BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 6(9)

Computational analysis reveals abundance of potential glycoproteins in Archaea, Bacteria and Eukarya

Sadia Zafar, Arshan Nasir, Habib Bokhari*

Department of Biosciences, COMSATS Institute of Information Technology, Park Road, ChakShahzad, Islamabad, Pakistan; Habib Bokhari - Email: habib@comsats.edu.pk; Phone: +92-300-5127684; Fax: 0092-051-4442805; *Corresponding author

Received July 06, 2011; Accepted July 14, 2011; Published July 19, 2011

Abstract:
Glycosylation is the most common type of post-translational modification (PTM) and is known to affect protein stability, folding and activity. Inactivity of enzymes mediating glycosylation can result in serious disorders including colon cancer and brain disorders. Out of five main types of glycosylation, N-linked glycosylation is most abundant and characterized by the addition of a sugar group to an Asparagine residue at the N-X-S/T motif. Enzyme mediating such transfer is known as oligosaccharyl transferase (OST). It has been hypothesized before that a significant number of proteins serve as glycoproteins. In this study, we used programming implementations of Python to statistically quantify the representation of glycoproteins by scanning all the available proteome sequence data at ExPASy server for the presence of glycoproteins and also the enzyme which plays critical role in glycosylation i.e. OST. Our results suggest that more than 50% of the proteins carry N-X-S/T motif i.e. they could be potential glycoproteins. Furthermore, approximately 28-36% (1/3) of proteins possesses signature motifs which are characteristic features of enzyme OST. Quantifying this bias individually reveals that both the number of proteins tagged with N-X-S/T motif and the average number of motifs per protein is significantly higher in case of eukaryotes when compared to prokaryotes. In the light of these results we conclude that there is a significant bias in the representation of glycoproteins in the proteomes of all species and is manifested substantially in eukaryotes and claim for glycosylation to be the most common and ubiquitous PTM in cells, especially in eukaryotes.

Keywords: Algorithm, ExPASy server, N-glycosylation, glycoprotein, glycosyltransferase.

Background:
Glycosylation is the most common type of post-translational modification (PTM) in proteomes and thought to be one factor contributing in enhancing the diversity of proteomes **[1]**. There are five main types of glycosylation: O-linked, N-linked, C-linked, P-linked and G-linked **[2]**. N-linked and O-linked glycosylation are the most abundant types of glycosylation **[3]**. O-linked glycosylation is characterized by the attachment of carbohydrate units to the hydroxyl (OH) group of Serine (S), Threonine (T), Tyrosine (Y), Hydroxyproline, or Hydroxylysine side chains **[4]**. In contrast, N-linked glycosylation generally means the attachment of carbohydrate units to the nitrogen group in Asp-Xaa-Ser/Thr (N-X-S/T) motif, where Xaa is any amino acid except for Proline **[5]**. Species of three domains of life possess unique biological characteristics that provides basis for discrimination among them, but some of the biological characteristics are common in them as well **[6]**. One of the common biological function is glycosylation which is known to help prokaryotes invade host cells and induce pathogenecity **[7]**. Glycosylation is a ubiquitous form of PTMs necessary for most of cellular organisms **[8]**. It is also know to affect the stability of proteins and some proteins need to be glycosylated in order to fold properly **[9]**. In the absence of glycosylation, immature proteins do not fold properly hence shows that it is an essential co-translational event for correct folding of proteins **[10]**. Furthermore, cells of immune system also employ glycosylation strategies for cell adhesion purposes **[11]**. Problems with glycosylation mechanism can result in serious disorders including colon cancer and brain diseases. Thus understanding glycosylation is central to our understanding **[12]**. In case of N-linked glycosylation, the N of N-X-S/T motif serves as the acceptor site for the addition of glycan chains. Xaa position can have any amino acid, but Proline, but it has also been shown that the presence of negatively charged residues at Xaa result in partial glycosylation and positively charged residues are favorable **[13]**. It is also an experimental fact that due to several structural constraints only 66% of the NXS/T motifs are glycosylated **[14]**. Enzyme responsible for charging the proteins with sugar groups is generally known as *oligosaccharyltransferase* (OST) in eukaryotes. Homologs of OST in bacteria are known as PglB whereas Archaeal enzymes are referred to as AglB. The C-terminal of these enzymes has a signature motif of 5 residues with a sequence of Trp-Trp-Asp-[Tyr-Asn-Phe-Trp]-Gly (WWD[YNFW]G) that is of central importance for the activity of OSTs. Any mutation in this motif results in deactivation of OSTs and consequently loss of glycosylation activity **[15]**. Other important motifs that have been identified in OSTs include Met-Xaa-Xaa-Ile/Val/Met(MxxI/V/M) and Asp-Xaa-Xaa-Lys (Dxxk) motifs **[16]**. MxxI motif is found to be involved in enhancing the functional activity of OSTs. DK motif in eukaryotes, especially yeast, revealed certain importance in the survival of the organism as it involves the metal ion binding to the OSTs **[17]**. In 1999, Rolf Apweiler analyzed protein sequence data and proposed that approximately half the

