370–372 *Nucleic Acids Research, 1999, Vol. 27, No. 1*

O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins

Ramneek Gupta*, Hanne Birch, Kristoffer Rapacki, Søren Brunak and Jan E. Hansen

Center for Biological Sequence Analysis, The Technical University of Denmark, DK-2800 Lyngby, Denmark

Received October 13, 1998; Revised October 16, 1998; Accepted October 22, 1998

ABSTRACT

O-GLYCBASE is a database of glycoproteins with O-linked glycosylation sites. Entries with at least one experimentally verified O-glycosylation site have been compiled from protein sequence databases and literature. Each entry contains information about the glycan involved, the species, sequence, a literature reference and http-linked cross-references to other databases. Version 4.0 contains 179 protein entries, an approximate 15% increase over the last version. Sequence logos representing the acceptor specificity patterns for GalNAc, GlcNAc, mannosyl and xylosyl transferases are shown. The O-GLYCBASE database is available through the WWW at http://www.cbs.dtu.dk/databases/OGLYCBASE/

INTRODUCTION

Glycosylation is an important post-translational modification of proteins and affects more than half of all the proteins in a cell (1). O-linked glycosylation refers to the event in which a carbohydrate is covalently linked to the hydroxyl group of serine or threonine, and is one of the three main types of post-translational modifications involving carbohydrates. This modification influences a number of properties of proteins (2,3) including proteolytic resistance, solubility, immunological properties, ligand binding and is also involved in recognition during sperm-egg binding. Glycosylation is important for the biotechnology industry (4,5) since altered glycosylation has been shown to impair protein function (6–8).

Mapping the exact sites of modification is important, both in understanding the glycosylation process per se and in inferring interactions with other processes such as phosphorylation (9). Additionally, it is important while designing constructs for recombinant glycoproteins to ensure that acceptor sites are placed within the correct sequence motif. While current experimental methods for mapping glycosylation sites (such as Edman degradation and mass spectrometry) are cumbersome, they are presently the only precise way of accurate determination. However, with the large amount of experimental data being made available, this data can be employed for constructing prediction methods for acceptor specificities. Thus, an added-value dataset of known and verified O-glycosylation sites would facilitate this research.

Table 1.

O-linked sugar	Abbreviation	Number of sites in O-GlycBase 4.0
N-Acetylgalactosamine	GalNAc	680 (194 Ser, 486 Thr)
N-Acetylglucosamine	GlcNAc	88
Mannose	Man	158
Xylose	Xyl	16
Glucose	Glc	11
Fucose	Fuc	7
Others/Unspecified	-	31

NEW FEATURES IN VERSION 4.0

Since version 3.0, 88 new O-glycosylation sites have been included in the database. A few earlier entries have been revised from recent literature and a few slight inconsistencies were corrected. Version 4.0 includes Medline references for the literature quoted, and database cross-references have been http-linked to sources available on the web.

O-GLYCOSYLATION TYPES AND SEQUENCE MOTIFS

The database includes entries with six types of sugars O-linked to the hydroxy amino acid (Table 1).

Only a minor fraction of serine and threonine residues are modified by glycosylation and no clear consensus sequence exists for acceptor sites in most types of O-glycosylation. While certain rules and acceptor motifs have been proposed (10–16), there are no definite rules which distinguish glycosylated hydroxy amino acid residues from non-glycosylated residues. Sequence logos displaying the fuzzy sequence context for mucin-type (GalNAc), GlcNAc, mannose and xylose glycosylation are shown in Figure 1.

DATA SOURCES

Proteins with a carbohydrate assigned either to a serine or a threonine residue were extracted from SWISS-PROT (17) (release 36) and PIR (18) (release 50) databases as well as directly from published reports. Only a few entries were submitted directly to the database. All proteins included in our database have at least one experimentally verified O-glycosylation site and consist largely of glycoproteins expressed *in vivo*.

