 2 The distribution of asparagine in unannotated and annotated NXS/T motifs

A. Distribution of unannotated NXS/T motifs in secondary structural elements. *P* values are calculated with χ^2 test by comparing the occurrence of asparagine in unannotated NXS/T motif to the distribution of all asparagines. **B**. Distribution of annotated NXS/T motifs in secondary structural elements. *P* values are calculated with χ^2 test by comparing the occurrence of asparagine in annotated NXS/T motif to the distribution of asparagines in unannotated NXS/T motifs.

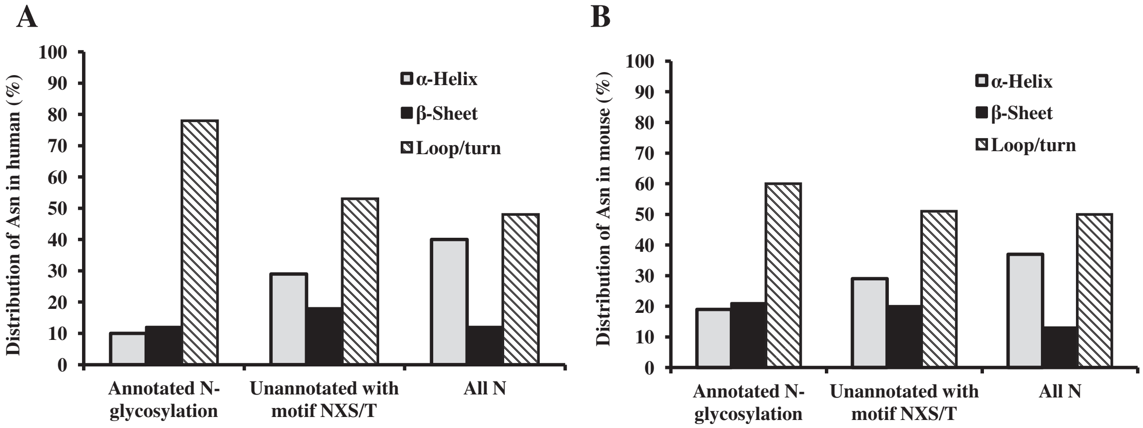


Figure 3 Distribution of asparagines in human and mouse proteins

A. Distribution of asparagines in human proteins. **B**. Distribution of asparagines in mouse proteins.

the species the percentages of asparagine that are part of the annotated motif is higher in loops/turns. The distribution of annotated NXS/T sites is significantly different from the distribution of the unannotated NXS/T in all the species. Representative data were shown for human and mouse in Figure 3 ($P < 0.0001$).

Our results in mouse show a noticeable difference of the distribution of NXS/T sites in loops/turns (60% for annotated and 51% for unannotated), compared to the analysis of Zielinska et al. (75% and 71%, respectively) [10]. This is potentially due to the fact that we have analyzed a larger set of structures. The occurrence of Asn-sequons in this type of secondary structure is favored for three reasons: (i) loops/turns represent spatial arrangements of the peptide chain which favor the hydrogen-bonded contact between the beta-amide of asparagines and the hydroxyl group of carbohydrates; (ii) turns constitute privileged conformations which guarantee accessibility of the sugar-acceptor sites due to their general location at the surface of proteins and (iii) these sites could be evolutionary selected because of functional importance.

Quantification and functional analysis of N-glycosylation sites in human and mouse proteome