proteins in a proteome are glycoproteins **[12]**. Presence of such a high number of potential glycoproteins is intriguing and points towards the importance of glycosylation for cells. In this study, we used programming implementations in Python to quantify the distribution of glycoproteins in the proteomes of all cellular organisms. We also explored the number of possible OST enzymes present in proteomes by searching for its signature motifs in target proteins. Our results support the findings Apweiler made 12 years ago and show that a significant proportion of all the cellular proteomes is devoted for the all-important function of glycosylation. Eukaryotes carry the most number of potential glycoproteins followed by Bacteria and Archaea. This tendency should not surprise us given the nature of rich protein repertoires of eukaryotes. The current study, based on the calculation of total number of glycoproteins, would be helpful in understanding the functional aspects of proteomes that have remained conserved in all the three domains of life and reveals exciting patterns in proteomes.

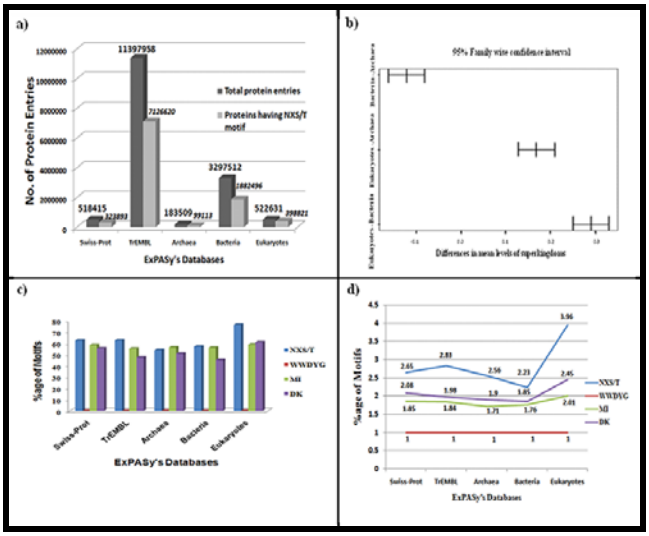

Figure 1: (a) Total Protein Entries in all Five ExPASy Databases, **(b)** Differences in mean levels of treatment for all three datasets using Tukey's HSD test, **(c)** Percentage of four defined motifs in ExPASy's databases, **(d)** Average number of defined motifs per protein in ExPASy's databases.

Methodology:

Proteome Data Collection:

The data of genome-encoded proteins till 13th July, 2010 was directly downloaded from the ExPASy databases namely Swiss-Prot (518415 total protein entries), TrEMBL (11397958), Eukaryotes (522631), Archaea (183509) and Bacteria (3297512).

A Computational Program: Motif Percentage Calculations:

In order to calculate statistics related to the characterization of glycoproteins and OSTs in proteomes a Python script was developed that first combines all the proteomes belonging to a superkingdom into a single file. For instance, the proteome data of all eukaryotes is represented by separate FASTA files and the first step in the execution of our script is to pool all the proteomes together and generate separate files for Archaea, Bacteria, and Eukarya. This step is not necessary when handling global statistics for Swiss-Prot and TrEMBL databases. Next step is to scan all the proteins for the presence of motifs linked to N-glycosylation. Specifically, the output reports the following statistics: total occurrences of N-X-S/T, MxxI, DxxK, and WWD[YNFW]G motifs; the number of proteins that are hit with these motif; number of times NXS\T, MxxI, DxxK and WWD[YNFW]G motif are present per protein; and, co-occurrences of pairs of motifs i.e. NXS/T + MI + WWD[YNFW]G, NXS/T + WWD[YNFW]G + DK, and NXS/T + WWD[YNFW]G + MI + DK motifs.

