Chapter 1

The Structural Assessment of Glycosylation Sites Database - SAGS – An Overall View on N-Glycosylation

Marius D. Surleac, Laurentiu N. Spiridon, Robi Tacutu, Adina L Milac, Stefana M. Petrescu and Andrei-J Petrescu

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/51690

1. Introduction

The expansion of high-throughput technologies have led over the past decade to an unprecedented increase of the pace of data accumulation in biological science at molecular level. However this increase was not even - as while over 2000 genomes are already completed (Pagani *et al*, 2012) the vast majority of encoded proteins have not yet an assigned 3D structure and the functions of these proteins are still under investigation. While next-generation massive parallel sequencers are heading toward a 1 genome per day threshold, with significant impact on medical and environmental research (Sastre 2011; Shokralla *et al* 2012), the overall number of known protein structures remains well below 10^5 as their determination still relies on techniques such as crystallography and NMR which are time consuming, expensive and do not always work, despite the efforts made to automate the experimental flow in crystallization factories (Stuart *et al*, 2006) and the emergence of over 20 consortiums of Structural Genomics (Chandonia & Brenner, 2006) aiming at large scale 3D protein structure solving in Europe, United States and Japan.

Similarly in glycomics and glycoproteomics the gap between compositional and structural information has also increased. By combining techniques such as liquid chromatography, capillary electrophoresis and mass spectrometry significant advances were made in the last couple of years related to glycan profiling and the assessment of glycan heterogeneity and site occupancy (Liu *et al*, 2011; Mittermayr *et al*, 2011; North *et al*, 2010; Song *et al*, 2011; Zaia, 2010, Zauner *et al*, 2010). For example by pairing glycosylation site-specific stable isotope tagging of lectin affinity-captured N-linked glycopeptides with mass spectrometry a number of 1465 N-glycosylated sites were identified on 829 proteins expressed in Caenorhabditis elegans (Kaji *et al*, 2007). More recently, using a "filter aided sample

preparation" (FASP) method in which glycopeptides are enriched by binding to lectins on the top of a filter, it was possible to map using mass spectrometry 6367 N-glycosylation sites on 2352 proteins in four mouse tissues and blood plasma (Zielinska *et al* 2010). By contrast the pace of data accumulation in structural glycobiology remained very low as less than 4% of the structural files deposited in the protein data bank (PDB) contain oligosaccharide structures.

In this context bioinformatics became crucial in the endeavor of interpreting the raw experimental data, in depositing, ordering and annotating the massive amount of accumulated information and also, when combined with molecular modeling, in getting more structural insights were experimental data is still unavailable.

For example to assist the interpretation of MS data in Glycobiology tools with different aims and flavors such as GlycoWorkbench, SysBioWare or SimGlycan were lately developed. GlycoWorkbench resulted from the EUROCarbDB initiative and was designed to assist the manual interpretation of MS data, hence this aims to evaluate user proposed structures by matching their theoretical list of fragment masses against the list of peaks derived from the MS spectrum (Ceroni *et al*, 2008). On the other hand SysBioWare performs isotopic grouping of detected peaks after de-noising and wavelet analysis and allows compositional assignment according to the tuned building block library (Vakhrushev *et al*, 2009); while SimGlycan predicts the glycan structure using a MS/MS database searching technique, and also facilitates novel glycans prediction by drawing a glycan and mapping it onto an experimental spectrum to check the degree of proximity between the theoretical and the experimental glycans (Apte & Meitei, 2010). Similarly other software platforms were developed for modeling the oligosaccharide composition starting from HPLC data - such as autoGU (Campbel *et al*, 2008) or from NMR Spectra - such as CASPER (Loss *et al*, 2006) or CCPN (Stevens *et al*, 2011). For depositing, ordering and annotating data in Glycobiology, a large number of web portals and frameworks appeared in the last couple of years such as KEGG (Hashimoto *et al*, 2006), CFG (Raman *et al*, 2006), GLYCOSCIENCES.de (Lütteke *et al*, 2006), RINGS (Akune *et al*, 2010) or UniCarbKB (Campbell *et al*, 2011). Finally, related to glycan and glycoprotein modeling we would mention here the results of Rob Woods group (Woods & Tessier, 2010) and the tools such as SWEET-II or GlyProt developed at GLYCOSCIENCES.de.

Structural information on glycans and glycoproteins is scarce. Intrinsic flexibility make oligosaccharides to be seldom resolved by crystallography as either they simply do not crystallize or they generate local disorder in glycoprotein crystals that make them unidentifiable in the electron density. In addition, the presence of glycoforms and glycan conformational heterogeneity usually prevent glycoprotein crystallization (Chang *et al*, 2007). Consequently structural glycoinformatics get relatively low attention and only recently a number of resources and services were developed such as GlycoCD (Lütteke, 2006), SAGS (Petrescu *et al*, 2006) or Glycoconjugate Data Bank (Nakahara *et al*, 2008).

We will concentrate here on SAGS - the Structural Assessment of Glycosylation Site Database (web: http://sags.biochim.ro) that contains information on N-linked

oligosaccharides and structural properties of the protein core around N-glycosylation sites, and was used over the past decade to fulfill statistical analysis on structural aspects of protein glycosylation.

2. Redundant and non-redundant sets in SAGS for structural analysis

Historically SAGS was initiated around 1997 aiming to gather information related to the structure of N- and O-linked oligosaccharides starting from crystallographic data, in order to derive from this the general conformational rules governing the glycosidic linkages for use in glycoprotein model refinement (Wormald *et al*, 2002; Paduraru *et al*, 2006; Cioaca *et al*, 2011). Compared to NMR or theoretical methods, crystallography has the advantage that delivers a complete deterministic model of the oligosaccharide structure from the experimental data, yet this is static in nature and gives no clue on the flexibility of glycosidic linkages - as can be inferred from some NMR or theoretical models. However in ergodic conditions the statistical ensemble of the overall set of static structures found in protein data bank, PDB, is expected to be equivalent to the configurational space sampled during dynamics by any given glycan and thus it gives information on the flexibility of glycosidic linkages. This is why the overall nonredundant set of glycan structures found in PDB was used to gauge various glycosidic linkage configurations (Petrescu *et al*, 1999, Wormald *et al*, 2002).

