Pacific Symposium on Biocomputing 7:310-322 (2002)

Prediction of glycosylation across the human proteome and the correlation to protein function

Ramneek Gupta and Søren Brunak
Center for Biological Sequence Analysis, Bldg-208, Bio-Centrum
Technical University of Denmark, DK-2800 Lyngby, Denmark.

1 Introduction

The addition of a carbohydrate moeity to the side-chain of a residue in a protein chain influences the physicochemical properties of the protein. Glycosylation is known to alter proteolytic resistance, protein solubility, stability, local structure, lifetime in circulation and immunogenicity[1,2].

Of the various forms of protein glycosylation found in eukaryotic systems, the most important types are N-linked, O-linked GalNAc (mucin-type) and O-β-linked GlcNAc (intracellular/nuclear) glycosylation. N-linked glycosylation is a co-translational process involving the transfer of the precursor oligosaccharide, $GlcNAc_2Man_9Glc_3$, to asparagine residues in the protein chain. The asparagine usually occurs in a sequon Asn-Xaa-Ser/Thr, where Xaa is not Proline. This is however, not a specific consensus since not all such sequons are modified in the cell. O-linked glycosylation involves the post-translational transfer of an oligosaccharide to a serine or threonine residue. In this case, there is no well-defined motif for the acceptor site other than the near vicinity of proline and valine residues.

We have developed glycosylation site prediction methods for these three types of glycosylation, using artificial neural networks that examine correlations in the local sequence context and surface accessibility. In this paper, we have used glycosylation site information on human proteins to illustrate the contribution of glycosylation to protein function and assess how widespread this modification is across the human proteome.

2 Methods

2.1 Data set

Analysis shown in this paper was derived on a set of human proteins obtained from the SWISS-PROT (rel. 38) database. This consisted of 5,795 well annotated proteins. We chose to work with proteins from a single organism i.e. humans, to restrict the diversity of oligosaccharyltransferase acceptor sites.

Glycosylation in simple organisms, such as yeast, is well studied [3,4], but their glycans are usually high mannosylated structures, and it is not clear how similar their mechanism of glycosylation is to that of humans. Combining data from different organisms would complicate the analysis: possible 'families' of acceptor specificities causing ambiguity in distinguishing acceptor (positive) sites from non-acceptor (negative) sites.

2.2 Functional categories for proteins

Defining protein function is a complicated task, and there are many different ways of describing the roles and functions of a protein in a cell. This is the topic of many on-going ontology projects [5]. Here we chose to use a cellular role descriptor and subcellular location as our categorisations:

15 categories (13 defined, 2 unknown) reflective of the 'cellular role' of the protein in the cell were employed (as shown in Figure 1). The automatic class assignment to sequences was made by an extension of the Euclid system performing a lingustic analysis and clustering of SWISS-PROT keywords [6,7]. Keywords were parsed for the human proteins in SWISS-PROT. For each functional class, the informative weight (Z-score) of each keyword was extracted from a dictionary [6]. Keyword sums gaves scores to all categories for a particular sequence. The central point of the Euclid system is the dictionary. The primary version of this dictionary was generated from an initial set of carefully, hand annotated proteins from different organisms spanning every kingdom of life. From this initial set, a first dictionary was defined which was used to assign all SWISS-PROT proteins and the process of dictionary definition and assignment was reiterated until convergence. This final dictionary obtained was used to assign functional classes to around 5,500 human proteins from SWISS-PROT.

The cellular role categories themselves [8] were derived from an earlier proposed scheme for *Escherichia coli* [9] which was later extended by the TIGR group for other complete genomes. These categories comprise 13 functional classes which are subsets of three superclasses: Energy, Communication and Information. Proteins which do not fit in the 13 categories are assigned to 'Other' (functionally undefined cluster) or to 'Unknown' (sequences which do not contain the relevant keywords needed for classification in the above system).

Subcellular locations of proteins were obtained from SWISS-PROT annotations and PSORT predictions [10] (where no parsable SWISS-PROT annotation was found).

