Prediction of O-glycosylation of mammalian proteins: specificity patterns of UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferase

Jan E. HANSEN,*†§ Ole LUND,*†‡ Jacob ENGELBRECHT,† Henrik BOHR,† Jens O. NIELSEN,* John-Erik S. HANSEN* and Søren BRUNAK†

*Laboratory for Infectious Diseases, Hvidovre Hospital, University of Copenhagen, DK-2650 Denmark, †Center for Biological Sequence Analysis, The Technical University of Denmark, DK-2800 Denmark, and ‡Department of Physics, The Technical University of Denmark, DK-2800 Denmark

The specificity of the enzyme(s) catalysing the covalent link between the hydroxyl side chains of serine or threonine and the sugar moiety *N*-acetylgalactosamine (GalNAc) is unknown. Pattern recognition by artificial neural networks and weight matrix algorithms was performed to determine the exact position of *in vivo* O-linked GalNAc-glycosylated serine and threonine residues from the primary sequence exclusively. The acceptor sequence context for O-glycosylation of serine was found to differ from that of threonine and the two types were therefore treated separately. The context of the sites showed a high abundance of proline, serine and threonine extending far beyond the previously reported region covering positions −4 through +4 relative to the glycosylated residue. The O-glycosylation sites were found to cluster and to have a high abundance in the N-terminal part of the protein. The sites were also found to have an increased preference for three different classes of β-turns. No simple consensus-like rule could be deduced for the complex glycosylation sequence acceptor patterns. The neural networks were trained on the hitherto largest data material consisting of 48 carefully examined mammalian glycoproteins comprising 264 O-glycosylation sites. For detection neural network algorithms were much more reliable than weight matrices. The networks correctly found 60–95% of the O-glycosylated serine/threonine residues and 88–97% of the non-glycosylated residues in two independent test sets of known glycoproteins. A computer server using E-mail for prediction of O-glycosylation sites has been implemented and made publicly available. The Internet address is NetOglyc@cbs.dtu.dk.

INTRODUCTION

Glycosylation serves a wide variety of functions in biology [1] and is involved in recognition [2], oncogenesis [3,4], and development [5]. Experimental biochemical analysis of potential O-glycosylation sites in glycoproteins at serine or threonine residues is difficult and time consuming, since it requires solid-phase Edman degradation [6]. The knowledge of the structure and function of O-glycosylated proteins could be improved if the exact location of their carbohydrate part could be deduced directly from the primary sequence. Computational prediction may therefore lead to a better understanding of O-glycosylated proteins in general. The sequence specificities of N- and O-linked glycosylation differ markedly. For N-linked glycosylation, the specificity of the enzyme can be formulated as the consensus triplet sequence Asn-Xaa-Ser/Thr, where Xaa can be any residue except proline [7]. Not all consensus triplets are actually glycosylated [8], but all known N-linked glycosylation sites are found within this triplet sequence [9]. This makes prediction of N-linked glycosylation from the primary sequence less problematic. The situation is far more complex for O-linked glycosylation since at least seven classes of O-linked glycosylation can be defined: (a) mucin type, where an *N*-acetylgalactosamine is α-1-linked to the side chain of serine or threonine residues in secreted or membrane-bound glycoproteins (for recent reviews see [10,11]); (b) intracellular type, where a single *N*-acetylglucosamine is β-1-linked to serine or threonine in nucleoplasmic and cytoplasmic glycoproteins (for a recent review see [12]); (c) proteoglycan type, where a xylose core (Gal-Gal-Xyl) is linked to serine in proteoglycan glycoproteins [13,14]; (d) collagen type, where galactose is linked to hydroxylysine in collagen [15]; (e) clotting factor type, where either fucose or a glucose core (Xyl-Glc or Xyl-Xyl-Glc) is linked to serine in factor VII and IX, protein Z and others [16,17]; (f) fungal type, where mannose is linked to serine or threonine in yeast and fungal glycoproteins [18]; and (g) plant type, where arabinose is linked to hydroxyproline and galactose to serine in plant glycoproteins [19].

The enzymes initiating the biosynthesis of O-linked glycosylation by transfer of monosaccharides to specific acceptor residues are type specific. Since it is likely that acceptor specificities of these glycosyltransferases differ [6], the present analysis has been limited to the first most common group. Mucin-type O-linked glycosylation has no unique consensus acceptor sequence [20,21]. This becomes apparent by inspection of the mucin-type O-glycosylated sequences in Table 1. However, a few rules of thumb have been formulated: proline occurs at increased frequency at positions −1 and +3 relative to the glycosylation site, as found by Wilson et al. [20]. Glycoproteins are, however, also rich in this motif at non-glycosylated sites [20]. Elhammer et al. [22] extended this by *in vitro* experiments using bovine colostrum transferase and matrix statistics on a data set of 196 glycosylation sites. They found that a high abundance of not only proline, but also serine and threonine, at all positions from −4 to +4 favoured glycosylation. Gooley et al. [23] have proposed the sequence motif Xaa-Pro-Xaa-Xaa, where one Xaa only is a glycosylated threonine. However, O-glycosylation of threonine which is not associated with proline also occurs [23].

Pisano et al. [24] have suggested three tetrapeptide sequence motifs for O-glycosylation of threonine: Xaa-Pro-Xaa-Xaa, Thr(g)-Xaa-Xaa-Xaa [where Thr(g) and one Xaa only are glycosylated threonine residues], and, Xaa-Xaa-Thr(g)-Xaa, (where Xaa represents lysine or arginine). For serine the proposed

§ To whom correspondence should be addressed at: Center for Biological Sequence Analysis, The Technical University of Denmark, Building 206, DK-2800 Denmark.

Table 1 (a), training set

Entry	O-site sequence	Ref.
Position	54321-0+12345	
GFHUC (P)	---MW S TRSPN	[55]
glycophorin C	--MWS T RSPNS	
- human	MWSTR S PNSTA	
	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	[24]
glycophorin A	---SS T TGVAM	
- human	--SST T GVAMH	
	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	[56]
glycophorin	PHQIS S KLPTQ	
- dog	SSKLP T QAGFI	
	QAGFI S TEDPS	
	AGFIS T EDPSF	
	DPSFN T PSTRE	
	TREDP S GTMYQ	
GFHOE (P)	----Q T IATGS	[57]
glycophorin HA	-QTIA T GSPPI	
- horse	TIATG S PPIAG	
	PPIAG T SDLST	
	GTSDL S TITSA	
	TSDLS T ITSAA	
	DLSTI T SAATP	
	LSTIT S AATPT	
	ITSAA T PTFTT	
	ATPTF T TEQDG	
GFPGE (P)	----- T ETPVT	[58]
Glycophorin	TETPV T GEQGS	
- pig	TGEQG S ATPGN	
	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	[59]
Leukosialin	LPNTM T MLPFT	
- rat	TMLPF T PNSES	
	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	[60]
Leukosialin	---ST T AVQTP	
- human	TTAVQ T PTSGE	
	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	[61]
Ig alpha-1 chain	TPPTP S PSTPP	
- human.	PTPSP S TPPTP	
	TPPTP S PSCCH	
	PTPSP S CCHPR	
ALC_MOUSE (S)	LDVNC S GPTPP	[62]
Ig Alpha		
chain C		
- mouse		
DHHU (P)	PKAQA S SVPTA	[63]
Ig delta	KAQAS S VPTAQ	
chain C	ASSVP T AQPQA	
- human	SLAKA T TAPAT	
	LAKAT T APATT	
	TTAPA T TRNTG	
	TAPAT T RNTGR	
FQHUGM (P)	-APAR S PSPST	[64]
granulocyte-macrophage	PARSP S PSTQP	
colony-stimulating	RSPSP S TQPWE	
- human	SPSPS T QPWEH	
EPO_HUMAN (S)	PPDAA S AAPLR	[65]
erythropoietin		
- human		
KTHUB(P)	QDSSS S KAPPP	[66, 67]
choriogonadotropin	KAPPP S LPSPS	
beta chain	SLPSP S RLPGP	
- human	RLPGP S DTPIL	
CTHUP(P)	DEQPL T ENPRK	[68, 69]
corticotropin		
- human		
A16604 (S)	KIIIP T INTIA	[70]
kappa casein	ATVEP T PAPAT	
- human	TPAPA T EPTVD	
	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	[71]
beta casein	PPLTQ T PVVVP	
- bovine	LSLSQ S KVLPV	
PLBO (P)	ESSPL S TERMD	[72]
plasmin		
- bovine		

Entry	O-site sequence	Ref.
Position	54321-0+12345	
PLHU (P)	EELAP T APPEL	[72]
plasmin		
- human		
PLMN_PIG (S)	TTPPP T SGPTY	[72]
plasmin		
- porcine		
ICHU2 (P)	---AP T SSSTK	[73]
interleukin 2		
- human		
KGHUH1 (P)	IKEET T VSPPH	[74]
kininogen HMW I	SEDST T PSAQT	
- human	QTQEK T EGPTP	
	KTEGP T PIPSL	
	AKPGV T VTFSD	
	SDLIA T MMPPI	
	MMPPI S PAPIQ	
	IPDIQ T DPNGL	
	SEINP T TQMKE	
KNH1_BOVIN (S)	EGPVV T AQYEC	[75]
kininogen	MKTEG S TTVSL	
HMW I - bovin	KTEGS T TVSLP	
	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	[76]
mucin	----S S VPGES	
(fragment)	SVPGE S ATPQQ	
- sheep	PGESA T PQQPG	
	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	[77]
Leucine rich		
alpha 2 gp		
- human		
LPHUC3 (P)	PEVRP T SAVAA	[78]
apolipoprotein		
C-III		
- human		
KQHU (P)	QFVHV S ESFPH	[79]
tissue kallikrein	PGFNM S LLENH	
- human	EPENF S FPDDL	
HEMO_HUMAN (S)	----- T PLPPT	[77]
hemopexin		
- human		
BOHUS (P)	RPVLP T QSAHD	[80]
sex steroid binding		
protein - human		
HCHU (P)	-GPVP T PPDNI	[81, 82]
a-1-microglobulin	QEEEG S GGGQL	
MBBOB (PIR)	IVTPR T PPPSQ	[83]
myelin basic		
protein - bovine		

tetrapeptide sequence motif is: Ser(g)-Xaa-Xaa-Xaa, (where Xaa represents one glycosylated serine residue). These motifs account for all glycosylations within glycophorin A [24] and 16 out of 18 sites in bovine casein [6]. However, these motifs are also quite frequent at non-glycosylated sites which limits their predictive value (see below).

