Glycosylation and Bioinformatics: current status for glycosylation prediction tools

REVIEW

✍ Yuliet Mazola, Glay Chinea, Alexis Musacchio

Department of Bioinformatics, Center for Genetic Engineering and Biotechnology, CIGB
Ave. 31 / 158 and 190, Playa, PO Box 6162, Havana, Cuba
E-mail: yuliet.mazola@cigb.edu.cu

ABSTRACT

Glycosylation is an important co- and post-translational modification involved in a variety of critical biological processes. The development of computational algorithms for protein glycosylation prediction has been propelled in the latest years. The localization of potential glycosylated sites facilitates the rational alteration of glycosylation-related functions in cells. This manuscript gives an overview of current available bioinformatics resources and databases for glycobiology, focusing on glycosylation predictors. As a complement, general features about the different glycosylation types are also exposed.

Keywords: glycosylation, post-translational modification, bioinformatics, prediction, glycobiology, carbohydrate, databases

Biotecnología Aplicada 2011;28:6-12

RESUMEN

Glicosilación y Bioinformática: estado actual de las herramientas para predecir glicosilación. El desarrollo de algoritmos computacionales para la predicción de sitios potenciales de glicosilación en las proteínas ha sido impulsado en los últimos años. La glicosilación constituye una modificación co- y post-traduccional involucrada en una gran variedad de procesos biológicos críticos. La localización de los sitios potenciales de glicosilación facilita la modificación racional de las funciones relacionadas con la glicosilación en las células. Este manuscrito resume el estado actual de las herramientas bioinformáticas y las bases de datos disponibles para la glicobiología, haciendo énfasis en los predictores de glicosilación. Además, como complemento se incluyen las principales características de los diferentes tipos de glicosilación.

Palabras clave: glicosilación, modificación post-traduccional, bioinformática, predicción, glicobiología, carbohidrato, bases de datos

*I*ntroduction

Glycosylation is an essential co- and post-translational modification occurring in cells. It involves the selective attachment of carbohydrate molecules (also referred as glycans, sugars or saccharides) to proteins and lipids. Glycans are composed of monosaccharides covalently linked by glycosidic bonds, adopting linear and branched structures. There are two major types of glycosylation: N-glycosylation and O-glycosylation. Besides, another three unusual glycosylation forms have been described, known as C-glycosylation (or C-mannosylation), Glypiation (or glycosylphosphatidylinositol (GPI) anchoring) and Glycation (or non-enzymatic glycosylation). All glycosylation types are enzyme-directed site-specific processes, with the exception of glycation. Glycation is a non-enzymatic reaction of glucose and other saccharide derivatives with proteins, nucleotides and basic phospholipids [1]. Since glycation is not an enzymatic post-translational modification process it will not be covered in this review.

The biological roles of glycosylation are diverse and influence both at cellular and protein levels, for example, protein folding and oligomerization [2], protein degradation [3], protein solubility and stability [4], epitope recognition [5], cell-cell interactions [6] and protein transport [7]. Hence, as may be expected, the glycoproteins are involved in the development and progression of several diseases, such as cancer [8], autoimmune diseases [9] and congenital disorders [10].

Among post-translational modifications, glycosylation may be considered the most complex of all [11]. First, carbohydrate structures are indirectly encoded in the genome. Both sequence and structure of glycan highly depends on the action of enzymes (*e.g.*, glycosyltransferases, carbohydrate-modifying enzymes and glycosidases) that create, modify or degrade glycosidic bonds [11]. Then, carbohydrate structure is well determined by the enzymes expressed in a particular cell or tissue. Second, glycoproteins can be modified with different carbohydrates in the same glycosylated site (leading to several glycoforms). Besides, not all potential glycosylation sites are simultaneously occupied [11].

Glycans bind to specific motifs within protein sequence depending on the glycosylation type, as will be described bellow. The recognition of glycan-occupied sites can be experimentally determined but it is an expensive and laborious process [12]. Thus, the number of verified glycosylated residues is still limited in relation with the growing number of known protein sequences [13]. The merely knowledge of glycosylated site locations may be a valuable tool. For example, to improve the 3D protein structure prediction, ensuring the appearance of glycosylated residues surface-exposed, as well as to modify the protein pharmacokinetic properties by changing protein-associated carbohydrate (glycoengineering). In this scenario, the development of bioinformatics tools to predict glyco-

✍ Corresponding author

1. Ramamurthy B, Hook P, Larsson L. An overview of carbohydrate-protein interactions with specific reference to myosin and ageing. Acta Physiol Scand. 1999; 167:327-9.

2. Mitra N, Sinha S, Ramya TN, Surolia A. N-linked oligosaccharides as outfitters for glycoprotein folding, form and function. Trends Biochem Sci. 2006;31:156-63.

3. Mbonye UR, Yuan C, Harris CE, Sidhu RS, Song I, Arakawa T, *et al*. Two distinct pathways for cyclooxygenase-2 protein degradation. J Biol Chem. 2008;283:8611-23.

4. Sola RJ, Griebenow K. Effects of glycosylation on the stability of protein pharmaceuticals. J Pharm Sci. 2009;98:1223-45.

5. Specks U, Fass DN, Finkielman JD, Hummel AM, Viss MA, Litwiller RD, *et al*. Functional significance of Asn-linked glycosylation of proteinase 3 for enzymatic activity, processing, targeting, and recognition by anti-neutrophil cytoplasmic antibodies. J Biochem. 2007;141:101-12.

6. Corthay A, Backlund J, Broddefalk J, Michaelsson E, Goldschmidt TJ, Kihlberg J, *et al*. Epitope glycosylation plays a critical role for T cell recognition of type II collagen in collagen-induced arthritis. Eur J Immunol. 1998;28:2580-90.