Results and Discussion:

Results compiled on 13/07/2010 suggest that 7126620 out of 11397958 protein entries in TrEMBL (62.5%), 323893 out of 518415 in Swiss-Prot (62.47%), 99113 out of 183509 in Archaea (54%), 1882496 out of 3297512 in bacteria (57%) and 398821 out of 522631 proteins in eukaryotes (76%) possess N-X-S/T motif i.e. these can potentially be glycoproteins **(Figure 1A)**. Eukaryotes appear to have the highest number of potential glycoproteins followed by Bacteria and then Archaea. In order to find whether the observed differences between the three superkingdoms (Archaea, Bacteria, and Eukarya) are statistically significant or not we conducted Analysis of Variance for Completely Randomized Design (ANOVA CRD) on a sample of 1,000 proteins each from three domains of life. The selected sample was parsed using Python and counts for number of times N-X-S/T motif is present per protein was calculated. The selected sample of 1,000 proteins each from the three superkingdoms had all their lengths in range of 1000-1500 amino acids. The computed statistics from ANOVA CRD are shown in **(Table 1 see Supplementary material)** The computed *p*-value for the differences between the three super kingdoms at 95% confidence interval using R was <2.2e-16 depicting that the results are significant at P<0.0002. Therefore, null hypothesis (means of count of NXS/T motif per protein for all superkingdoms are equal) is rejected. Tukey's Honesty Significant Differences (HSD) test was used to detect the pair-wise differences between all of the three superkingdoms of sample data. It appears that all the three kingdoms differ significantly to each other at 95% confidence intervals as shown in **Figure 1B**. The interval is not spanning zero hence implies that superkingdoms differ significantly. Variance between the eukaryotes and bacteria pair is more than the rest of the two. MI motif follows the similar pattern i.e. eukaryotes represents the higher number. However, the numbers for DK motif deviate from the general pattern a little. Eukaryotes have still the highest percentage for DK motifs (60.89%), followed by Archaea (50.88%) and then Bacteria (47.33%) as shown in **(Table 2 see Supplementary material)** Global analysis of Swiss-Prot and TrEMBL revealed that nearly 50.88% of the proteins carry DK motif for Swiss-Prot whereas 45.25% of proteins are tagged with DK motif in TrEMBL **(Figure 1C)**. **Figure 1D** reports the average number of times a motif is present in an individual protein. The mean values are reported for all the 4 motifs. On average more than one site of NXS/T sequon was recorded in each of the potential acceptor protein in all three domains. According to our results tendency of eukaryotes to produce glycans is higher, ~4, than the rest of the other two domains of life i.e., archaea and bacteria, 2.56 and 2.23 respectively. However, one to two sites of MxxI and DxxK motifs could be present per protein, eukaryotes being the highest one. It can be observed that eukaryotes possess higher content of glycoproteins. It has also been cleared from the literature study that highest percentage of glycosylated proteins is present in eukaryotes as compared to prokaryotes **[18]**. Hence implying that glycosylation is being an important mechanism for the healthy survival of eukaryotic organisms. Furthermore, co-occurrence of the motif pairs ((WWD[YNFW]G+DK), (WWD[YNFW]G+MI) and (MI+DK)) was calculated in order to find out the total number and percentage of proteins that can potentially act as OSTs. According to the results, the highest percentages for the WWD[YNFW]G+DK (0.056%) and WWD[YNFW]G+MI (0.057%) motif pairs were observed in case of archaea database suggesting that these proteins can be potential OSTs. Whereas, it was observed that percentage of DK+MI motif co-occurrence pair was relatively higher than that of the WWD[YNFW]G+DK and WWDYG+MI pairs in all the five databases, especially in case of eukaryotes i.e. 40.44% (211381/522631) **(Table 3 see Supplementary material)**. But the proteins possessing MI+DK motif pairs cannot be declared as OSTs as they lack the catalytic motif WWD[YNFW]G. Therefore, overall 28-36% of proteins can act as potential OSTs. There is a chance that these potential OSTs may possess NXS/T sequon i.e. self glycosylation site as well. Signature motifs in catalytic center help differentiating such OSTs from those of the glycoproteins. In order to filter out the number of OSTs, that possess catalytic center motifs and additionally the self glycosylation site (NXS/T), three different pairs for the co-occurrences of catalytic center motifs also containing NXS/T motif were made. WWD[YNFW]G+MI+DK+NXS/T was the first pair whose ratio was calculated. It was shown that archaeal proteomes have higher percentage i.e. 0.052314 % for co-occurrence of all of these four motifs **(Table 4)**. Percentage calculation of DK motif in **(Table 2 see Supplementary material)** and the literature study have shown that it might be one of the components of the catalytic center of OSTs but MI motif have higher chance of occurrence. So two sub pairs of the WWD[YNFW]G+MI+DK+NXS/T signature co-occurrence pair were made. One was WWD[YNFW]G+NXS/T+MI (DK motif eliminated) and the other was WWD[YNFW]G+NXS/T+DK (MI motif eliminated) co-occurrence sub pair. Total number and percentage calculation of both of these sub pairs have shown more or less similar results except in case of bacteria whose difference was little bit higher than the rest of the four databases. An interesting observation was made from the computed result that the percentage occurrence of WWD[YNFW]G+MI as shown in **Table 3 (see Supplementary material)** is similar to the percentage of WWD[YNFW]G+MI+NXS/T co-occurrence pair therefore implying that all of the potential OSTs with WWD[YNFW]G+MI catalytic signature motifs