O'Connell et al. found that positions −6, −1 and +3 were of particular significance [25]. By *in vitro* glycosylation of model

Table 1 (b), test set

	Low Sim. test set	
Entry	O-site sequence	Ref.
Position	54321-0+12345	
NBHUIA (P) platelet gp. Ib a.chain - human	DKVRA T RTVVK	[84]
WOHU (P) a 2 HS gp. - human	QTQPV T SQPQP NEAVP T PVVDP TVVQP S VGAAA	[85, 86]
ITHUC1 (P) complement C1 inhibitor - human	GKVAT T VISKM ILEVS S LPTTN PTTNS T TNSAT ITANT T DEPTT TDEPT T QPTTE TTQPT T EPTTQ TTEPT T QPTIQ	[87, 88]
LITH_HUMAN (S) lithostathine - human	-PEAQ T ELPQA	[89]
APE_HUMAN (S) apolipoprotein E - human	RVRAA T VGSLA	[90]
JXHU (P) transferrin receptor - human	ERLAG T ESPVR	[91, 92]
IVHUA2 (S) interferon a2 - human	QGVGV T ETPLM	[93]
TNFB_HUMAN (S) lymphotoxin (TNF)	PGVGL T PSAAQ	[94]
MMMSND (P) nidogen - mouse	DYDLV T SHLGL VPRIL S PGYEA PGYEA T ERPRG PRGVP T ERTRS VPTER T RSFQL PPCLS T VAPPI GPVVP T AVIPL	[95]
A25093 (P) granulocyte colony stim. f. - human	PALQP T QGAMP	[96]
ENV_MLVFR (S) knob protein gp71 - mouse	PVSNS T PTMIS TPTMI S PSPTP TMISP S PTPTQ ISPSP T PTQPP PSPTP T QPPPA QALNL T NPDKT TNPDK T QECWL	[97]
IL5_HUMAN (S) interleukin 5 - human	---IP T EIPTS	[98]
IGHU2 (P) insulin-like growth factor II	VSTPP T VLPDN	[99]
TTHUAP (P) thyrotropin chain A	SRAYP T PLRSK	[100]

	High Sim. test set	
Entry	O-site sequence	Ref.
Position	54321-0+12345	Ref.
LMP1_HUMAN (S) lysosome associated membrane glycoprotein	EQDRP S PTTAP DRPSP T TAPPA RPSPT T APPAP PPAPP S PSPSP APPSP S PSPVP PSPSP S PVPKS	[101]
LMP2_HUMAN (S) lysosome associated membrane glycoprotein	DKDKT S TVAPT KDKTS T VAPTI STVAP T IHTTV APTIH T TVPSP PTIHT T VPSPT HTTVP S PTTTP TVPSP T TTPTP VPSPT T TPTPK PSPTT T PTPKE PTTTP T PKEKP	[101]
ITAB_HUMAN (S) platelet M. glycoprotein IIB	DWGLP S PSPSP PSPSP S PIHPA	[102]
GLP_MACFU (S) glycophorin macaca fuscata	----S S TTVPA ---SS T TVPAT --SST T VPATH TTVPA T HTSSS 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	[103]

peptides from von Willebrand factor using bovine colostrum GalNAc-transferase, O'Connell et al. [26] found that amino acid substitutions with alanine, proline, isoleucine, asparagine or glutamic acid at positions +3, −2, −3 reduced O-glycosylation, as did substitution with a charged residue at position −1. O'Connell et al. could not glycosylate a serine residue in a peptide derived from the *in vivo* glycosylated human erythropoietin. If serine was substituted with threonine, the peptide was glycosylated.

From these findings O'Connell et al. speculated that separate serine- and threonine-specific GalNAc-transferases may exist [26]. However, GalNAc-transferases from bovine colostrum [22] and pig submaxillary glands [27] were recently shown to be able to glycosylate serine as well as threonine residues.

Using a purified GalNAc-transferase from pig submaxillary glands on model peptides from erythropoietin or pig mucin, Wang et al. [27] found that the residues adjacent to the serine or threonine strongly influenced the glycosylation. Testing the effects of all possible substitutions at all sites, by synthesizing putative acceptor peptides and testing their activity with a purified GalNAc-transferase, would be a formidable task [27]. This motivates the development of a sequence-based tool to point out likely acceptor peptides.

Originally our aim was to use a modified version of a recently developed statistical method for analysing broad specificities of enzymes [28]. The goal was to predict, from the primary sequence alone, whether specific serine and threonine residues in a glycoprotein are O-glycosylated or not. This matrix method, however, used the sequence context around glycosylated residues only. Information in sequence contexts of non-glycosylated serines and threonines is consequently lost.

In addition we have therefore applied artificial neural network algorithms, which have recently been used with success in predicting protein secondary structure from primary sequence [29–34] and human mRNA donor and acceptor sites from the pre-mRNA sequence [35]. Neural networks are capable of including the negative information from the non-glycosylated sequences and can reproduce even highly complex and non-linear sequence patterns [36].

Thus, two independent statistical tools were applied, together with a carefully selected data set of *in vivo* mucin-type O-glycosylated proteins. The validity of the prediction methods was

assessed by tests on two different data sets of novel glycoproteins with known sequence and glycosylation assignments, which were not used when generating the methods.

MATERIALS AND METHODS

Sequence data selection criteria

All proteins with a carbohydrate assigned to either serine or threonine were extracted from the 1994 versions of SWISS-PROT (Rel. 29) and PIR (Rel. 41.00) databases [37,38].

In SWISS-PROT all protein entries with a CARBOHYD assignment for a serine or threonine in the feature table were extracted. Sites with feature table information listed as POTENTIAL, PARTIAL, BY SIMILARITY, PROBABLE were discarded. Entries with FUCOSE, GLCNAC, GLUCOSAMINE, GLYCOSAMINOGLYCAN, MANNOSE, OLIGOMANNOSIDIC, N-ACETYLGLUCOSAMINE, N-ACETYLLACTOSAMINIC, XYLOSE(2)-GLUCOSE or XYLOSE-GLUCOSE were also excluded.

In PIR all protein entries with a 'binding-site carbohydrate (Ser) status Experimental' or 'binding-site carbohydrate (Thr) status Experimental' were extracted. Assignments with 'status predicted' were excluded.

This crude extract constituted 221 glycoproteins with known sequence and O-glycosylation sites. In the case of copies of the same sequence only one was kept. All non-mammalian glycoproteins were excluded. Furthermore, all proteins which did not have a GalNAc sugar moiety linked to serine or threonine were excluded. Nuclear and cytoplasmic glycoproteins, coagulation factors and proteoglycans were thereby removed from the data set.

Importantly, all examples were extensively checked for sequence and assignment errors by consulting the original references listed in Table 1. The errors found can be divided into two groups. One group comprises sites described in the literature as being glycosylated, but with no assignments in the databases. Another group contains sites in the databases which could not be verified by the original references. A small group consists of glycoproteins which have been revised recently.

One example is glycophorin A, which earlier was thought to contain 15 O-glycosylation sites [39] and do so in PIR (entry GFHUE). However, glycophorin A has recently been revised to contain 16 O-glycosylation sites [24].

A list of the glycoprotein entries, which were revised according to the literature, is given in Table 2.

In general the information from the databases was reliable for glycosylated residues, but caution should be taken before concluding that a given serine or threonine was not glycosylated if not listed in the feature table as being so. An example is the leukosialin family where the two SWISS-PROT entries: LEUK_HUMAN and LEUK_RAT, have 25 and 24 'CARBOHYD' assignments respectively, but probably contain approx. 80 O-glycosylation sites [2]. In order to eliminate sites which are O-glycosylated, but not assigned in the database, we therefore presented these proteins as the smaller fragments used during sequencing.

Though it still is debated whether the initial addition of GalNAc to the peptide is located late in the transitional elements of the endoplasmic reticulum, or in the *cis*-Golgi cisternae [10,11], it is generally believed that O-glycosylation is a post-translational event. The signal peptide, transmembrane and cytoplasmic domains of membrane proteins will consequently never be presented to the luminally located GalNAc-transferase. Therefore the serines and threonines in these domains can never be mucin-type O-glycosylated. The glycoprotein sequences used were truncated accordingly, in order not to dilute the true non-glycosylated sequence contexts with irrelevant information.

A few glycosylation sites in recently published glycoproteins were extracted directly from the literature. The corresponding sequences were found in the database and assigned according to the reference (Table 2).

In order not to bias the data set with identical sequence contexts, the sequence similarity between the glycosylation sites was quantified by computing the number of identical residues between all pairs in a 9-residue window. Examples with identical residues from position -4 to $+4$ were excluded. Finally, all non-glycosylated sequence windows were compared with the glycosylated sequences; no conflicts were found.