7. Vagin O, Kraut JA, Sachs G. Role of N-glycosylation in trafficking of apical membrane proteins in epithelia. Am J Physiol Renal Physiol. 2009;296:F459-F469.

sylation sites is playing an increasing important role. Recently, bioinformatics resources for glycomics and glycobiology-related databases have been nicely reviewed elsewhere [14-16]. The application of bioinformatics tools in any biology field certainly demands a general understanding of the biological processes involved. Hence, this manuscript gives an overview of available glycosylation prediction methods, supported by a description of essential features for the different known glycosylation types.

N-glycosylation

N-glycosylation consists in the attachment of a sugar moiety to the amide side chain of an Asn residue within any of the following consensus sequences: Asn-X-Ser and Asn-X-Thr (and in some rare cases, Asn-X-Cys), where X could be any amino acid except Pro [17]. These tripeptide sequences are known as sequon. For many years, it was thought that N-glycosylation was present just in eukaryotes. But today, N-glycosylated proteins in prokaryotes is a fact [18]. Several differences have been observed in the biosynthesis of eukaryotic N-glycans compared with bacteria and archaea [19, 20]. In eukaryotes, N-linked protein glycosylation occurs in the endoplasmic reticulum (ER) [21]. It begins by transferring the oligosaccharide portion ($Glc_3Man_9GlcNAc_2$) from a lipid-linked precursor (dolichol phosphate) to the protein that is being translated in the ribosome (Figure 1) [22]. Next, the oligosaccharide immediately undergoes trimming and processing. First, two terminal glucose residues are removed leading to Glc_1Man_9-Glc-NAc_2-Asn linked protein. In this form, the newly synthesized glycoprotein enters to the calnexin/calreticulin cycle [23]. Calnexin (membrane-bound) and calreticulin (soluble) are lectin proteins residing in the ER. They specifically interact with the monoglucosylated glycoproteins to assist their folding and quality control. Once the glycoprotein has acquired its native conformation, it exits calnexin/calreticulin cycle and continues along the secretory pathway. Instead, the glycoprotein is re-glucosylated and re-sent to the calnexin/calreticulin cycle. Upon the deletion of the remaining glucose residue, one mannose is trimmed leading to $Man_9GlcNAc_2$-Asn linked protein (Figure 1). This emerging N-glycosylated protein is transported to the Golgi apparatus for other mannose trimmings [22]. Finally, the $Man_5GlcNAc_2$-Asn linked protein is the starting point to generate a huge repertory of N-glycan types in the Golgi apparatus [24].

In bacteria, the transference of the glycan portion to the nascent protein is a reaction similar to that occurring in eukaryotes [19]. However, both the enzyme catalyzing such reaction, named as oligosaccharyltransferase enzyme (OST) and the initial lipid-linked precursor differ in prokaryotes and eukaryotes [19]. The eukaryotic OST is a complex containing several membrane-associated protein subunits anchored in the lumen of the ER. The OST complex is involved in other functions besides oligosaccharide transfer reaction [23]. For example: (1) scanning of the polypeptide for possible N-linked glycosylation sites bearing the tripeptide sequon Asn-X-Ser/Thr, (2) directing the nascent polypeptide chain to the OST active site in the proper conformation, (3) positioning the active site of the OST complex near the retrotranslocon complex, and (4) recognizing and moving the lipid-linker precursor to the OST active site. Indeed, the availability and location of the OST complex relative to the nascent glycoprotein affects the N-glycan site occupancy [23]. However, in both bacteria and archaea, the OST enzyme is a single subunit homologous to the catalytic subunit of the multimeric OST eukaryotic complex [19]. There are other differences in the N-linked glycosylation process between eukaryotes and prokaryotes. For example, in bacteria, the N-glycosylation occurs in the periplasm and seems to be only a post-translational process, although it has not been confirmed yet [19]. Instead, N-glycosylation in eukaryotes is a co- and post-translational process [21].

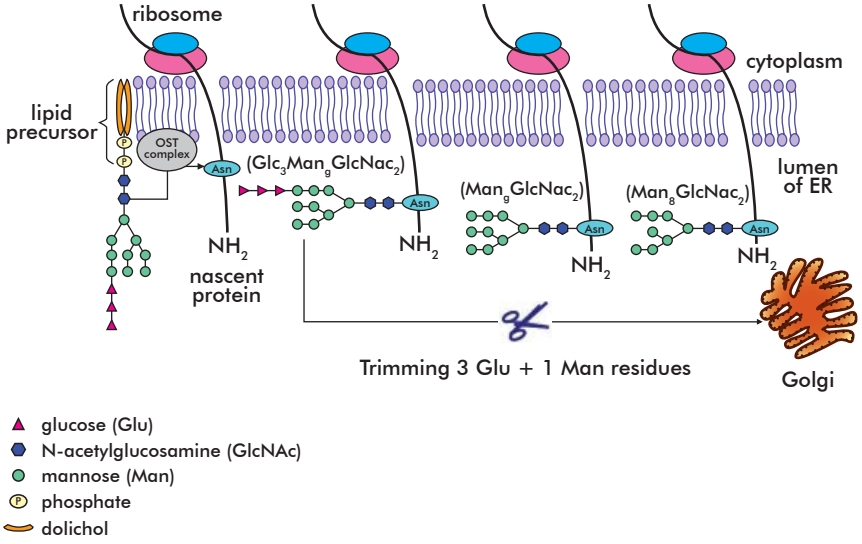

Figure 1. Schematic representation of the steps involved in the N-glycosylation process while protein resides in the endoplasmic reticulum (ER). OST- Oligosaccharyltransferase complex.