UniProtKB contains 15,828 sites for the entire human proteome where the sequon NXS/T is annotated as the N-linked glycosylation site. Among them, 15,168 sites are found in "secreted and membrane proteins" and 747 sites in "cytoplasm and nucleus proteins" (there are cases where the same protein can be found in two different places). Thus, when comparing annotated NXS/T motifs found in the human proteome, approximately 96% of sites were found in "secreted and membrane proteins" and only 4.7% sites in "cytoplasm, nucleus proteins." For mouse, 95% of the proteins with annotated NXS/T motifs are "secreted and membrane proteins", while only 3.8% are "cytoplasm and nucleus" proteins (in this case some proteins do not have location information). In large-scale analysis of NGS sites, it was found that none of the identified glycoproteins are located in the mitochondria, cytosol or nucleus [11]. Kung et al. [19] had identified mitochondrial glycoproteins using protein microarrays and it is possible that they are either errors in the curated data or mitochondrial glycoproteins could not be captured because of the experimental protocols used. It is also possible that some proteins could be present in more than one compartment at different time points, which might explain the differences in the results in UniProtKB and the aforementioned studies. Looking at the proportion of NXS/T sites that are annotated and the total number of NXS/T sites in the human proteome, the UniProt data suggest that only 27% of all NXS/T sites are N-glycosylated. Among "secreted and membrane proteins", the number increases to 53% of all NXS/T sites (**Table 2**). When comparing annotated NXS/T motifs to total NXS/T motifs for mouse, we find 21% of NXS/T motifs is annotated, which is lower than 27% for the human proteome, and this number increases to 36% if only "secreted and membrane proteins" are considered. This could be due to the fact that not all of mouse proteome has been manually curated by UniProtKB/Swiss-Prot curators and extreme caution is employed by UniProtKB/Swiss-Prot curators to ensure close to zero false positives. For cytoplasm/nucleus/mitochondria proteins, the numbers are similar for human and mouse (Table 2).

These findings show that there exist a very high number of NXS/T motifs in the proteome, but less than one third of them have so far have been documented as glycosylated for both human and mouse. A higher percentage of annotated NXS/T motifs are present in "secreted and membrane proteins" than "cytoplasm and nucleus proteins", which is consistent with previously reported preferential N-glycosylation of proteins in the secretory pathways [10]. Comparison of the NXS/T and the reverse S/TXN site in the human proteome, which is not expected to carry glycans, reveals 58,781 NXS/T sites and 50,577 S/TXN sites ($P = 0.9$) (for additional details please see Table S1). Similar distribution is observed in the mouse proteome (data not shown). Additionally, for the human proteome we noticed that there are 28,527 and 22,568 NXS/T sites in secreted and membrane proteins and in cytoplasmic and nuclear proteins, respectively, which is significantly different ($P < 0.0001$). However, the number of the reverse motifs (S/TXN) in secreted and membrane proteins and in cytoplasmic and nuclear proteins is comparable, which is 21,213 and 21,375, respectively. It is important to note that similar results were obtained in terms of the relative cellular distribution of NXS/T and S/TXN sites, if only the proteins with PDB structural information were considered (Table S1), which strongly supports the inclusion of just protein sequences with structural information for the type of analysis performed in this study.

Non-synonymous single nucleotide variation and polymorphic glycoproteins

The structural analysis shows that NGS are over-represented on the surface of proteins. We find that 91% of the annotated sites in humans and 93% in mouse are solvent-accessible, compared to 67% and 70% of the unannotated sites. We used informatic tools to extract information on polymorphic N-glycosylation variants from the UniProtKB/Swiss-Prot database and dbSNP [20]. Previously we have shown using pathway and function enrichment analysis that a significant number of proteins that gain or lose the glycosylation motif are involved in kinase activity, immune response and blood coagulation [4]. However, it remains to be investigated whether a polymorphic site can indeed be glycosylated when there is gain of the NGS motif. Our current analysis shows that of the 20,238 proteins in the complete human proteome (based on UniProtKB/Swiss-Prot), 3328 proteins contain polymorphic sites that create or abolish existing glycosylation sites (**Figure 4**). We employed machine learning techniques based on the rules developed from this study to examine the proteins that have crystallographic structure information available at NGS. As a result, we identified 108 out of 221 polymorphic proteins with structures (Table S2), which have one or more gain of glycosylation that are expected to have some impact on protein function. Based on UniProtKB/Swiss-Prot and Gene Ontology (GO) analysis, several of these proteins are involved in blood coagulation, cell adhesion, host-pathogen interaction, immunity and transport (Table S3). The major molecular functions represented are hydrolases, receptors and transferases. Out of 2299 proteins that do not have structures, 2247 proteins are predicted to be glycosylated at the gain of glycosylation site by the SFAT tool (total NXS/T sites: 12,623; sites predicted as yes: 11,651 and sites predicted as no: 972). Based on GO analysis using Panther tools [21], the over-represented GO biological processes in this dataset include immune response, response to stimulus, signaling and blood coagulation, which agrees with the our results obtained previously [4].

