Glycoconjugate Journal (1998) **15:** 115–130

NetOglyc: Prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility

Jan E. Hansen[1]*, Ole Lund[1], Niels Tolstrup[1], Andrew A. Gooley[2], Keith L. Williams[2] and Søren Brunak[1]

[1]*Center for Biological Sequence Analysis, The Technical University of Denmark, Building 206 Lyngby, DK-2800 Denmark*
[2]*School of Biological Sciences, Macquarie University, Sydney, 2109 NSW, Australia*

The specificities of the UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferases which link the carbohydrate GalNAc to the side-chain of certain serine and threonine residues in mucin type glycoproteins, are presently unknown. The specificity seems to be modulated by sequence context, secondary structure and surface accessibility. The sequence context of glycosylated threonines was found to differ from that of serine, and the sites were found to cluster. Non-clustered sites had a sequence context different from that of clustered sites. Charged residues were disfavoured at position – 1 and +3. A jury of artificial neural networks was trained to recognize the sequence context and surface accessibility of 299 known and verified mucin type O-glycosylation sites extracted from O-GLYCBASE. The cross-validated NetOglyc network system correctly found 83% of the glycosylated and 90% of the non-glycosylated serine and threonine residues in independent test sets, thus proving more accurate than matrix statistics and vector projection methods. Predictions of O-glycosylation sites in the envelope glycoprotein gp120 from the primate lentiviruses HIV-1, HIV-2 and SIV are presented. The most conserved O-glycosylation signals in these evolutionary-related glycoproteins were found in their first hypervariable loop, V1. However, the strain variation for HIV-1 gp120 was significant. A computer server, available through WWW or E-mail, has been developed for prediction of mucin type O-glycosylation sites in proteins based on the amino acid sequence. The server addresses are http://www.cbs.dtu.dk/services/NetOGlyc/ and netOglyc@cbs.dtu.dk.

***Keywords*: mucin type O-glycosylation, specificity, GalNAc transferase, prediction, neural networks, gp120, SIV, HIV**

Introduction

A number of biological functions of glycoproteins, including solubility, lifetime in circulation and proteolytic resistance, are modulated by O-glycosylation [1–4]. These oligosaccharides serve as ligands for selectins and participate in specific recognition events [5] such as sperm-egg binding [6] and leukocyte adhesion to endothelia. Aberrant short chained O-glycosylation such as the Tn antigen is abnormally expressed in number of cancers and may be involved in the metastatic spread of tumour cells [7].

As these functions are of potential pharmacological importance, there is an interest in designing competitive inhibitors of glycosyltransferases. This requires a detailed understanding of the fuzzy rules which determine glycosyltransferase specificity. Mucin type O-glycosylation, where an *N*-acetylgalactosamine is α-1 linked to the side-chain of serine or threonine residues in secreted or membrane bound glycoproteins, is the most frequent type of O-glycosylation, (for recent reviews see [4, 8–11]). In contrast to N-linked glycosylation no simple consensus acceptor sequence has been found for mucin type O-linked glycosylation [12–14]. The acceptor sequence patterns are highly dependent on the amino acid sequences flanking the serine and threonine [12,14–17]. These patterns are rich in serine, threonine, proline, valine, alanine and glycine and have only few charged residues especially at position −1 relative to the glycosylated residue [14, 18, 19]. Clustering and partial glycosylation of mucin type O-glycosylation sites are frequently found. As O-linked glycosylation is a post-translational event taking place in the cis-Golgi compartment [20] after N-glycosylation, folding and oligomerization [21], acceptor motifs must be exposed on the glycoprotein surface to be accessible to a UDP-GalNAc:polypeptide *N*-acetylgalactosaminyl (GalNAc) transferase. This is in fact the case for all the O-linked glycoproteins for which crystallographic coordinates have been deposited in the Brookhaven Protein Data Bank (PDB). O-glycosylation is therefore also dependent on local conformation [22] and overall tertiary structure [23–25] of the glycoprotein. The lack of understanding

*To whom correspondence should be addressed. Tel.: +45 4525 2485; Fax: +45 4593 4808; E-mail: janhan@cbs.dtu.dk.

0282–0080

of mucin type O-glycosylation site selection and occupancy heterogeneity, prompted us to investigate whether the two known determinant features of mucin type O-glycosylation, sequence context and surface exposure, could be utilized to reliably predict O-glycosylation sites exclusively from the sequence.

Several highly homologous GalNAc transferases from different species, preferentially expressed in different tissues, have been cloned and designated GalNAc-T1 to GalNAc-T4 transferase [10, 26–36]. More probably exist. These enzymes may have distinct but overlapping acceptor specificities [10]. However, currently there is no *in vivo* information linking a specific enzyme and individual sites. Secondly, if the sites glycosylated by each of the GalNAc transferases are equally distributed in the database used, our statistical method will pick up a compound specificity for these enzymes. Therefore, the prediction method is likely to assign positive sites which correspond to the *in vivo* specificity of the entire transferase family, and not to that of individual enzymes. We recently developed a predictive tool [14] based on neural networks and a more limited data set. The method presented here is based on a carefully selected enlarged database of 299 O-glycosylation sites extracted from O-GLYCBASE [37], an averaging of eight independently trained networks and an additional variable threshold feature based on the surface accessibility. The integration of these features resulted in significantly increased performance. The validity of the prediction method was assessed by four-fold cross-validation on independent test sets. We compare the predictive performance of the new NetOglyc algorithm with the Elhammer *et al.* [38] matrix statistics method and the vector projection method of Chou *et al.* published recently [39, 40].

With this new revised NetOglyc method we also predict O-glycosylation sites in the envelope glycoproteins from Human Immunodeficiency Virus type 1 and 2 (HIV-1, HIV-2) and Simian Immunodeficiency Virus (SIV). These glycoproteins mediating specific viral adhesion, co-receptor binding and triggered fusion with the target cells are known to be O-glycosylated [41–44], but the exact sites have not yet been mapped experimentally.

Materials and methods

Sequence data selection criteria

All glycoprotein sequences were extracted from O-GLYCBASE, which is a revised database of O-glycosylated proteins [37]. Only mammalian sequences with a GalNAc sugar moiety linked to serine or threonine were included. Coagulation factors and proteoglycans and nuclear and cytoplasmic glycoproteins were excluded. All examples of O-glycosylation were extensively cross-checked for glycosylation site assignment errors and sequence errors. In order not to bias the data set with identical sequence contexts, the sequence similarity between the glycosylation sites was quantified by aligning all glycosylated 11 (5 + 1 + 5) amino acid sequence windows with each other secondly with all non-glycosylated sequence windows. In total 335,592 alignments were performed with the FASTA package [45] using a high gap penalty score. Examples with identical residues from positions −5 to +5 were excluded. Finally, all non-glycosylated sequence windows were compared to the glycosylated sequences; no conflicts were found. However, a small number of glycosylated sequence windows with a high similarity to non-glycosylated sequence windows were found (Table 1). This high similarity between glycosylated acceptor sequences and non-glycosylated non-acceptor sequences demonstrate the difficulties in elucidating the specificity of the GalNAc transferases. The sequences were presented to the network without the signal sequence. For the sequences where only a fraction of the O-glycosylation sites were known, the sequences were truncated so only the sequence regions with verified O-glycosylation sites were presented. The final yield was 60 unique mammalian glycoproteins with 186 mucin type O-glycosylated threonine residues, 113 O-glycosylated serine residues and 2516 non-glycosylated threonine and serine residues. This data set (Table 2), which to our knowledge is the most extensive presented to date, was the empirical basis for the present analysis.

Prediction of surface accessibility based on known 3D protein structures

The network used to predict surface accessibility was trained on a non-homologous data set of 134 globular

Table 1. The high sequence identity between glycosylated and non-glycosylated sequences may demonstrate the fine tuned specificity of the GalNAc transferases. It also demonstrates the level of complexity in prediction of mucin type O-glycosylation sites.

O-glycosylated			*Not O-glycosylated*			
Glycoprotein	Swiss-prot entry	Oglyc + sequence	Glycoprotein	Swiss-prot entry	Oglyc − sequence	Similarity
Kininogen, bovine	KNH_i_BOVIN	EGPVV-T-AQYEC	Kininogen, human	KNH_HUMAN	EGPVV-T-AQYDC	90.9%
Kininogen, human	KNH_HUMAN	SEDST-T-PSAQT	Kininogen, bovine	KNH_BOVIN	YEDST-T-SSAQT	81.8%
Leukosialin, rat	LEUK_RAT	TMATG-S-LGPSK	Leukosialin, human	LEUK_HUMAN	TMATV-S-LETSK	72.7%

Table 2. The glycosylated data set extracted from O-GLYCBASE [37]. References and entry names to sequence databases PIR [76] and SWISS-PROT [77] are given. Entry numbers in O-GLYCBASE are noted in the brackets (OGB:entry.no)

Entry	O-site sequence	Ref.
Position	54321-0+12345	
GFHUC (P)	---MW S TRSPN	[78]
glycophorin	--MWS T RSPNS	
C - human	MWSTR S PNSTA	
(OGB:009)	TRSPN S TAWPL	
	RSPNS T AWPLS	
	TAWPL S LEPDP	
	DPGMA S ASTTM	
	GMASA S TTMHT	
	MASAS T TMHTT	
	ASAST T MHTTT	
	STTMH T TTIAE	
	TTMHT T TIAEP	
	TMHTT T IAEPD	
	PDPGM S GWPDG	
GFHUE (P)	----S S TTGVA	[79]
glycophorin	---SS T TGVAM	
A- human	--SST T GVAMH	
(OGB:008)	GVAMH T STSSS	
	VAMHT S TSSSV	
	AMHTS T SSSVT	
	MHTST S SSVTK	
	TSSSV T KSYIS	
	SSVTK S YISSQ	
	TKSYI S SQTND	
	YISSQ T NDTHK	
	THKRD T YAATP	
	DTYAA T PRAHE	
	RAHEV S EISVR	
	EVSEI S VRTVY	
	EISVR T VYPPE	
A05273 (S)	IPHQI S SKLPT	[80]
glycophorin	PHQIS S KLPTQ	
- dog	SSKLP T QAGFI	
(OGB:021)	QAGFI S TEDPS	
	AGFIS T EDPSF	
	DPSFN T PSTRE	
	TREDP S GTMYQ	
GFHOE (P)	----Q T IATGS	[81]
glycophorin	-QTIA T GSPPI	
HA	TIATG S PPIAG	
- horse	PPIAG T SDLST	
(OGB:012)	GTSDL S TITSA	
	TSDLS T ITSAA	
	DLSTI T SAATP	
	LSTIT S AATPT	
	ITSAA T PTFTT	
	ATPTF T TEQDG	
GFPGE (P)	----- T ETPVT	[82]
Glycophorin	TETPV T GEQGS	
- pig	TGEQG S ATPGN	
(OGB:013)	EQGSA T PGNVS	
	NVSNA T VTAGK	
	SNATV T AGKPS	
	GKPSA T SPGVM	
	KPSAT S PGVMT	
	TIKNT T AVVQK	
	VVQKE T GVPES	
LEUK_RAT (S)	ENLPN T MTMLP	[83]
Leukosialin	LPNTM T MLPFT	
- rat	TMLPF T PNSES	
(OGB:005)	PFTPN S ESPST	
	TPNSE S PSTSE	
	NSESP S TSEAL	
	SESPS T SEALS	
	ESPST S EALST	
	TSEAL S TYSSI	
	SEALS T YSSIA	
	ALSTY S SIAT-	
	LSTYS S IAT--	
	YSSIA T -----	
	----- S SGVAS	
	SSGVA S DPPVT	
	PPVTI T NPATS	
	ITNPA T SS---	
	TNPAT S S---	
	NPATS S -----	
	----A T GSLGP	
	--ATG S LGPSK	
	GSLGP S KETHG	
	ETHGL S ATIA-	
	HGLSA T IA---	