N-glycosylation consensus sequence

The existence of the above described consensus sequon does not guarantee the occurrence of N-glycosylation [25-33]. For example, the N-glycan occupied sequences usually appear at points of change in secondary structure and on hydrophobic exposed patches in protein surface [30]. Instead, non-occupied asparagines residues are generally located on non-accessible surface areas and close to 60 residues from the C-terminal protein end [30, 34]. However, the influence of sequon distance to the protein C-terminal end in N-glycosylation site occupancy is still controversial. Since it was demonstrated that the same sequons in similar positions, located fewer than 60 residues of the C-terminal end, from two different proteins can be differentially utilized by OST complex in the same cell line [35]. Besides, the nature of amino acids both at position X and surrounding the sequon strongly modulate the occurrence of N-glycosylation. For example, occupied sequons from eukaryotes, bacteria or archaea never contain Pro residues at position X (also referred as position +1) [20, 30-33, 36]. In case of bacteria, the existence of Pro residues at position -1 also inhibits N-glycosylation [32]. Besides, the frequency of Pro residue is very low at position +3 in eukaryal

8. Couldrey C, Green JE. Metastases: the glycan connection. Breast Cancer Res. 2000;2:321-3.

9. Corthay A, Backlund J, Holmdahl R. Role of glycopeptide-specific T cells in collagen-induced arthritis: an example how post-translational modification of proteins may be involved in autoimmune disease. Ann Med. 2001;33:456-65.

10. Freeze HH. Update and perspectives on congenital disorders of glycosylation. Glycobiology. 2001;11:129R-43R.

11. Varki A, Cummings RD, Esko JD, *et al.*, editors. Essentials of glycobiology. 2nd ed. New York: Cold Spring Harbor Laboratory press; 2008.

12. Zaia J. Mass spectrometry and the emerging field of glycomics. Chem Biol. 2008;15:881-92.

13. Chen YZ, Tang YR, Sheng ZY, Zhang Z. Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. BMC Bioinformatics. 2008;9:101.

14. Aoki-Kinoshita KF. An introduction to bioinformatics for glycomics research. PLoS Comput Biol. 2008;4:e1000075.

15. der Lieth CW, Bohne-Lang A, Lohmann KK, Frank M. Bioinformatics for glycomics: status, methods, requirements and perspectives. Brief Bioinform. 2004;5:164-78.

16. Frank M, Schloissnig S. Bioinformatics and molecular modeling in glycobiology. Cell Mol Life Sci. 2010;67:2749-72.

17. 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.

18. Schaffer C, Graninger M, Messner P. Prokaryotic glycosylation. Proteomics. 2001;1:248-61.

19. Weerapana E, Imperiali B. Asparagine-linked protein glycosylation: from eukaryotic to prokaryotic systems. Glycobiology. 2006;16:91R-101R.

N-glycosylated sequons [30]. An acidic residue (Asp or Glu) at position -2 is required for N-glycosylation in bacteria [32]. Hence, bacterial N-linked sequon was extensive to Asp/Glu-Y-Asn-X-Ser/Thr, where X and Y are not Pro residues [32]. However, negatively charged amino acids are disfavored at position -2 in both eukaryal and archaeal N-glycosylated proteins. This position is preferred for non-polar residues, particularly for aromatic amino acids in the case of eukaryotes [30]. The position X contains a high incidence of small amino acids (Gly, Ala, Val) in eukaryal modified sequons [20]. By contrast, Ser or Thr residues are found at such position in archaeal N-linked sequences [20]. Large hydrophobic residues (Ile, Leu, Met, Phe, Trp or Tyr) are located at position +3 in eukaryal occupied sites [30]. However, such position is occupied by small amino acids (Ala and Gly) in archaeal modified sequons [20]. Moreover, in both archaeal and eukaryal N-glycoproteins, basic residues are poorly represented at position +3. The frequency of N-glycosylation also varies between sequons types; Asn-X-Thr sequon is the most glycosylated one [30]. The roles of Ser and Thr residues at position +2 in Asn-X-Ser/Thr sequons have been already discussed [36, 37]. The hydroxyl group of Ser and Thr residues interacts with the amide side chain of the Asn residue via hydrogen bond, accepting a hydrogen atom [17, 37]. This interaction is facilitated when sequons are placed on beta-turn or other loops because the hydroxyl group of Ser or Thr residues may be oriented close to the Asn amide group [38]. It was also noted that, the side chain of the amino acid at position X is opposite to the hydroxyl and amide groups from Ser/Thr and Asn residues, respectively. Thus, it was suggested that Pro residue is not favored at position X because its unusual rigid structure disrupts the turn structure [38]. Additional studies correlating the frequencies of N-glycosylation with other sequon characteristics have been done. For example, it was demonstrated that overlapping sequons in the yeast invertase (*e.g.*, Asn-Asn-Ser-Ser sequons) can be both clearly glycosylated [39]. Such evidence discharged the idea that steric hindrance might prevent the N-glycosylation of overlapping sequon, at least in yeast invertase [39].

N-glycosylation prediction tools

Currently, three softwares are capable of predicting N-glycosylation; they are known as NetNGlyc [40], EnsembleGly [41] and GPP (Glycosylation Prediction Program) [42]. EnsembleGly and GPP were recently developed and can be used not only for N-glycosylation prediction, but also for the prediction of other glycosylation types (Table 1). However, the web-online NetNGlyc server is still the most used predictor for N-glycosylation [43-45]. All prediction methods use machine learning techniques trained on amino acid sequences. These methods examine the sequon vicinity to discriminate between possible modified and non-modified asparagines, since the amino acid composition flanking potential N-glycosylation sequon is determinant [30].

O-glycosylation

O-linked glycosylation involves the binding of glycans to hydroxyl side chains of serine and threonine residues. There is no well defined motif for the O-glycan acceptor site. The glycans bind to serine and threonine residues which are usually found in a beta conformation and in close vicinity to proline residues [46]. The O-linked glycosylation occurs in bacteria, archaea and eukaryotes [47, 48]. O-glycosylation is a stepwise process where one monosaccharide is added at a time, rather than N-glycosylation where the high-mannose oligosaccharide is transfer en bloc to the target protein. Examples of O-glycans include: O-N-acetylgalactosamine (O-GalNAc), O-N-acetylglucosamine (O-GlcNAc), O-Fucose, O-Glucose, O-Mannose, O-Hexose, O-Xylose. The most abundant and better characterized O-glycosylation type is mucin-type glycosylation [49]. This reaction is catalyzed by the enzymes UDP-N-acetyl-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase which adds an N-acetyl-galactosamine molecule to serine and threonine residues. Mucin-type glycans are found on many secreted and membrane-bound mucin proteins, which are the mucus main components. The function of such proteins is to protect epithelial surfaces [50, 51]. Mucin-type O-glycosilation occurs in the ER and the Golgi apparatus after N-glycosylation, folding, and oligomerization [52].