The final yield was 48 unique mammalian glycoproteins containing 161 *in vivo* O-glycosylated threonine residues, 103 *in vivo* O-glycosylated serine residues and 2065 non-glycosylated threonine/serine residues. This data set, which to our knowledge is the largest presented to date, was the empirical basis for the analysis.

Similarity within the glycosylated sequences

The predictive performance of any data-driven algorithm will depend on the degree of similarity between the test sequence, and the sequences used when developing the method. Therefore, in

Table 2 Corrected and revised glycoproteins in the database

Entry	Protein	Revision performed	Ref.
GFHUC (P)	Glycophorin C	12 O-glyc. sites, revised to 14 sites	[55]
KNH1-BOVIN (S)	Kininogen HMW I – bovine	1 O-glyc. site, revised to 12 sites	[75]
GFHUE (P)	Glycophorin A – human	15 O-glyc. sites, revised to 16 sites	[24]
JXHU (P)	Transferrin receptor	No O-glyc. sites, revised to 1 site	[91]
IVHUA2 (S)	Interferon alpha 2 – human	No O-glyc. sites, revised to 1 site	[93]
PLMN-PIG (S)	Plasmin – porcine	No O-glyc. sites, revised to 1 site	[72]
TTHUAP (P)	Thyrotropin/gonadotropin alpha chain – human	No O-glyc. sites, revised to 1 site	[100]
LMP1-HUMAN (S)	Lysosome-associated membrane glycoprotein	No O-glyc. sites, revised to 6 sites	[101]
LMP2-HUMAN (S)	Lysosome-associated membrane glycoprotein	No O-glyc. sites, revised to 10 sites	[101]
MBBOB (P)	Myelin basic protein – bovine	No O-glyc. sites, revised to 1 site	[83]
IGHU2 (P)	Insulin-like growth factor II – human	No O-glyc. sites, revised to 1 site	[99]
MMMSND (P)	Nidogen – mouse	No O-glyc. sites, revised to 7 sites	[95]
FQHUGM (P)	Granulocyte–macrophage colony-stimulating factor	No O-glyc. sites, revised to 4 sites	[64]
A25093 (P)	Granulocyte colony-stimulating factor – human	No O-glyc. sites, revised to 1 site	[96]

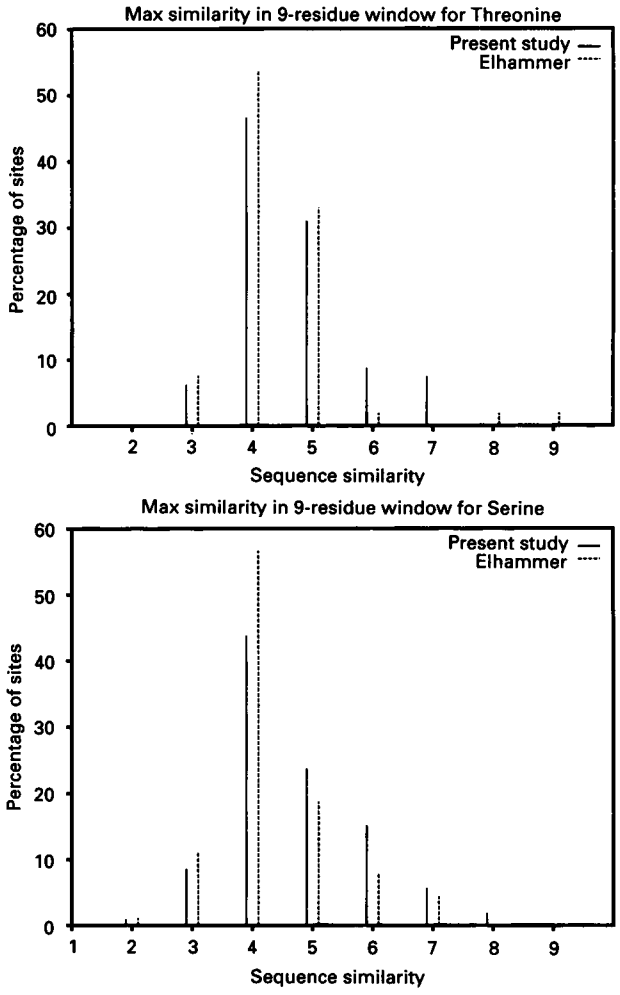

Figure 1 Distribution of the maximal local similarity within the present data set, computed as the number of identical residues in 9-residue windows of the glycosylated sequences

(**a**) Serine, (**b**) threonine. The data set of Elhammer et al. [22] is shown for comparison (dashed lines).

order to compare our predictive performance with the earlier method of Elhammer et al. [22], we compared the degree of similarity within our data set to the degree of similarity within the data used by Elhammer et al. Comparing the complete sequences is not relevant, since it is the local similarity around glycosylated sites which is of primary importance in determining whether or not the sites are glycosylated. Instead, we counted the number of identical residues around glycosylation sites in the 9-residue windows used by Elhammer et al. This local similarity was computed for glycosylated serine and threonine sequence windows in our data set and compared with the data set of Elhammer et al. [22]. The majority of sites had four identical residues in a 9-residue window, see Figure 1. Around 2% of the glycosylated threonine windows in the data set of Elhammer and colleagues were identical, biasing this data set slightly. Otherwise the similarity distributions are comparable.

Selection of independent test sets

In order to test the accuracy of our prediction schemes, 18 of the most recently published O-glycosylated proteins, which were not included in the data used by Elhammer et al., were selected as test proteins. These were used neither during training of the networks, nor during calculation of the weights of the matrices, and were consequently unknown to both methods. The test set was divided into two subsets with 14 proteins (34 sites) with low similarity to the training set, and a set of four proteins (36 sites) with high similarity to the training set, see Table 1. This separation was made in order to determine the degree of prediction accuracy that could be expected on a given novel protein sequence.

Quantification of sequence information content

When a large set of sequences is aligned, the Shannon information measure [40,40a] can be used to quantify the randomness of each column. The information content I was computed by the formula:

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

where p_i^L was the frequency of a particular amino acid L at position i. The unit of information was bits/amino acid. The information content may be displayed in the form of sequence logos [41]. Instead of histograms or curves showing the variation of the information content, the amino acid symbols themselves are used to represent the value of I at a given position. The sum of the height 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 much easier to comprehend the numerical variation in comparison with many more conventional methods.

Neural network algorithms

We used a data-driven neural network algorithm, with one layer of hidden units, and adjusted the weights according to the method described by Rumelhart et al. [42]. Hence each neuron (unit), besides those in the input layer, calculates a weighted sum of its inputs and passes this sum through a sigmoidal function to produce the output:

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

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

We trained the network on experimentally determined relations between amino acid compositions surrounding glycosylated and non-glycosylated serine and threonine residues. Using a steepest descent method, the purpose of the training is to adjust the weights and thresholds in the network to quantitatively minimize the error between the prediction and the experimentally determined assignment.

Each amino acid was represented as a binary string of 21 bits using 21 input neurons. Alanine and cysteine were, for example, represented as 10000000000000000000 and 01000000000000000000 respectively.

We used a symmetric input window of amino acids, ranging from three amino acids (covering the two amino acids flanking the serine/threonine site) to 49 (24 amino acids on each side of the glycosylated serine/threonine). Neural networks with 0 to 15 hidden units were evaluated for prediction performance. A small network with a window of three residues, and with three hidden units, is shown in Figure 2. Other details of the training procedure may be found elsewhere [35].

To find the best network we tested the performance of different

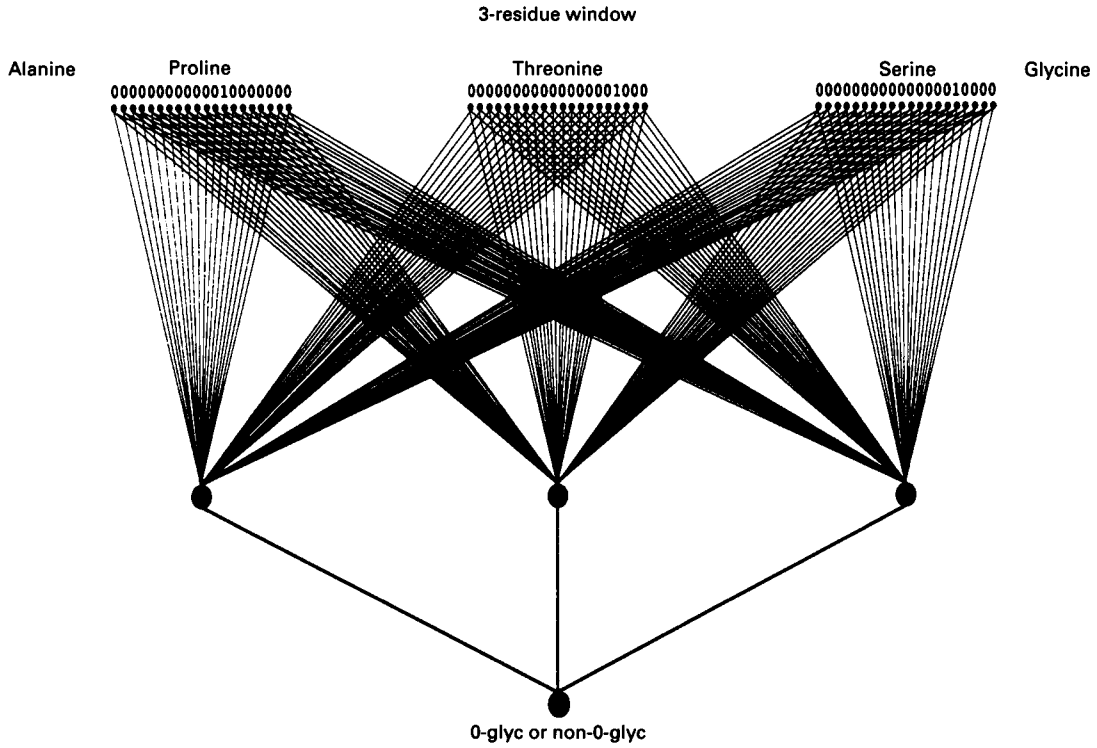

Figure 2 A neural network with three hidden units covering a window of three residues (proline, threonine, serine) is shown

If the value of the output neuron is above the threshold, the central residue is predicted to be O-glycosylated. Network architectures with window sizes from 3 to 49 and 0 to 15 hidden units were evaluated.

network architectures (different numbers of units in the input and in the hidden layer) by training on 34 glycosylated proteins as described above and testing the performance of the network on the remaining proteins, using the correlation coefficient C [35,43] as quality measure. This measure is more relevant than merely measuring the percentage of correctly predicted residues, as approximately 90% of the residues are non-glycosylated. Hence a network which classifies all threonines or serines as non-glycosylated will be 90% correct, but not of much use.