3 Results

3.1 N-Glycosylation

N-linked glycosylation modifies membrane and secreted proteins. This co-translational process occurs in the endoplasmic reticulum and is known to influence protein folding. The modification attributes various functional properties to a protein. To examine if certain categories of proteins were more prone to glycosylation than others, we studied the spread of known glycosylation sites across different categories.

N-glycosylation may also display some positional preferences in the protein chain. Specifically, it has been shown that sites need to be 12-14 residues away from the N-terminus [11] and that glycosylation efficiency is reduced within 60 residues of the C-terminus [12].

In our data set of approximately 6,000 human proteins, only 189 proteins (at 453 *confirmed* sites) were annotated in SWISS-PROT as N-glycosylated (not considering proteins with only POTENTIAL or PROBABLE sites). Figure 1 illustrates the spread of human glycosylation sites along the protein chain and across predicted subcellular locations and keyword based assignment of cellular role categories. Relative positions of sites on proteins were calculated with respect to normalised sequence lengths. The sequence length, divided into tenths is shown along the x-axis, from the N-terminal start on the left to the C-terminal end on the right.

N-glycosylated proteins appeared to almost exclusively belong to the functional category, 'Transport and binding'. This may not be too surprising considering that this category consists largely of membrane and secreted proteins. Only a few proteins belonged to any other cellular role category and most of these appeared involved in central intermediary metabolism. Subcellularly, extracellular proteins were the most favoured and others occurred in membrane proteins and in the endoplasmic reticulum or Golgi.

A clear positional preference for glycosylation sites on protein chains was apparent. The terminal ends of proteins seemed unfavourable and most sites seemed to occur N-terminal to the centre of the protein chain (20 to 40% along the length from the N-terminal start). The frequency of sites smoothly tapered off on both ends from this peak with a longer C-terminal tail. This statistical observation agrees with specific experimental indications of a 12-14 residue distance from the N-terminal and a 60 residue distance from the C-terminal end [11,12]. One peculiar observation from the figure was the C-terminal sites in nuclear proteins. On examination, these turned out to be around 10 proteins which were indeed annotated to be N-glycosylated in the C-terminal. However, this seems to be an anomaly of the sub-cellular prediction by PSORT. For