O-glycosylation target sequence

There is no clear consensus sequence for O-glycosylation. Some studies have confirmed a higher frequency of residues like Pro, Ser, Thr and Ala neighboring mucin-type glycosylated sites [53, 54]. Such sites are preferentially found in coil, turn or linker regions connecting domains. Besides, the experimentally verified O-glycosylated sites are more surface exposed than the non-glycosylated ones [55].

O-glycosylation prediction methods

Several methods for O-glycosylation prediction have been developed (Table 1). Among them, NetOGlyc [55] and Oglyc [56] predictors have been the most

Table 1. Summary of available web-online glycosylation predictors

Server	Glycosylation	URL
NetNGlyc	N-glycosylation	http://www.cbs.dtu.dk/services/NetNGlyc/
EnsembleGly	N-glycosylation O-glycosylation C-glycosylation	http://turing.cs.iastate.edu/EnsembleGly
GPP	N-glycosylation O-glycosylation	http://comp.chem.nottingham.ac.uk/glyco/
NetOglyc	O-glycosylation	http://www.cbs.dtu.dk/services/NetOGlyc/
Oglyc	O-glycosylation	http://www.biosino.org/Oglyc
CKSAAP_OGlySite	O-glycosylation	http://bioinformatics.cau.edu.cn/zzd_lab/CKSAAP_OGlySite/
YinOYang	O-glycosylation	http://www.cbs.dtu.dk/services/YinOYang/
DictyOGlyc	O-glycosylation	http://www.cbs.dtu.dk/services/DictyOGlyc/
Big-PI	GPI-anchor	http://mendel.imp.ac.at/gpi/gpi_server.html
DGPI*	GPI-anchor	http://129.194.185.165/dgpi/
GPI-SOM	GPI-anchor	http://gpi.unibe.ch/
FragAnchor	GPI-anchor	http://navet.ics.hawaii.edu/~fraganchor/NNHMM/NNHMM.html
MemType-2L	GPI-anchor	http://www.csbio.sjtu.edu.cn/bioinf/MemType/
PredGPI	GPI-anchor	http://gpcr.biocomp.unibo.it/predgpi/
NetCGlyc	C-mannosylation	http://www.cbs.dtu.dk/services/NetCGlyc/

* The url for this database was unavailable at the time of publishing.

20. Yurist-Doutsch S, Chaban B, VanDyke DJ, Jarrell KF, Eichler J. Sweet to the extreme: protein glycosylation in Archaea. Mol Microbiol. 2008;68:1079-84.

21. Helenius A, Aebi M. Intracellular functions of N-linked glycans. Science. 2001;291:2364-9.

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

23. Jones J, Krag SS, Betenbaugh MJ. Controlling N-linked glycan site occupancy. Biochim Biophys Acta. 2005;1726:121-37.

24. Munro S. What can yeast tell us about N-linked glycosylation in the Golgi apparatus? FEBS Lett. 2001;498:223-7.

25. Roitsch T, Lehle L. Structural requirements for protein N-glycosylation. Influence of acceptor peptides on cotranslational glycosylation of yeast invertase and site-directed mutagenesis around a sequon sequence. Eur J Biochem. 1989;181:525-9.

26. Gavel Y, von Heijne G. Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. Protein Eng. 1990;3:433-42.

27. Shakin-Eshleman SH, Spitalnik SL, Kasturi L. The amino acid at the X position of an Asn-X-Ser sequon is an important determinant of N-linked core-glycosylation efficiency. J Biol Chem. 1996;271:6363-6.

28. Kasturi L, Chen H, Shakin-Eshleman SH. Regulation of N-linked core glycosylation: use of a site-directed mutagenesis approach to identify Asn-Xaa-Ser/Thr sequons that are poor oligosaccharide acceptors. Biochem J. 1997;323 (Pt 2):415-9.

29. Christlet TH, Biswas M, Veluraja K. A database analysis of potential glycosylating Asn-X-Ser/Thr consensus sequences. Acta Crystallogr D Biol Crystallogr. 1999;55:1414-20.

widely used. The former considers primary amino acid sequence, secondary structure and evolutionary information of both mucin-type glycosylated and non-mucin type glycosylated sites [55]. The Oglyc server also predicts mucin-type O-glycosylation in mammalians, and it is based on a combination of physical properties of amino acids [56]. Recently, three new O-glycosylation predictors have been available: EnsembleGly [41], CKSAAP_OGlySite [13] and GPP [42]. CKSAAP_OGlySite server only predicts mucin-type O-glycosylation, like NetOGlyc and Oglyc servers [13]. It considers the local conformation and the short-range interactions of amino acids close to the mucin-type glycosylated sites. The latest published software, GPP, is based on pairwise sequence patterns combined with prediction of protein secondary structure, surface accessibility and hydrophobicity [42]. Actually, GPP is the most accurate program for both N- and O-glycosylation prediction available [42].