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

In this equation P_x is the number of true positives (experimentally determined glycosylated, predicted glycosylated), N_x the number of true negatives (experimentally 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). The correlation coefficients were obtained as the average over 10 networks with different initialization of the weights.

Matrix method

A matrix method was also used to predict glycosylation sites. We investigated different revisions of the method developed by Poorman et al. [28] and used by Elhammer et al. [22], to predict O-glycosylation sites. The method uses the relative abundance of amino acids surrounding serine and threonine sites $s_{i,j} = (N^P_{i,j}/N^{TP})/(N^T_{i,j}/N^T)$, where $N^T_{i,j}$ and $N^P_{i,j}$ are the total number of amino acids i at site j and the number at glycosylated sites, respectively. The total numbers of amino acids and glycosylated sites are given by N^T and N^{TP}, respectively. The predicted probability h that a given site is glycosylated is calculated (using the relevant value of i for each j) as:

$$h = \frac{\Pi_{j=1}^{W} 1 + (s_{i,j} - 1)(1 - q_{i,j})}{42 + \Pi_{j=1}^{W} 1 + (s_{i,j} - 1)(1 - q_{i,j})} \quad (4)$$

where W is the size of the window, $q_{i,j}$ the probability that the frequency of amino acid i at site j in all windows and the corresponding frequency in the glycosylated windows are equal. The probabilities $q_{i,j}$ are calculated by the library routine 'betai' [44] taking as the standardized normal deviate [45]:

$$\frac{N^P_{i,j}/N^{TP} - N^T_{i,j}/N^T}{\sqrt{fg(1/N^{TP} + 1/N^T)}} \quad (5)$$

where $f = (N^P_{i,j} + N^T_{i,j})/(N^{TP} + N^P)$ and $g = 1 - f$.

We revised the method used by Elhammer et al. [22] in three ways. (a) We weighted the relative frequencies by the probability that the difference is in fact genuine. This was done because some of the frequencies calculated were based on very few examples, and random fluctuations in the number of observed examples could thus bias the method. Poorman et al. [28] deal with this problem by setting $N^P_{i,j} = 1/2$ or $1/4$ if $N^P_{i,j}$ vanishes, depending on the frequency of the amino acid at the given position in the reference distribution. (b) To estimate the frequency of the different amino acids around non-glycosylated sites we used the amino acid distribution in the complete training set, or the distribution around serine or threonine residues only, or the frequencies used by Elhammer et al. of the different amino acids determined by Nakashima et al. [46] for a broad set of globular,

but not necessarily glycosylated proteins. (c) We used relatively large window sizes also, in order to determine the window size with the best predictive performance. The correlation coefficient was used to choose the best cutoff value for h, separating the positive sites from the negative.

RESULTS

Mucin-type O-linked glycosylation is clustered and frequent near the N-terminus

Sequence logos of the glycosylated sequences were prepared *ad modus* Schneider [41], see Figure 3. The logos reflect the bias in the distribution of residues surrounding glycosylation sites. The size of the symbols reflects the patterns of *in vivo* acceptor specificity of the GalNAc-transferase(s) catalysing O-linked glycosylation. The logos demonstrate that the high abundance of threonine, serine and proline in the context of O-glycosylation extends upstream beyond position −4 and downstream beyond position +4. This abundance pattern was specific for glycosylated sites, since it was not found in the non-glycosylated sequences (seen in lower part of graphs). The abundance of proline at positions −1 and +3, originally reported by Wilson et al. [20], is especially evident for glycosylated threonine residues. It may be observed that the glycosylation signal for threonine differs from the signal of serine glycosylation.

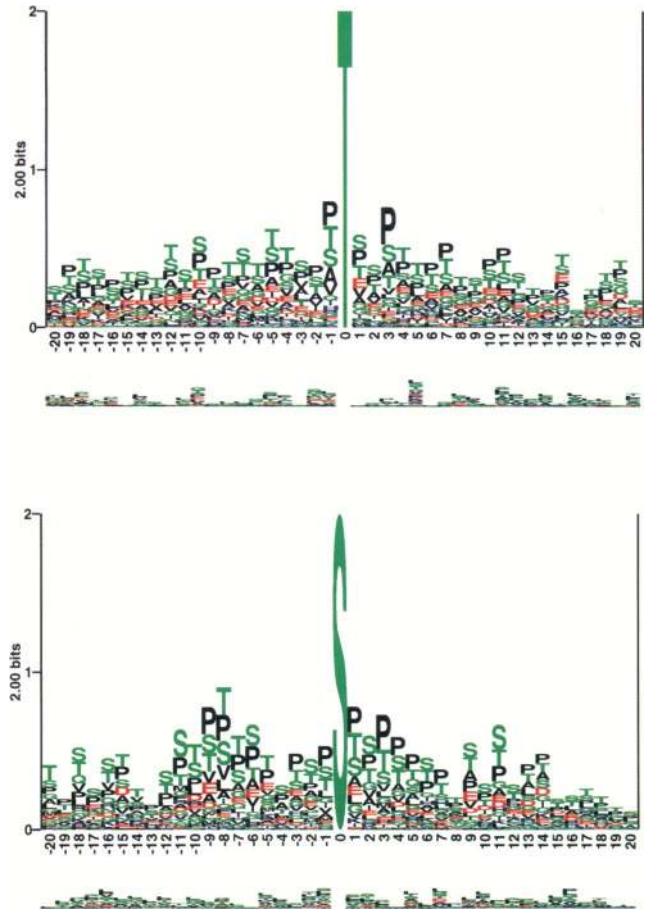

Figure 3 Shannon information content shown in logo form for O-glycosylation sites: (a) serine, (b) threonine

All sites were aligned at the glycosylated/non-glycosylated serine and threonine residues (position 0). Logos for glycosylated (upper) and non-glycosylated (lower) serine sites and threonine sites are shown. The height of each column reflects the bias in the distribution of residues surrounding glycosylation sites. The size of a residue letter reflects the frequency of that residue on that position. The height of the central serine/threonine has been scaled to magnify the context and is thus non-informative. The neutral and polar amino acids are shown in green, the basic in blue, the acidic in red and the neutral and hydrophobic in black. The sequence context around non-glycosylated serine and threonine sites in the lower part of the graphs is seen to be much more random.

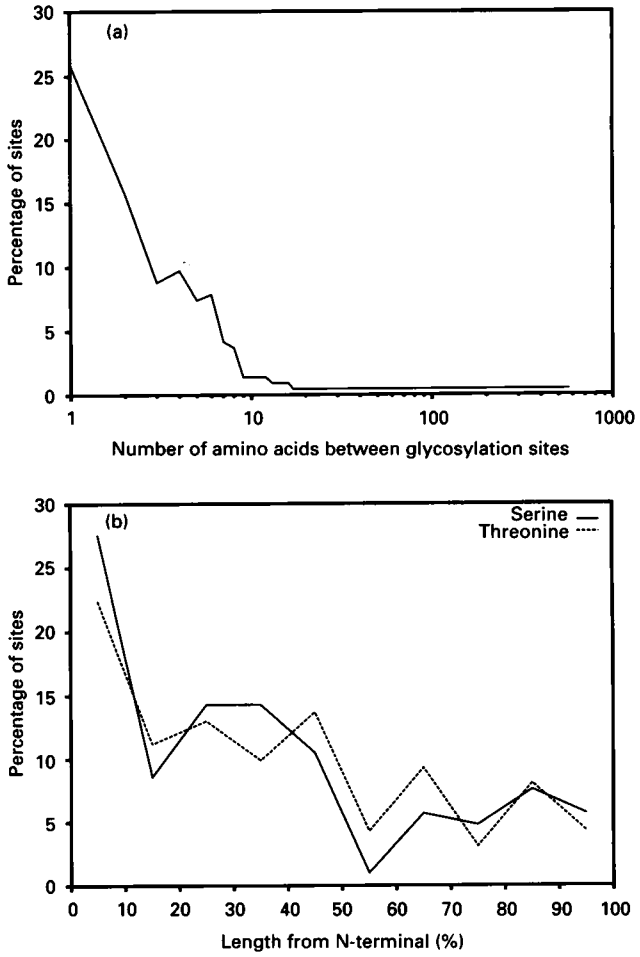

Figure 4 (a) The linear sequence distance in residues between O-glycosylation sites and (b) the distribution of distances from the N-terminal to the O-glycosylation site in percent of the full length of the protein

(**a**) Note that 25% of the sites have another glycosylation site as neighbour, and the majority of sites was a distance of less than 10 residues to the next O-glycosylation site, demonstrating the high degree of clustering of O-glycosylation sites. (**b**) 25% of the O-glycosylation sites are located within the first 10% of the protein.