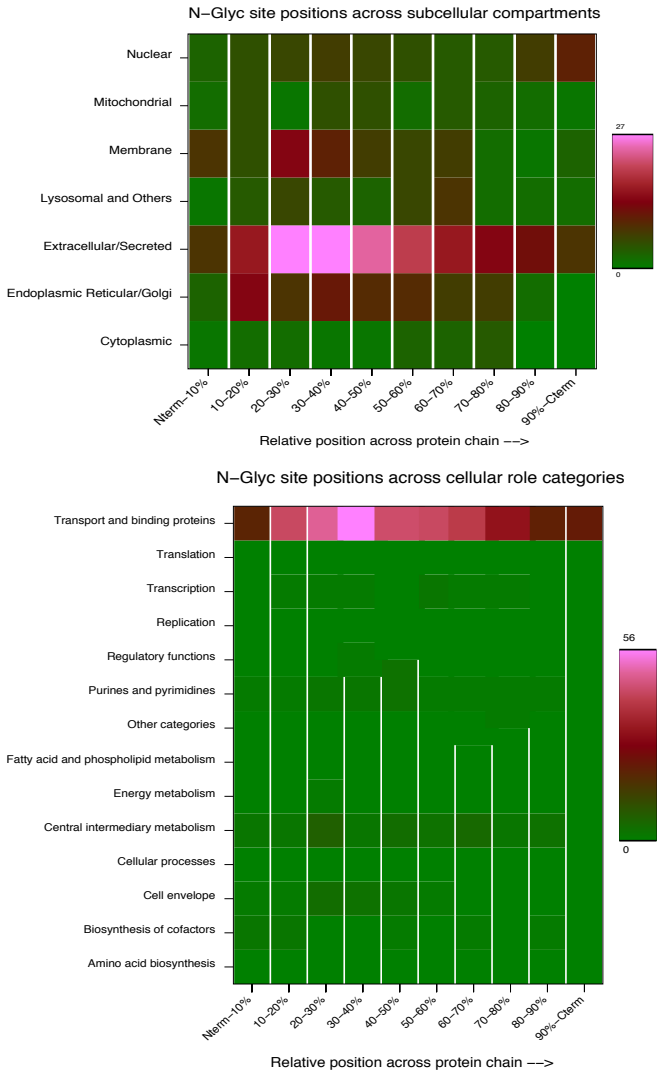

Figure 1: **Categorical distribution of known N-glycosylation sites across the protein chain.** Colour indicates frequency of sites (green to pink in increasing order). Protein chains, normalised in length, are represented across the x-axis from N-terminal to C-terminal. Subcellular locations (top) were predicted using PSORT, and cellular role classification (bottom) by lexical analysis of SWISS-PROT keywords (Alfonso Valencia *et al.*). Most N-glycosylation sites were clustered in the first half of all protein chains, and mainly occurred in extracellular transport and binding proteins.

instance, some secreted proteins among these were Vasopressin-Neurophysin 2-Copeptin precursor, Von Willebrand Factor Precursor and Immunoglobulin Delta Chain C.

Experimental determination of glycosylation sites is difficult to achieve as large amounts of purified protein are needed for the analysis of glycosylation sites. In addition, glycosylation can be an organism- and tissue specific event. Therefore only a few glycoproteins have been characterised so far as reflected in the low percentage of glycoprotein entries in SWISS-PROT (approx. 10% of human proteins, *see also*[13]). This motivates the need for developing theoretical means of predicting the glycosylation potential of sequons.

3.2 O-linked GalNAc Glycosylation

The addition of GalNAc linked to serine or threonine residues of secreted and cell surface proteins, and further addition of Gal/GalNAc/GlcNAc residues[2], is also known as mucin type glycosylation and is catalysed by a family of UDP-N-acetylgalactosamine: polypeptide N-acetylgalactosaminyltransferases (GalNAc-transferases). The modification, a post-translational event, takes place in the cis-Golgi compartment[14] after N-glycosylation and folding of the protein, and affects secreted and membrane bound proteins.

There is no acceptor motif defined for O-linked glycosylation. The only common characteristic among most O-glycosylation sites is that they occur on serine and threonine residues in close vicinity to proline residues, and that the acceptor site is usually in a beta-conformation. A prediction method[15,16] for this type of glycosylation on mammalian proteins has been built earlier and made available as a web server[a]. A database of O-glycosylated sequences is also available[b] and was used in constructing the O-glycosylation site prediction methods[17].

Figure 2 shows the spread of predicted glycosylation sites (O-GalNAc, mucin-type) across different categories and across the protein chain. To construct this plot, sequence lengths were normalised, and relative position expressed on a percent (0-100) scale. Glycosylation sites were binned (10 bins across each sequence), and their frequency plotted across different categories. Sites tend to cluster towards the C- and N-termini of proteins for some categories. This figure also shows that O-glycosylation acceptor sites occur in a wide range of proteins, though glycosylation patterns (frequency, positions across chain) may differ for different types of proteins.

[a] http://www.cbs.dtu.dk/services/NetOGlyc/
[b] http://www.cbs.dtu.dk/databases/OGLYCBASE/