Entry	O-site sequence	Ref.
Position	54321-0+12345	
LEUK_HUMAN (S)	----S T TAVQT	[84]
Leukosialin	---ST T AVQTP	
- human	TTAVQ T PTSGE	
(OGB:004)	AVQTP T SGEPL	
	VQTPT S GEPLV	
	GEPLV S TSEPL	
	EPLVS T SEPLS	
	PLVST S EPLSS	
	TSEPL S SKMYT	
	SEPLS S KMYTT	
	SSKMY T TSITS	
	SKMYT T SITSD	
	KMYTT S ITSDP	
	YTTSI T SDPKA	
	PKADS T GDQTS	
	ALPPS T SINEG	
	TYQEV S IKMSS	
	VSIKM S SVPQE	
	SVPQE T PHATS	
	ETPHA T SHPAV	
	TPHAT S HPAVP	
	TGGTI T TNSPE	
	GGTIT T NSPET	
	----A T VSLET	
	TVSLE T SKGTS	
ALC1_HUMAN (S)	PCPVP S TPPTP	[85]
Ig alpha-1 chain	TPPTP S PSTPP	
- human.	PTPSP S TPPTP	
(OGB:027)	TPPTP S PSCCH	
	PTPSP S CCHPR	
ALC_MOUSE (S)	LDVNC S GPTPP	[86]
Ig Alpha		
chain C		
- mouse		
(OGB:041)		
DHHU (P)	PKAQA S SVPTA	[87]
Ig delta	KAQAS S VPTAQ	
chain C	ASSVP T AQPQA	
- human	SLAKA T TAPAT	
(OGB:020)	LAKAT T APATT	
	TTAPA T TRNTG	
	TAPAT T RNTGR	
FQHUGM (P)	-APAR S PSPST	[88]
granulocyte-macrophage	PARSP S PSTQP	
colony-stimulating	RSPSP S TQPWE	
- human	SPSPS T QPWEH	
(OGB:029)		
EPO_HUMAN (S)	PPDAA S AAPLR	[89]
erythropoietin		
- human		
(OGB:059)		
KTHUB(P)	QDSSS S KAPPP	[90, 91]
choriogonadotropin	KAPPP S LPSPS	
beta chain	SLPSP S RLPGP	
- human	RLPGP S DTPIL	
(OGB:028)		
CTHUP(P)	DEQPL T ENPRK	[92, 93]
corticotropin		
- human		
(OGB:061)		
A16604 (S)	KIIIP T INTIA	[94]
kappa casein	ATVEP T PAPAT	
- human	TPAPA T EPTVD	
(OGB:015)	PATEP T VDSVV	
	VDSVV T PEATT	
	TPEAT T ESIIT	
	TESII T STPET	
	SIITS T PETPT	
	TVAVP T TSA--	
	VAVPT T SA---	
CASB_BOVIN (S)	FAQTQ S LVYPF	[95]
beta casein	PPLTQ T PVVVP	
- bovine	LSLSQ S KVLPV	
(OGB:030)		
PLMN_BOVIN (P)	ESSPL S TERMD	[96]
plasmin		
- bovine (OGB:072)		

Entry	O-site sequence	Ref.
Position	54321-0+12345	
PLHU (P)	EELAP T APPEL	[96]
plasmin		
- human		
(OGB:071)		
PLMN_PIG (S)	TTPPP T SGPTY	[96]
plasmin		
- porcine		
(OGB:089)		
ICHU2 (P)	---AP T SSSTK	[97]
interleukin 2		
- human		
(OGB:057)		
KGHUH1 (P)	IKEET T VSPPH	[98]
kininogen HMW I	SEDST T PSAQT	
- human	QTQEK T EGPTP	
(OGB:017)	KTEGP T PIPSL	
	AKPGV T VTFSD	
	SDLIA T MMPPI	
	MMPPI S PAPIQ	
	IPDIQ T DPNGL	
	SEINP T TQMKE	
KNH1_BOVIN (S)	EGPVV T AQYEC	[99]
kininogen	MKTEG S TTVSL	
HMW 1 - bovin	KTEGS T TVSLP	
(OGB:010)	TEGST T VSLPH	
	VSLPH S AMSPV	
	EDSTT S SAQTQ	
	QTQEK T EETTL	
	KTEET T LSSLA	
	PGVAI T FPDFQ	
	SDLIA T VMPNT	
	TVMPN T LPPHT	
	IPDIQ T EPNSL	
A29789 (S)	----- S SVPGE	[100]
mucin	----S S VPGES	
(fragment)	SVPGE S ATPQQ	
- sheep	PGESA T PQQPG	
(OGB:007)	QPGAL S ESTTQ	
	GALSE S TTQLP	
	ALSES T TQLPG	
	LSEST T QLPGV	
	QLPGV T GTSAV	
	PGVTG T SAVTG	
	GVTGT S AVTGS	
	GTSAV T GSEPG	
	SAVTG S EPGLP	
	EPGLP S TGVSG	
	PGLPS T GVSGL	
	PSTGV S GLPGT	
	SGLPG T -----	
NBHUA2 (P)	----V T LSPKD	[101]
Leucine rich		
alpha 2 gp		
- human		
(OGB:055)		
LPHUC3 (P)	PEVRP T SAVAA	[102]
apolipoprotein		
C-III - human		
(OGB:064)		
KQHU (P)	QFVHV S ESFPH	[103]
tissue kallikrein	PGFNM S LLENH	
- human	EPENF S FPDDL	
(OGB:031)		
HEMO_HUMAN (S)	----- T PLPPT	[101]
hemopexin		
- human		
(OGB:058)		
BOHUS (P)	RPVLP T QSAHD	[104]
sex steroid binding		
protein - human		
(OGB:053)		
HCHU (P)	-GPVP T PPDNI	[105, 106]
a-1-microglobulin	QEEEG S GGGQL	
(OGB:047)		
MBBOB (PIR)	IVTPR T PPPSQ	[107]
myelin basic		
protein - bovine		
(OGB:092)		

Table 2. (Continued)

Entry	O-site sequence	Ref.
Position	54321-0+12345	
NBHUIA* (P)	DKVRA T RTVVK	[108]
platelet gp.		
Ib a.chain		
- human (OGB:054)		
WOHU* (P)	QTQPV T SQPQP	[109, 110]
a 2 HS gp.	NEAVP T PVVDP	
- human	TVVQP S VGAAA	
(OGB:032)		
ITHUC1* (P)	GKVAT T VISKM	[111, 112]
complement	ILEVS S LPTTN	
C1 inhibitor	PTTNS T TNSAT	
- human	ITANT T DEPTT	
(OGB:018)	TDEPT T QPTTE	
	TTQPT T EPTTQ	
	TTEPT T QPTIQ	
LITH_HUMAN* (S)	-PEAQ T ELPQA	[113]
lithostathine		
- human (OGB:076)		
APE_HUMAN* (S)	RVRAA T VGSLA	[114]
apolipoprotein E		
- human (OGB:088)		
JXHU* (P)	ERLAG T ESPVR	[115, 116]
transferrin receptor		
- human (OGB:094)		
IVHUA2* (S)	QGVGV T ETPLM	[117]
interferon a2		
- human (OGB:091)		
TNFB_HUMAN* (S)	PGVGL T PSAAQ	[118]
lymphotoxin (OGB:075)		
MMMSND* (P)	DYDLV T SHLGL	[119]
nidogen	VPRIL S PGYEA	
- mouse	PGYEA T ERPRG	
(OGB:022)	PRGVP T ERTRS	
	VPTER T RSFQL	
	PPCLS T VAPPI	
	GPVVP T AVIPL	
A24573* (P)	PALQP T QGAMP	[120]
granulocyte colony		
stim. f. - human (OGB:082)		
ENV_MLVFR* (S)	PVSNS T PTMIS	[75]
knob protein	TPTMI S PSPTP	
gp71 - mouse	TMISP S PTPTQ	
(OGB:019)	ISPSP T PTQPP	
	PSPTP T QPPPA	
	QALNL T NPDKT	
	TNPDK T QECWL	
IL5_HUMAN* (S)	---IP T EIPTS	[121]
interleukin 5		
- human (OGB:079)		
IGHU2* (P)	VSTPP T VLPDN	[122]
insulin-like		
growth factor II (OGB:090)		
TTHUAP* (P)	SRAYP T PLRSK	[123]
thyrotropin		
chain A (OGB:095)		
A2HS_BOVIN*	EAEAP S AVPDA	[124]
Fetuin	DAAGP T PSAAG	
-bovin	AGPTP S AAGPP	
(OGB:130)	IVGQP S IPGGP	
CD8A_RAT*	VNSII T KPVTR	[125]
(OGB:129)	ITKPV T RAPTP	
	VTRAP T PVPPP	
	PVPPP T GTPRL	
	PPPTG T PRLLR	