The O-glycosylation signal is seen to be considerably less well defined than the consensus triplet of N-linked glycosylation. Certain residues are rarely seen in close proximity of the glycosylation sites. We never found cysteine in position −2 to +2, nor tryptophan in position −1 or +1 relative to glycosylated threonines, although they were found frequently in the non-glycosylated sequences. This indicates that O-glycosylation at threonine is disfavoured in proximity to disulphide bridges and very large side chains, possibly because of steric hindrance. Methionine, aspartic acid and asparagine were also very rarely found juxtaposed to the O-glycosylated residue.

We also computed the distance between the sites in order to estimate the degree of clustering. As shown in Figure 4 more than

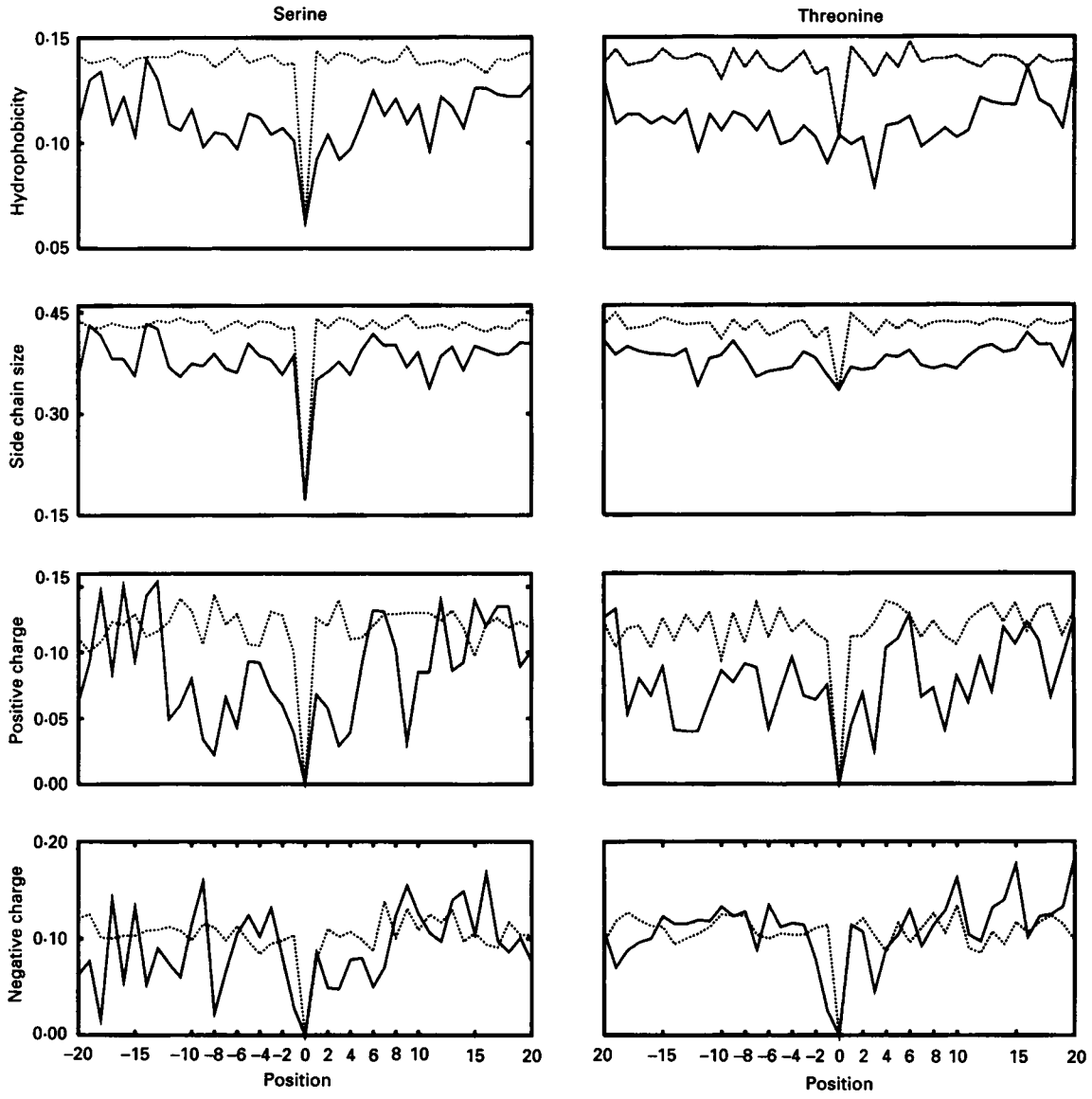

Figure 5 Physico-chemical characteristics of the O-glycosylation sites

They were aligned with the glycosylated threonine or serine at the central position 0. The distribution of hydrophobicity [54], normalized side-chain size (normalized van der Waals volume; glycine = 0, tryptophan = 1), and charge. Averaged values were computed for both glycosylated and non-glycosylated (dashed) sequence contexts.

25% of the sites have another O-glycosylation site as neighbour, and for 15% of the sites only one residue is interposed. We found that 25% of the sites were located within the first 10% of the protein length, meaning that O-glycosylation is abundant near the N-terminus of the protein.

Physicochemical properties of O-glycosylation sites

We examined the difference in averaged hydrophobicity, side-chain size, and charge between the O-glycosylated versus the non-glycosylated sequences, see Figure 5.

On average the hydrophobicity of the glycosylated sites was lower than for the non-glycosylated sites, which confirms that O-glycosylation is unambiguously found on the protein surface. The side-chain size was generally smaller in the glycosylated sites, probably due to steric hindrance of the GalNAc-transferase by very large side chains.

The average values of positive charge (arginine, histidine, and leucine) were lower for the glycosylated sites when compared with the non-glycosylated sites. If the GalNAc-transferase itself has a net positive charge in or near the active site, electrostatic repulsion could be one of the factors determining the acceptor specificity. The presence of negative charge (aspartate and glutamate) was only significantly lowered at position −1, most clearly in the case of threonine. No difference was found at position +1. O'Connell et al. found that substitution at position −1 by a charged residue decreased O-glycosylation [26].

Conformational preference of O-glycosylation sites

We examined the secondary structure preference of glycosylated sequences compared with non-glycosylated sequences. The α-helix preference was lower for the glycosylated sequences com-

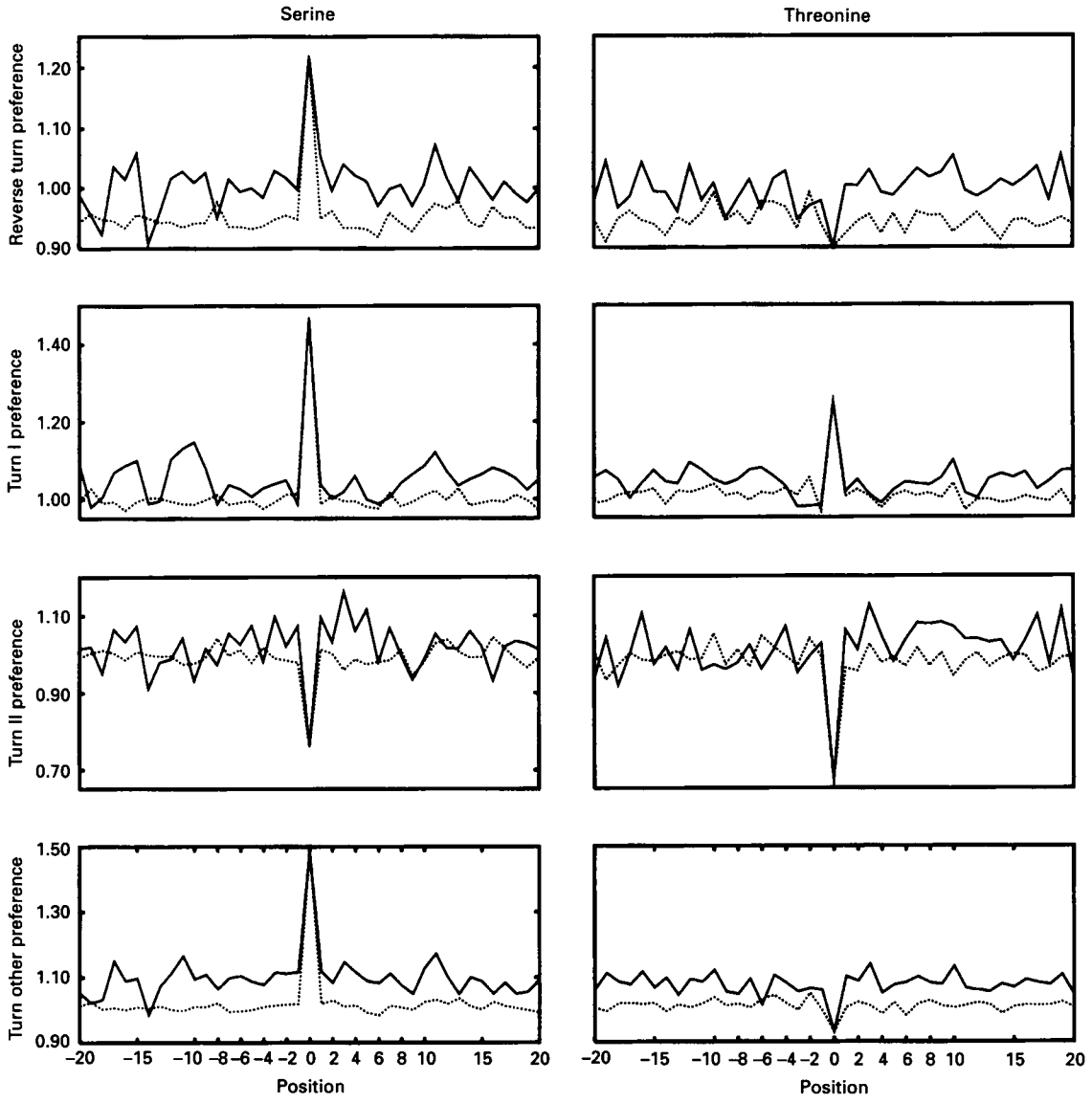

Figure 6 Turn preference of O-glycosylation sites

They were aligned with the glycosylated threonine or serine at the central position 0. Averaged values were computed for both glycosylated and non-glycosylated (dashed) contexts.

pared with the non-glycosylated, as was also the case for β-strand (data not shown). However, the reverse-turn preference was clearly enhanced for the glycosylated sequences. We therefore examined the preference for several different types of turns, see Figure 6.