Glypiation

Glypiation is ubiquitous in eukaryotes and also possible in a reduced subset of *archaea* species [57], but probably absent in other Archaea and Eubacteria [58]. It involves the addition of a GPI molecule to the C-terminal end of the target protein. The GPI molecule is composed by a phosphatidylinositol group and a sugar moiety. The sugar moiety comprises a non-acetylated glucosamine attached to three mannosyl residues and to the phosphatidylinositol group. A phosphoethanolamine residue connected to the terminal mannose mediates the binding of the GPI to the C-terminal end of the mature protein by an amide linkage. Two long-chain fatty acids included in the phosphatidylinositol group anchor the protein to the cell membrane. Although all GPI molecules share a common core, some variability have been observed depending on the organism and the cell type in which they are synthesized [59]. Such differences correspond to substitutions in the oligosaccharidic and lipidic portions of the GPI residue. Proteins susceptible to glypiation contain a C-terminal signal sequence which is sufficient for GPI attachment [60]. Glypiation begins with the recognition of the ω-site in the C-terminal protein end embedded in the ER membrane. Then, about 20-30 residues downstream from the ω-site (propeptide) are removed and replaced by the GPI molecule via an amide bond between the phosphoethanolamine group and the new C-terminal residue (ω-site). The transamidase complex is responsible for the above mentioned events that occur during glypiation [61]. After glypiation process, proteins pass through the secretory pathway in vesicles to Golgi apparatus and finally, most of them are translocated to the cell membrane [62]. Then, glypiated proteins reside attached to the cell membrane facing the extracellular environment, where they perform different important functions [61]. Figure 2 schematizes the steps involved in the glypiation process. Glypiated proteins may operate as enzymes [63], membrane receptors [64], surface antigens [65] and adhesion molecules [66]. However, some GPI-anchored proteins are released by enzymatic cleavage of their anchor to achieve other functions [67]. Glypiation process is relevant for cell functions and development. Indeed, mice lacking the GPI synthesis were unable to grow at embryonic stages [68].

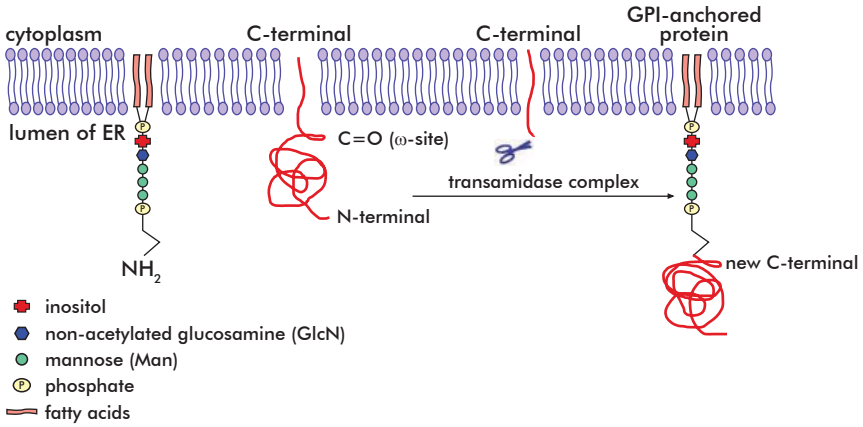

Figure 2. Schematic representation of the steps involved in the GPI-anchor event. GPI: glycosylphosphatidylinositol.

GPI-anchor signal sequence

None consensus sequence characterizes the amino acid at the ω-site [69]. Some residues experimentally verified in ω-site include: cysteine, aspartic acid, glycine, asparagine, and serine [69]. However, some generally features can be found in the non-cleaved C-terminal end of glypiated proteins like: an unfolded linker region comprising about 11 residues (upstream from position ω-1), a region comprised by small amino acids surrounding the cleavage site (from positions ω-1 to ω+2) which is followed by a moderately polar spacer region (from positions ω+3 to ω+9) and finally, a hydrophobic tail extended from the position ω+10 up to the C-terminal end [70].

GPI-anchor prediction

Various computational methods have been developed for the prediction of glypiated proteins (Table 1). Such predictors identify the C-terminal GPI-anchor signals and most of them also distinguish the ω-site. Methods like Big-PI [70] and DGPI [71] are based on the rules concerning the amino acid composition around the ω-site observed in glypiated proteins. The Big-PI predictions are divided in relation to life-kingdoms, for example, for animals, fungi and plants. The performance of these softwares is better for the already verified glypiated proteins compared to the non-confirmed ones [72]. More recently, the GPI-SOM program was developed [72]. This one achieved better results than Big-PI and DGPI [71]. The improvement of GPI-SOM is related with its power to discriminate false positives GPI-anchor proteins. The main source of false positive predictions is the existence of integral membrane proteins having a transmembrane domain at the C-terminal end. GPI-SOM overcomes this difficulty by the simulation of mutagenesis experiments at the C-terminus leading to the identification of key residues that discriminates GPI-anchored proteins from nonGPI-anchored proteins. This procedure follows an experimental demonstration that a single point mutation may be enough to convert a GPI-anchor signal into a transmembrane domain [73]. The latest prediction tools are FragAnchor [74] and PredGPI [75].

30. Petrescu AJ, Milac AL, Petrescu SM, Dwek RA, Wormald MR. Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. Glycobiology. 2004; 14:103-14.

31. Ben Dor S, Esterman N, Rubin E, Sharon N. Biases and complex patterns in the residues flanking protein N-glycosylation sites. Glycobiology. 2004;14:95-101.

32. Kowarik M, Young NM, Numao S, Schulz BL, Hug I, Callewaert N, *et al*. Definition of the bacterial N-glycosylation site consensus sequence. EMBO J. 2006; 25:1957-66.

33. Abu-Qarn M, Eichler J. An analysis of amino acid sequences surrounding archaeal glycoprotein sequons. Archaea. 2007;2:73-81.

34. Nilsson I, von Heijne G. Glycosylation efficiency of Asn-Xaa-Thr sequons depends both on the distance from the C terminus and on the presence of a downstream transmembrane segment. J Biol Chem. 2000;275:17338-43.

35. Walmsley AR, Hooper NM. Glycosylation efficiency of Asn-Xaa-Thr sequons is independent of distance from the C-terminus in membrane dipeptidase. Glycobiology. 2003;13:641-6.

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

37. 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.

38. Yan A, Lennarz WJ. Unraveling the mechanism of protein N-glycosylation. J Biol Chem. 2005;280:3121-4.

39. Reddy A, Gibbs BS, Liu YL, Coward JK, Changchien LM, Maley F. Glycosylation of the overlapping sequons in yeast external invertase: effect of amino acid variation on site selectivity *in vivo* and *in vitro*. Glycobiology. 1999;9:547-55.