Entry	O-site sequence	Ref.
Position	54321-0+12345	Ref.
LMP1_HUMAN* (S)	EQDRP S PTTAP	[126]
lysosome	DRPSP T TAPPA	
associated	RPSPT T APPAP	
membrane	PPAPP S PSPSP	
glycoprotein	APPSP S PSPVP	
(OGB:024)	PSPSP S PVPKS	
LMP2_HUMAN* (S)	DKDKT S TVAPT	[126]
lysosome	KDKTS T VAPTI	
associated	STVAP T IHTTV	
membrane	APTIH T TVPSP	
glycoprotein	PTIHT T VPSPT	
(OGB:016)	HTTVP S PTTTP	
	TVPSP T TTPTP	
	VPSPT T TPTPK	
	PSPTT T PTPKE	
	PTTTP T PKEKP	
ITAB_HUMAN* (S)	DWGLP S PSPSP	[127]
platelet M.	PSPSP S PIHPA	
glycoprotein IIB (OGB:070)		
GLP_MACFU* (S)	----S S TTVPA	[128]
glycophorin	---SS T TVPAT	
macaca	--SST T VPATH	
fuscata	TTVPA T HTSSS	
(OGB:006)	VPATH T SSSSL	
	PATHT S SSSLG	
	ATHTS S SSLGP	
	THTSS S SLGPE	
	HTSSS S LGPEQ	
	PEQYV S SQSND	
	EQYVS S QSNDK	
	SNDKH T SDSHP	
	SDSHP T PTSAH	
	SHPTP T SAHEV	
	SAHEV T TEFSG	
	AHEVT T EFSGR	
	VTTEF S GRTHY	
	EFSGR T HYPPE	
TPO_HUMAN*	RLTLS S PAPPA	
Megakaryocyte	RRAPP T TAVPS	
colony stim.	RAPPT T AVPSR	
factor	TTAVP S RTSLV	
(OGB:097)	LFPGP S RRTLG	
CASK_BOVIN*(S)	KTEIP T INTIA	[129]
Kappa casein	ASGEP T STPTT	
-bovine	GEPTS T PTTEA	
(OGB:096)	TSTPT T EAVES	
	EAVES T VATLE	
	NTVQV T STAV-	
HGF_HUMAN*	GDTTP T IVNLD	[130]
Hepatocyte		
growth f.(OGB:117)		
LCAT_HUMAN*	PPASP T ASPEP	[131]
phospholipid-cholesterol	ASPTA S PEPPP	
acyltransferase (OGB:107)		
MUC1_HUMAN*	PAHGV T SAPDT	[132]
MUC-1 repeat	RPAPG S TAPPA	
(OGB:104)	PAPGS T APPAH	
BAL_HUMAN*	PPVPP T GDSGA	[133]
Bile-salt-activated		
lipase (OGB:102)		
FA10_BOVIN*	ADLSP T ESSLD	[134, 135]
coagulation	APATW T VPPPL	
factor X (OGB:042)		
FA10_HUMAN*	APDSI T WKPYD	[134]
(OGB:036)	ADLDP T ENPFD	

proteins extracted from the PDB database using the Hobohm-1 procedure [46] for redundancy reduction. This ensured that no aligned sub-sequences had more than 25% sequence identity for alignment lengths of 80 or more residues. All protein structures were resolved to a resolution better than 2.5 Å. No NMR structures were included. The Connolly Molecular Surface procedure [47] was used with a probe radius of 1.4 Å, corresponding to the molecular radius of water, to label all residues in the 134 protein structures as either surface exposed or buried. Surface assignment was defined as having more than 20% of the normalized standard maximal surface area [48] exposed to the solvent. We thus obtained a data set of 134 protein sequences, based on high resolution protein structures, for which all residues had buried/surface assignment. This data set was used to train neural networks to recognize the relation between sequence and surface accessibility. Overall the method correctly predicted surface accessibility for 74% of the residues. Ninety percent of the glycosylated residues were correctly predicted to be surface accessible. Details of the network used, training procedures and performance will be reported elsewhere. This surface prediction method has

recently been used advantageously for the joint prediction of post-translational cleavage of picornaviral polyproteins [49]. With the knowledge that glycosylation sites are surface exposed, the surface prediction was used to modulate the threshold of the O-glycosylation networks. If a sequence region was predicted buried the threshold for O-glycosylation was increased. On the contrary if the sequence region was predicted surface exposed, the threshold for O-glycosylation was reduced. This increased the sensitivity as well as the specificity of the final prediction scheme (Figure 1 and Table 3).

Quantification of sequence information content

When a set of sequences is aligned, the Shannon information measure [50] can be used to quantify the conservation at each position. The Shannon information content was computed by the formula

$$I(i) = \log_2 20 + \sum_{L=1}^{20} p_i^L \log_2 p_i^L, \qquad (1)$$

where p_i^L was the probability of occurrence of a particular amino acid L at position i. The unit of information was bits/amino acid. The information content can be displayed as sequence logos [51], where the amino acid symbols themselves are used to represent the value of I at a given position. The sum of the heights of the letters indicates the value of I and the height of each letter its frequency at the position. This powerful visualization approach makes it straightforward to comprehend the overall statistics of the complex acceptor patterns of the GalNAc transferases as one can directly see which residues are favoured at particular positions.

Neural network algorithms

We have utilized artificial neural network algorithms due to their ability to classify even highly complex and non-linear biological sequence patterns where correlations between positions are important [52]. They have been used extensively for predicting of protein secondary structure from primary sequence [53–58], surface accessibility [59], transmembrane helices [60], cleavage sites in viral polyproteins [49] and human and plant intron donor and acceptor sites from the pre-mRNA sequence [61, 62]. We used a standard feed forward neural network algorithm, with one layer of hidden units, and adjusted the weights by backpropagation [63]. Hence each neuron (unit), except those in the input layer, calculates a weighted sum of its inputs and passes this sum through a sigmoidal function to produce the output

$$O = \sigma\left(\sum_{n=1}^{N} w_n I_n - t\right), \qquad (2)$$

where N is the number of neurons in the previous layer, I_n the nth input to the neuron, w_n its weight, and t its threshold. σ was the sigmoidal function $\sigma(x) = 1/(1 + \exp(-x))$.

We used the slightly more powerful error function suggested by McClelland,

$$E = -\sum_{\alpha} \log(1 - (O^{\alpha} - T^{\alpha})^2),$$

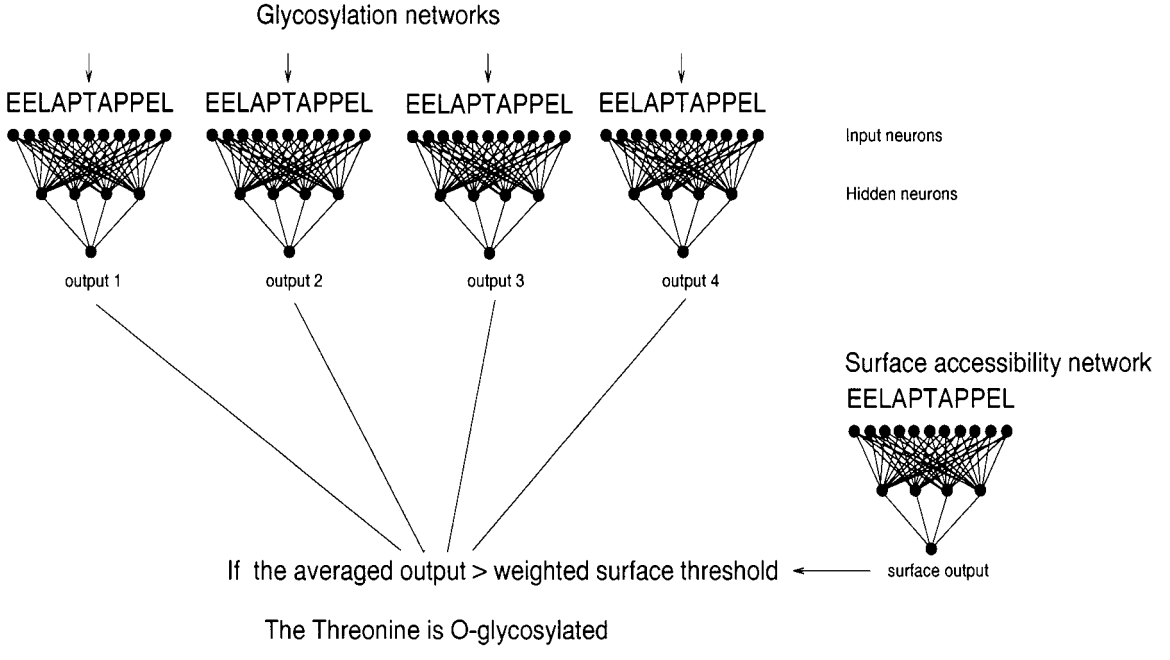

Figure 1. A simplified illustration of the four glycosylation networks and the surface accessibility network equipped with 11 amino acid windows and four hidden neurons. Each amino acid was encoded by 21 units. Sequence windows from 3 to 25 residues, and up to 25 hidden units were tested. The arrows indicate the central threonine in the sequence window being evaluated for its mucin type glycosylation potential.

Table 3. Predictive performance of the methods for prediction of mucin type O-glycosylation. Values for the neural network method were cross-validated over all positive and negative examples in the data set. In this way the method was evaluated against the largest possible test data set not used during training. The vector projection method by Chou [39] and the matrix method of Elhammer [38] were evaluated using the available data marked by * in Table 1 not used for generating these methods. The averaged performance refers to the final NetOglyc method available on the WWW server evaluated on all 60 glycoproteins (training performance). If the sequence from the user has high similarity to the sequences in the data set this averaged performance can be expected. If the sequence have no similarity to the data set sequences the cross-validated performance (in bold face) can be expected.

Method	*Correlation coefficient*	*Sensitivity %*	*Specificity %*
Vector Chou threonine [39]	0.15	73.6	15.5
Vector Chou serine [39]	0.09	79.3	5.5
Matrix Elhammer threonine [38]	0.52	73.6	46.9
Matrix Elhammer serine [38]	0.35	75.9	20.6
Neural threonine	0.58	85.0	50.2
Neural serine	0.49	73.4	40.0
Neural+surface threonine	**0.60**	**87.6**	**51.1**
Neural+surface serine	**0.54**	**75.2**	**41.5**
Neural+surface threonine (averaged)	0.74	95.6	60.3
Neural+surface serine (averaged)	0.76	96.8	63.2

replacing the conventional error function [63]

$$E = \sum_{\alpha} (O^{\alpha} - T^{\alpha})^2$$

where T^{α} is the training target value and O^{α} the actual output value for input window α. This reduced the convergence time and allowed a given network architecture to learn more complex tasks without increasing the network size. Each amino acid was represented by sparse encoding [55] as a binary string of 21 bits. Alanine and cysteine, for example, were represented as: 100000000000000000000 and 010000000000000000000, respectively. The 21st bit encoded C- and N-terminal ends of the sequence.

We used a symmetric input window of amino acids ranging from 3 amino acids (covering the two amino acids flanking the serine or threonine site) up to 25 amino acids. Neural networks with up to 25 hidden units were evaluated. Details of the training procedure may be found elsewhere [62]. The training was balanced by presenting the non-glycosylated sequence windows in random order with a lower frequency such that glycosylated and non-glycosylated sequence windows were presented equally often.