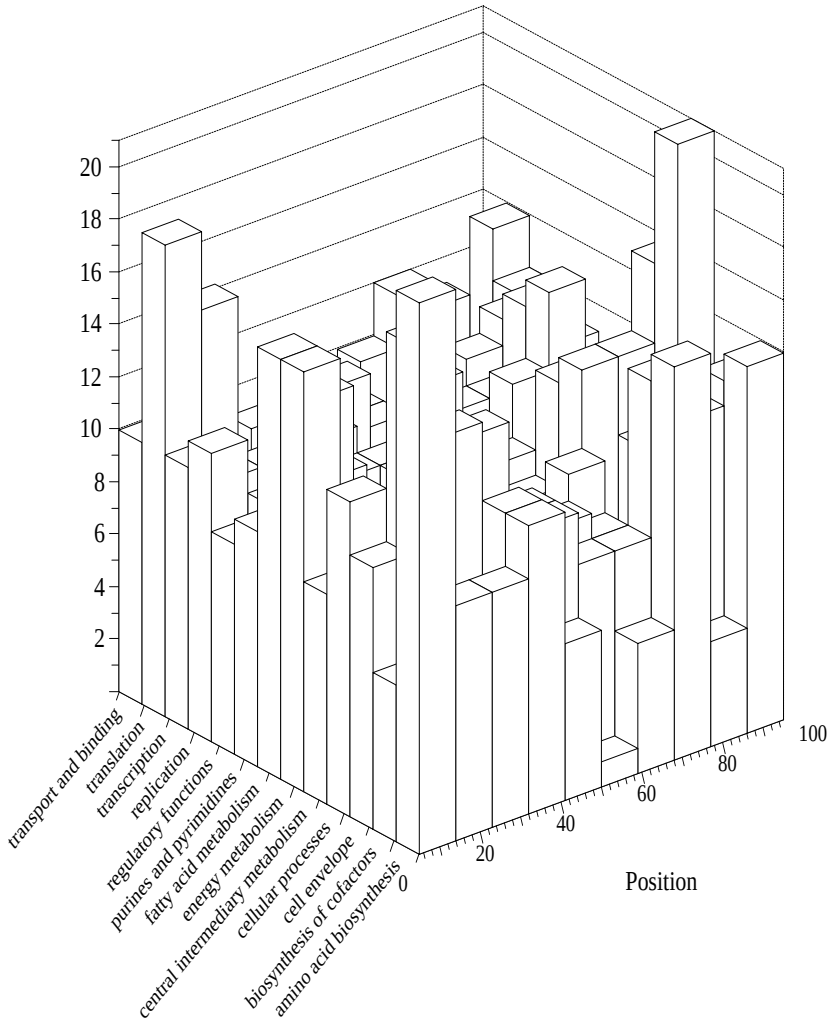

Figure 2: **Postional O-GalNAc glycosylation.** O-GalNAc (mucin type) glycosylation displays preference for position across a protein chain which could be significant across different categories. The *Position* axis reflects normalised protein chain length from N-terminal (0 on the axis) to C-terminal (100). The height of the bars indicates the number of predicted O-GalNAc sites (in $\sim 5,500$ human proteins) for a particular category in a particular *position bin*.

3.3 O-linked GlcNAc Glycosylation

Glycosylation of cytosolic and nuclear proteins by single N-acetylglucosamine (GlcNAc) monosaccharides is known to be highly dynamic and occurs on proteins with wide-ranging functions and cellular roles[18,19]. N-acetylglucosamine, donated by the nucleotide precursor UDP-N-acetylglucosamine, is attached in a beta-anomeric linkage to the hydroxyl group of serine or threonine residues.

So far, all proteins with O-β-GlcNAc linked residues, are also known to be phosphorylated. Evidence suggests that at least in some cases, these two post-translational modification events may share a reciprocal relationship[18,20]. This peculiar behaviour strongly suggests a regulatory role for this modification. Sites which can be both glycosylated and alternatively phosphorylated are also known as 'yin-yang' sites[18].

The acceptor site for O-β-GlcNAc glycosylation does not display a definite consensus sequence, nor are there many annotated sites in public databases. However, the fuzzy motif is marked by the close proximity of Proline and Valine residues, a downstream tract of Serines and an absence of Leucine and Glutamine residues in the near vicinity (data not shown). A prediction method for this type of glycosylation on human proteins has been built and made available[c] as a web server (*in preparation*).