patterns also possess self glycosylation site (NXS/T) as shown in **(Table 4 see Supplementary material)**.

Conclusion:

This study highlights the importance of glycosylation for the cellular life. It appears that potentially half of the proteome is devoted to perform functions related to glycosylation. Our results highlight the bias in the representation of glycoproteins in all the proteomes. Eukaryotes carry most number of potential glycoproteins and greater than 70% of the eukaryotic proteomes are proposed to be glycosylated, followed by Bacteria and then Archaea. This tendency is not surprising as eukaryotes have the most advanced and rich protein repertoires compared to the prokaryotes. Such high representation advocates for the importance of glycosylation as the crucial most post-translational modification that is central to cellular life in the form we see it today. In addition, prokaryotic proteomes are also enriched with potential glycoproteins. Finally, the strategy based on calculations of co-occurrences of signature motifs in OSTs is good for fishing out potential OSTs in available proteomes.

References:

[1] Sparbier K *et al*. *J Biomol Tech.* 2005 **16**: 407 [PMID: 16522863]
[2] Pandhal J & Wright PC. *Biotechnol lett.* 2010 **32**: 1189 [PMID: 20449632]
[3] Dube DH *et al*. *Proc Natl Acad Sci U S A.* 2006 **103**: 4819 [PMID: 16549800]
[4] Werner RG *et al*. *Acta Paediatr Suppl.* 2007 **96**: 17 [PMID: 17391433]
[5] Shelikoff M *et al*. *Biotechnol Bioeng.* 1996 **50**: 73 [PMID: 18626901]
[6] Calo D *et al*. *Glycobiology* 2010 **20:** 1065 [PMID: 20371512]
[7] Zhou M & Wu H. *Microbiology* 2009 **155**: 317 [PMID: 19202081]
[8] Hitchen PG & Dell A. *Microbiology* 2006 **152**: 1575 [PMID: 16735721]
[9] Shental-Bechor D & Levy Y. *Proc Natl Acad Sci U S A.* 2008 **105**: 8256 [PMID: 18550810]
[10] Imperiali B & O'Connor SE. *Pure and Applied chemistry* 1998 **70**: 33
[11] Yang L *et al*. *Mol Cell Proteomics.* 2011 **10**: M110.007294 [PMID: 21447706]
[12] Apweiler R *et al*. *Biochim Biophys Acta.* 1999 **1473**: 4 [PMID: 10580125]
[13] Yan B *et al*. *J Protein Chem.* 1999 **18**: 511 [PMID: 10524769]
[14] Pandhal J & Wright PC. *Biotechnol lett.* 2010 **32**: 1189 [PMID: 20449632]
[15] Weerapana E & Imperiali B. *Glycobiology* 2006 **16**: 91R [PMID: 16510493]
[16] Maita N *et al*. *J Biol Chem*. 2010 **285**: 4941 [PMID: 20007322]
[17] Igura M & Kohda D. *J Biol Chem.* 2011 **286**: 13255 [PMID: 21357684]
[18] Dell A *et al*. *Int J Microbiol*. 2010 **2010**: 148178 [PMID: 21490701]