To find the best network we tested the performance of different network architectures by training them on 75% of the data and testing the performance of each network on the remaining proteins, using the correlation coefficient C [62, 64] as performance measure.

$$C = \frac{P_x N_x - N_{fx} P_{fx}}{\sqrt{(N_x + N_{fx})(N_x + P_{fx})(P_{fx} N_{fx})(P_x + P_{fx})}} \quad (3)$$

Here P_x is the number of true positives (experimentally verified glycosylated, predicted glycosylated), N_x the number of true negatives (experimentally verified non-glycosylated, predicted non-glycosylated), P_{fx} the number of false positives (experimentally non-glycosylated, predicted glycosylated), and N_{fx} the number of false negatives (experimentally glycosylated, predicted non-glycosylated.

This measure is more relevant than calculating the percentage of correct assignments, as more than 90% of the residues are non-glycosylated. Hence a network which predicts all residues as non-glycosylated will be 90% correct, but obviously not of any use. For such a prediction the correlation coefficient C will be zero. For a completely perfect prediction the correlation coefficient C will equal 1 and for completely imperfect prediction -1.

The sensitivity S_n was computed as:

$$S_n = \frac{P_x}{P_x + N_{fx}} \quad (4)$$

and the specificity Sp by

$$Sp = \frac{P_x}{P_x + P_{fx}}. \quad (5)$$

Four-fold cross-validation on 60 glycoproteins

A fair evaluation of any sequence driven prediction method has to take into account that using a test data set with high similarity to the sequences used for generating the method lead to overestimated predictive performance. In our earlier study, we therefore reported the accuracy of our method [14] on two independent test sets with both high and low similarity to the training data set. However, using small test data sets might not be representative of the overall performance on new glycoprotein sequences. Here we therefore applied fourfold cross-validation [53] on the full set of 60

glycoproteins. This involved dividing the data set into 45 glycoproteins to be used for training and 15 proteins used for testing. This division was repeated four times so that each glycoprotein was used for testing once. Each of the four divisions was carried out in such a way that none of the proteins used for testing had significant similarity with any of the glycoproteins in the corresponding training set.

Combination of O-glycosylation and surface accessibility networks

The output from the networks range from zero to unity. Traditionally a threshold k of 0.5 is used such that glycosylation is assigned if the network output O^g is larger than this fixed cutoff,

$$O^g > k \tag{6}$$

However as O-glycosylation sites are found exclusively on the surface of proteins, the probability for O-glycosylation should be larger if the site is surface exposed and smaller if buried. The output from the surface network O^s was therefore used to derive a modulated threshold k for O-glycosylation. If the site and surroundings were predicted surface accessible the cutoff was lowered. This surface score O^s was weighted with a factor ε and an off-set constant c equivalent to the normal cutoff. Using this variable cutoff, glycosylation was assigned if

$$O^g > c - eO^s \tag{7}$$

The following optimal values for e and c were found by systematically screening the combinations of e and c.

Glycosylation type	e	c
Serine	0.51	0.90
Threonine	0.64	0.97

To maximize the generalization ability and minimize the danger of overfitting in the final method, the outputs O^g from four differently trained networks were averaged before combination with the surface network (Figure 1). This combination procedure is similar to the one used in the NetGene and SignalP prediction methods [61, 62, 65].

Results

The sequence context surrounding mucin type O-linked glycosylation sites

The distribution of residues around mucin type O-glycosylation sites preferred by the GalNAc transferases is illustrated in Figure 2. The abundance of favoured amino acids extends far beyond the positions -4 and $+4$ previously reported [38] and no clear consensus pattern is observable. This might indicate that a specific conformation or an ensemble of closely related conformations, probably exposed β-turns [14], is recognized by the GalNAc transferases rather than specific sequences. The sequence context differs for serine and threonine and for single and multiple sites. The sequence rule of proline in position -1 and $+3$ deduced by Wilson *et al.* [12] holds only for threonine as proline in positions -6 to -9 and $+3$ to $+7$ also are frequent in the serine context. A general feature is the electrostatic gradient. Negatively charged residues, especially glutamic acid (E) are more frequent than positively charged residues. Charged residues are disfavoured at positions -1 and $+3$ but accepted at position $+1$. This feature was also found by O'Connell *et al.* [18] by analysis of *in vitro* glycosylated peptides and recently confirmed by Nehrke *et al.* [19] by mutations in an *in vivo* glycosylated reporter protein. Altogether these findings suggest that docking of the GalNAc transferase on the substrate peptide may depend on charge. Further, that the glycosylation process appears to have a polarity as the acceptor sequence is asymmetric with regard to charge. The high frequency of serine and threonine in the context around glycosylated sites reflects that glycosylated sites often are clustered. This high hydroxy amino acid frequency is not seen in the context of non-clustered sites. Valine at position -4 and $+2$ is relatively frequent in the context of single serine sites.

Conformational preference of O-glycosylation sites

The secondary structure of the 60 glycoproteins was predicted using the PHD method [53, 66, 67]. The glycosylated sites were predominantly predicted to be in coil regions (Figure 3). Of the glycosylation sites 68% were predicted to have no secondary structure, 30% were predicted to be β-strand and only 2% were predicted to be in helical regions. This is in accordance with our earlier finding that O-glycosylation sites predominantly adopt specific β-turns [14], which then is correctly classified as coil by the three state prediction PHD method. Synthetic glycopeptides also adopt β-turns as found by Hollosi *et al.* [68].

Predictive performance of neural network algorithms

The performance measures are summarized in Table 3. The final method correctly finds 83% of the glycosylated residues and 90% of the non-glycosylated residues in all glycoproteins. The addition of surface derived threshold increases specificity as well as sensitivity. Compared to the vector projection method of Chou [39] and the matrix method of Elhammer [38] we obtain considerably higher correlation coefficients. The vector projection method of Chou [39] and the matrix method of Elhammer [38] were evaluated using the available data marked by * in Table 2 not used for generating these methods. The vector projection method, which originally was tested against a test set of glycoproteins highly homologous to the training set and a negative set of only four non-glycoproteins, seem less robust to the choice of test set than the matrix method originally

a.

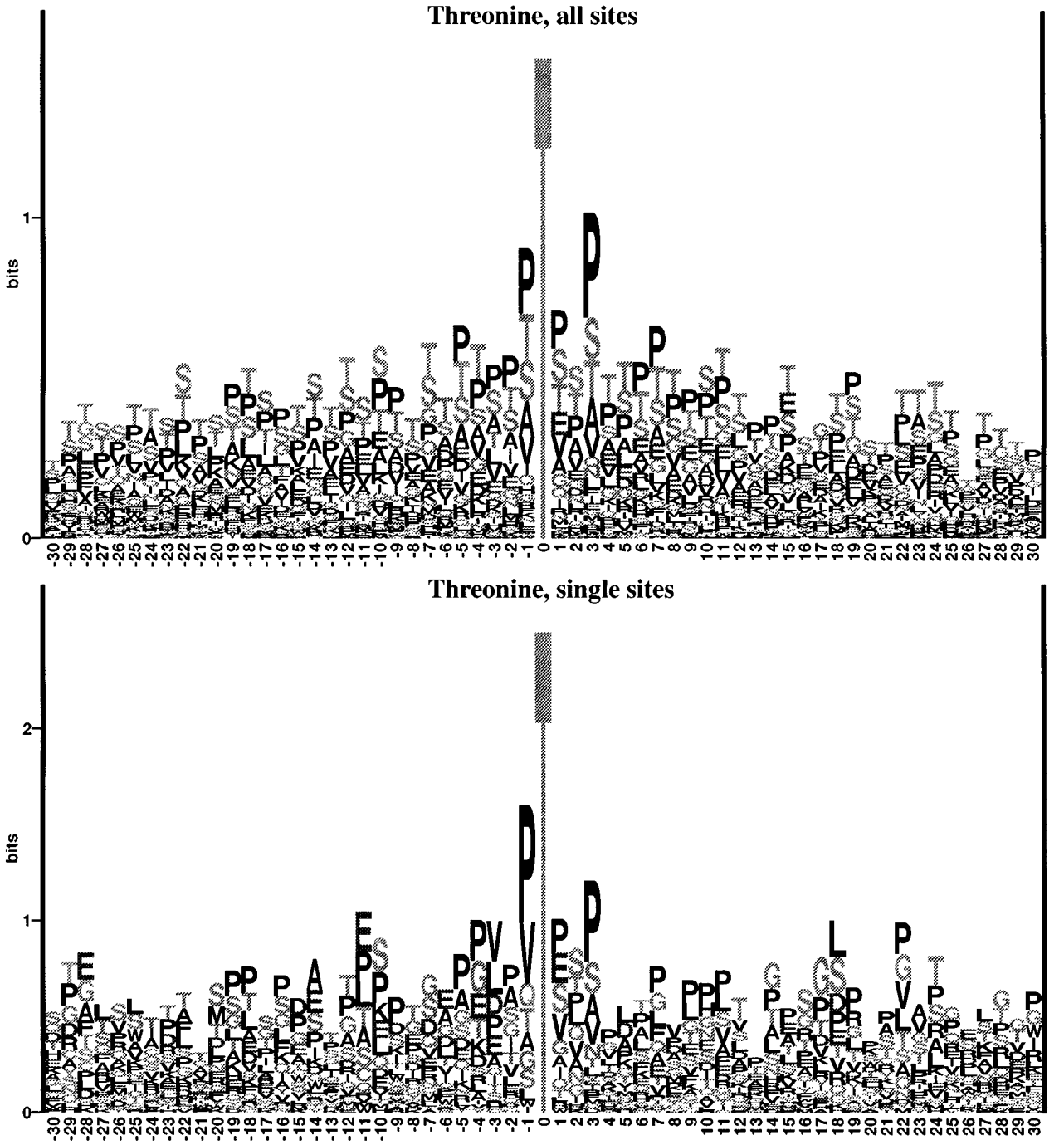

Figure 2. Shannon information content for O-glycosylation sites shown as sequence logos [51]. All sites were aligned with the glycosylated serine or threonine residues at position 0. Logos for threonine sites (a) and serine (b) are shown. Lower panels are logos for non-clustered single sites, defined as not having any glycosylated site 10 residues upstream or downstream. There were 23 non-clustered serine and 51 non-clustered threonine sites. The statistics for single serine sites are at present poor. The logos reflect the residues favoured on specific positions by the GalNAc transferase. The height of the central serine/threonine has been rescaled to magnify the context and is thus non-informative. The neutral and polar amino acids are shown in green, the charged basic in blue, the charged acidic in red and the neutral and hydrophobic in black.