*To whom correspondence should be addressed. Tel: +45 4525 2472; Fax: +45 4593 1585; Email: ramneek@cbs.dtu.dk

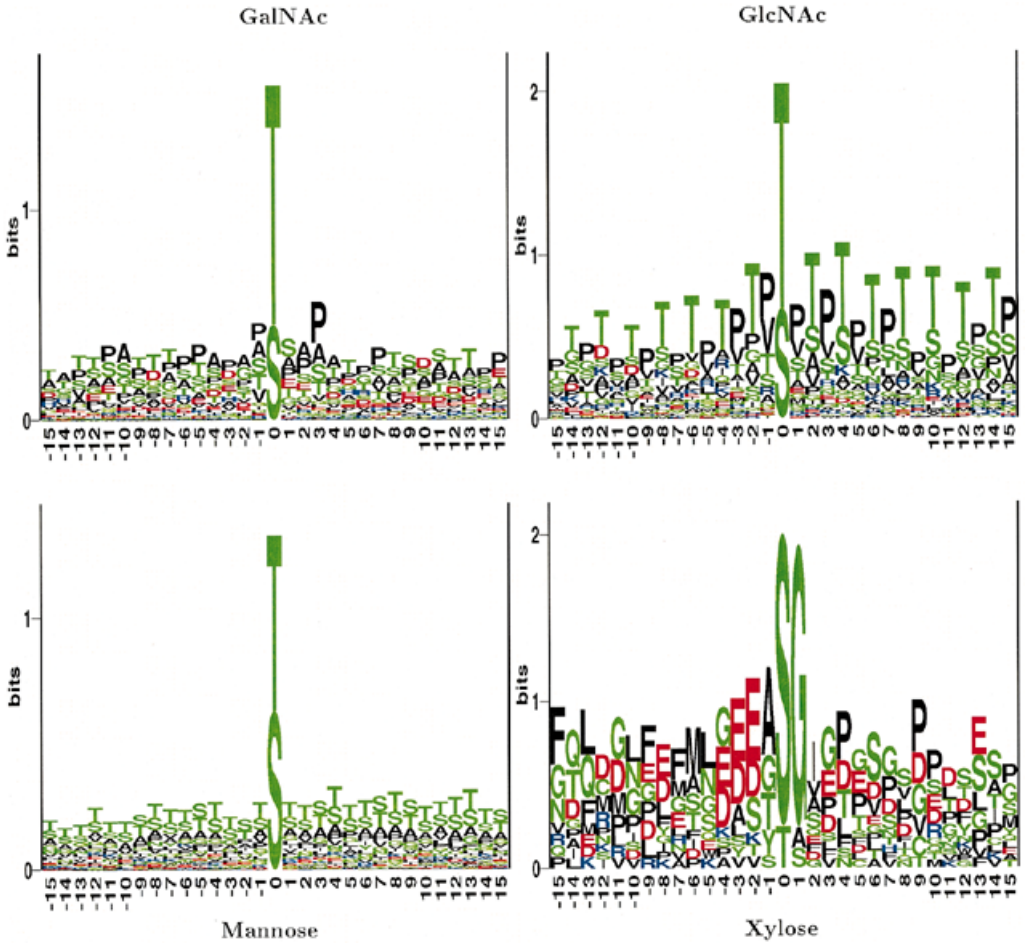


Figure 1. Acceptor site specificities are shown in the form of Shannon information content (21) represented as sequence logos (22). Logos are shown for GalNAc, GlcNAc, mannose and xylose-linked glycosylations and incorporate all sites available in OGLYCBASE 4.0 (Table 1). Sites were aligned with the glycosylated residue at position 0. The height of each column reflects the bias from random of particular residues surrounding the glycosylation site. The size of each residue letter reflects the frequency of the residue on that position. The height of the central serine/threonine has been rescaled to magnify the context of other positions, and is thus non-informative. Neutral and polar amino acids are shown in green, acidic and basic charged residues in red and blue, respectively, and the neutral and hydrophobic in black. The high frequency of serine and threonine in the vicinity of the glycosylated position 0 is largely indicative of the close clustering of O-glycosylation sites in general (11,12,20,23). All four acceptor patterns display a low frequency of charged amino acids at position –1 as originally found for the GalNAc transferase (24). GalNAc and GlcNAc acceptor sites are influenced to a large extent by proline vicinity (23–26). Further, the clustering of GlcNAc acceptor sites indicates an even positioning of acceptor sites and an odd positioning of proline/valine sites (Gupta *et al*., in preparation). For mannose modified sites, the most frequent hydrophobic residue is alanine on all positions. Xylose acceptor sites agree with the characteristic 'SGXG' motif suggested for glycosaminoglycan attachment (16).