Gradually the focus of the database has extended to cover also the assessment of structural properties of the protein core around N-glycosylation sites. These are critical in understanding the function of protein glycosylation in general, while in particular we were interested to see if from an overall view on glycosylation patterns some lessons could be retrieve related to site occupancy and GP folding which was our main interest at that stage. Whether N-glycans are used only to anchor glycoproteins to the quality control system during their folding and processing by the plethora of lectins, chaperones and enzymes (Petrescu *et al*, 1997, Zapun *et al*, 1997, Petrescu *et al*, 2000, Branza-Nichita *et al*, 2000), or do have a more direct effect onto the structure of the protein core is a fundamental problem in which a survey of the location of N-glycosylation sites in proteins is critical.

Nevertheless, investigating properties related to the (glyco)protein core implies a different approach from that used in glycosidic linkage evaluation. The overall set of glycoproteins found in PDB comprise in many cases multiple copies of the same protein or very close homologues crystallized or co-crystallized in various conditions, for instance there are no less than 50 copies of various variants of Hemagglutinin in PDB. While atomic level parameters such as the dihedral angles of various glycosidic linkages at a given site may vary drastically from crystal to crystal (Fig. 1) higher level properties such as the glycosylation site occupancy or accessibility, the sequence and secondary structure upstream and downstream a given site do not change so much and have to be counted only once in the statistics as otherwise multiple copies induce a bias into the statistics. To overcome this bias SAGS was provided with a system of clustering glycoproteins in general, according to their fold, but also the individual N-glycosylation sites.

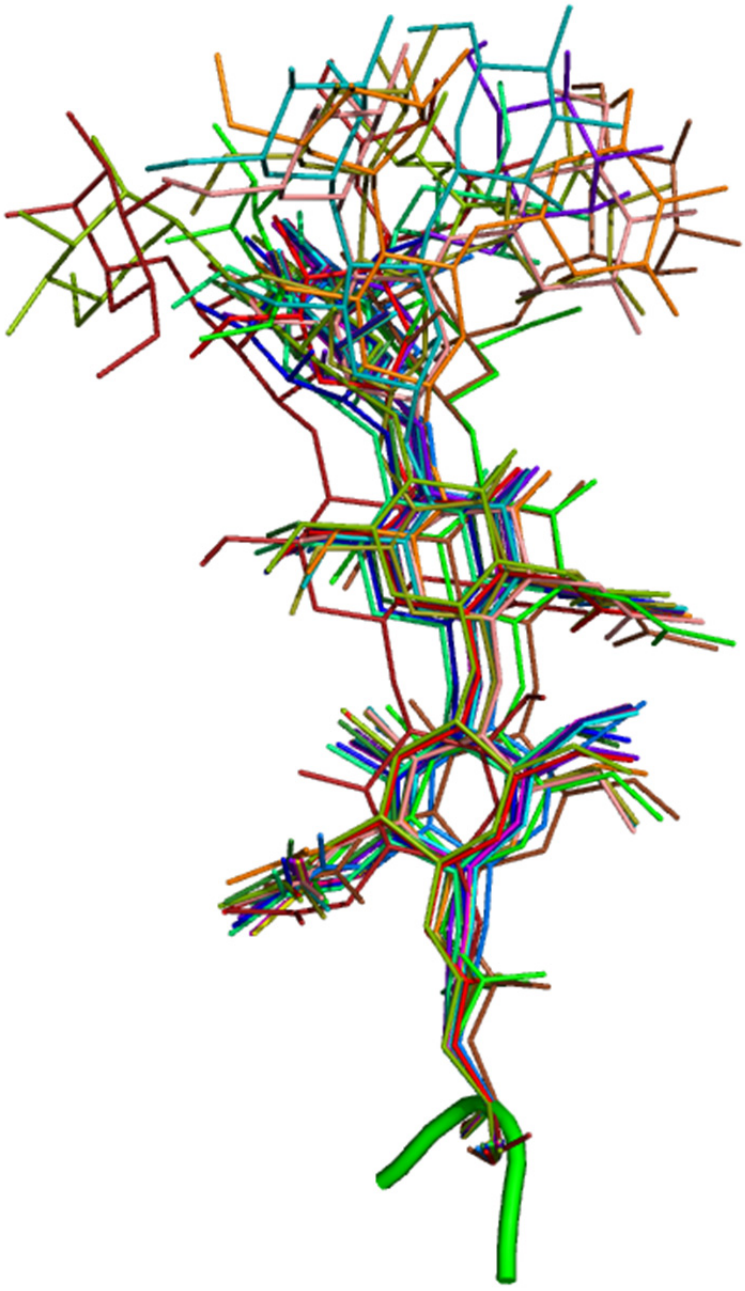


Figure 1. Structure variability of the glycan linked at N42 in all 19 crystal structures of the alpha subunit of the high affinity IG ε-receptor showing SAGS cluster ID: b.060.0040.ETT.h.12.

No simple N-site clustering system based entirely on either the sequence or the 3D structure was found to be adequate. For example very different sequences may adopt identical configurations, equally identical sequences may be found in significantly different configurations (Fig. 2). Tests have shown that an optimal separation between groups of similar N-sites is obtained by a two step procedure. First the sequence between -15 to +15 around a new N-site is compared to that of the existing clusters and in the case the identity is less than 60% the site is considered a new entry in SAGS. Secondly, if the identity is higher than 60% but lower than 90% a structural superposition onto the existing sites is further performed and again the site is retained as a new entry if the RMS to the rest of the sites is higher than 2Å, otherwise it is clustered to the closest group. Hence SAGS was automated to cluster groups of N-glycosylation sites that have over 60% sequence similarity and their structure is in less than 2Å from each other. Representatives for each cluster are further selected for representation and statistical analysis of higher scale properties based on non-redundant sets. Selection of the cluster representative relies on two criteria. First, sites with the highest number of glycosidic units are selected and secondly at equal glycan size the site in the crystal structure at the highest resolution is retained.