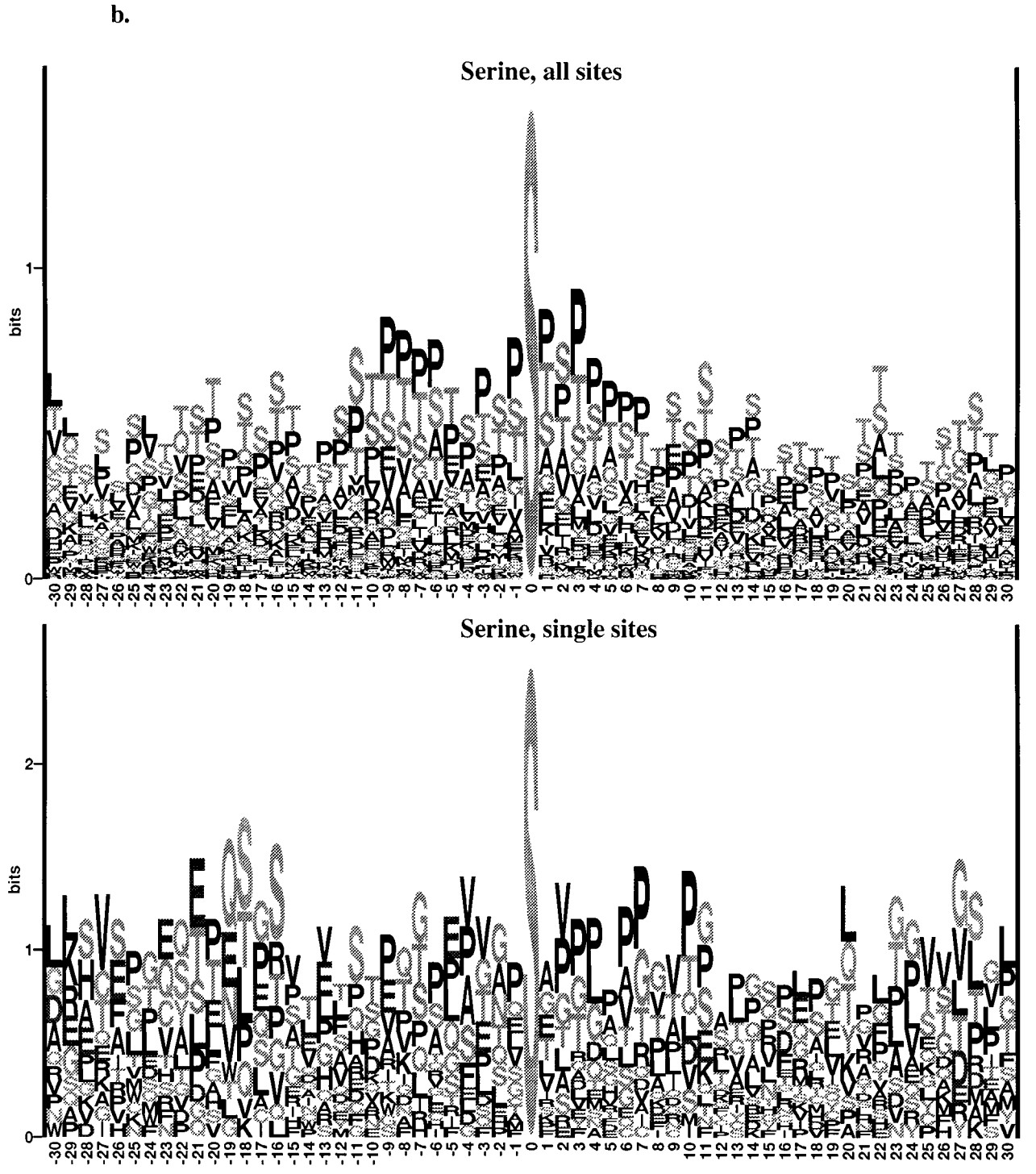

Figure 2. (Continued)

presented by Elhammer *et al.* [38]. This may be caused by overfitting. The vector projection method has a reasonable sensitivity, but a very low specificity as five times as many false positives compared to true positives are predicted.

Testing this new algorithm and our previous version [14] against the complete data set (training performance) revealed a significant increase in sensitivity (96% vs. 84%) as well as correlation coefficient (0.74 vs 0.69).

It is considerably less problematic to predict threonine than serine glycosylation sites, which may be a result of the poorer representation on serine sites in the data base (113 vs 186 sites). This difference may also reflect a genuine feature

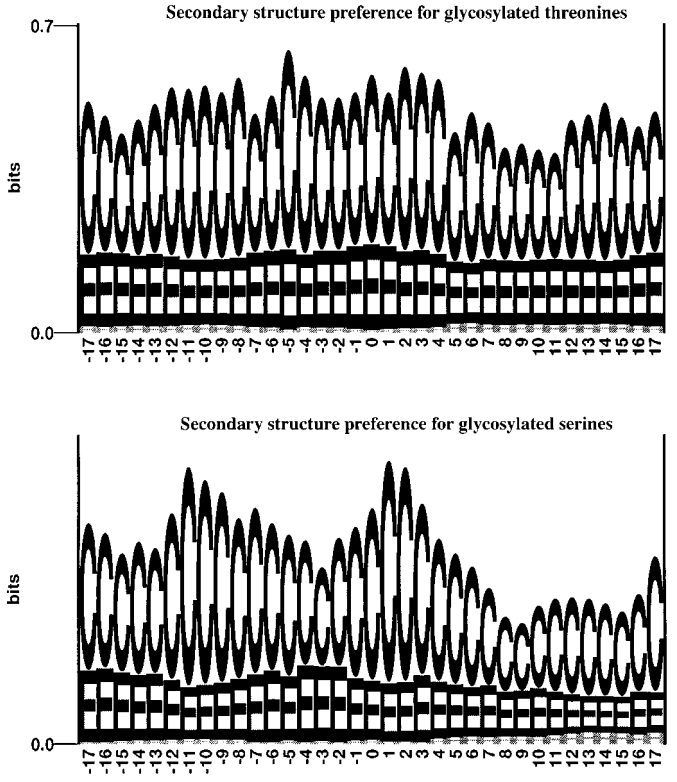

Figure 3. The secondary structures of the glycosylated residues at position 0 as predicted by the PHD method [53, 66, 67]. 68% of the sites were predicted to be in coil (C, black) regions, while 30% were predicted to be in extended β-strand conformation (E, blue). Only 2% were predicted to be in an α-helical structure (H, green). As seen in the logos the preference for coil extends upstream and downstream relative to the glycosylated site indicating that the O-glycosylation sites are situated in surface exposed loops with no particular secondary structure. The heights of the letters reflect the predicted frequency of coil, β-strand and α-helix at that position.

of the GalNAc transferases which *in vitro* glycosylates serine acceptor motifs much less efficiently than threonine motifs [16, 26, 38, 69, 70].

Prediction of the O-glycosylation sites in HIV-1, HIV-2 and SIV envelope glycoproteins

Glycosidase, lectin and antibody binding studies have shown that HIV-1 gp120 is modified by O-linked glycosylation [42, 43] and it has been suggested that these carbohydrates may act as neutralization epitopes [41, 71]. We therefore predicted the O-glycosylation sites in 29 strains of HIV-1 gp120, 9 strains of HIV-2 gp120 and 9 strains of SIV gp10. As seen in Figure 4, we consistently found strong O-glycosylation signals in the first variable region V1 of all three distinct but evolutionary related viruses. Despite comparable content of hydroxy amino acids, SIV gp110 was predicted to be much more O-glycosylated (11.3 site per strain; range 4–18) than HIV-1 (1.2 site per strain; range 0–5). It seems that HIV-1 gp120 O-glycosylation is much less conserved than O-glycosylated, while all SIV strains contained predicted O-glycosylation sites. The SIV strain with fewest O-glycosylation sites (4) was SIV-chimpanzee which is most closely related to HIV-1. The relative lack of O-glycosylation in the V1 region of HIV-1 gp120 may evolutionarily be replaced by N-linked glycosylation sites as most HIV-1 gp120 strains contain four N-linked glycosylation consensus signals in the V1 region, while most SIV gp110 strains contain only two N-linked glycosylation consensus signals in the V1 region. Overbaugh *et al.* [44] have reported that progression to Simian AIDS in macaques is followed by accumulation of regions rich in clustered serine and threonines and thereby putative O-glycosylation motifs in the V1 region of SIV gp110. Many of these Ser/Thr rich motifs have a very high predicted potential for being O-glycosylated (see Table 4). Recently, Overbaugh *et al.* [72] have by MALDFI-TOF mass spectrometry and differences in electrophoretic mobility after O-glycanase treatment directly demonstrated that these Ser/Thr rich domains in SIV *env* V1 are in fact O-glycosylated. These findings directly confirm the statistical trend in our predictions. The accumulation of O-glycosylation signals in the V1 region made these viruses more resistant to antibody neutralization suggesting that their emergence was a result of antibody driven selection, which makes the virus capable of escaping recognition by the humoral immune system [72]. No sites are predicted in the V3 loop (the single site in the V3 region depicted in Figure 4 with a potential of 0.6416 is below the local surface derived threshold). Neither are any sites predicted in the V3 loop by the Elhammer algorithm [38]. Testing 385 different V3 loop sequences extracted from GenBank revealed that only 14 (4%) were predicted to be O-glycosylated. As the *in vitro* neutralization of HIV-1 infection by O-linked carbohydrate specific monoclonal antibodies is not abrogated by mutating the possible sites in the V3 loop to alanine thereby deleting the possible signals for O-glycosylation [73], both prediction and experiment indicate that the O-glycosylation site in gp120 mediating neutralization is located outside the V3 loop.