40. NetNGlyc 1.0 Server [Internet; updated 2007 Mar 09; cited 2011 Jan 17]. Available from: http://www.cbs.dtu.dk/services/NetNGlyc/

Both are able to recognize a higher number of GPI-anchored proteins with a lower rate of false positive errors, in comparison with other earlier described algorithms. But, PredGPI outperforms all available prediction methods [75]. MemType-2L is another resource capable to predict not only GPI-anchored protein, but also other seven types of membrane proteins [76]. However, both FragAnchor and MemType-2L are not able to detect the ω-site [74, 76].

C-mannosylation

C-mannosylation was originally found in human ribonuclease protein (RNase 2) from urine [77]. This post-translational modification involves the attachment of a mannopyranosyl residue to the C2 atom indole moiety of a tryptophan residue via a C-C bond [77, 78]. The transfer of the mannose residue to the target protein is catalyzed by the enzyme C-mannosyltransferase [77]. It has been suggested that this reaction probably occurs in the ER since already folded proteins are poor substrates *in vitro* [79]. Although C-glycosylation appears to be common in mammalian proteins, it has not been observed in yeast and bacteria [80]. At present, the knowledge related with the C-mannosylation functions and disorders is still very limited. However, at least three functions have been recently described. For example, C-mannosylation is required for an adequate folding of Cys subdomains contained in two mucin proteins (MUC5AC and MUC5B)[79]. Also, C-glycosylation appears to control the secretion of the puntion-1 protein [81] and it may be involved in the development of diabetic complications under hyperglycemic conditions [82].

C-mannosylation consensus sequence

C-mannosylation generally occurs at the first tryptophan residue (position 0) contained in the Trp-X-X-Trp sequence motif, where X could be any amino acid [77]. However, other studies revealed that mannosylated tryptophan has also been detected in other sequences motifs, for example in Trp-X-X-Phe, Trp-X-X-Tyr and Trp-X-X-Cys [83-85]. Since some sequence motifs having another aromatic residue instead of Trp at the position +3 can also be mannosylated, the Trp-X-X-Trp pattern seems to be sufficient but not strictly required for C-mannosylation. Indeed, it was demonstrated that only two-thirds of known mannosylated sites are found in Trp-X-X-Trp motifs [83]. Besides, small and/or polar residues (Ala, Gly, Ser and Thr) are preferred for occupying X positions within the sequence motif while Phe and Leu residues are not well tolerated at the mentioned position [83]. Other general features related with the structure of C-mannosylation motif, have also been described. For example, modified tryptophan residues instead of non-modified ones, are partly solvent exposed [83]. Besides, a particular interaction takes place at the Trp-X-X-Trp motif, where both tryptophan residues interact via an aromatic stacking. It was suggested that this type of interaction may account for the recognition of the C-mannosyltransferase enzyme [83].

C-mannosylation prediction

There is only one computational method reserved for the solely prediction of C-mannosylation and it is called NetCGlyc [83] (Table 1). It predicts not only the typical mannosylated site (Trp-X-X-Trp), but others like those previously mentioned (Trp-X-X-Phe, Trp-X-X-Tyr and Trp-X-X-Cys). However, better predictions are achieved for the first case [83]. Recently, EnsembleGly could be also used for C-mannosylation prediction [41] (Table 1).

Databases

Undoubtedly, the development of accuracy glycosylation prediction methods requires the existence of databases with experimentally verified glycosylated sites. Usually, the data is extracted from O-GlycBase database [86], but in some cases, the Swiss-Prot database has been also used [86, 87]. O-GlycBase contains proteins having at least one experimentally verified O- or C-glycosylation site. Glycosylation data may also be found in databases including different types of post-translational modifications, like dbPTM [88], SysPTM [89], RESID [90] and others (Table 2). Besides glycoprotein databases, carbohydrate molecules databases also exists. For example, the major available databases of complex carbohydrates are: Complex Carbohydrate Structure Database (CCSD or CarbBank) [91], Glycosciences.de [92], KEGG GLYCAN [93], GlycomeDB [94], Carbohydrate DB from the Consortium Functional Genomic [95] and others (Table 2). Other available databases include the carbohydrate tertiary structures (*e.g.*, 3D Disaccharides [96], GDB:Structures [97] and GlycoMapsDB [98]). Others like GlyTorsionDB, GlySeqDB and GlyVicinityDB are integrated in the Carbohydrate Structure Suite [99]. Each database comprises the carbohydrate torsion angles, the glycoprotein sequences and the carbohydrate-protein interactions, respectively [99].

Concluding remarks

Currently, about 15 web-online softwares for the prediction of all glycosylation types are available. The possibility to perform prediction analysis using at least two or three different methods in parallel now exists. However, in our opinion, with upcoming prediction methods, a critical comparative study would

41. 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.

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

43. Liu Y, Nguyen A, Wolfert RL, Zhuo S. Enhancing the secretion of recombinant proteins by engineering N-glycosylation sites. Biotechnol Prog. 2009;25:1468-75.

44. Radoslavov G, Jordanova R, Teofanova D, Georgieva K, Hristov P, Salomone-Stagni M, *et al*. A novel secretory poly-cysteine and histidine-tailed metalloprotein (Ts-PCHTP) from Trichinella spiralis (Nematoda). PLoS One. 2010;5:e13343.

45. Lu C, Walker WH, Sun J, Weisz OA, Gibbs RB, Witchel SF, *et al*. Insulin-like peptide 6: characterization of secretory status and posttranslational modifications. Endocrinology. 2006;147:5611-23.

46. Gupta R, Brunak S. Prediction of glycosylation across the human proteome and the correlation to protein function. Pac Symp Biocomput. 2002;7:310-22.

47. Nothaft H, Szymanski CM. Protein glycosylation in bacteria: sweeter than ever. Nat Rev Microbiol. 2010;8:765-78.

48. Gentzsch M, Tanner W. Protein-O-glycosylation in yeast: protein-specific mannosyltransferases. Glycobiology. 1997;7:481-6.