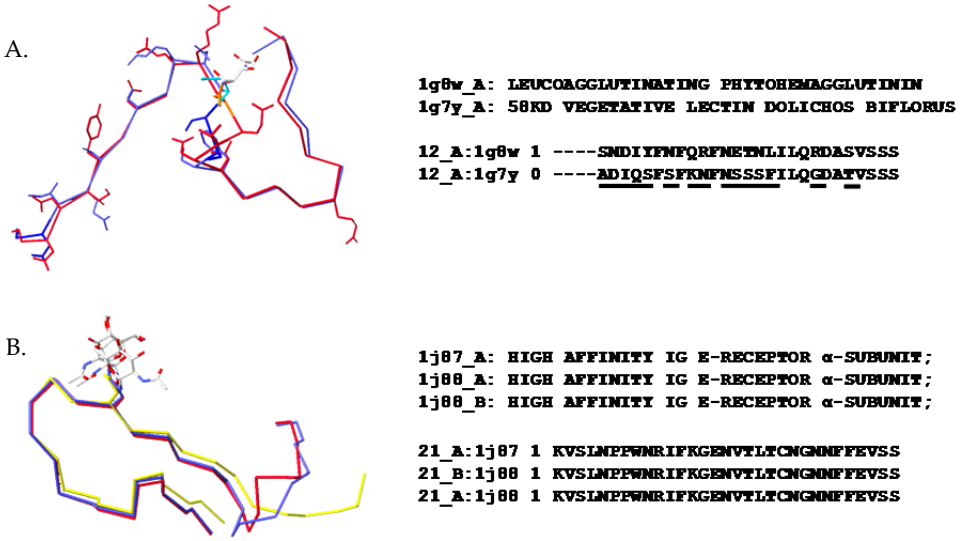


Figure 2. (A) Very different sequences can adopt almost identical structures.
12_A_1g8w - SAGS cluster ID: b.060.0120.CCG.h.53
12_A_1g7y - SAGS cluster ID: b.060.0120.CCS.m.55
(B) The same sequence can be found in different main chain configurations in various crystals.
SAGS cluster ID: b.060.0040.CCE.h.87

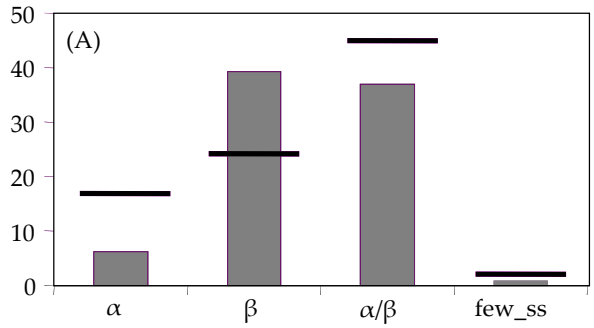
Each cluster is described by an identifier that captures the main structural properties of the group such that closely related groups to be close into the list. The identifier describes the class, architecture and topology of the glycosylated protein according the CATH taxonomy

(Greene *et al*, 2007), it then incorporates information on the secondary structure (ss) in positions N-1, N and N+1 based on DSSP (Kabsch & Sander, 1983) followed by a three state accessibility quantification: l, m, h (low, medium and high) and a counter. As CATH involves expert assessment of the protein fold and it is not automatically updated while SAGS is now fully automated we have had to introduce five states describing the class: a, b, c, d or x - corresponding to all alpha (a), all beta (b), alpha/beta (c), few secondary (d) and not yet assigned (x). An example is given in Table 1:

c.020.0020.HHH.m.36:							
class	architecture	topology	N-1_ss	N_ss	N+1_ss	accessibility	counter
alpha/beta	*barrel*	*Tim Barrel*	*Helix*	*Helix*	*Helix*	*medium*	*36*

Table 1. Example for the structural significance of an identifier in SAGS

Introduction of glycoprotein and N-site non-redundant sets with their identifiers was particularly instrumental not only in ordering the data, with consequences on a more accurate statistical assessment of various structural aspects of protein N-glycosylation, but also in generating and overall map of N-glycosylation onto the protein fold taxonomy. Even if ~40% of recently accumulated structures are yet to be assigned by CATH it is still striking to note that ~50% of all assigned glycoproteins in SAGS are of 'sandwich' type architecture. In particular, over 35% of all assigned structures are β-sandwiches which is two times more than their occurrence into the overall set of known protein structures present in CATH (Fig. 3). Therefore sandwich type architecture seems particularly fit to accommodate N-glycosylation either due to its prevalence in the secretome either due to its stability or other features favoring N-glycosylation. Further investigation is needed to address questions related to this striking prevalence.



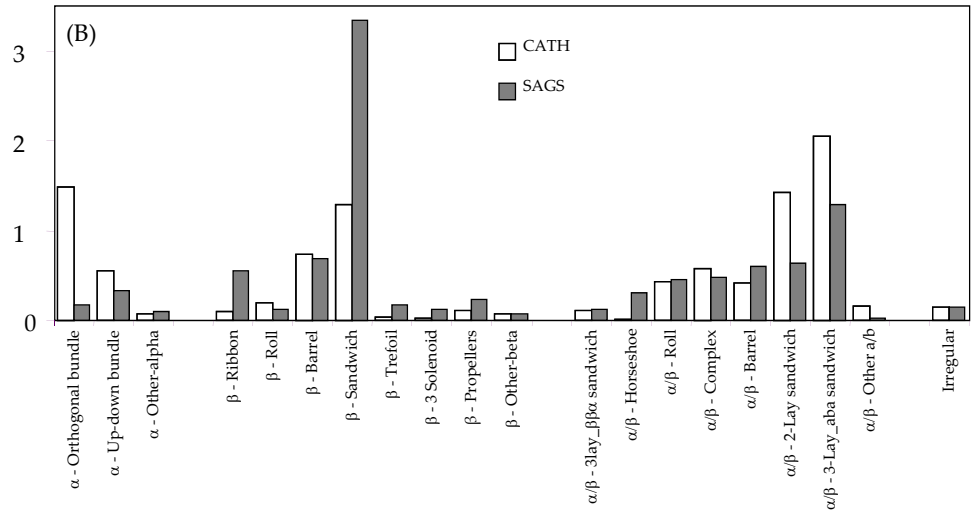


Figure 3. Occurrence of N-glycosylation into the four CATH classes (A) and more resolved architectures (B).

3. Evolution of SAGS and the robustness of the structural assessments over time

During the years SAGS increased steadily. For instance the first results on glycosidic linkage statistics rely on a set of only 639 representatives (Petrescu *et al* 1999); to date this set increased more than an order of magnitude to 9848.

In 2003, the analysis based on a non-redundant set of 386 entries revealed a number of patterns related to the properties of the sequence, secondary structure, shape and composition of the surface around occupied site with consequence on better understanding N-site occupancy and the relation between glycosylation and folding (Petrescu *et al* 2004). As the set was quite small it was important to assess how robust these findings would stand with data accumulation over the time. This was one of the main reasons that led to further developing SAGS and fully automate it for the foreseeable increase of structural calculations. Table 2 shows the evolution of N-glycosylation sites included in SAGS over the past decade.