Discussion

As several highly homologous isoforms of the GalNAc transferase have been cloned [10, 26–36] from different mammalian species and tissues, and as these may have different specificities, it is remarkable that as many as 83% of the sites can be reliably predicted. This may reflect that the sites these enzymes glycosylate are uniformly distributed in the data base, but it may also reflect a large degree of overlapping specificities for this enzyme family [10, 27]. More accurate prediction requires that these enzymes are expressed and tested selectively against a large panel of acceptor peptides *in vivo* using reporter proteins as pioneered by Nehrke *et al.* [19] in a system where all but one

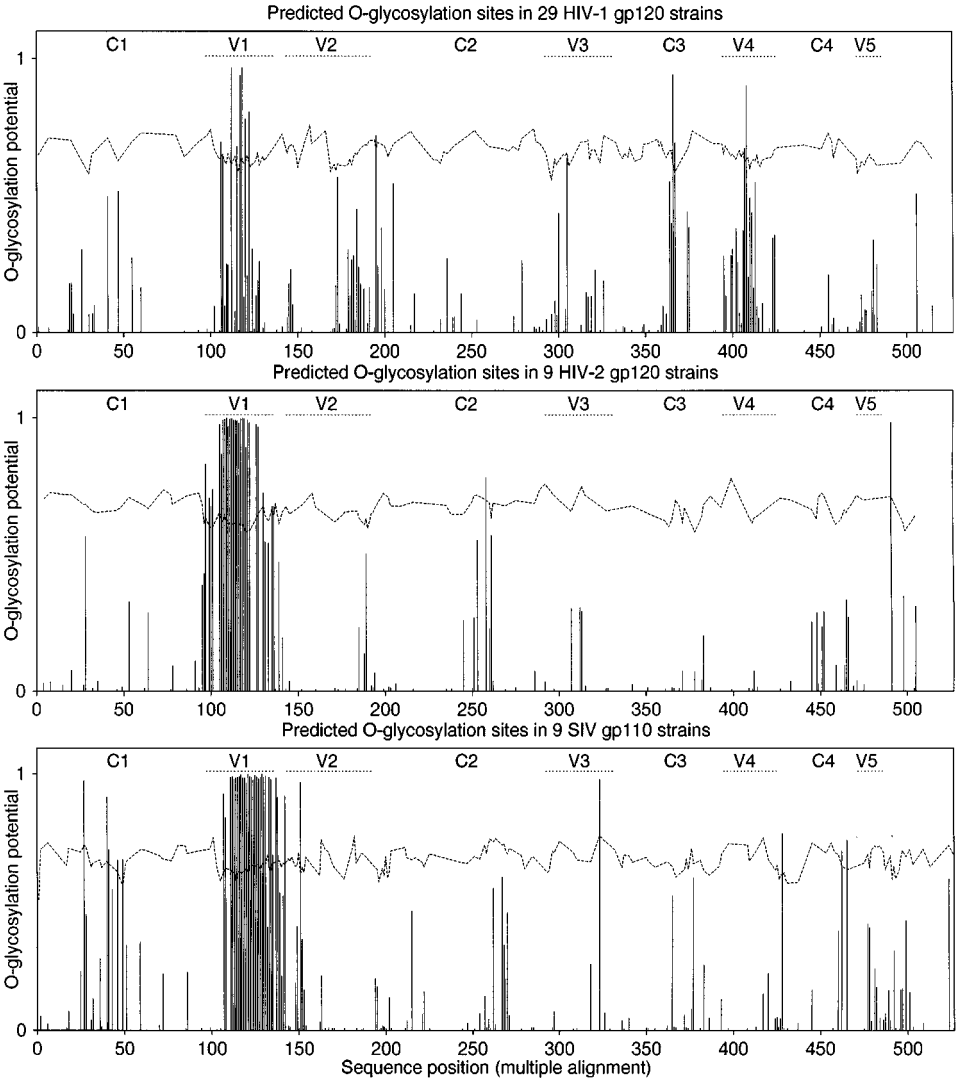

Figure 4. Predicted O-glycosylation sites in multiple aligned gp120 sequences from 29 different HIV-1 strains superimposed on the same graph (upper panel). The X-axis represents the gp120 sequence position in the multiple alignment (without signal peptide) and the Y-axis the predicted O-glycosylation potential at that position. All values above the averaged surface accessibility derived threshold (dotted line) can be regarded as O-glycosylated. Below is plotted the predicted mucin type O-glycosylation sites in 9 HIV-2 gp120 sequences (middle) and in 9 SIV gp110 sequences (lower panel). Variable regions in HIV-1 gp120 are indicated by bars. For HIV-1 the gp120 serine/threonine content was 15.6% and 34 sites were predicted (1.6%). For HIV-2 the gp120 serine/threonine content was 16.3% and 71 sites were predicted (10.1%). For SIV the gp110 serine/threonine content was 16.1% and 102 sites were predicted (14.0%).

GalNAc transferase have been silenced in order to delineate their subtle differences in specificity.

The fact that the acceptor motifs for serine were less recognizable by the networks and less efficiently glycosylated *in vitro* [16, 26, 38, 69, 70, 74] by purified GalNAc transferase, may indicate that this enzyme family originally was 'designed' for glycosylation of threonine residues. One may speculate that glycosylation of serine sites originally was a stereo-chemical failure of the enzyme to distinguish between the very similar hydroxyl side chains of serine and threonine.

Our finding of conserved O-glycosylation sites in V1 of the SIV *env* glycoprotein is experimentally confirmed by the recent work of Overbaugh *et al.* [72]. The O-glycosylation sites in V1 of HIV-2 also seems to be conserved. However O-glycosylation of HIV-1 is not predicted to be conserved as we find at least one positive signal in only 18 of the 29 HIV-1 gp120 strains examined. In these the O-glycosylation sites were not all located in the V1 but also in the C3, V4 and C2 domains (Figure 4 and Table 4). For these 18 strains 1 to 5 sites are predicted which correspond to the experimentally estimated 3–8 O-linked oligosaccharides in HIV-1 gp120 [42, 43]. However, in these studies only three laboratory adapted strains have been examined. Conservation of O-glycosylation sites in gp120 from an array of primary HIV-1 isolates cultured in human leukocytes have yet to be

Table 4. The 20 unique sites with highest potential for mucin type O-glycosylation in HIV-1, HIV-2 and SIV envelope glycoproteins. The strain name refers to the Swiss-prot entry code. Sites with identical sequence to the shown have been omitted. Numbering are without signal peptide.

	Predicted sites	*Sub-sequence*	*Potential*	*Region*
HIV-1 strain				
ENV_HV1RH	Thr-112	NGTNV-T-SSSGG	0.9689	V1
ENV_HV1SC	Thr-110	RNDTS-T-NATNT	0.9678	V1
ENV_HV1ND	Ser-331	ITFKP-T-SGGDP	0.9441	C3
ENV_HV1SC	Thr-116	NATNT-S-SSNRG	0.9398	V1
ENV_HV1Y2	Thr-110	RNATN-T-TSSSW	0.9376	V1
ENV_HV1MA	Thr-378	RLSNS-T-ESTGS	0.9030	V4
ENV_HV1J3	Thr-t14	PNATN-S-TSSGG	0.9027	V1
ENV_HV1Z2	Ser-332	IIFKP-T-SGGDP	0.8553	C3
ENV_HV1SC	Thr-115	TNATN-T-TSSNR	0.8252	V1
ENV_HV1Y2	Thr-111	NATNT-T-SSSWE	0.8210	V1
ENV_HV1BR	Thr-115	TNSSN-T-NSSSG	0.8081	V1
ENV_HV1EL	Ser-333	IKFKP-S-SGGDP	0.7887	C3
ENV_HV1Z6	Ser-334	IIFKP-S-SGGDA	0.7834	C3
ENV_HV1JR	Ser-109	TNTTS-S-SEGMM	0.7818	V1
ENV_HV1JR	Thr-106	VNATN-T-TSSSE	0.7813	V1
ENV_HV1OY	Thr-106	CTDVN-T-TSSSL	0.6981	V1
ENV_HV1ND	Ser-332	TFKPS-S-GGDPE	0.6959	C3
ENV_HV1OY	Thr-115	SLRNA-T-NTTSS	0.6821	V1
ENV_HV1H2	Thr-372	NSTWS-T-EGSNN	0.6752	V4
ENV_HV1B1	Ser-169	TSCNT-S-VITQA	0.6695	C2
HIV-2 strain				
ENV_HV2RO	Ser-108	NTTSK-S-TSTTT	0.9993	V1
ENV_HV2BE	Ser-109	NPRTS-S-STTSR	0.9990	V1
ENV_HV2BE	Ser-110	PRTSS-S-TTSRP	0.9988	V1
ENV_HV2ST	Ser-103	AKNTT-S-TPTTT	0.9968	V1
ENV_HV2RO	Ser-110	TSKST-S-TTTTT	0.9963	V1
ENV_HV2G1	Ser-105	TTTTG-S-TTGMS	0.9960	V1
ENV_HV2D2	Ser-102	PGNAS-S-TTTTK	0.9956	V1
ENV_HV2RO	Thr-113	STSTT-T-TTPTD	0.9922	V1
ENV_HV2BE	Thr-112	TSSST-T-SRPPT	0.9922	V1
ENV_HV2RO	Ser-106	GNNTT-S-KSTST	0.9919	V1
ENV_HV2CA	Ser-110	RTTTP-S-TAKEA	0.9918	V1
ENV_HV2ST	Thr-102	TAKNT-T-STPTT	0.9910	V1
ENV_HV2D1	Thr-102	SGTTA-T-PSPPN	0.9909	V1
ENV_HV2CA	Thr-106	TTMIR-T-TTPST	0.9909	V1
ENV_HV2D1	Thr-100	ITSGT-T-ATPSP	0.9907	V1
ENV_HV2ST	Thr-108	STPTT-T-TTANT	0.9893	V1
ENV_HV2D2	Thr-105	ASSTT-T-TKPTT	0.9893	V1
ENV_HV2ST	Thr-106	TTSTP-T-TTTTA	0.9891	V1
ENV_HV2CA	Thr-107	TMIRT-T-TPSTA	0.9888	V1
ENV_HV2BE	Thr-102	QGNTT-T-PNPRT	0.9884	V1
SIV strain				
ENV_SIVSP	Ser-118	TTTQA-S-TTPTS	0.9996	V1
ENV_SIVAT	Ser-108	TTTPK-S-TGLPC	0.9990	V1
ENV_SIVAG	Ser-104	TTTPK-S-TTAST	0.9984	V1
ENV_SIVAG	Ser-108	KSTTA-S-TTNIT	0.9970	V1
ENV_SIVM1	Ser-110	STTTA-S-TTTTT	0.9952	V1
ENV_SIVAG	Thr-99	NSSEP-T-TTPKS	0.9917	V1
ENV_SIVS4	Thr-109	TAITT-T-ATPSV	0.9913	V1
ENV_SIVM1	Thr-107	TKSST-T-TASTT	0.9913	V1
ENV_SIVAT	Thr-103	RATTP-T-TTPKS	0.9908	V1

Table 4. (Continued)