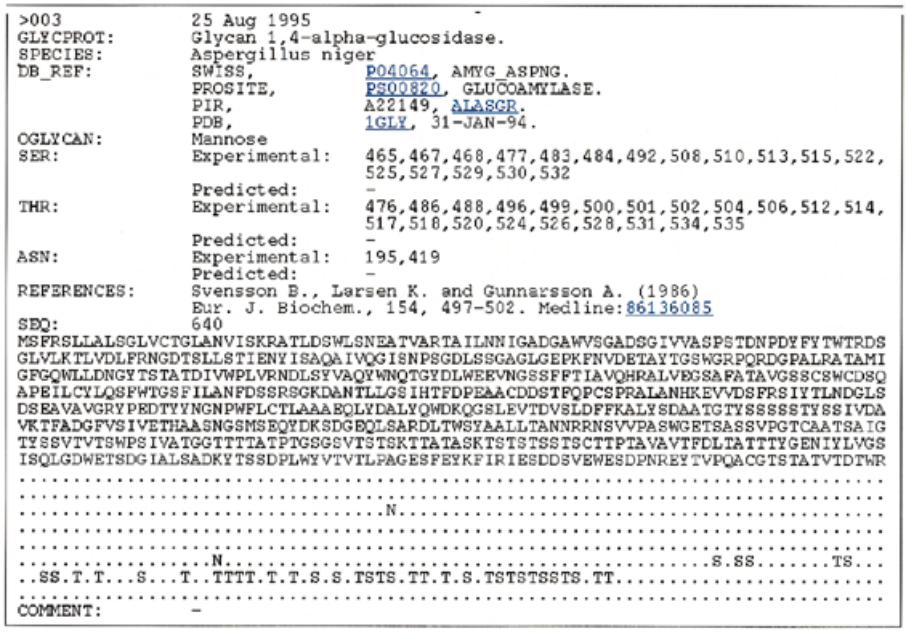
DESCRIPTION

Version 4.0 (7 Oct, 1998) of O-GLYCBASE contains 179 glycoprotein entries with 991 experimentally determined O-glycosylation sites. The entries consist of 11 fields (Fig. 2) including database and literature references, the O-linked glycan, sequence and a list of O-linked sites. N-glycosylated sites have also been marked on the included proteins where information was readily available. The database is extensively cross-referenced to sequence and structure databases including SWISS-PROT, PROSITE, TREMBL, PIR, PDB, EMBL, HSSP, GenPept and GenBank. The database is available in a plain-text as well as an html format with http links to Medline abstracts (maintained at PubMed, NCBI) and database sources on the internet. The database will be continually updated with the availability of more information. For a more comprehensive description of the database, see Hansen *et al.* (19).

PREDICTION OF O-GLYCOSYLATION

We have earlier reported a neural network based prediction method of mucin type O-glycosylation on mammalian proteins, NetOGlyc 2.0 (20). The method is available at the URL: http://www.cbs.dtu.dk/services/NetOGlyc/ both as a mail and web server. We have recently also made available a prediction method for α-linked *N*-acetylglucosamine (GlcNAc) acceptor sites on *Dictyostelium discoideum* proteins at http://www.cbs.dtu.dk/services/DictyOGlyc/ (Gupta *et al*., in preparation). Further prediction servers for other types of O-glycosylation will be developed when a statistically sufficient number of acceptor sites is available.