	2003	**2005**	**2007**	**2008**	**2010**	**2011**	**2012**
PDB	23596	34140	47814	54778	70066	78157	82679
PDBs with N-glycans	711	1027	1436	1716	2006	2359	2605

	2003	2005	2007	2008	2010	2011	2012
% from PDB	*3.0*	*3.0*	*3.0*	*3.1*	*2.8*	*3.0*	*3.2*
glycosylated polypeptides	763	1443	2546	3010	3574	4349	4842
potential N-sites (sequons)	2219	5043	9157	11184	13895	16979	19164
occupied N-sites	1362	3013	5774	6985	8425	10142	11267
occupation [%]	*61*	*60*	*63*	*62*	*61*	*60*	*59*
non-redundant sequons	626	1055	1497	1825	2150	2580	2910
non-redundant N-sites	386	622	990	1184	1389	1666	1853
occupation [%]	*62*	*59*	*66*	*65*	*65*	*65*	*64*

Table 2. The evolution of the PDB data processed in SAGS over the past decade

As can be seen since 2003 the amount of data increased more than four fold. Interestingly, the main patterns proved surprisingly consistent over the time while new, more refined patterns shaped up as a result of the significant accumulation of data.

For example the ratio between NxT and NxS occupied sequons remained consistently 65% vs 35%. Similarly the occurrence of aromatic aminoacids in position N-2 and N-1 remained consistently in the range 15÷17% - namely 2-3 standard deviations (σ) over the expected percentage of ~10%; while the occurrence of acidic aminoacids in the same positions is >2σ lower than the 11% expected value, namely 5÷6%. Similar to the initial set the main deviations from the expected values remained located close to the N-site, in the region N-2 ÷ N+3. For the 2012 set these are shown in Fig. 4.

As the set of occupied sites increased over 4 fold more refined correlations are now possible to investigate. For instance when looking more closely by sequons types it is striking to see that in NxS sequons the occurrences of aromatic aminoacids raise to 19% in position N-2 (i.e. 4σ) and for acidic aminoacids these fall below 4% (i.e. 3σ) suggesting that the 'signature' for occupation was enhanced to compensate the presence of serine in position N+2. Similarly more subtle correlations between occurrence and accessibility shape up in occupied sites, e.g. in low accessibility sites aromatics are highly preferred in position N-1 (>3.6σ) while in highly accessible sites aromatics are far more frequent in position N-2 (>3.2σ).

Over the years the secondary structure statistics has also varied only marginally in SAGS, with a slight decrease of Turns, Bends and Coils in favor of β-structures that rise at glycosylation site (position N) from 20% to 27% (Fig. 5A). On the other hand these slight variations do not affect in any way the rate of change measured as the probability of having one type of structure in a given position followed by a different type for the next amino acid, suggesting that glycosylation remains strongly correlated with changes in secondary

structure that make N-sites frequently landmarks for starting or ending regular secondary structure landmarks (Fig. 5B).

Turning now to the properties of the protein surface around N-glycosylation sites some of them remained remarkably stable while new other emerged during the analysis of the new 5 fold larger set of nonredundant entries in SAGS. For example the proportion of N-sites deeply buried in the surface remained surprisingly high as in the old, reduced set. Figure 6A shows the change in accessibility to a water molecule probe (1.5Å) when the glycan is taken into account at occupied sites. Here again, over 15% of the N-sites see their Asn side-chain accessibility reduced to less than 5%, showing that the glycan's first NAcGlc unit acts as part of the protein surface and completely obstructs the access of water to the Asn side-chain. Also very robust proved to be the distribution of contacts between the first two NAcGlc glycan units and the aminoacids brought close in space by the folding process. The percentage of contacts with aromatic aminoacids increased slightly to more than 3 fold than expected by chance (Fig. 6B). On the other hand as the non-redundant set and number of documented contacts increased significantly during time new details emerged from their analysis. For instance there are now 450 N-sites, from 1853, in which were identified 592 very close glycan-protein contacts, laying within less than 3Å. Interestingly over 30% of such hydrogen-bond contacts are formed with acidic aminoacids brought close in space by the folding process, which is twice the percentage expected by chance.

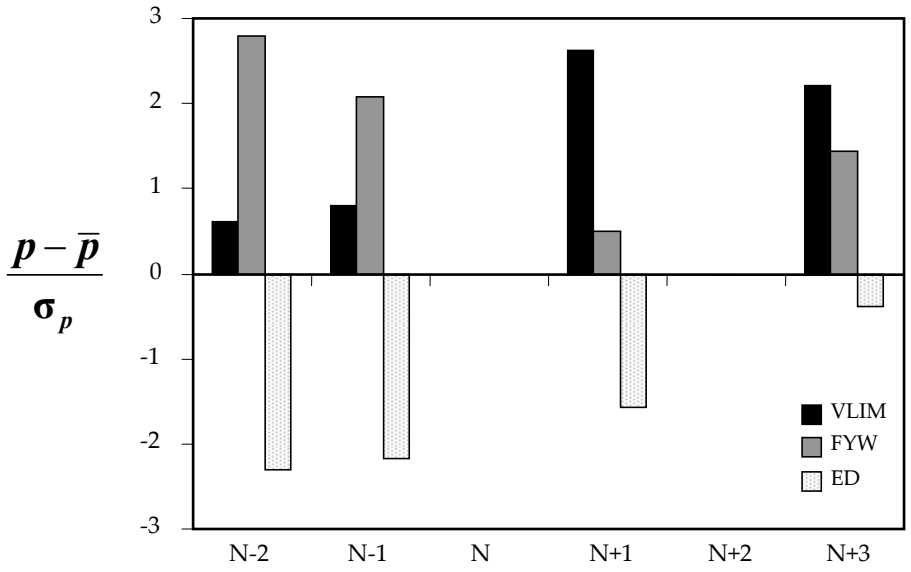


Figure 4. The main deviations from the expected occurrence around occupied N-glycosylation involve hydrophobic (VLIM), aromatic (F,Y,W) and acidic (E,D) aminoacids. These are shown here measured in standard deviations from the expected values.