N-linked glycosylation prediction tool

The analysis that we have performed here represents an efficient way to explore the glycosylation potential of protein if the structure is known. We have also extended the tool to work on proteins without structure, albeit with lower accuracy: 74% without structure compared to 93% with structure. It is expected that within the next decade, majority of proteins with NGS will have their structures solved, or it will be possible to generate high-quality homology models for these proteins based on related protein structures. To provide easy comparison of these structures and sequences we have web-enabled our tools developed for this study. The tool can perform the following tasks: (i) predict N-linked glycosylation sites, (ii) determine the secondary structural elements of any site of interest (such as active site, metal binding site, N-linked glycosylation site or any other sited based on user-defined motif) and (iii) map UniProtKB and PDB sequence features. The tool, Structure Feature Analysis Tool (SFAT), is expected to be useful for

Table 2 Subcellular distribution of annotated NXS/T motifs in human and mouse proteome

Species	Entire proteome (%)	Secreted/membrane (%)	Cytoplasm/nucleus/mitochondria (%)
Human	27	53	3
Mouse	21	36	2.6

Note: Percentage of annotated NXS/T motifs against all NXS/T motifs in respective categories is shown.

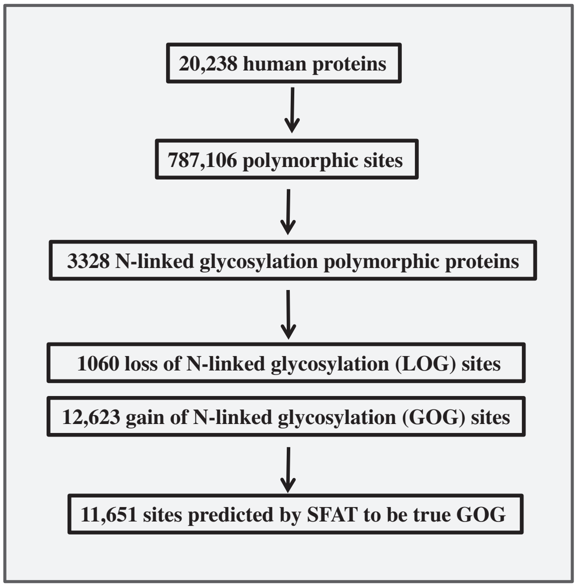

20,238 human proteins
787,106 polymorphic sites
3328 N-linked glycosylation polymorphic proteins
1060 loss of N-linked glycosylation (LOG) sites
12,623 gain of N-linked glycosylation (GOG) sites
11,651 sites predicted by SFAT to be true GOG

Figure 4 Identification of N-linked glycosylation (NLG) sites using SFAT

researchers interested in site-specific quantitative structural analysis. **Figure 5** shows a snapshot of the SFAT interface.

Prediction of N-linked glycosylation sites

The prediction is performed using the following four basic rules: (i) presence of Endoplasmic Reticulum targeting sequence; (ii) not nuclear, mitochondrial or cytoplasmic; (iii) present in a loop or turn and (iv) exposed. Data for the first two rules are extracted from the UniProtKB flat file. The third piece of information is extracted from the PDB file and the relative solvent accessibility is obtained from Define Secondary Structure of Proteins (DSSP) database [22]. When structure is not available, the last two pieces of information is predicted. We identified 96 new (currently not annotated in UniProt) NGS in the human proteome that matches all four rules (3.6% of the total unannotated NXS/T sites). Instead of using a set of strict rules for prediction which can potentially lead to large numbers of false negatives, we implemented these rules into a machine learning framework for better prediction accuracy. Using this approach our cross validation prediction model showed the overall accuracy of 93%, and the precision for true positive of 90% for proteins with structures. The accuracy and precision is 74% and 70%, respectively, when the tool is applied to proteins without structures. The model was then used to predict NGS in the polymorphic glycoproteins. The results showed that for the gain of N-glycosylation, around 40% are predicted to be glycosylated.