Out of approximately 5,500 human sequences from SWISS-PROT (rel. 38), over 4,600 had at least one predicted O-GlcNAc site. 1,535 of these proteins had at least one high scoring O-GlcNAc site prediction (with 3,154 high scoring Ser/Thr sites). A number of these were DNA-binding proteins and involved in transcriptional regulation. When ranked according to scores, a large fraction at the top of this list were found to be nuclear proteins (as annotated in SWISS-PROT). The O-GlcNAc transferase itself (P100 subunit) was found to have predicted O-GlcNAc sites.

To study if the O-β-GlcNAc modification was specific for certain types of proteins, we classified the potentially modified proteins into cellular role categories and subcellular locations. Figure 3 illustrates the spread of proteins with at least one high-scoring O-β-GlcNAc site, across different categories. Also shown in this figure is the spread of phosphorylated proteins (as predicted[21] by NetPhos[d]), 'Yin-yang' proteins, proteins with PEST regions[22] and proteins with O-β-GlcNAc (+++) sites which fall within PEST regions.

[c] http://www.cbs.dtu.dk/services/YinOYang/
[d] http://www.cbs.dtu.dk/services/NetPhos/

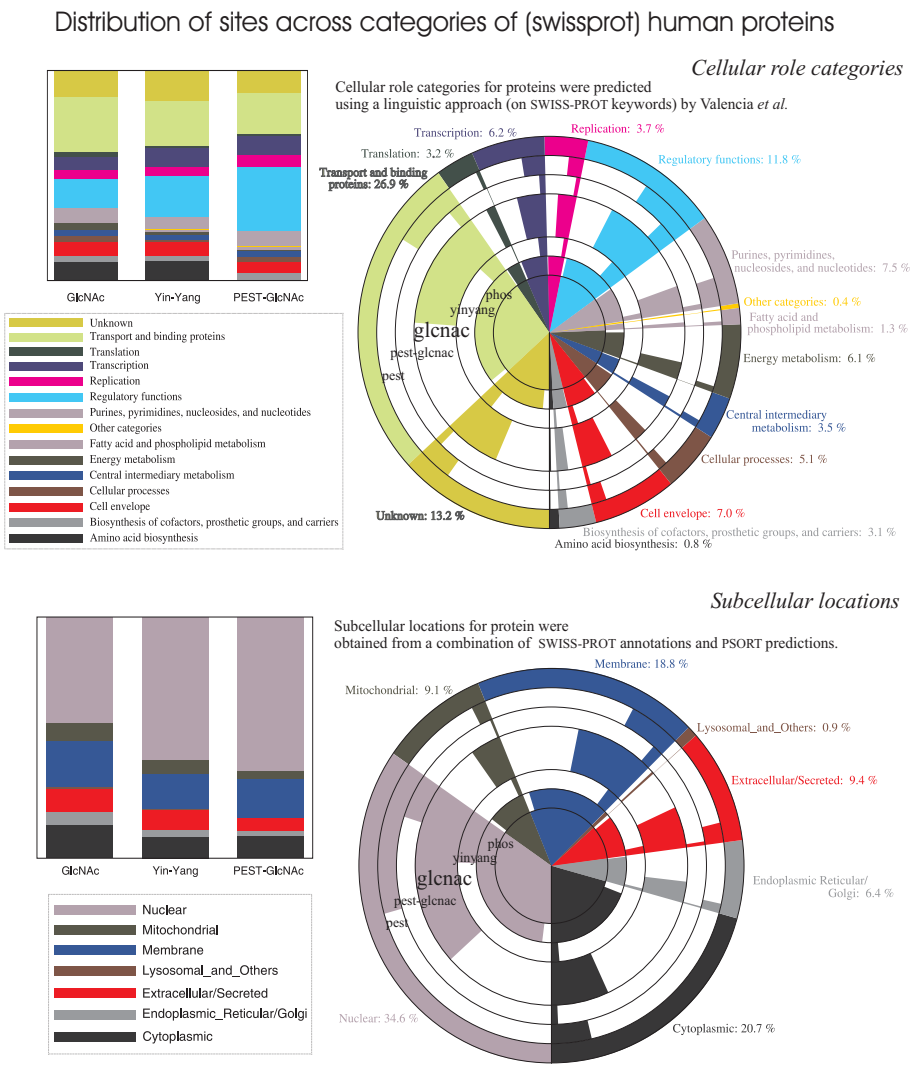

Figure 3: **Predicted O-β-GlcNAc sites across the human proteome.** The two panels (top, bottom) indicate different categorisations of proteins as depicted in the innermost and outmost circles of the pies. Individual rings represent different post-translational modifications and their occurrence in the corresponding category. E.g., phosphorylation occurs widely across all categories of proteins. Potential O-GlcNAc sites occur in half of all nuclear proteins and regulatory proteins. They also occur widely in replication and transcription proteins. Proteins with PEST regions and O-GlcNAc sites are mostly regulatory although PEST regions themselves also occur in other categories.