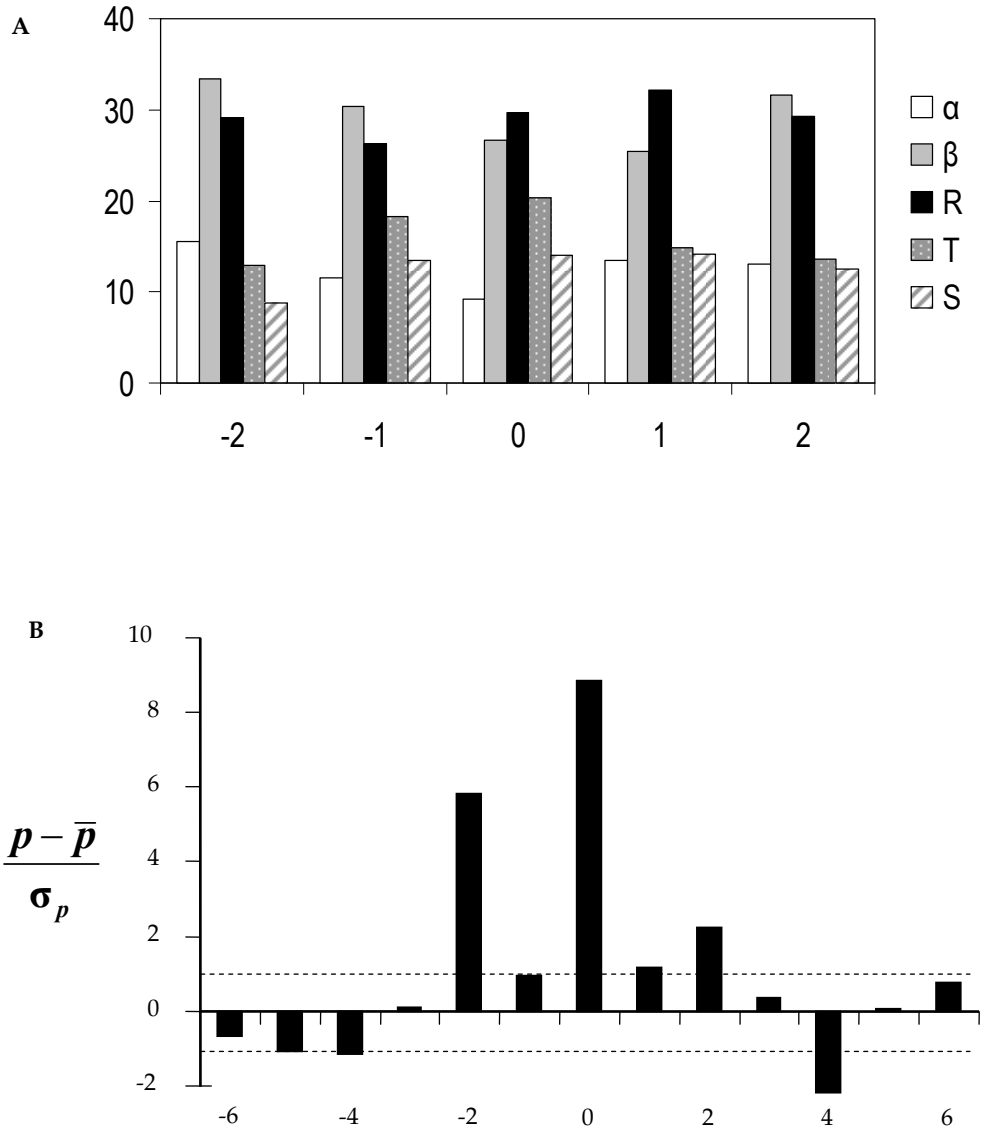


Figure 5. (A) Secondary structure distribution in positions N-2 to N+1.
(B) Rate of secondary structure change measured as the probability of having one type of structure in a given position followed by a different type for the next amino acid.

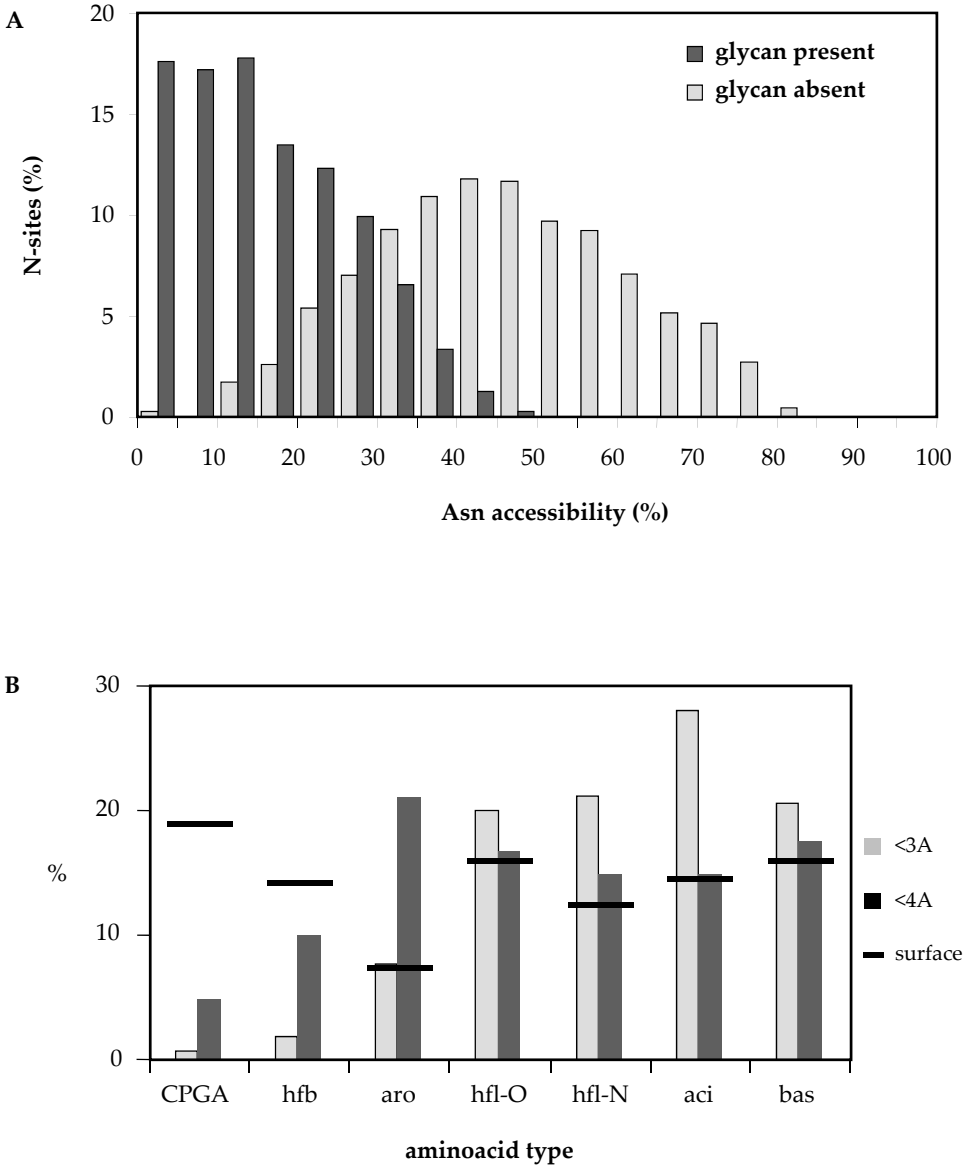


Figure 6. (A) Change of Asn side-chain accessibility over glycosylation.
(B) Percentage of glycan-protein contacts induced by folding.