The analysis was based on the turn preferences of different amino acids computed in [47,48]. They used the classification of Richardson [49], who defined seven categories: type I, II, I′, II′, VIa, VIb and a miscellaneous category IV. Wilmot and Thornton [47] also defined an additional class VIII. However, they had only sufficient data to perform a statistical analysis on type-I and -II turns. The remaining turns were combined to give a non-specific turn data set (other turns). We observed an increased preference of the glycosylated sequences for all three of these turn categories when compared with the non-glycosylated sequences. The most prominent difference was observed for class I and for other turns in which the most prevalent turn types were IV and VIII.

We conclude that there is a significant difference in the secondary structure preference of the glycosylated sites versus the non-glycosylated sites. Glycosylated sequences can be characterized as having low preference for α-helix and β-strand and high preference for reverse turns.

Predictive performance of neural network algorithms

The predictive performance of the networks was strongly dependent on the window size, as shown in Figure 7. We observed correlations at a longer range far beyond what has been reported previously. The other main observation was that a separation of the threonine and serine examples was beneficial as an increase in

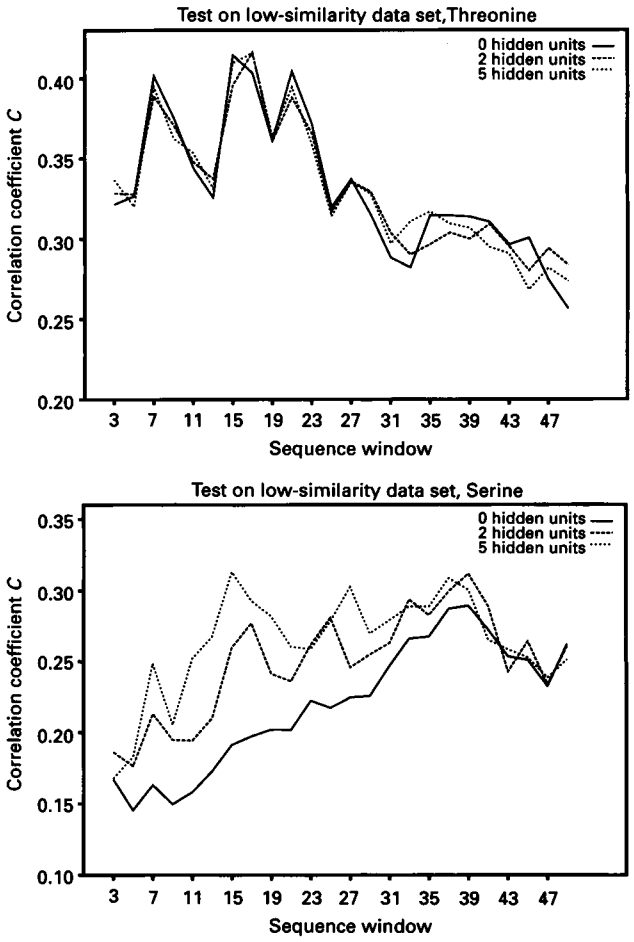

Figure 7 Predictive performance of the neural network method, quantified by the correlation coefficient as function of window size

The figure displays two parts: (**a**) threonine, and (**b**) serine. On each plot the performance of several network configurations are shown.

performance occurred when predictions were carried out separately.

For threonine, the performance variation as a function of window size on the low-similarity test set is shown in Figure 7. The optimal window size was 17 residues with eight residues symmetrically on each side of the glycosylated threonine. This indicates that correlations exist between certain amino acids eight residues away from the glycosylated residue. This is also indicated by the high abundance of serine, threonine, proline and alanine at these distances (Figure 3). In the work of Elhammer et al. an optimal window size of nine residues was found; in Figure 7 this is clearly seen to be suboptimal.

The problem of predicting glycosylated threonines is not highly non-linear, as networks with 0, 3, 5, 7, 9, 11, and 13 hidden units have comparable performance (data not shown). A network with five hidden units was slightly better than the other configurations. With this network architecture we obtained a maximal correlation coefficient of 0.42, corresponding to finding 67% of the glycosylated threonines and 88% of the non-glycosylated threonines in the low-similarity test set (Table 3). On this novel test set the matrix method of Elhammer et al. [22] had a lower correlation coefficient of 0.35, and percentages of 59% and 88% respectively. For the high-similarity test set the maximal network correlation coefficient was 0.88, corresponding to finding 95% of the glycosylated threonines and 96% of the non-glycosylated threonines. For these values the matrix method of Elhammer gave 0.73, 85% and 93% respectively.

Table 3 Predictive performance of the methods for prediction of O-glycosylation

Values for the network method were computed as the average over 10 training runs with different initializations. Two independent test sets are evaluated, a low-similarity test set of 34 O-glycosylation sites and a high-similarity test set of 36 O-glycosylation sites.

Method	Test set...	Low serine	Low threonine	High serine	High threonine
Neural network					
	Correlation coefficient *C*	0.33	0.42	0.70	0.88
	Glyc-sites (%)	60.0	66.8	75.0	95.0
	Non-glyc-sites (%)	96.6	88.4	96.7	96.6
Elhammer matrix					
	Correlation coefficient *C*	0.14	0.35	0.63	0.73
	Glyc-sites (%)	60.0	58.6	75.0	85.0
	Non-glyc-sites (%)	83.9	87.8	94.7	93.3
Our matrix					
	Correlation coefficient *C*	0.33	0.33	0.64	0.80
	Glyc-sites (%)	40.0	48.3	62.5	85.0
	Non-glyc-sites (%)	98.6	90.7	97.4	96.2

For serine, a window of 15 residues (seven on each side of the glycosylated serine) gave the optimal correlation coefficient. However, a window of 39 residues was nearly as good. This indicates the existence of correlations between distant residues in the sequence, and reaffirms that the acceptor sequence pattern for serine differs from that of threonine.

Table 3 reports the performance values (*C*, percentages of correctly predicted glycosylated residues and non-glycosylated residues) on the low-similarity test set: 0.33, 60% and 97% respectively. The Elhammer matrix gave 0.14, 60% and 84%. On the high-similarity test set we obtained 0.70, 75% and 97%, respectively; and (Elhammer) 0.63, 75% and 95%.

The problem of predicting glycosylated serine residues is more non-linear, as networks with hidden units performed significantly better than linear networks without hidden units. The linear network with 0 hidden units had a performance comparable with the matrix method of Elhammer et al.

Overall, it was more difficult to predict glycosylated serines than glycosylated threonines. This could either reflect the higher degree of non-linearity in the serine prediction, or it could be due to the poorer statistical representation of glycosylated serines compared with threonines (103 sites versus 161 sites). For both cases the window size of nine residues used by Elhammer et al. [22] is seen to be suboptimal for the predictive performance (Figure 7).

Prediction performance of weight matrix algorithms

For the matrices, increased performance was also found when treating serine and threonine residues separately. Curves of similar shape as in Figure 7, were obtained with the matrix method, albeit with a somewhat lowered performance in comparison with the networks equipped with hidden units. This means that we have confirmed the long-range correlations found, especially for serine, by two independent methods.

Besides varying the window size, we investigated a number of other possible ways of increasing the performance of the method,

Table 4 Predictive performance of the proposed sequence motifs for GalNAc-transferase

The occurrences of the motifs at glycosylated sites (Thr/Ser-glyc) and non-glycosylated (Thr/Ser non-glyc) sites were counted in the complete data set of 48 O-glycosylated proteins. Some sites fit more than one motif.

A. Proposed sequence motif (ref.)	Thr-glyc	Thr non-glyc
Xaa-Pro-Xaa-Xaa, where one Xaa is a glycosylated threonine [23]	79	161
Thr(g)-Xaa-Xaa-Xaa, where one Xaa and Thr(g) is a glycosylated threonine [24]	36	224
Xaa-Xaa-Thr(g)-Xaa, where one Xaa is either arginine/lysine [24]	18	191
B. Proposed sequence motif (ref.)	**Ser-glyc**	**Ser non-glyc**
Ser(g)-Xaa-Xaa-Xaa, where Ser(g) is a glycosylated serine and Xaa is a serine [24]	46	327

but none of the variations described in the Materials and methods section led to a substantial increase in performance.