49. Hanisch FG. O-glycosylation of the mucin type. Biol Chem. 2001;382:143-9.

50. Carraway KL, Hull SR. Cell surface mucin-type glycoproteins and mucin-like domains. Glycobiology. 1991;1:131-8.

51. Strous GJ, Dekker J. Mucin-type glycoproteins. Crit Rev Biochem Mol Biol. 1992;27:57-92.

52. Asker N, Baeckstrom D, Axelsson MA, Carlstedt I, Hansson GC. The human MUC2 mucin apoprotein appears to dimerize before O-glycosylation and shares epitopes with the 'insoluble' mucin of rat small intestine. Biochem J. 1995;308(Pt 3):873-80.

Table 2. Summary of available web-online glycosylation predictors

Databases	URL
Carbohydrates	
CarbBank	http://www.boc.chem.uu.nl/sugabase/carbbank.html
Glycosciences.de	http://www.glycosciences.de
KEGG GLYCAN	http://www.genome.jp/kegg/glycan/
GlycomeDB	http://www.glycome-db.org/
Carbohydrate Binding Protein DB	http://web.mit.edu/glycomics/cbp/cbpdb.shtml
SWEET-DB	http://www.glycosciences.de/modeling/sweet2
EuroCarbDB	EuroCarbDB http://www.ebi.ac.uk/eurocarb/
Glycoproteins	
BCSD	http://www.glyco.ac.ru/bcsdb/start.shtml
dbPTM	http://dbPTM.mbc.nctu.edu.tw/
GlycoProtDB (GPDB)	http://riodb.ibase.aist.go.jp/rcmg/glycodb/
GlycoSuiteDB	http://glycosuitedb.expasy.org/glycosuite/glycodb
O-GlycBase	http://www.cbs.dtu.dk/databases/OGLYCBASE/
RESID	http://www.ncifcrf.gov/RESID/
SysPTM	http://www.biosino.org.cn/SysPTM/

be required to define the most accurate one. The combination of such prediction methods with other non-carbohydrate tools facilities the characterization and rational modification of the native protein glycosylation pattern.

Acknowledgments

We thank Dr. Jose Alberto Cremata, from the CIGB Carbohydrate Department, for his useful comments and suggestions on the manuscript.

53. Thanka Christlet TH, Veluraja K. Database analysis of O-glycosylation sites in proteins. Biophys J. 2001;80:952-60.

54. Wilson IB, Gavel Y, von Heijne G. Amino acid distributions around O-linked glycosylation sites. Biochem J. 1991;275(Pt 2):529-34.

55. Julenius K, Molgaard A, Gupta R, Brunak S. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. Glycobiology. 2005;15:153-64.

56. Li S, Liu B, Zeng R, Cai Y, Li Y. Predicting O-glycosylation sites in mammalian proteins by using SVMs. Comput Biol Chem. 2006;30:203-8.

57. Kobayashi T, Nishizaki R, Ikezawa H. The presence of GPI-linked protein(s) in an archaeobacterium, Sulfolobus acidocaldarius, closely related to eukaryotes. Biochim Biophys Acta. 1997;1334:1-4.

58. Eisenhaber B, Bork P, Eisenhaber F. Post-translational GPI lipid anchor modification of proteins in kingdoms of life: analysis of protein sequence data from complete genomes. Protein Eng. 2001;14:17-25.

59. Menon AK. Structural analysis of glycosylphosphatidylinositol anchors. Methods Enzymol. 1994;230:418-42.

60. Caras IW, Weddell GN, Davitz MA, Nussenzweig V, Martin DW Jr. Signal for attachment of a phospholipid membrane anchor in decay accelerating factor. Science. 1987;238:1280-3.

61. Udenfriend S, Kodukula K. How glycosylphosphatidylinositol-anchored membrane proteins are made. Annu Rev Biochem. 1995;64:563-91.

62. Chatterjee S, Mayor S. The GPI-anchor and protein sorting. Cell Mol Life Sci. 2001;58:1969-87.

63. Koelsch R, Gottwald S, Lasch J. Release of GPI-anchored membrane aminopeptidase P by enzymes and detergents has some peculiarities. Biochim Biophys Acta. 1994;1190:170-2.

64. Bergelson JM, Chan M, Solomon KR, St John NF, Lin H, Finberg RW. Decay-accelerating factor (CD55), a glycosylphosphatidylinositol-anchored complement regulatory protein, is a receptor for several echoviruses. Proc Natl Acad Sci USA. 1994;91:6245-8.

65. Chan CH, Stanners CP. Recent advances in the tumour biology of the GPI-anchored carcinoembryonic antigen family members CEACAM5 and CEACAM6. Curr Oncol. 2007;14:70-3.

66. Cervello M, Matranga V, Durbec P, Rougon G, Gomez S. The GPI-anchored adhesion molecule F3 induces tyrosine phosphorylation: involvement of the FNIII repeats. J Cell Sci. 1996;109(Pt 3):699-704.

67. Frieman MB, Cormack BP. Multiple sequence signals determine the distribution of glycosylphosphatidylinositol proteins between the plasma membrane and cell wall in Saccharomyces cerevisiae. Microbiology. 2004;150:3105-14.

68. Nozaki M, Ohishi K, Yamada N, Kinoshita T, Nagy A, Takeda J. Developmental abnormalities of glycosylphosphatidylinositol-anchor-deficient embryos revealed by Cre/loxP system. Lab Invest. 1999;79:293-9.

69. Orlean P, Menon AK. Thematic review series: lipid posttranslational modifications. GPI anchoring of protein in yeast and mammalian cells, or: how we learned to stop worrying and love glycophospholipids. J Lipid Res. 2007;48:993-1011.

70. Eisenhaber B, Bork P, Eisenhaber F. Prediction of potential GPI-modification sites in proprotein sequences. J Mol Biol. 1999;292:741-58.