The SAGS increase over time led also to a 10 fold increase of documented structures of glycans larger than the manosidic core. In the non-redundant set there are now available 29 complex glycan structures and 28 high mannose structures which makes now possible to investigate statistically not only the relation between N-glycosylation and folding but also if and how the structure of the protein core influence glycan processing into the Golgi.

These examples suggest that the analysis of ever increasing number of SAGS entries will soon make possible to develop more refined tools for glycoprotein modeling and predicting site occupancy based on the assessment of multiple sequence properties.

4. Significance and cross-validation of some results derived from SAGS analysis

Statistical analysis on SAGS suggests the existence of several sequence signatures that favor occupation. It is hence interesting to see if these correlate in any way with some different statistics performed on other data sets. Of particular importance is to use an as large as possible data set of sequences with known subcellular localization. On this line we found most useful LOCATE which curates all documented proteins sequences from mouse and human (Sprenger *et al*, 2008) and CYGD that comprise the documented localization in yeast (Güldener *et al*, 2005).

As cytosolic proteins are not subject to glycosylation, aminoacid occurrences is expected to be consistent with an independent probabilistic model while in the secretome or transmembrane proteins evolutionary pressures related to glycosylation, if any, are expected to influence the occurrences by inducing positive or negative correlations. Calculations are strikingly consistent with the statistical data from SAGS. Indeed in cytosolic proteins in all cases joint probabilities are within less than 5% from an independent probabilistic model. On the other hand within the secretome and membrane proteins significant correlations are shaping up. For example, in both secretome and TM proteins the occurrence of threonine in NxT sequences is 50% ÷ 60% higher than expected from independent probabilities in mouse and human respectively. Interestingly in yeast the deviation from independency is only 15%. Results on the occurrences of aromatic aminoacids in position N-2 and N-1 are even more striking. They reach an >80% increase in human secretome while in mouse and TM proteins the increase is >60%. In addition sequences of type Aro-N-x-S occur 70% (human) ÷ 80% (mouse) more frequently in the secretome than in the cytosol while for those of type Aro-N-x-T the difference in occurrence peaks even higher levels: 120% ÷ 150% in mouse and human respectively. In yeast these correlations still exists but again they are far less significant as they do not exceed 20-30%. Similar positive correlations shape up for hydrophobic bulky aminoacids (VLIM) in position N+1. These are on average 40% ÷ 60% more frequent than expected from independent probabilities in the secretome and 30%÷90% more frequent when compared to N(VLIM)S/T stretches in cytosolic proteins of mouse and humans. By contrast in yeast correlations are again dumped to less than 15%.

Hence an evolutionary pressure enhancing the occurrence of sequences known from SAGS to favoring occupation is obvious in the secretome and TM proteins. Besides this is far more significant in higher eukaryotes than in yeast.

N-site occupation is a direct consequence of the co- and post-translational interaction of the nascent polypeptide with the oligosaccharyltransferase (OST). Hence the yield of N-glycosylation at any given site depends on both the structure of OST and the sequence and local structure of the polypeptide at the time the transfer takes place.

In Eukaryotes OST is a multimeric complex located in the endoplasmic reticulum (ER), partly buried within the membrane of this compartment. Eukaryal OST transfers en-block a $Glc_3Man_9GlcNAc_2$ (G3M9) oligosaccharide from a dolichol pyrophosphate carrier to the side-chain amide nitrogen of only those asparagines located in sequons. The OST complex is composed from seven subunits of which two: the catalytic subunit and the thioredoxin-like subunit, come each in two isoforms: Stt3A/Stt3B and N33-Tusc3/IAP respectively. This results in four possible forms of the complex which are presumed to be optimized for co- vs. post-translational events and/or different types of substrates (Mohorko *et al*, 2011). In lower eukaryotes OST is a single-polypeptide membrane protein consisting alone in a Stt3 homologue (Maita *et al*, 2010) and similar monomeric Stt3 homologues performing OST functions were even identified in Bacteria and Archaea (Wacker *et al* 2002, Calo *et al*, 2010, Dell *et al*, 2010). In this diversity it is hence expected that peculiarities of each OST dictate the occupation rules, as indicated by the increasing number of studies reporting glycosylation at N-sites breaking the general accepted NxS/T sequon rule (Schwarz *et al*, 2011). Even in higher eukaryotes were the occupation control by OST is highly sophisticated deviations from the generally accepted NxS/T rules were recently reported (Schwarz & Aebi, 2011). Interestingly even the occurrence of N-glycosylation at an NPS site was recently identified in SAGS (ID: x.000.0000.ECC.m.77), in viral envelope GP160 (pdb code: 2qad-A362, Huang *et al*, 2007). This brings the absence of proline in position N+1 into the realm of statistical rare events, rather than a compulsory rule - as previously considered.

The mechanism by which OST transfers the G3M9 glycan to the asparagine is not yet fully understood but most likely this involves a nucleophilic attack to the amide nitrogen of the Asn side-chain. Historically there were two mechanisms proposed for this transfer both of which imply the involvement of β-OH group of the serine/threonine found in position N+2 of the sequon (NxS/T). However significant distance constraints have to be fulfilled in order that such an interaction between the amide of the Asn side-chain and the OH group of Ser/Thr to take place. These are satisfied only in a local geometry of the main chain consistent with β-turn or Asx turn configurations.

On this line SAGS provided a good opportunity to assess the proportion of sites in folded glycoproteins that actually fulfill the constraints imposed by the historic models of N-glycosylation. Both the distribution of N-sites in secondary structures and the distance measurements between the ND2 atom of Asn and the OG atom of Ser/Thr have shown that less than 20% of the N-sites found in SAGS are consistent with the two mechanisms. This indicates that either a different mechanism involving the nucleophilic attack provided by

OST is actually in place, or the polypeptide adopts transiently, at each site, a β-turn configuration before folding (Petrescu *et al*, 2006). Yet this later possibility is highly improbable in a process in which both the initial unfolded state and the final native state are locally far more extended than a β-turn, at over 80% of the N-sites.