	Predicted sites	*Sub-sequence*	*Potential*	*Region*
ENV_SIVMK	Thr-110	STTIT-S-AAPTS	0.9895	V1
ENV_SIVM1	Thr-108	KSSTT-S-ASTTT	0.9888	V1
ENV_SIVSP	Thr-104	WGLTG-T-PAPTT	0.9884	V1
ENV_SIVM1	Ser-105	GLTKS-S-TTTAS	0.9883	V1
ENV_SIVSP	Thr-120	TQAST-T-PTSPI	0.9881	V1
ENV_SIVM1	Thr-114	ASTTT-T-TTAKS	0.9877	V1
ENV_SIVSP	Thr-122	ASTTP-T-SPITA	0.9875	V1
ENV_SIVS4	Thr-108	TTAIT-T-TATPS	0.9875	V1
ENV_SIVA1	Thr-99	LKGSA-T-STPAT	0.9871	V1
ENV_SIVM1	Thr-113	TASTT-T-TTTAK	0.9858	V1
ENV_SIVSP	Thr-115	TQTTT-T-QASTT	0.9857	V1

unambiguously demonstrated. If O-glycosylation is not a conserved feature of HIV-1 gp120, the hope of utilizing O-linked carbohydrates as a conserved immunogen in a HIV-1 vaccine [41] is somewhat reduced. In agreement with the Elhammer method [38] we do not find any potential O-glycosylation signals in the V3 loop, which is the principal neutralizing determinant of HIV-1. Secondly, sensitivity to neutralization to O-linked carbohydrate specific antibodies is not abrogated by deletion of all possible O-glycosylation sites in the V3 loop, which indicates that these carbohydrate neutralization epitopes are located outside the V3 loop of gp120 [73]. These studies are in conflict with the study of Bennet *et al.* [10, 35], who found that the human GalNAc-T3 transferase *in vitro* could O-glycosylate a 15-mer peptide with a sequence derived from a part of the V3 loop. However, there might exist large differences between *in vitro* O-glycosylation of peptides by a partly purified recombinant transferase and *in vivo* O-glycosylation of a real protein in a cell as recently shown by Nehrke *et al.* [19]. Secondly, the GalNAc-T3 transferase is preferentially expressed in pancreas and testis but not in leukocytes, nor spleen or thymus [35], which are the only tissues where HIV is produced. This makes GalNAc-T3 transferase mediated O-glycosylation of HIV-1 gp120 in the V3 loop *in vivo* less likely. Other retroviral *env* glycoproteins such as gp71 Friend murine leukaemia virus have also been shown to be O-glycosylated [75]. For this glycoprotein our method correctly identifies the exact location of 5 of the 7 O-linked glycosylation sites.

By using surface modulated threshold and updating the data set with new experimentally verified glycosylation sites the accuracy of our method has increased significantly. In future, specific neural networks have to be developed for each isoform of the GalNAc transferase if these isoforms prove to have marked differences in specificity patterns. Even when each transferase has been selectively expressed in an *in vivo* system, a complete description of its specificity requires testing 5×10^{11} different sequences assuming a nine-mer as acceptor. Another more realistic approach is to create a database of two hundred specific GalNAc transferase acceptors and non-acceptors and a primary neural network could be trained to recognize good acceptors among random sequences. These could then be synthesized and tested and the data used for a second network system, which then will gain accuracy. Such an 'intelligent iterative data sampling' of the otherwise experimentally intractable sequence space could significantly reduce the time used for a reasonably accurate description of the specificity of the GalNAc transferases. We therefore hope that such a close cooperation between bioinformatics and experimental glycobiology will be realized in the near future for the benefit of both. Until then we hope that our tool will prove useful for identifying putative mucin type O-glycosylation sites in recently sequenced glycoproteins and for designing efficient GalNAc transferase acceptor peptides.

Publicly available E-mail and WWW server

The method is publicly available as a WWW tool at http://www.cbs.dtu.dk/services/NetOGlyc/. Using E-mail forward sequences to netOglyc@cbs.dtu.dk, send the word 'help' to receive information on input and output formats. The users are encouraged to feedback any experimental confirmation or falsification of the predictions. Any new information regarding new glycoproteins with verified sites are also highly welcome. Both can be used to retrain the networks to increased performance.

Acknowledgements

This work was supported by The Danish National Research Foundation, The Danish 1991 Pharmacy Foundation, and the Danish Medical Research Council. Kristoffer Rapacki is thanked for excellent technical assistance.

References

1 Hounsell E, Davies M, Renouf D (1996) *Glycoconj J* **13**: 19–26.

2 Varki A (1993) *Glycobiology* **3**: 97–130.
3 Jentoft N (1990) *Trends in Biochemical Sciences* **15**: 291–4.
4 Hart GW (1992) *Curr Opin Cell Biol* **4**: 1017–23.
5 Fukuda M (1991) *Glycobiology* **1**: 337–356.
6 Kinloch RA, Sakai Y, Wassarman PM (1995) *Proc Natl Acad Sci USA* **92**: 263–67.
7 Muramatsu T (1993) *Glycobiology* **3**: 294–6.
8 Strous GJ, Dekker J (1992) *Crit Rev Biochem Mol Biol* **27**: 57–92.
9 Carraway K, Hull SR (1991) *Glycobiology* **1**: 131–8.
10 Clausen H, Bennett E (1996) *Glycobiology* **6**: 635–46.
11 Gooley AA, Williams KL (1994) *Glycobiology* **4**: 413–17.
12 Wilson IBH, Gavel Y, Heijne GV (1991) *Biochem J* **275**: 529–34.
13 Wang Y, Abernethy JL, Eckhardt AE, Hill RL (1992) *J Biol Chem* **267**: 12709–16.
14 Hansen JE, Lund O, Engelbrecht J, Bohr H, Nielsen JO, Hansen JES, Brunak S (1995) *Biochem J* **308**: 801–13.
15 O'Connell BC, Hagen FK, Tabak LA (1992) *J Biol Chem* **267**: 25010–18.
16 Wang Y, Agwral N, Eckhard AE, Stevens RD, Hill R (1993) *J Biol Chem* **268**: 22979–83.
17 Nishimori I, Johnson NR, Sanderson SD, Perini F, Mountjoy K, Cerny R, Gros ML, Hollingsworth MA (1994) *J Biol Chem* **269**: 16123–30.
18 O'Connell BC, Tabak LA, Ramasubbu N (1991) *Biochem Biophys Res Com* **180**: 1024–30.
19 Nehrke K, Hagen FK, Tabak LA (1996) *J Biol Chem* **271**: 7061–65.
20 Roth J, Wang Y, Eckhardt AE, Hill RL (1994) *Proc Natl Acad Sci USA* **91**: 8935–9.
21 Asker N, Baeckstrom D, Axelsson MA, Carlstedt I, Hansson GC (1995) *Biochem J* **308**: 873–80.
22 Hansen JE, Lund O, Rapacki K, Clausen H, Mosekilde E, Nielsen JO, Hansen JES (1994) In *Protein Structure by Distance Analysis* (Bohr H. and Brunak S eds), pp 247–54. Amsterdam: IOS Press.
23 Dahms NM, Hart GW (1986) *J Biol Chem* **261**: 13186–96.
24 Yamada T, Uyeda A, Takao T, Shimonishi Y, Matsushima M, Kikuchi M (1995) *Eur J Biochem* **230**: 965–70.
25 Elliott S, Bartley T, Delorme E, Derby P, Hunt R, Lorenzini T, Parker V, Rohde M, Stoney K (1994) *Biochemistry* **33**: 11237–45.
26 Hagen FK, Van-Wuyckhuyse B, Tabak LA (1993) *J Biol Chem* **268**: 18960–5.
27 Zara J, Hagen FK, Hagen KGT, Van-Wuyckhuyse BC, Tabak LA (1996) *Biochem Biophys Res Commun* **228**: 38–44.
28 Sorensen T, White T, Wandall HH, Kristensen AK, Roepstorff P, Clausen H (1995) *J Biol Chem* **270**: 24166–73.
29 White T, Bennett EP, Takio K, Sorensen T, Bonding N, Clausen H (1995) *J Biol Chem* **270**: 24156–65.
30 Homa FL, Hollander T, Lehman DJ, Thomsen DR, Elharmmer AP (1993) *J Biol Chem* **268**: 12609–16.
31 Meurer JA, Naylor JM, Baker CA, Thomsen DR, Homa FL, Elhammer AP (1995) *J Biochemistry* **118**: 568–74.
32 Hennet T, Hagen F, Tabak L, Marth J (1995) *Proc Natl Acad Sci USA* **92**: 12070–74.
33 Hagen FK, Gregoire CA, Tabak LA (1995) *Glycoconj J* **12**: 901–9.
34 Yoshida A, Hara T, Ikenaga H, Takeuchi M (1995) *Glycoconj J* **12**: 824–82.
35 Bennett E, Hassan H, Clausen H (1996) *J Biol Chem* **271**: 17006–12.
36 Meurer J, Drong R, Homa F, Slightom J, Elhammer A (1996) *Glycobiology* **6**: 231–41.
37 Hansen JE, Lund O, Rapacki K, Brunak S (1997) *Nucleic Acid Research* **25**: 278–82.
38 Elhammer AP, Poorman RA, Brown E, Maggiora LL, Hoogerheide JG, Kzdy FD (1993) *J Biol Chem* **268**: 10029–38.
39 Chou KC (1995) *Protein Sci* **4**: 1365–83.
40 Chou KC, Zhang CT, Kezdy FJ, Poorman RA (1995) *Proteins* **21**: 118–26.
41 Hansen JE, Nielsen C, Arendrup M, Olofsson S, Mathiesen L, Nielsen JO, Clausen H (1991) *J Virol* **65**: 6461–7.
42 Hansen JE, Clausen H, Hu SL, Nielsen JO, Olofsson S (1992) *Arch Virol* **126**: 11–20.
43 Bernstein HB, Tucker SP, Hunter E, Schutzbach JS, Compans RW (1994) *J Virol* **68**: 463–8.
44 Overbaugh J, Rudensey LM (1992) *J Virol* **66**: 5937–48.
45 Pearson WR (1994) *Methods Mol Biol* **25**: 365–89.
46 Hobohm U, Scharf M, Schneider R, Sander C (1992) *Protein Sci* **1**: 409–17.
47 Connolly ML (1983) *Science* **221**: 709–13.
48 Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH (1985) *Science* **229**: 834–8.
49 Blom N, Hansen JE, Blaas D, Brunak S (1996) *Protein Sci* **5**: 2203–16.
50 Shannon CE (1948) *Bell System Tech J* **27**: 379–23, 623–56.
51 Schneider TD, Stephens RM (1990) *Nucleic Acids Res* **18**: 6097–100.
52 Presnell SR, Cohen FE (1993) *Annu Rev Biophys Biomol Struct* **22**: 283–98.
53 Rost B, Sander C (1994) *Proteins* **19**: 55–72.
54 Bohr H, Bohr J, Brunak S, Cotterill RMC, Lautrup B, Nørskov L, Olsen O, Hsen SB, Peters B (1988) *FEBS Lett* **241**: 223–8.
55 Qian N, Sejnowski TJ (1988) *J Mol Biol* **202**: 865–84.
56 Holley LH, Karplus M (1989) *Proc Natl Acad Sci* **86**: 152–6.
57 MacGregor MJ, Flores TP, Sternberg MJE (1989) *Protein Engineering* **2**: 521–6.
58 Kneller DG, Cohen FE, Langridge R (1990) *J Mol Biol* **214**: 171–82.
59 Rost B, Sander C (1994) *Proteins* **20**: 216–26.
60 Rose B, Casadio R, Fariselli P, Sander C (1995) *Protein Sci* **4**: 521–33.
61 Korning PG, Hebsgaard S, Tolstrup N, Engelbrecht J, Rouzé P, Brunak S (1996) *Nucl Acids Res* **24**: 3439–52.
62 Brunak S, Engelbrecht J, Knudsen S (1991) *J Mol Biol* **220**: 49–65.
63 Rumelhart DE, Hinton GE, Williams RJ (1986) *Nature* **323**: 533–6.
64 Mathews BW (1975) *Biochim Biophys Acta* **405**: 442–51.
65 Nielsen H, Engelbrecht J, Brunak S, von Heijne G (1997) *Protein Eng* **10**: 1–6.
66 Rost B, Sander C (1993) *J Mol Biol* **232**: 584–99.
67 Rost B, Sander C, Schneider R (1994) *Comput Appl Biosci* **10**: 53–60.