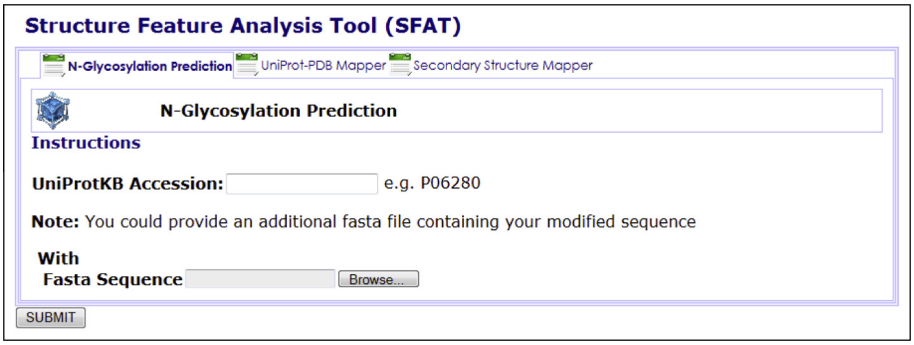
There are several other NGS prediction tools currently available. However, none of them use a rule-based method that is dependent on structural information. EnsembleGly [23], a sequence-based method using ensembles of support vector machine classifiers, has 94% accuracy; NetNGlyc (http://www.cbs.dtu.dk/services/NetNGlyc/) uses artificial neural networks that examine the sequence context of Asn-X-Ser/Thr sequons with an overall accuracy of 76%. In addition, GPP [24] uses the random forest algorithm and pairwise patterns to predict glycosylation sites with an accuracy of 90.8% for Ser sites, 92.0% for Thr sites and 92.8% for Asn sites. It is important to note here that the authors used their own training and test datasets to determine the accuracy of their tools. A direct comparison between different tools is thus difficult because the tools were developed and tested on different training and test datasets. Furthermore, the definition of negative NGS is an open discussion, because it is difficult to prove definitively that a particular residue is not glycosylated under any conditions, although experiments can verify that a particular residue can be glycosylated.

Determining the secondary structure elements of any amino acid site

Knowledge of the distribution of a specific motif in the secondary structure elements can be useful to predict the functional relevance. To facilitate studies similar to the one described here, we implement within SFAT an option that can provide a distribution report of any motif of interest. The input file is the UniProtKB flat file. The tool gives the user multiple options about different feature information in UniProtKB such as: N-linked glycosylation, active site or metal-binding site. Alternatively, the user can define their motif of interest. Instructions for how to determine the correct pattern can be found in the Help document. The results are given in a downloadable table. A pie chart and plot graph are generated to illustrate the obtained results.

Mapping of UniProtKB and PDB features

Sequences from PDB and UniProt may not be identical (PDB sequences can be shorter or longer compared to the


Structure Feature Analysis Tool (SFAT)
N-Glycosylation Prediction
UniProt-PDB Mapper
Secondary Structure Mapper
N-Glycosylation Prediction
Instructions
Browse...
SUBMIT

UniProtKB Accession: e.g. P06280

Note: You could provide an additional fasta file containing your modified sequence

With
Fasta Sequence

Figure 5 Home page for N-linked glycosylation prediction tool SFAT

User can either predict N-linked glycosylation sites, find the distribution of a motif in secondary structural elements or map UniProtKB and PDB sequence features.

UniProtKB sequence). Therefore, it is important to align them to explore the different feature annotations that are available on the sequence from UniProt or PDB. Often this task is done manually and can lead to errors. We have implemented a simple alignment tool and the features are extracted from UniProtKB flat files and PDB records and the user can easily identify the features of interest in the amino acid sequence. This tool was used in this study to further analyze the polymorphic glycosylation sites.