It is only recently that the inference on the existence of a mechanism involving the intervention of OST in the nucleophilic attack was substantiated by the work performed at ETH in Zurich by the groups of Markus Aebi and Kaspar Locher (Lizak *et al*, 2011). From their seminal work on a bacterial homologue of OST, the PglB protein from *Campylobacter lari* - which was co-crystallized with an acceptor polypeptide - it results that the nucleophilic attack is facilitated by the aspartic acid D56 and glutamic acid E319 of Pg1B and not by the threonine in position N+2 of the glycosylated polypeptide. The crystal structure shows also, in agreement with what is an expected state along the folding pathway, that the polypeptide complexed to Pg1B is in fact in an extended configuration. The structure also indicates that the threonine/serine in position N+2 lays at the interface between the periplasmic and TM domain of Pg1B and that its side chain β-OH group is anchored by hydrogen bonds to the WWD motif 462-464 of Pg1B. In addition threonine (N+2) based complexes are further stabilized by hydrophobic interactions with I572 and V575 of Pg1B, as their side-chains are in contact with threonine gamma carbon (CG2) which in addition prevents free rotation of its side chain. Conversely these stabilization factors lack in serine (N+2) based complexes due to the absence of the gamma carbon. The additional stability induced by CG2 methyl group might explain the higher transfer yield to threonine as compared to serine sequons, shown by SAGS.

As seen many aspects of the statistics preformed on SAGS have a firm structural basis and the details of higher eukaryote OST structure, which are not yet known, will probably explain the other aspects of occupation signatures.

Author details

Marius D. Surleac, Laurentiu N. Spiridon, Robi Tacutu,
Adina L Milac, Stefana M. Petrescu and Andrei-J Petrescu*
Institute of Biochemistry of the Romanian Academy, Bucharest, Romania

Acknowledgement

This work was supported by the Romanian Academy Research Plan, CNCS grant PCE-ID-3-0342/2011 and the POSDRU/89/1.5/S/60746 Program.

5. References

Akune Y, Hosoda M, Kaiya S, Shinmachi D, Aoki-Kinoshita KF. *The RINGS resource for glycome informatics analysis and data mining on the Web*. *OMICS*.,14(4), 475-486 (2010)

* Corresponding Author

Apte A, Meitei NS. *Bioinformatics in glycomics: glycan characterization with mass spectrometric data using SimGlycan*. *Methods Mol Biol*. 600, 269-281 (2010)

Branza-Nichita N, Petrescu A-J, Negroiu G, Dwek RA, Petrescu SM, *N-glycosylation processing and glycoprotein folding-lessons from the tyrosinase- related proteins*, *Chem. Rev.*, 100, 4697-4711 (2000)

Calo D, Kaminski L, Eichler J, *Protein glycosylation in Archaea: sweet and extreme*, *Glycobiology*, 20(9), 1065–1076 (2010).

Campbell MP, Hayes CA, Struwe WB, Wilkins MR, Aoki-Kinoshita KF, Harvey DJ, Rudd PM, Kolarich D, Lisacek F, Karlsson NG, Packer NH. *UniCarbKB: putting the pieces together for glycomics research. Proteomics*. 11(21), 4117-21 (2011)

Campbell MP, Royle L, Radcliffe CM, Dwek RA, Rudd PM. *GlycoBase and autoGU: tools for HPLC-based glycan analysis*. *Bioinformatics*, 24, 1214–1216 (2008)

Ceroni A, Maass K, Geyer H, Geyer R, Dell A, Haslam SM. *GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. J Proteome Res*. 7(4):1650-1659 (2008)

Chandonia JM, Brenner SE. "*The impact of structural genomics: expectations and outcomes*", *Science* 311, 347-351 (2006)

Chang VT, Crispin M, Aricescu AR, Harvey DJ, Nettleship JE, Fennelly JA, Yu C, Boles KS, Evans EJ, Stuart DI, Dwek RA, Jones EY, Owens RJ, Davis SJ. *Glycoprotein structural genomics: solving the glycosylation problem*. Structure, 15(3), 267-273 (2007).

Cioaca D, Ghenea S, Spiridon LN, Marin M, Petrescu AJ, Petrescu SM. *C-terminus glycans with critical functional role in the maturation of secretory glycoproteins. PLoS One.* 6(5):e19979, 2011.

Dell A, Galadari A, Sastre F, Hitchen P. *Similarities and differences in the glycosylation mechanisms in prokaryotes and eukaryotes. Int J Microbiol*. 2010,148178 (2010)

Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA, *The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution*. *Nucleic Acids Research* 35, D291-D297 (2007).

Güldener U, Münsterkötter M, Kastenmüller G, Strack N, van Helden J, Lemer C, Richelles J, Wodak SJ, Garcia-Martinez J, Perez-Ortin JE, Michael H, Kaps A, Talla E, Dujon B, Andre B, Souciet JL, De Montigny J, Bon E, Gaillardin C, Mewes HW. *CYGD: the Comprehensive Yeast Genome Database*. *Nucleic Acids Res*, 33, D364-D368 (2005)

Hashimoto K, Goto S, Kawano S, Aoki-Kinoshita KF, Ueda N, Hamajima M, Kawasaki T, Kanehisa M. *KEGG as a glycome informatics resource. Glycobiology*. 16(5), 63R-70R (2006)

Huang CC, Lam SN, Acharya P, Tang M, Xiang SH, Hussan SS, Stanfield RL, Robinson J, Sodroski J, Wilson IA, Wyatt R, Bewley CA, Kwong PD. *Structures of the CCR5 N terminus and of a tyrosine-sulfated antibody with HIV-1 gp120 and CD4. Science*. 317(5846), 1930-1934 (2007)

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

Kaji H, Kamiie J, Kawakami H, Kido K, Yamauchi Y, Shinkawa T, Taoka M, Takahashi N, Isobe T. *Proteomics reveals N-linked glycoprotein diversity in Caenorhabditis elegans and suggests an atypical translocation mechanism for integral membrane proteins*. *Mol Cell Proteomics*. 6(12), 2100-2109 (2007)

Liu L, Telford JE, Knezevic A, Rudd PM. *High-throughput glycoanalytical technology for systems glycobiology*. *Biochem Soc Trans*. 38(5), 1374-1377 (2010).

Loss A, Stenutz R, Schwarzer E, von der Lieth CW. *GlyNest and CASPER: two independent approaches to estimate 1H and 13C NMR shifts of glycans available through a common web-interface*. *Nucleic Acids Res*. 34:W733-W737 (2006)

Lütteke T, Bohne-Lang A, Loss A, Goetz T, Frank M, von der Lieth CW. *GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research*. *Glycobiology*. 16(5):, 71R-81R (2006)

Lizak C, Gerber S, Numao S, Aebi M, Locher KP. *X-ray structure of a bacterial oligosaccharyl-transferase*. *Nature* 474(7351), 350-355 (2011)

Maita N, Nyirenda J, Igura M, Kamishikiryo J, Kohda D. *Comparative structural biology of eubacterial and archaeal oligosaccharyltransferases*. *J Biol Chem*. 285(7), 4941-4950 (2010)

Mittermayr S, Bones J, Doherty M, Guttman A, Rudd PM. *Multiplexed analytical glycomics: rapid and confident IgG N-glycan structural elucidation*. *J Proteome Res*. 10(8), 3820-3829 (2011).