Limited predictive performance of proposed motifs

The proposed motifs of Gooley et al. [23] and Pisano et al. [24] fit a reasonable fraction of the glycosylated residues in the 48 O-glycosylated proteins (Table 4), but the predictive value of the motifs is limited since the motifs are also found at many non-glycosylated sites. A maximal correlation coefficient of 0.27 was found for the first motif in Table 4, Xaa-Pro-Xaa-Xaa. No other motif gave better results. This is well below the performance level of Elhammer's method.

DISCUSSION

If, as proposed recently [6], more than one GalNAc-transferase exists with different but overlapping specificities, how can one be sure that only one enzyme with a well-defined specificity is isolated during the purification process? This may explain the diversity of findings using *in vitro* glycosylation of synthetic model peptides with purified GalNAc-transferases.

All O-linked glycosylations have been predicted to be located within β-turns either in the second or the third position of the turn [50]. Synthetic peptides, which are O-glycosylated *in vitro*, adopt a β-turn as shown by circular dichroism by Hollosi et al. [51]. The specificity of the GalNAc-transferase could therefore be defined as the amino acid sequence which is able to adopt and/or stabilize an exposed β-turn.

If O-glycosylation is a post-translational process [11] taking place after N-glycosylation, folding and oligomerization, O-glycosylation is limited to residues present at the surface of the protein. This is in fact the case for the few examples of O-glycosylated proteins present in the Brookhaven Protein Data Bank. O-glycosylation may therefore be strongly influenced by the local conformation [52] and tertiary structure [53] of the protein.

The sequences fulfilling these structural demands may well be so complex that no simple consensus-like rule accounting for all known sites can be defined. Solving the conundrum of acceptor specificity for UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferase therefore probably involves giving up the simplistic view of finding simple sequence rules. This does not necessarily mean that the location of O-glycosylation cannot be predicted from the primary sequence, provided the proper statistical tool and a sufficient amount of examples of *in vivo* glycosylation sites are available.

The earlier proposed motifs [23,24] account for a fraction of O-glycosylated threonines, but are non-specific as they are also present at many non-glycosylated sites, thus limiting their predictive value.

Analysing 264 O-glycosylation sites, we found a high frequency of serine, threonine, proline, alanine and valine residues extending up to 20 residues upstream and downstream relative to the O-glycosylated threonine or serine. This extends the result of Elhammer et al. [22] who found an increased frequency of serine, threonine and proline in the region covering positions -4 through $+4$.

This does not mean that proline, serine and threonine favour glycosylation at all these positions, but probably reflects that these residues have different positional preferences in the various types of reverse turns. When all glycosylated sequences are aligned relative to the glycosylated residue, proline, serine and threonine will be abundant on nearly all positions, as shown in Figure 3.

We have shown that the acceptor sequence context of glycosylated serine residues is much harder to recognize computationally than that of threonine residues. Using neural networks we have shown that the former problem is non-linear, while the latter is linearly separable. This is consistent with the experimental difficulties in glycosylation of serine sites as compared with threonine sites, e.g. Elhammer et al. found that bovine GalNAc-transferase glycosylates threonine 35 times faster than serine [22].

In agreement with Wilson et al. [20] we found an increased frequency of proline at positions -1 and $+3$ relative to both glycosylated serine and threonine residues. However, for glycosylated serine residues, positions $+1$, -8 and -9 also had a significantly increased frequency of proline, threonine and serine. One may speculate that proline in positions -1 and $+1$ functionally acts as a 'gating' residue favouring O-glycosylation and inhibiting N-glycosylation, because it contradicts the extended consensus sequence Asn-Xaa-Ser/Thr-Xaa, where Xaa can be any residue except proline.

The glycosylation patterns can, with advantage, be recognized using neural networks and a large and verified database of *in vivo* O-glycosylated proteins. Using the neural network technique we were able to predict the position of 60–75% of the glycosylated serines and 67–95% of the glycosylated threonines, exclusively from the primary sequence. This is better than any other available method. A key problem is the presence of many buried negative sites in the data set, which would probably be glycosylated if they were exposed on the surface of the protein and were thereby accessible to the GalNAc-transferase. Adding a reliable prediction of surface exposure and of β-turn positioning will presumably lead to enhanced prediction performance.

E-mail server publicly available

Sequences in single-letter amino acid code of 80 characters per line should be put into a file and mailed to the Internet address NetOglyc@cbs.dtu.dk. Mail the word 'help' to receive information about sequence format. The prediction of the networks will be returned immediately.

This work was supported by the Danish 1991 Pharmacy Foundation and the Danish National Research Foundation. Kristoffer Rapacki and Hans Henrik Stærfeldt are thanked for excellent technical assistance.