Conclusion

Comparative structural study of asparagines in human, mouse, fly, plant and yeast showed that a high percentage of asparagines in NXS/T motifs implicated in N-glycosylation are localized within a turn/loop and are solvent-exposed at the protein surface. The N-glycosylated proteins are typically not cytoplasmic, nuclear or mitochondrial. We have incorporated these observations into an N-glycosylation prediction tool which combines structure- and sequence-based rules that significantly improve sequence-based prediction methods. The tool was used to predict glycosylation sites of a set of polymorphic human proteins.

Materials and methods

By ‘annotated’ glycosylation sites, we mean NXS/T sites that are indicated as the N-linked glycosylation site in UniProtKB/Swiss-Prot [25] protein record, while unannotated glycosylation sites include all other NXS/T sites. We consider these annotated glycosylation sites as potential NGS, as they have been manually curated by UniProtKB/Swiss-Prot curators based on experimental evidence, similarity to experimentally-validated NGS in homologous proteins and/or in-depth sequence and functional analysis. UniProtKB/Swiss-Prot human proteome is considered the gold standard set of manually-curated human proteins and in our opinion provides the best positive and negative datasets, since UniProtKB/Swiss-Prot curators have manually curated all entries for the human proteome. All predictions are checked manually by curators to ascertain if there are any homologous sites in related proteins. Predictions that are dubious are not included in the sequence feature annotation.

Datasets

All data were collected from UniProtKB and Protein Data Bank (PDB) [26]. For the human proteome, the complete proteome available from UniProtKB/Swiss-Prot was used. For mouse, the file rp-seqs-15.fasta.gz from http://pir.georgetown.edu/rps/data/current/15/ was downloaded and parsed to obtain the complete proteome. This was done because the complete proteome of mouse from UniProtKB has several potential splice variants as separate entries. All other proteomes were obtained from UniProtKB using the ‘complete proteome’ keyword tag. The access dates for all data retrievals are between 15th February and 15th June, 2012. Experimentally-validated datasets were obtained from UniProtKB and supplementary materials in Zielinska et al. [10,11].

Data analysis

Python scripts were used to extract information from UniProtKB flat file feature (FT) lines, cross-reference (DR) lines for PDB database and sequence (SQ) lines. Information was also extracted from the PDB files in order to get secondary structure information. For annotated N-linked glycosylation sites, all UniProt FT lines annotated as “N-linked” whether “confirmed”, “potential” or “by similarity” were retained. The positions of these sites were retrieved and the corresponding sequences were checked for the NXS/T motif. PDB IDs were extracted from UniProtKB flat files. As there can be more than one PDB file mapped to any UniProtKB protein, the PDB structure which meets the following criteria was selected: the structure was determined by x-ray diffraction, highest resolution, and the site of interest is contained within the solved structure. Once selected, PDB files were downloaded from PDB and the positions aligned to the PDB sequence and the secondary structure was determined based on the secondary structure assignment in the PDB file.

Relative solvent accessibility was calculated using the information in DSSP database [22]. Based on the value of the relative accessible surface area (ASA), the residues were grouped as buried (0.0–0.25) or exposed (0.25–1.0). The choice of the threshold was based on previous studies [10,18,27]. For machine learning using classification and regression tree (CART), we allow CART to automatically select the relative ASA based on the test set. MUSCLE was used to perform pairwise alignment to map UniProtKB protein sequences to PDB [28]. Prediction of the N-terminal targeting sequences was performed using Predotar [29]. Subcellular location analysis is performed based on UniProtKB keywords. For sequences that do not have structural information, secondary structure and surface accessibility of the individual amino acid was predicted using NetSurfp [30]. *P* value was calculated using a binomial statistic based on methodology described earlier [31,32]. *P* value of 0.05 or less was considered significant.

Identification of polymorphic glycosylation sites

Gain of glycosylation sites were identified by using variation data from UniProtKB FT lines and dbSNP [20,33]. Entries were first mapped to UniProtKB accessions using the ID mapping service [34] followed by sequence mapping. This resulted in a table with UniProtKB accession numbers and position of variation, the variation and the data source (Table S2).