Predictions indicate that more than one-third of all SWISS-PROT entries may be O-glycosylated, which is in sharp contrast to the very small number of entries annotated as O-glycosylated.



```
>003              25 Aug 1995
GLYCPROT:        Glycan 1,4-alpha-glucosidase.
SPECIES:         Aspergillus niger
DB_REF:          SWISS,          P04064, AMYG_ASPNG.
                 PROSITE,        PS00820, GLUCOAMYLASE.
                 PIR,            A22149, ALASGR.
                 PDB,            1GLY, 31-JAN-94.
OGLYCAN:         Mannose
SER:             Experimental:   465,467,468,477,483,484,492,508,510,513,515,522,
                                 525,527,529,530,532
                 Predicted:      -
THR:             Experimental:   476,486,488,496,499,500,501,502,504,506,512,514,
                                 517,518,520,524,526,528,531,534,535
                 Predicted:      -
ASN:             Experimental:   195,419
                 Predicted:      -
REFERENCES:      Svensson B., Larsen K. and Gunnarsson A. (1986)
                 Eur. J. Biochem., 154, 497-502. Medline:86136085
SEQ:             640
MSFRSLLALSGLVCTGLANVISKRATLDSWLSNEATVARTAILNNIGADGAWVSGADSGIVVASPSTDNPDYFYTWTRDS
GLVLKTLVDLFRNGDTSLLSTIENYISAQAIVQGISNPSGDLSSGAGLGEPKFNVDETAYTGSWGRPQRDGPALRATAMI
GFGQWLLDNGYTSTATDIVWPLVRNDLSYVAQYWNQTGYDLWEEVNGSSFFTIAVQHRALVEGSAFATAVGSSCSWCDSQ
APEILCYLQSFWTGSFILANFDSSRSGKDANTLLGSIHTFDPEAACDDSTFQPCSPRALANHKEVVDSFRSIYTLNDGLS
DSEAVAVGRYPEDTYYNGNPWFLCTLAAAEQLYDALYQWDKQGSLEVTDVSLDFFKALYSDAATGTYSSSSSTYSSIVDA
VKTFADGFVSIVETHAASNGSMSEQYDKSDGEQLSARDLTWSYAALLTANNRRNSVVPASWGETSASSVPGTCAATSAIG
TYSSVTVTSWPSIVATGGTTTTATPTGSGSVTSTSKTTATASKTSTSTSSTSCTTPTAVAVTFDLTATTTYGENIYLVGS
ISQLGDWETSDGIALSADKYTSSDPLWYVTVTLPAGESFEYKFIRIESDDSVEWESDPNREYTVPQACGTSTATVTDTWR
.....
.....
.....N.....
.....
.....
.....N.....S.SS.....TS...
..SS.T.T...S...T..TTTT.T.T.S.S.TSTS.TT.T.S.TSTSTSSTS.TT.....
.....
COMMENT:         -
```

Figure 2. A typical database entry from O-GLYCBASE 4.0. Underlining indicates html-links. A plaintext version is also available. The O-GLYCBASE WWW home page is at http://www.cbs.dtu.dk/databases/OGLYCBASE/

DATABASE ACCESS

Version 4.0 of O-GLYCBASE is available in both plain-text and html versions via the internet at http://www.cbs.dtu.dk/databases/OGLYCBASE/ or by anonymous FTP from ftp.cbs.dtu.dk/pub/Oglyc/Oglyc.base

Users of the database are encouraged to provide corrections, comments, or new information for the database; unpublished material will be held in confidence at the request of the authors. The administrator of the database can be contacted by Email at ramneek@cbs.dtu.dk. We encourage users of the database to cite this paper.

ACKNOWLEDGEMENT

This work was supported by the Danish National Research Foundation.