71. Kronegg J, Buloz D. Detection/prediction of GPI cleavage site (GPI-anchor) in a protein (DGPI). 1999. Available from: http://129.194.185.165/dgpi/index_en.html

72. Fankhauser N, Maser P. Identification of GPI anchor attachment signals by a Kohonen self-organizing map. Bioinformatics. 2005;21:1846-52.

73. Dalley JA, Bulleid NJ. The endoplasmic reticulum (ER) translocon can differentiate between hydrophobic sequences allowing signals for glycosylphosphatidylinositol anchor addition to be fully translocated into the ER lumen. J Biol Chem. 2003;278:51749-57.

74. Poisson G, Chauve C, Chen X, Bergeron A. FragAnchor: a large-scale predictor of glycosylphosphatidylinositol anchors in eukaryote protein sequences by qualitative scoring. Genomics Proteomics Bioinformatics. 2007;5:121-30.

75. Pierleoni A, Martelli PL, Casadio R. PredGPI: a GPI-anchor predictor. BMC Bioinformatics. 2008;9:392.

76. Chou KC, Shen HB. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun. 2007;360:339-45.

77. Hofsteenge J, Muller DR, de Beer T, Loffler A, Richter WJ, Vliegenthart JF. New type of linkage between a carbohydrate and a protein: C-glycosylation of a specific tryptophan residue in human RNase Us. Biochemistry. 1994;33:13524-30.

78. de Beer T, Vliegenthart JF, Loffler A, Hofsteenge J. The hexopyranosyl residue that is C-glycosidically linked to the side chain of tryptophan-7 in human RNase Us is alpha-mannopyranose. Biochemistry. 1995;34:11785-9.

79. Perez-Vilar J, Randell SH, Boucher RC. C-Mannosylation of MUC5AC and MUC5B Cys subdomains. Glycobiology. 2004;14:325-37.

80. Zanetta JP, Pons A, Richet C, Huet G, Timmerman P, Leroy Y, *et al*. Quantitative gas chromatography/mass spectrometry determination of C-mannosylation of tryptophan residues in glycoproteins. Anal Biochem. 2004;329:199-206.

81. Wang LW, Leonhard-Melief C, Haltiwanger RS, Apte SS. Post-translational modification of thrombospondin type-1 repeats in ADAMTS-like 1/punctin-1 by C-mannosylation of tryptophan. J Biol Chem. 2009;284:30004-15.

82. Ihara Y, Manabe S, Kanda M, Kawano H, Nakayama T, Sekine I, *et al*. Increased expression of protein C-mannosylation in the aortic vessels of diabetic Zucker rats. Glycobiology. 2005;15:383-92.

83. Julenius K. NetCGlyc 1.0: prediction of mammalian C-mannosylation sites. Glycobiology. 2007;17:868-76.

84. Krieg J, Hartmann S, Vicentini A, Glasner W, Hess D, Hofsteenge J. Recognition signal for C-mannosylation of Trp-7 in RNase 2 consists of sequence Trp-x-x-Trp. Mol Biol Cell. 1998;9:301-9.

85. Hofsteenge J, Blommers M, Hess D, Furmanek A, Miroshnichenko O. The four terminal components of the complement system are C-mannosylated on multiple tryptophan residues. J Biol Chem. 1999;274:32786-94.

86. Gupta R, Birch H, Rapacki K, Brunak S, Hansen JE. O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. Nucleic Acids Res. 1999;27:370-2.

87. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003;31:365-70.

88. Lee TY, Hsu JB, Chang WC, Wang TY, Hsu PC, Huang HD. A comprehensive resource for integrating and displaying protein post-translational modifications. BMC Res Notes. 2009;2:111.

89. Li H, Xing X, Ding G, Li Q, Wang C, Xie L, *et al*. SysPTM: a systematic resource for proteomic research on post-translational modifications. Mol Cell Proteomics. 2009;8:1839-49.

90. Garavelli JS. The RESID Database of Protein Modifications as a resource and annotation tool. Proteomics. 2004;4:1527-33.

91. Doubet S, Bock K, Smith D, Darvill A, Albersheim P. The complex carbohydrate structure database. Trends Biochem Sci. 1989;14:475-7.

92. Lutteke T, Bohne-Lang A, Loss A, Goetz T, Frank M, der Lieth CW. GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. Glycobiology. 2006;16:71R-81R.

93. Hashimoto K, Goto S, Kawano S, Aoki-Kinoshita KF, Ueda N, Hamajima M, *et al*. KEGG as a glycome informatics resource. Glycobiology. 2006;16:63R-70R.

94. Ranzinger R, Herget S, Wetter T, der Lieth CW. GlycomeDB - integration of open-access carbohydrate structure databases. BMC Bioinformatics. 2008;9:384.

95. Consortium for Functional Glycomics [Internet]. Consortium for Functional Glycomics funded by NIGMS [updated 2007 Apr 05; cited 2011 Jan 17]. Available from: http://www.functionalglycomics.org/static/consortium/consortium.shtml

96. Imberty A, Delage MM, Bourne Y, Cambillau C, Perez S. Data bank of three-dimensional structures of disaccharides: Part II, N-acetyllactosaminic type N-glycans. Comparison with the crystal structure of a biantennary octasaccharide. Glycoconj J. 1991;8:456-83.

97. Lutteke T, der Lieth CW. pdb-care (PDB carbohydrate residue check): a program to support annotation of complex carbohydrate structures in PDB files. BMC Bioinformatics. 2004;5:69.

98. Frank M, Lutteke T,der Lieth CW. GlycoMapsDB: a database of the accessible conformational space of glycosidic linkages. Nucleic Acids Res. 2007;35:287-90.

99. Lutteke T, Frank M, der Lieth CW. Carbohydrate Structure Suite (CSS): analysis of carbohydrate 3D structures derived from the PDB. Nucleic Acids Res. 2005;33:D242-D246.

Received in February, 2011. Accepted for publication in March, 2011.