Machine learning using CART

This method consists of creating a model that predicts the value or a class for a predictive variable based on several input variables. The algorithms of this decision tree usually work top-down by choosing a variable at each step that is the next best variable to use in splitting the set of items. “Best” is defined by how well the variable splits the set into homogeneous subsets that have the same value of the predictive variable [35]. As the number of unannotated sites is higher than that of the annotated sites in our analysis of the human proteome (2:3 ratio), the training dataset for our prediction model contains 200 experimentally-validated sites (noted as positive) and 300 unannotated sites (noted as negative). Training classifiers are challenging because the positive and negative datasets are unbalanced with more negative sites than positive sites, which can result in poor classification of the minority class (in this case positive sites). One solution is to change the distribution of major and minor classes during training by randomly selecting a subset of the training data. However, this approach does not then take into consideration all available information in the real dataset. Therefore, we chose our training dataset to contain original data in a ratio that reflects the natural dataset.

First, a set of patterns is generated from the training data for each of the glycosylation sites and then used to generate a value for each instance. Multiple runs are performed with each instance collecting weights to determine the positive or negative NGS class. Each of the runs used dataset comprising randomly chosen positive and negative instances from the cross validation fold. The accuracy of the prediction was evaluated by cross validation (described below).

The variables that describe the dataset are selected based on the rules derived from this study: (i) ER targeting sequence (ii) sub-cellular location, (iii) secondary structure and (iv) exposed/buried. To build the classification model, the function 'rpart' in statistical language R was used. Twenty-fold internal cross validation was performed to validate the training dataset as described previously [24]. More specifically, the dataset was partitioned randomly into 20 sections and the training procedure was carried out using 19 of these while the 20th section provides a test dataset. This was repeated 20 times on each occasion with a different section of the data acting as the test set. The evaluation of the model is based on the number of true positives (TP), false positives (FP), true negatives (TN), false negatives (FN). The model was then used to predict NGS in the test dataset.

Authors' contributions

RM conceived, designed and coordinated the study and developed a general outline for the algorithm. PL developed the specific algorithm, was responsible for software design and implementation and drafted the manuscript. TN participated in the design and evaluation of the tool interface. KK generated the variation data. RG tested the tool and helped design the study. VSN and VSK performed the statistical analysis. All authors read and approved the final manuscript.

Competing interests

We declare that we have no competing interests.

Acknowledgements

We want to thank Hayley Hamilton of the George Washington University and Nagarajan Pattabiraman of MolBox LLC for critical reading of the manuscript and providing useful comments. All computations were performed at High-performance Integrated Virtual Environment (HIVE) located at The George Washington University and co-developed by Drs. Raja Mazumder and Vahan Simonyan. Support for this work came from the George Washington University funds to RM. RG's participation is supported by RO1 CA135069 and U01 CA168926. This project is also supported in part by an appointment to the Research Participation Program at the Center for Biologics Evaluation and Research administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration.

Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gpb. 2012.11.003.

References

[1] Helenius A, Aebi M. Roles of N-linked glycans in the endoplasmic reticulum. Annu Rev Biochem 2004;73:1019–49.

[2] Varki A. Biological roles of oligosaccharides: all of the theories are correct. Glycobiology 1993;3:97–130.

[3] Woods RJ, Edge CJ, Dwek RA. Protein surface oligosaccharides and protein function. Nat Struct Biol 1994;1:499–501.

[4] Mazumder R, Morampudi KS, Motwani M, Vasudevan S, Goldman R. Proteome-wide analysis of single-nucleotide variations in the N-glycosylation sequon of human genes. PLoS One 2012;7:e36212.

[5] Ohtsubo K, Marth JD. Glycosylation in cellular mechanisms of health and disease. Cell 2006;126:855–67.

[6] Li H, d'Anjou M. Pharmacological significance of glycosylation in therapeutic proteins. Curr Opin Biotechnol 2009;20:678–84.

[7] Kawasaki N, Itoh S, Hashii N, Takakura D, Qin Y, Huang X, et al. The significance of glycosylation analysis in development of biopharmaceuticals. Biol Pharm Bull 2009;32:796–800.