68 Hollosi M, Perczel A, Fasman GD (1990) *Biopolymers* **29**: 1549–64.
69 Wragg S, Hagen FK, Tabak LA (1995) *J Biol Chem* **270**: 16947–54.
70 Brockhausen I, Toki D, Brockhausen J, Peters S, Bielfeldt T, Kleen A, Paulsen H, Meldal M, Hagen F, Tabak LA (1996) *Glycoconj J* **13**: 849–56.
71 Hansen JE, Clausen H, Nielsen C, Teglbjaerg LS, Hansen LL, Nielsen CM, Dabelsteen E, Mathiesen L, Hakomori SI, Nielsen JO (1990) *J Virol* **64**: 2833–40.
72 Chackerian B, Rudensey L, Overbaugh J (1997) *J Virol* **71**: 7719–27.
73 Hansen JES, Jansson B, Gram GJ, Clausen H, Nielsen JO, Olofsson S (1996) *Arch Virol* **141**: 291–300.
74 O'Connell B, Tabak L (1993) *J Dent Res* **72**: 1554–8.
75 Geyer R, Dabrowski J, Dabrowski U, Linder D, Schlueter M, Schott HH, Stirm S (1990) *Eur J Biochem* **187**: 95–110.
76 Barker WC, George DG, Mewes HW, Pfeiffer F, Tsugita A (1993) *Nucleic Acids Res* **21**: 3089–96.
77 Bairoch A, Apweiler R (1996) *Nucleic Acids Res* **24**: 21–5.
78 Dahr W, Beyreuther K (1985) *Biol Chem Hoppe-Seyler* **366**: 1067–70.
79 Pisano A, Redmond JW, Williams KL, Gooley AA (1993) *Glycobiology* **3**: 429–35.
80 Murayama J, Yamashita T, Tomita M, Hamada A (1983) *Biochim Biophys Acta* **742**: 477–83.
81 Murayama JI, Tomita M, Hamada A (1982) *J Membr Biol* **64**: 205–15.
82 Honma K, Tomita M, Hamada A (1980) *J Biochem* **88**: 1679–91.
83 Killeen N, Barclay AN, Willis AC, Williams AF (1987) *EMBO J* **6**: 4029–34.
84 Schmid K, Heidiger MA, Brossmer R, Collins JH, Haupt H, Marti T, Offner GD, Schaller J, Takagaki K, Walsh MT, Schwick HG, Rose FS, Remold-O'Donnell E (1992) *Proc Natl Acad Sci* **89**: 663–7.
85 Putnam FW, Liu YSV, Low TLK (1979) *J Biol Chem* **254**: 2865–74.
86 Robinson EA, Appella E (1979) *J Biol Chem* **254**: 11418–30.
87 Takayasu T, Suzuki S, Kametani F, Takahashi N, Shinoda T, Okuyama T, Munekata E (1982) *Biochem Biophys Res Commun* **105**: 1066–71.
88 Kaushansky K, Lopez JA, Brown CB (1992) *Biochemistry* **31**: 1881–6.
89 Takeuchi M, Kobata A (1991) *Glycobiology* **1**: 337–46.
90 Birken S, Agosto G, Amr S, Nisula B, Cole L, Lewis J, Canfield R (1988) *Endocrinology* **122**: 2054–6.
91 Morgan FJ, Birken S, Canfield RE (1975) *J Biol Chem* **250**: 5247–58.
92 Seidah NG, Chretien M (1981) *Proc Natl Acad Sci* **78**: 4236–40.
93 Bennett HP, Seidah NG, Benjannet S, Solomon S, Chretien M (1986) *Int J Pept Protein Res* **27**: 306–13.
94 Fiat AM, Jolles J, Aubert JP, Loucheux-Lefebvre MH, Jolles P (1980) *Eur J Biochem* **111**: 333–9.
95 Yan SB, Wold F (1984) *Biochemistry* **23**: 3759–65.
96 Schaller J, Marti T, Rosselet SJ, Kampfer U, Rickli EE (1987) *Fibrinolysis* **1**: 91–102.
97 Robb RJ, Kutny RM, Panico M, Morris HR, Chowdhry V (1984) *Proc Natl Acad Sci* **81**: 6486–90.
98 Lottspeich F, Kellermann J, Henschen A, Foertsch B, Muller-Esterl W (1985) *Eur J Biochem* **152**: 307–14.
99 Kellermann J, Lottspeich F, Henschen A, Muller-Esterl W (1986) *Adv Exp Med Biol* **198**: 85–9.
100 Hill HD, Schwyzer M, Steinman HM, Hill RL (1977) *J Biol Chem* **252**: 3799–804.
101 Takahashi N, Takahashi Y, Putnam FW (1985) *Proc Natl Acad Sci* **82**: 1906–10.
102 Brewer Jr HB, Shulman R, Herbert P, Ronan R, Wehrly K (1974) *J Biol Chem* **249**: 4975–84.
103 Kellerman J, Lottspeich F, Geiger R, Deutzmann R (1989) *Adv Exp Med Biol* **247A**: 519–25.
104 Walsh KA, Titani K, Takio K, Kumar S, Hayes R, Petra PH (1986) *Biochemistry* **25**: 7584–90.
105 Lopez Otin C, Grubb A, Mendez E (1984) *Arch Biochem Biophys* **228**: 544–54.
106 Hochstrasser K, Schonberger OL, Rossmanith I, Wachter E (1981) *Hoppe-Seyler's Z Physiol Chem* **372**: 1357–62.
107 Young JD, Tsuchiya D, Sandlin DE, Holroyde MJ (1979) *Biochemistry* **18**: 4444–8.
108 Titani K, Takio K, Handa M, Ruggeri ZM (1987) *Proc Natl Acad Sci* **84**: 5610–14.
109 Gejyo F, Chang JL, Burgi W, Zand Schmid K, Offner GD, Troxler R, Van Halbeek H, Dorland L, Gerwig G, Vliegenthart F (1983) *J Biol Chem* **258**: 4966–71.
110 Watzlawick H, Walsh MT, Yoshioka Y, Schmid K, Brossmer R (1992) *Biochemistry* **31**: 12198–203.
111 Perkins SJ, Smith KF, Amatayakuul S, Ashford D, Rademacher TW, Dwek RA, Lachmann PJ, Harrison RA (1990) *J Mol Biol* **214**: 751–63.
112 Bock SC, Skriver K, Nielsen E, Thogersen HC, Wiman B, Donaldson VH, Eddy RL, Marrinan J, Radziejwska E, Huber R (1986) *Biochemistry* **25**: 4292–301.
113 De Caro AM, Adrich Z, Fournet B, Capon C, Bonicel JJ, De Caro JD, Rovery M (1989) *Biochim Biophys Acta* **994**: 281–84.
114 Wernette-Hammond ME, Lauer S, Corsini A, Walker D, Taylor J, Rall SC (1989) *J Biol Chem* **264**: 9094–101.
115 Hayes GR, Enns CA, Lucas JJ (1992) *Glycobiology* **2**: 355–9.
116 Do SI, Cummings RD (1992) *Glycobiology* **2**: 345–53.
117 Adolf GR, Kalsner I, Ahorn H, Maurer Fogy I, Cantell K (1991) *Biochem J* **276**: 511–18.
118 Voigt CG, Maurer-Fogy I, Adolf GR (1992) *FEBS Lett* **314**: 85–8.
119 Fujiwara S, Shinkai H, Mann K, Timpl R (1993) *Matrix* **13**: 215–22.
120 Clogston CL, Hu S, Boone TC, Lu HS (1993) *J Chromatogr* **637**: 55–62.
121 Minamitake Y, Kodama S, Katayama T, Ada H, Tanaka S, Tsujtmoto M (1990) *J Biochem* **107**: 292–7.
122 Daughaday WH, Trivedi B, Baxter RC (1993) *Proc Natl Acad Sci* **90**: 5823–7.
123 Peters B, Krzesicki R, Perini F, Ruddon R (1989) *Endocrinology* **124**: 1602–12.
124 Pisano A, Jardine DR, Packer NH, Farnsworth V, Carson W, Cartier P, Redmond JW, Williams KL, Gooley AA, (1996) In *Techniques in Glycobiology* (Townsend R and Hotchkiss A eds). New York: Marcel Dekker Inc.
125 Gooley AA, Classon BJ, Marshcalek R, Williams KL (1991) *Biochem Biophys Res Com* **178**: 1194–1200.

126 Carlsson SR, Lycksell PO, Fukuda M (1993) *Arch Biochem Biophys* **304**: 65–73.
127 Calvete JJ, Muniz-Diaz E (1993) *FEBS Lett* **328**: 30–4.
128 Murayama JI, Utsumi H, Hamada A (1989) *Biochim Biophys Acta* **999**: 273–80.
129 Pisano A, Packer NH, Redmond JW, Williams KL, Gooley AA (1994) *Glycobiology* **4**: 837–44.
130 Shimizu N, Hara H, Sogabe T, Sakai H, Ihara I, Inoue H, Nakamura T, Shimizu S (1992) *Biochem Biophys Res Commun* **189**: 1329–35.
131 Schindler PA, Settineri CA, Collet X, Fielding CJ, Burlingame AL (1995) *Protein Sci* **4**: 791–803.
132 Stadie TR, Chai W, Lawson AM, Byfield PG, Hanisch FG (1995) *Eur J Biochem* **229**: 140–7.
133 Wang CS, Dashti A, Jackson KW, Yeh JC, Cummings RD, Tang J (1995) *Biochemistry* **34**: 10639–44.
134 Inoue K, Morita T (1993) *Eur J Biochem* **218**: 153–63.
135 Mizuochi T, Yamashita K, Fujikawa K, Titani K, Kobata A (1980) *J Biol Chem* **255**: 3526–31.

Received 21 February 1997, revised 26 May 1997, accepted June 1997