Genomics Proteomics Bioinformatics 11 (2013) 96–104

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

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.