[8] Hecht ML, Stallforth P, Silva DV, Adibekian A, Seeberger PH. Recent advances in carbohydrate-based vaccines. Curr Opin Chem Biol 2009;13:354–9.

[9] Hart GW. Glycosylation. Curr Opin Cell Biol 1992;4:1017–23.

[10] Zielinska DF, Gnad F, Wisniewski JR, Mann M. Precision mapping of an *in vivo* N-glycoproteome reveals rigid topological and sequence constraints. Cell 2010;141:897–907.

[11] Zielinska DF, Gnad F, Schropp K, Wisniewski JR, Mann M. Mapping N-glycosylation sites across seven evolutionarily distant species reveals a divergent substrate proteome despite a common core machinery. Mol Cell 2012;46:542–8.

[12] Bause E, Legler G. The role of the hydroxy amino acid in the triplet sequence Asn-Xaa-Thr(Ser) for the N-glycosylation step during glycoprotein biosynthesis. Biochem J 1981;195:639–44.

[13] Wyss DF, Choi JS, Li J, Knoppers MH, Willis KJ, Arulanandam AR, et al. Conformation and function of the N-linked glycan in the adhesion domain of human CD2. Science 1995;269:1273–8.

[14] Bause E. Structural requirements of N-glycosylation of proteins. Studies with proline peptides as conformational probes. Biochem J 1983;209:331–6.

[15] Junker VL, Apweiler R, Bairoch A. Representation of functional information in the SWISS-PROT data bank. Bioinformatics 1999;15:1066–7.

[16] Beeley JG. Peptide chain conformation and the glycosylation of glycoproteins. Biochem Biophys Res Commun 1977;76:1051–1055.

[17] Bause E, Hettkamp H, Legler G. Conformational aspects of N-glycosylation of proteins. Studies with linear and cyclic peptides as probes. Biochem J 1982;203:761–8.

[18] Park C, Zhang J. Genome-wide evolutionary conservation of N-glycosylation sites. Mol Biol Evol 2011;28:2351–7.

[19] Kung LA, Tao SC, Qian J, Smith MG, Snyder M, Zhu H. Global analysis of the glycoproteome in *Saccharomyces cerevisiae* reveals new roles for protein glycosylation in eukaryotes. Mol Syst Biol 2009;5:308.

[20] Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res 2010;38:D5–16.

[21] Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res 2010;38:D204–10.

[22] Joosten RP, te Beek TA, Krieger E, Hekkelman ML, Hooft RW, Schneider R, et al. A series of PDB related databases for everyday needs. Nucleic Acids Res 2011;39:D411–9.

[23] Caragea C, Sinapov J, Silvescu A, Dobbs D, Honavar V. Glycosylation site prediction using ensembles of Support Vector Machine classifiers. BMC Bioinformatics 2007;8:438.

[24] Hamby SE, Hirst JD. Prediction of glycosylation sites using random forests. BMC Bioinformatics 2008;9:500.

[25] UniProt-Consortium. Reorganizing the protein space at the universal protein resource (UniProt). Nucleic Acids Res 2012;40:D71–5.

[26] Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, et al. The RCSB protein data bank: redesigned web site and web services. Nucleic Acids Res 2011;39:D392–401.

[27] Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. Proteins 2004;56:753–67.

[28] Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004;5:113.

[29] Small I, Peeters N, Legeai F, Lurin C. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 2004;4:1581–90.

[30] Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol 2009;9:51.

[31] Mi H, Thomas P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. Methods Mol Biol 2009;563:123–40.

[32] Cho RJ, Campbell MJ. Transcription, genomes, function. Trends Genet 2000;16:409–15.

[33] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. DbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;29:308–11.

[34] Huang H, McGarvey PB, Suzek BE, Mazumder R, Zhang J, Chen Y, et al. A comprehensive protein-centric ID mapping service for molecular data integration. Bioinformatics 2011;27:1190–1.

[35] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. New York: Chapman and Hall; 1984.