Mohorko E, Glockshuber R, Aebi M. *Oligosaccharyltransferase: the central enzyme of N-linked protein glycosylation. J Inherit Metab Dis*. 34(4), 869-878 (2011)

Nakahara T, Hashimoto R, Nakagawa H, Monde K, Miura N, Nishimura S. *Glycoconjugate Data Bank: Structures--an annotated glycan structure database and N-glycan primary structure verification service*. *Nucleic Acids Res*. 36, D368-71 (2008).

North SJ, Hitchen PG, Haslam SM, Dell A. *Mass spectrometry in the analysis of N-linked and O-linked glycans*. *Curr Opin Struct Biol*. 19(5), 498-506 (2009).

Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. *The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata*. *Nucleic Acids Res*., 40, D571-9 (2012)

Paduraru C, Spiridon L, Yuan W, Bricard G, Valencia X, Porcelli S, Besra G, Petrescu SM, Petrescu A-J, Cresswell P. "*An N-linked glycan modulates the interaction between the CD1d heavy chain and beta 2-microglobulin*.", *J Biol Chem*., 281(52), 40369-78 (2006)

Petrescu A-J, Butters TD, Petrescu SM, Platt FM, Dwek RA, Wormald MR, *The solution NMR structure of Glc3Man9 unit in Glc3Man7GlcNAc2*, *Embo J*. 16, 4302-4310 (1997)

Petrescu A-J, Wormald MR, Dwek RA."*Structural aspects of glycomes with a focus on N-glycosylation and glycoprotein folding*.", *Curr Opin Struct Biol*. 16(5): 600-607 (2006)

Petrescu A-J, Milac A-L, Petrescu SM, Dwek RA, Wormald M.R."*Statistical analysis of the protein core around N-glycosylation sites. Implications on occupancy, folding and function*", *Glycobiology*, 14: 103-114 (2004)

Petrescu A-J, Petrescu S.M., Dwek R.A., Wormald M.R., "*A Statistical Analysis of N- and O-glycan linkages from crystallographyc data*" *Glycobiology*, 9, 343-352 (1999)

Petrescu SM, Branza-Nichita N, Negroiu G, Petrescu A-J, Dwek RA, *Tyrosinase and glycoprotein folding: roles of chaperones that recognize glycans*, *Biochemistry*, 39, 5229-5237 (2000)

Raman R, Venkataraman M, Ramakrishnan S, Lang W, Raguram S, Sasisekharan R *Advancing glycomics: Implementation strategies at the consortium for functional glycomics*. *Glycobiology* 16, 82R–90R (2006)

Sastre L., *New DNA sequencing technologies open a promising era for cancer research and treatment*. *Clin Transl Oncol*. 3(5):301-6 (2011).

Schwarz F, Aebi M. *Mechanisms and principles of N-linked protein glycosylation*. *Curr Opin Struct Biol*. 21(5), 576-82 (2011).

Schwarz F, Lizak C, Fan YY, Fleurkens S, Kowarik M, Aebi M. *Relaxed acceptor site specificity of bacterial oligosaccharyltransferase in vivo*. *Glycobiology*. 21(1), 45-54 (2011).

Shokralla S, Spall JL, Gibson JF, Hajibabaei M. Next-generation sequencing technologies for environmental DNA research. Mol Ecol. 21(8):1794-805 (2012).

Song W, Henquet MG, Mentink RA, van Dijk AJ, Cordewener JH, Bosch D, America AH, van der Krol AR. *N-glycoproteomics in plants: perspectives and challenges*. *J Proteomics*. 74(8), 1463-1474 (2011).

Sprenger J, Lynn Fink J, Karunaratne S, Hanson K, Hamilton NA, Teasdale RD, *LOCATE: a mammalian protein subcellular localization database*. *Nucleic Acids Res*. 36, D230-D233 (2008).

Stevens TJ, Fogh RH, Boucher W, Higman VA, Eisenmenger F, Bardiaux B, van Rossum BJ, Oschkinat H, Laue ED. *A software framework for analysing solid-state MAS NMR data*. *J Biomol NMR*. 51(4), 437-47 (2011)

Stuart DI, Jones EY, Wilson KS, Daenke S "*SPINE: Structural Proteomics IN Europe - the best of both worlds*", *Acta Cryst*, D62, 10 (2006).

Vakhrushev SY, Dadimov D, Peter-Katalinić *Software platform for high-throughput glycomics*. *J.Anal Chem*. 81(9), 3252-3260 (2009)

Wacker M, Linton D, Hitchen PG, Nita-Lazar M, Haslam SM, North SJ, Panico M, Morris HR, Dell A, Wren BW, Aebi M. *N-linked glycosylation in Campylobacter jejuni and its functional transfer into E. coli*. *Science*. 298(5599):1790-1793 (2002).

Woods RJ, Tessier MB. *Computational glycoscience: characterizing the spatial and temporal properties of glycans and glycan-protein complexes. Curr Opin Struct Biol*. 20(5), 575-83 (2010)

Wormald M, Petrescu A-J, Pao Y-L, Glythero A, Elliot T, Dwek RA, "*Conformational Studies of Oligosacharides and Glycopeptides: Complementarity of NMR, X-Ray Crystallogrphy and Molecular Modelling*", *Chem.Rev*., 102, 371-387 (2002)

Zaia J. *Mass spectrometry and glycomics*. *OMICS* 14(4), 401-418 (2010).

Zapun A, Petrescu SM, Rudd PM, Dwek RA, Thomas DY, Bergeron JJM, *Conformation - independent binding of monoglucosylated ribonuclease B to calnexin*, *Cell*, 88, 29-38 (1997)

Zauner G, Deelder AM, Wuhrer M. *Recent advances in hydrophilic interaction liquid chromatography (HILIC) for structural glycomics*. *Electrophoresis*. 32(24), 3456-3466. (2011).

Zielinska DF, Gnad F, Wisniewski JR, Mann M, *Precision Mapping of an In Vivo N-Glycoproteome Reveals Rigid Topological and Sequence Constraints*, *Cell* 141, 897–907 (2010).