While the O-β-GlcNAc modification seems to potentially affect almost all types of proteins, most O-GlcNAcylated proteins were either regulatory proteins or 'transport and binding' proteins. A large fraction of unclassified proteins ('unknown' in role categories) were also predicted to contain this modification. Over half of all nuclear proteins contained a high ranking O-β-GlcNAc modified site. Cytoplasmic proteins, membrane proteins and secreted proteins also contained potential sites.

Phosphorylation is a very wide-spread modification [23]. This is reflected in our graphs as phosphorylation sites (> 0.9 potential by NetPhos) appeared well represented in all protein categories. However, Yin-yang sites appeared to exist largely in regulatory proteins, transcription related proteins or 'transport and binding proteins', and were mostly nuclear. O-GlcNAcylated PEST regions were also mostly nuclear, though a large membrane fraction also existed. Around half of all these proteins were involved in regulatory functions.

In an additional study, the number of potential O-β-GlcNAc sites in proteins was studied with respect to function and cellular location. Figure 4 illustrates the number of predicted (high-scoring) sites per 100 Ser/Thr residues (per protein). Proteins with 1-2 predicted GlcNAc sites (per 100 Ser/Thr) were predominantly nuclear, cytoplasmic or membrane proteins. Nuclear and cytoplasmic proteins carried the highest densities of sites, a few cytoplasmic proteins having as many as 50 high-scoring O-GlcNAc sites among 100 Ser/Thr residues. With respect to cellular roles, proteins belonging to the category 'Purines, pyrimidines, nucleosides and nucleotides' contained well spaced out sites (only a few sites among 100 Ser/Thr residues). Proteins with a wider distribution of sites included regulatory, transcription, replication, 'transport and binding', cell envelope and the 'unknown' category proteins. The highest density of sites (30-40 per 100 Ser/Thr) was found in transcription and regulatory proteins, though some 'unknown' proteins had over 40 sites (per 100 Ser/Thr). In general, the intracellular O-β-GlcNAc modification does not seem to cluster among close residues or display any characteristic spacing as was evident for the O-α-GlcNAc modification affecting surface and membrane proteins of *Dictyostelium discoideum* [24].