Edited by P Kangueane
Citation: Zafar *et al*. Bioinformation 6(9): 352-355 (2011)

BIOINFORMATION

open access

Supplementary material:

Table 1: Analysis of Variance (ANOVA) on a sample dataset of 1,000 proteins each from the three super kingdoms

Source	Degree of Freedom (df)	Sum of Squares (SS)	Mean Squares (MS)	F-ratio	Computed P-value
Kingdoms	2	42.59	21.2950	141.73	<2.2e-16
Residuals	2997	450.30	0.1502		
Total	2999	492.89	0.1643		

Table 2: Statistics for all of the five databases at ExPASy: (a) Total number of proteins. (b) Total number of proteins with N-X-S/T, M-xx-I/V/M, D-xx-K and WWD[YNFW]G sequins. (c) Total occurrence of all these four motifs. (d) Individual percentage of proteins with all these four sequins. (e) Time these motifs present per protein.

ExPASy's Databases	Total Entries of Proteins	Motifs	Proteins with Motifs	Total Occurrence of Motifs	Percentage of Proteins with Motifs (%)	Average Number of Motifs per Protein
		NXS/T	323893	860953	62.47	2.65
Swiss-Prot	518415	MXXI	302955	560933	58.43	1.85
		DXXK	287322	598613	55.42	2.08
		WWD[YNFW]G	72	72	0.013	1.0
		NXS/T	7126620	20189746	62.52	2.83
TrEMBL	11397958	MXXI	6300288	11624974	55.27	1.84
		DXXK	5394755	10684046	47.33	1.98
		WWD[YNFW]G	1095	1101	0.009	1.0
		NXS/T	99113	254589	54.00	2.56
Archaea	183509	MXXI	103388	177118	56.33	1.71
		DXXK	93373	178042	50.88	1.90
		WWD[YNFW]G	111	111	0.06	1.0
		NXS/T	1882496	4199802	57.08	2.23
Bacteria	3297512	MXXI	1848201	3257755	56.04	1.76
		DXXK	1492167	2772787	45.25	1.85
		WWD[YNFW]G	96	96	0.002	1.0
		NXS/T	398821	1582393	76.31	3.96
Eukaryotes	522631	MXXI	307572	621097	58.85	2.01
		DXXK	318277	781550	60.89	2.45
		WWD[YNFW]G	93	93	0.017	1.0

Table 3: Co-occurrences of WWD[YNFW]G+DK, WWD[YNFW]G+MI and MI+DK motifs in all five databases of ExPASy

ExPASy's Databases	Total sequences with WWD[YNFW]G + DK motifs/ Total protein entries in database	Total sequences with WWD[YNFW]G + MI motifs/ Total protein entries in database	Total sequences with MI + DK motifs/ Total protein entries in database
Archaea	103/183509$_{(0.056\%)}$	106/183509$_{(0.057\%)}$	59573/183509$_{(32.46\%)}$
Eukaryotes	81/522631$_{(0.015\%)}$	70/522631$_{(0.013\%)}$	211381/522631$_{(40.44\%)}$
Bacteria	70/3297512$_{(0.002\%)}$	80/3297512$_{(0.002\%)}$	947589/3297512$_{(28.73\%)}$
Swiss-Prot	64/518415$_{(0.012\%)}$	44/518415$_{(0.008\%)}$	189322/518415$_{(36.51\%)}$
TrEMBL	956/11397958$_{(0.008\%)}$	775/11397958$_{(0.006\%)}$	3447117/11397958$_{(30.24\%)}$

Table 4: Self glycosylation sites (NXS/T) in potential OSTs in all five ExPASy's databases by the mid of 2010

Database	Swiss-Prot	TrEMBL	Archaea	Bacteria	Eukaryotes
Entries containing NXS/T+ MxxI/V/M+ WWD[YNFW]G+ DxxK sequins	28 (0.005401 %)	637 (0.005589 %)	96 (0.052314 %)	56 (0.001698 %)	61 (0.011672 %)
Entries containing NXS/T+ DxxK+ WWD[YNFW]G	42 (0.008102 %)	799 (0.00701 %)	101 (0.055038 %)	60 (0.00182 %)	72 (0.013776 %)
Entries containing NXS/T+ MxxI+ WWD[YNFW]G	44 (0.008487 %)	775 (0.006799 %)	106 (0.056673 %)	80 (0.002426 %)	70 (0.013394 %)