REFERENCES

1 Van den Steen,P., Rudd,P.M., Dwek,R.A. and Opdenakker,G. (1998) *Crit. Rev. Biochem. Mol. Biol.*, **33**, 151–208.
2 Hounsell,E.F., Davies,M.J. and Renouf,D.V. (1996) *Glycoconjugate J.*, **13**, 19–26.
3 Hart,G.W. (1997) *Annu. Rev. Biochem.*, **66**, 315–335.
4 Jenkins,N., Parekh,R.B. and James,D.C. (1996) *Nature Biotechnol.*, **14**, 975–981.
5 Jung,E. and Williams,K.L. (1997) *Biotechnol. Appl. Biochem.*, **25**, 3–8.
6 Chou,S.M. and Wang,H.S. (1997) *J. Neurosci.*, **152**, 198–209.
7 Iino,M., Foster,D.C. and Kisiel,W. (1998) *Arch. Biochem. Biophys.*, **352**, 182–192.
8 Tomita,S., Kirino,Y. and Suzuki,T. (1998) *J. Biol. Chem.*, **273**, 6277–6284.
9 Hart,G.W., Greis,K.D., Dong,L.Y., Blomberg,M.A., Chou,T.Y., Jiang,M.S., Roquemore,E.P., Snow,D.M., Kreppel,L.K. and Cole,R.N. (1995) *Adv. Exp. Med. Biol.*, **376**, 115–123.
10 Yoshida,A., Suzuki,M., Ikenaga,H. and Takeuchi,M. (1997) *J. Biol. Chem.*, **272**, 16884–16888.
11 Gooley,A.A., Classon,B.J., Marshcalek,R. and Williams,K.L. (1991) *Biochem. Biophys. Res. Com.*, **178**, 1194–1200.
12 Pisano,A., Redmond,J.W., Williams,K.L. and Gooley,A.A. (1993) *Glycobiology*, **3**, 429–435.
13 Haltiwanger,R.S., Kelly,W.G., Roquemore,E.P., Blomberg,M.A., Dong,L.Y.D., Kreppel,L., Chou,T.Y. and Hart,G.W. (1992) *Biochem. Soc. Trans.*, **20**, 264–269.
14 Haynes,P.A., Gooley,A.A., Ferguson,M.A., Redmond,J.W. and Williams,K.L. (1993) *Eur. J. Biochem.*, **216**, 729–737.
15 Harris,R.J. and Spellman,M.W. (1993) *Glycobiology*, **3**, 219–224.
16 Bourdon,M.A., Krusius,T., Campbell,S., Schwartz,N.B. and Ruoslahti,E. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 3194–3198.
17 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res.*, **26**, 38–42.
18 Barker,W.C., George,D.G., Mewes,H.W., Pfeiffer,F. and Tsugita,A. (1993) *Nucleic Acids Res.*, **21**, 3089–3096.
19 Hansen,J.E., Lund,O., Nilsson,J.O., Hansen,J.-E.S. and Brunak,S. (1996) *Nucleic Acids Res.*, **24**, 248–252.
20 Hansen,J.E., Lund,O., Tolstrup,N., Gooley,A.A., Williams,K.L. and Brunak,S. (1998) *Glycoconjugate J.*, **15**, 115–130.
21 Shannon,C.E. (1948) *Bell System Tech. J.*, **27**, 379–423 and 623–656.
22 Schneider,T.D. and Stephens,R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.
23 Wilson,I.B.H., Gavel,Y. and Heijne,G.V. (1991) *Biochem. J.*, **275**, 529–534.
24 O'Connell,B.C., Tabak,L.A. and Ramasubbu,N. (1991) *Biochem. Biophys. Res. Com.*, **180**, 1024–1030.
25 Elhammer,A.P., Poorman,R.A., Brown,E., Maggiora,L.L., Hoogerheide,J.G. and Kzdy,F.D. (1993) *J. Biol. Chem.*, **268**, 10029–10038.
26 Jung,E., Gooley,A.A., Packer,N.H., Slade,M.B., Williams,K.L. and Dittrich,W. (1997) *Biochemistry*, **36**, 4034–4040.