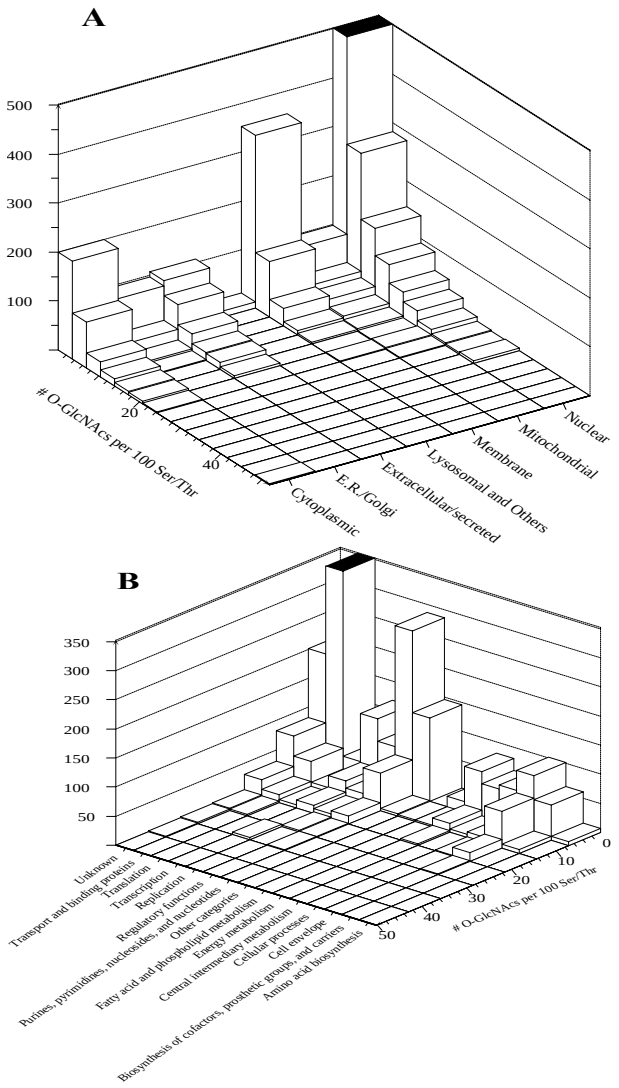

Figure 4: **Number of predicted O-β-GlcNAc sites per 100 Ser/Thr, in different categories of human proteins.** (A) shows proteins in different subcellular locations and (B) indicates cellular role categories. The z-scale (0-500 in A or 0-350 in B) is a frequency count for a particular bin; e.g. 0-2 O-GlcNAcs (per 100 Ser/Thr) occur most frequently for nuclear proteins in (A). These modifications usually do not occur in clusters. Although potential acceptor sites are largely found in nuclear/cytoplasmic proteins (usually regulatory), they also surprisingly occur in membrane proteins (mostly transport and binding proteins).

Human proteome-wide scans revealed that the O-β-GlcNAc acceptor pattern occurs across a wide range of functional categories and subcellular compartments. For humans, the most populated functional categories were regulatory proteins and transport and binding proteins. Nuclear and cytoplasmic proteins were prominent, though membrane and secreted proteins were surprisingly also in high numbers. It is interesting to know that acceptor patterns exist on these proteins too, but the cellular machinery defines protein targetting and consequently influences their modifications. The prediction server guards against this possibility by generating a warning when a potential signal peptide is detected by SignalP[e].

PEST regions, rich in the amino acids Proline (P), Glutamic acid (E), Serine (S) and Threonine (T), are hypothesised to be degradative signals for constitutive of conditional protein degradation[22]. Phosphorylation, a common mechanism to activate the PEST-mediated degradation pathway, may be signalled by deglycosylation in the same region. Our scans revealed that a small fraction of O-GlcNAc sites appeared in PEST regions. Such sites were mostly found in proteins involved in regulatory functions.

4 Final Remarks

Glycosylation is clearly a modification affecting a wide range of proteins, and is now known to affect both intracellular and secreted proteins. Different types of glycosylation have varying site preferences on proteins, and occur in different patterns across the protein chain.

In a project (*in preparation*) predicting protein function solely from protein chain global properties (molecular weight, length, etc.) and potential post-translational modifications, glycosylation was one of the most important determinants for functional classification.

Since characterising glycoproteins experimentally is a tedious and time-consuming task, it is worthwhile at this juncture to develop tools for predicting glycosylation sites. This is essential information for deciphering protein function and characterising complete proteomes.

5 Acknowledgements

The Danish National Research Foundation is acknowledged for support.