REFERENCES

1 Hart, G. (1992) Curr. Opin. Cell Biol. **4**, 1017–1023
2 Fukuda, M. (1991) Glycobiology **1**, 337–356
3 Muramatsu, T. (1993) Glycobiology **3**, 294–296
4 Dennis, J. (1991) Semin. Cancer Biol. **2**, 411–420
5 Wing, D. R., Rademacher, T. M., Schmidtz, B., Schachner, M. and Dwek, R. A. (1992) Biochem. Soc. Trans. **20**, 386–390
6 Gooley, A. and Williams, K. (1994) Glycobiology **4**, 413–417
7 Bause, A. (1983) Biochem. J. **209**, 331–336
8 Hunt, L. and Dayhoff, M. (1970) Biochem. Biophys. Res. Commun. **39**, 757–765
9 Opdenakker, G., Rudd, R. M., Pomting, C. P. and Dwek, R. A. (1993) FASEB J. **7**, 1330–1337
10 Strous, G. and Dekker, J. (1992) Crit. Rev. Biochem. Mol. Biol. **27**, 57–92
11 Carraway, K. and Hull, S. (1991) Glycobiology **1**, 131–138
12 Haltiwanger, R., Kelly, W., Roquemore, E., Blomberg, M., Dong, L., Kreppel, L., Chou, T. and Hart, G. (1992) Biochem. Soc. Trans. **20**, 264–269
13 Hardingham, T. and Fosang, A. (1992) FASEB. J. **6**, 861–870
14 Yanagishita, M. and Hascall, V. (1992) J. Biol. Chem. **267**, 9451–9454
15 Spiro, R. (1973) Adv. Protein Chem. **27**, 349–467
16 Nishimura, H., Kawabata, S., Kisiel, W., Hase, S., Ikenaka, T., Takao, T., Shimonishi, Y. and Iwanaga, S. (1989) J. Biol. Chem. **264**, 20320–20325
17 Nishimura, H., Takao, T., Hase, S., Shimonishi, Y. and Iwanaga, S. (1992) J. Biol. Chem. **267**, 17520–17525
18 Hausler, A., Ballou, L., Ballou, C. and Robbins, P. (1992) Proc. Natl. Acad. Sci. U.S.A. **89**, 6846–6850
19 Allen, A. K. Desai, N. N., Neurberger, N. and Creeth, J. M. (1978) Biochem. J. **171**, 665–674
20 Wilson, I. B. H., Gavel, Y. and Heijne, G. (1991) Biochem. J. **275**, 529–534
21 Wang, Y., Abernethy, J., Eckhardt, A. and Hill, R. (1992) J. Biol. Chem. **267**, 12709–12716
22 Elhammer, A. P., Poorman, R., Brown, E., Maggiora, L., Hoogerheide, J. G. and Kzdy, F. D. (1993) J. Biol. Chem. **268**, 10029–10038
23 Gooley, A., Classon, B., Marshcalek, R., and Williams, K. L. (1991) Biochem. Biophys. Res. Commun. **178**, 1194–1200
24 Pisano, A., Redmond, J., Williams, K. and Gooley, A. (1993) Glycobiology **3**, 429–435
25 O'Connell, B. C., Tabak, L. A. and Ramasubbu, N. (1991) Biochem. Biophys. Res. Commun. **180**, 1024–1030
26 O'Connell, B. C., Hagen, F. K. and Tabak, L. A. (1992) J. Biol. Chem. **267**, 25010–25018
27 Wang, Y., Agwral, N., Eckhard, A., Stevens, R. and Hill, R. (1993) J. Biol. Chem. **268**, 22979–22983
28 Poorman, R. A., Tomasselli, A. G., Heinrikson, R. L. and Kezdy, F. J. (1991) J. Biol. Chem. **266**, 14554–14561
29 Rost, B. and Sander, C. (1994) Proteins **19**, 55–72
30 Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. C., Lautrup, B., Nørskov, L., Olsen, O. H. and Petersen, S. B. (1988) FEBS Lett. **241**, 223–228
31 Qian, N. and Sejnowski, T. J. (1988) J. Mol. Biol. **202**, 865–884
32 Holley, L. H. and Karplus, M. (1989) Proc. Natl. Acad. Sci. U.S.A. **86**, 152–156
33 MacGregor, M. J., Flores, T. P. and Sternberg, M. J. E. (1989) Protein Eng. **2**, 521–526
34 Kneller, D., Cohen, F. and Langridge, R. (1990) J. Mol. Biol. **214**, 171–182
35 Brunak, S., Engelbrecht, J. and Knudsen, S. (1991) J. Mol. Biol. **220**, 49–65
36 Presnell, S. and Cohen, F. (1993) Annu. Rev. Biophys. Biomol. Struct. **22**, 283–298
37 Bairoch, A. and Boeckman, B. (1993) Nucleic Acids Res. **21**, 3093–3096
38 Barker, W., George, D., Mewes, H., Pfeiffer, F. and Tsugita, A. (1993) Nucleic Acids Res. **21**, 3089–3092
39 Gahmberg, C., Ekblom, M. and Andersson, L. (1984) Proc. Natl. Acad. Sci. U.S.A. **81**, 6752–6756
40 Shannon, C. E. I. (1948) Bell System Tech. J. **27**, 379–423
40a Shannon, C. E. I. (1948) Bell System Tech. J. **27**, 623–656
41 Schneider, T. D. and Stephens, R. M. (1990) Nucleic Acids Res. **18**, 6097–6100
42 Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986) Nature (London) **323**, 533–536
43 Mathews, B. W. (1975) Biochim. Biophys. Acta **405**, 442–451
44 Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1989) Numerical Recipies In Pascal, The Art Of Scientific Computing, Cambridge University Press, Cambridge
45 Armitage, B. and Berry, G. (1987) Statistical Methods in Medical Research, Blackwell, Oxford
46 Nakashima, H., Nishikawa, K. and Ooi, T. (1986) J. Biochem. (Tokyo) **99**, 153–162
47 Wilmot, C. and Thornton, J. (1988) J. Mol. Biol. **203**, 221–232
48 Williams, R. W. (1987) Biochim. Biophys. Acta **916**, 200–204
49 Richardson, J. (1981) Adv. Protein Chem. **34**, 167–339
50 Aubert, J., Biserte, G. and Loucheaux-Lefebvre, M. (1976) Arch. Biochem. Biophys. **175**, 410–418
51 Hollosi, M., Perczel, A. and Fasman, G. (1990) Biopolymers **29**, 1549–1564
52 Hansen, J., Lund, O., Rapacki, K., Clausen, H., Mosekilde, E., Nielsen, J. O. and Hansen, J.-E. S. (1994) in Protein Structure by Distance Analysis, (Bohr, H. and Brunak, S., eds.), pp. 247–254, IOS Press, Amsterdam
53 Dahms, N. and Hart, G. (1986) J. Biol. Chem. **261**, 13186–13196
54 Fauchere, J. and Pliska, V. (1983) Eur. J. Med. Chem. **18**, 369–375
55 Dahr, W. and Beyreuther, K. (1995) Biol. Chem. Hoppe-Seyler **366**, 1067–1070
56 Murayama, J., Yamashita, T., Tomita, M. and Hamada, A. (1983) Biochim. Biophys. Acta **742**, 477–483
57 Murayama, J., Tomita, M. and Hamada, A. (1982) J. Membr. Biol. **64**, 205–215
58 Honma, K., Tomita, M. and Hamada, A. (1980) J. Biochem. (Tokyo) **88**, 1679–1691
59 Killeen, N., Barclay, A., Willis, A. and Williams, A. (1987) EMBO J. **6**, 4029–4034
60 Schmid, K., Heidiger, M., Brossmer, R., Collins, J., Haupt, H., Marti, T., Offner, G., Schaller, J., Takagaki, K., Walsh, M., Schwick, H., Rose, F. and Remold-O'Donnell, E. (1992) Proc. Natl. Acad. Sci. U.S.A. **89**, 663–667
61 Putnam, F., Liu, Y.-S. and Low, T. (1979) J. Biol. Chem. **254**, 2865–2874
62 Robinson, E. and Appella, E. (1979) J. Biol. Chem. **254**, 11418–11430
63 Takayasu, T., Suzuki, S., Kametani, F., Takahashi, N., Shinoda, T., Okuyama, T. and Munekata, E. (1982) Biochem. Biophys. Res. Commun. **105**, 1066–1071
64 Kaushansky, K., Lopez, J. and Brown, C. (1992) Biochemistry **31**, 1881–1886
65 Takeuchi, M. and Kobata, A. (1991) Glycobiology **1**, 337–346
66 Birken, S., Agosto, G., Amr, S., Nisula, B., Cole, L., Lewis, J. and Canfield, R. (1988) Endocrinology **122**, 2054–2056
67 Morgan, F., Birken, S. and Canfield, R. (1975) J. Biol. Chem. **250**, 5247–5258
68 Seidah, N. and Chretien, M. (1981) Proc. Natl. Acad. Sci. U.S.A. **78**, 4236–4240
69 Bennett, H., Seidah, N., Benjannet, S., Solomon, S. and Chretien, M. (1986) Int. J. Pept. Protein Res. **27**, 306–313
70 Fiat, A., Jolles, J., Aubert, J., Loucheux-Lefebvre, M. and Jolles, P. (1980) Eur. J. Biochem. **111**, 333–339
71 Yan, S. and Wold, F. (1984) Biochemistry **23**, 3759–3765
72 Schaller, J., Marti, T., Rosselet, S. J., Kampfer, U. and Rickli, E. E. (1987) Fibrinolysis **1**, 91–102
73 Robb, R., Kutny, R., Panico, M., Morris, H. and Chowdhry, V. (1984) Proc. Natl. Acad. Sci. U.S.A. **81**, 6486–6490
74 Lottspeich, F., Kellermann, J., Henschen, A., Foertsch, B. and Muller-Esterl, W. (1985) Eur. J. Biochem. **152**, 307–314
75 Kellermann, J., Lottspeich, F., Henschen, A. and Muller-Esterl, W. (1986) Adv. Exp. Med. Biol. **198**, 85–89
76 Hill, H., Schwyzer, M., Steinman, H. M. and Hill, R. L. (1977) J. Biol. Chem. **252**, 3799–3804
77 Takahashi, N., Takahashi, Y. and Putnam, F. (1985) Proc. Natl. Acad. Sci. U.S.A. **82**, 1906–1910
78 Brewer, H., Jr., Shulman, R., Herbert, P., Ronan, R. and Wehrly, K. (1974) J. Biol. Chem. **249**, 4975–4984
79 Kellerman, J., Lottspeich, F., Geiger, R. and Deutzmann, R. (1989) Adv. Exp. Med. Biol. **247A**, 519–525
80 Walsh, K., Titani, K., Takio, K., Kumar, S., Hayes, R. and Petra, P. (1986) Biochemistry **25**, 7584–7590
81 Lopez Otin, C., Grubb, A. and Mendez, E. (1984) Arch. Biochem. Biophys. **228**, 544–554
82 Hochstrasser, K., Schonberger, O., Rossmanith, I., Wachter, E. (1981) Hoppe-Seylers Z. Physiol. Chem. **362**, 1357–1362
83 Young, J., Tsuchiya, D., Sandlin, D. and Holroyde, M. (1979) Biochemistry **18**, 4444–4448
84 Titani, K., Takio, K., Handa, M. and Ruggeri, Z. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 5610–5614
85 Gejyo, F., Chang, J., Burgi, W., Zand Schmid, K., Offner, G., Troxler, R., Van Halbeek, H., Dorland, L., Gerwig, G. and Vliegenthart, F. (1983) J. Biol. Chem. **258**, 4966–4971
86 Watzlawick, H., Walsh, M., Yoshioka, Y., Schmid, K. and Brossmer, R. (1992) Biochemistry **31**, 12198–12203
87 Perkins, S., Smith, K., Amatayakuul, S., Ashford, D., Rademacher, T., Dwek, R., Lachmann, P. and Harrison, R. (1990) J. Mol. Biol. **214**, 751–763
88 Bock, S. C., Skriver, K., Nielsen, E., Thogersen, H. S., Wiman, B., Donaldson, V. H., Eddy, R. L., Marrinan, J., Radziejwska, E. and Huber, R. (1986) Biochemistry **25**, 4292–4301
89 De Caro, A., Adrich, Z., Fournet, B., Capon, C., Bonicel, J., De Caro, J. and Rovery, M. (1989) Biochim. Biophys. Acta **994**, 281–284
90 Wernette-Hammond, M., Lauer, S., Corsini, A., Walker, D., Taylor, J. and Rall, S. (1989) J. Biol. Chem. **264**, 9094–9101
91 Hayes, G., Enns, C. and Lucas, J. (1992) Glycobiology **2**, 355–359
92 Do, S. and Cummings, R. (1992) Glycobiology **2**, 345–353

93 Adolf, G., Kalsner, I., Ahorn, H., Maurer Fogy, I. and Cantell, K. (1991) Biochem. J. **276**, 511–518
94 Voigt, C., Maurer-Fogy, I. and Adolf, G. (1992) FEBS Lett. **314**, 85–88
95 Fujiwara, S., Shinkai, H., Mann, K. and Timpl, R. (1993) Matrix **13**, 215–222
96 Clogston, C., Hu, S., Boone, T. and Lu, H. (1993) J. Chromatogr. **637**, 55–62
97 Geyer, R., Dabrowski, J., Dabrowski, D. U., Linder, D., Schlueter, M., Schott, H. and Stirm, S. (1990) Eur. J. Biochem. **187**, 95–110
98 Minamitake, Y., Kodama, S., Katayama, T., Ada, H., Tanaka, S. and Tsujtmoto, M. (1990) J. Biochem. (Tokyo) **107**, 292–297
99 Daughaday, W. H., Trivedi, B. and Baxter, R. C. (1993) Proc. Natl. Acad. Sci. U.S.A. **90**, 5823–5827
100 Peters, B., Krzesicki, R., Perini, F. and Ruddon, R. (1989) Endocrinology **124**, 1602–1612
101 Carlsson, S. R., Lycksell, P. and Fukuda, M. (1993) Arch. Biochem. Biophys. **304**, 65–73
102 Calvete, J. and Muniz-Diaz, E. (1993) FEBS Lett. **328**, 30–34
103 Murayama, J.-I., Utsumi, H. and Hamada, A. (1989) Biochim. Biophys. Acta **999**, 273–280

Received 21 November 1994/25 January 1995; accepted 31 January 1995