[e] http://www.cbs.dtu.dk/services/SignalP/

6 References

1. H Lis and N Sharon. Protein glycosylation: Structural and functional aspects. *Cur. J. Biochem.*, 218:1–27, 1993.
2. EF Hounsell, MJ Davies and DV Renouf. O-linked protein glycosylation structure and function. *Glycoconjugate J.*, 13:19–26, 1996.
3. MA Kukuruzinska, ML Bergh and BJ Jackson. Protein glycosylation in yeast. *Annu. Rev. Biochem.*, 56:915–944, 1987.
4. TR Gemmill and RB Trimble. Overview of N- and O-linked oligosaccharide structures found in various yeast species. *Biochim. Biophys. Acta*, 1426:227–237, 1999.
5. M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin and G Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25:25–29, 2000.
6. J Tamames, C Ouzounis, G Casari, C Sander and A Valencia. EUCLID: automatic classification of proteins in functional classes by their database annotations. *Bioinformatics*, 14:542–543, 1998.
7. C Blaschke, MA Andrade, C Ouzounis and A Valencia. Automatic extraction of biological information from scientific text: protein-protein interactions. In *Proc., Intelligent Systems for Molecular Biology*, pages 60–67, Menlo Park, CA, 1999. AAAI Press.
8. MA Andrade, C Ouzounis, C Sander, J Tamames and A Valencia. Functional classes in the three domains of life. *J. Mol. Evol.*, 49:551–557, 1999.
9. M Riley. Functions of the gene products of Escherichia coli. *Microbiol. Rev.*, 57:862–952, 1993.
10. K Nakai and P Horton. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, 24:34–36, 1999.
11. IM Nilsson and G von Heijne. Determination of the distance between the oligosaccharyltransferase active site and the endoplasmic reticulum membrane. *J. Biol. Chem.*, 268:5798–5801, 1993.
12. I Nilsson and G von Heijne. 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.*, 275:17338–17343, 2000.
13. R Apweiler, H Hermjakob and N Sharon. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database.

Biochim. Biophys. Acta., 1473:4–8, 1999.

14. J Roth, Y Wang, AE Eckhardt and RL Hill. Subcellular localization of the UDP-N-acetyl-D-galactosamine: polypeptide N-acetylgalactosaminyltransferase-mediated O-glycosylation reaction in the submaxillary gland. *Proc. Natl. Acad. Sci. USA*, 91:8935–8939, 1994.
15. JE Hansen, O Lund, J Engelbrecht, H Bohr, JO Nielsen, JES Hansen and S Brunak. Prediction of O-glycosylation of mammalian proteins: specificity patterns of UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferase. *Biochem. J.*, 308:801–813, 1995.
16. JE Hansen, O Lund, N Tolstrup, AA Gooley, KL Williams, and S Brunak. NetOglyc: Prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconjugate J.*, 15:115–130, 1998.
17. R Gupta, H Birch, K Rapacki, S Brunak, and JE Hansen. O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res.*, 27:370–372, 1999.
18. GW Hart, KD Greis, LY Dong, MA Blomberg, TY Chou, MS Jiang, EP Roquemore, DM Snow, LK Kreppel and RN Cole. O-linked N-acetylglucosamine: the 'yin-yang' of Ser/Thr phosphorylation? Nuclear and cytoplasmic glycosylation. *Adv. Exp. Med. Biol.*, 376:115–123, 1995.
19. DM Snow and GW Hart. Nuclear and Cytoplasmic Glycosylation. *Int. Rev. Cytol.*, 181:43–74, 1998.
20. FI Comer and GW Hart. O-Glycosylation of Nuclear and Cytosolic Proteins: Dynamic Interplay Between O-GlcNAc and O-Phosphate. *J. Biol. Chem.*, 275:29179–29182, 2000.
21. N Blom, S Gammeltoft, and S Brunak. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, 294:1351–1362, 1999.
22. M Rechsteiner and SW Rogers. PEST sequences and regulation by proteolysis. *Trends Biochem. Sci.*, 21:267–271, 1996.
23. EG Krebs. The growth of research on protein phosphorylation. *Trends Biochem. Sci.*, 19:439, 1994.
24. R Gupta, E Jung, AA Gooley, KL Williams, S Brunak, and J Hansen. Scanning the available *Dictyostelium discoideum* proteome for O-linked GlcNAc glycosylation sites using neural networks. *Glycobiology*, 9:1009–1022, 1999.