Glycobiology vol. 15 no. 2 pp. 153–164, 2004
doi:10.1093/glycob/cwh151
Advance Access publication on September 22, 2004

Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites

Karin Julenius[1], Anne Mølgaard, Ramneek Gupta, and Søren Brunak

Center for Biological Sequence Analysis, BioCentrum, Building 208, Technical University of Denmark, DK-2800 Lyngby, Denmark

Received on January 9, 2004; revised on September 15, 2004; accepted on September 15, 2004

O-GalNAc-glycosylation is one of the main types of glycosylation in mammalian cells. No consensus recognition sequence for the O-glycosyltransferases is known, making prediction methods necessary to bridge the gap between the large number of known protein sequences and the small number of proteins experimentally investigated with regard to glycosylation status. From O-GLYCBASE a total of 86 mammalian proteins experimentally investigated for *in vivo* O-GalNAc sites were extracted. Mammalian protein homolog comparisons showed that a glycosylated serine or threonine is less likely to be precisely conserved than a nonglycosylated one. The Protein Data Bank was analyzed for structural information, and 12 glycosylated structures were obtained. All positive sites were found in coil or turn regions. A method for predicting the location for mucin-type glycosylation sites was trained using a neural network approach. The best overall network used as input amino acid composition, averaged surface accessibility predictions together with substitution matrix profile encoding of the sequence. To improve prediction on isolated (single) sites, networks were trained on isolated sites only. The final method combines predictions from the best overall network and the best isolated site network; this prediction method correctly predicted 76% of the glycosylated residues and 93% of the nonglycosylated residues. NetOGlyc 3.1 can predict sites for completely new proteins without losing its performance. The fact that the sites could be predicted from averaged properties together with the fact that glycosylation sites are not precisely conserved indicates that mucin-type glycosylation in most cases is a bulk property and not a very site-specific one. NetOGlyc 3.1 is made available at www.cbs.dtu.dk/services/netoglyc.

Key words: machine learning/mucin-type/neural networks/O-glycosylation/prediction

Introduction

Protein glycosylation is more abundant and structurally diverse than all other types of posttranslational modifications combined (Hart, 1992; Seitz, 2000). More than half of all proteins are glycosylated according to estimates based on the SWISS-PROT database (Apweiler *et al.*, 1999). There are two principally different roles for extracellular protein-bound glycans: specific carbohydrate epitopes can serve as ligands for receptors that mediate recognition events or glycan structures can be employed to change the biophysical properties of a protein, such as charge, solubility, folding, or sensitivity toward proteases (Varki, 1993). At the present stage of knowledge, an impressive variety of carbohydrate–peptide linkages have been described that are distributed among glycoproteins found in essentially all living organisms, ranging from eubacteria to eukaryotes. Thirteen different monosaccharides and 8 amino acid types participate in these bonds so that at least 31 sugar–amino acid combinations exist (Spiro, 2002).

One of the most abundant types of mammalian glycosylation is when an N-acetylgalactosamine (GalNAc) is α-1 linked to the hydroxyl group of a serine or threonine residue. This type of glycosylateion is also called mucin-type. Mucin-type glycans are found on many secreted and membrane-bound mucins, but also on other glycoproteins. Mucins typically have very high carbohydrate content (>50% of the dry weight) and are the principal component of mucus, the gel that protects epithelial surfaces from dehydration, mechanical injury, proteases, and pathogens (Carraway and Hull, 1991; Strous and Dekker, 1992). The protein backbone of a mucin contains a number of repetitive sequences, including virtually all the O-linked oligosaccharide attachment sites. Although these differ in terms of length and sequence from mucin to mucin, they all have a high serine, threonine, and proline content and are sometimes referred to as Ser/Thr/Pro-rich domains. Due to the steric hindrance introduced by the glycans, these domains adopt a stiff extended conformation, with an average length of 2.5 Å per amino acid residue (Coltart *et al.*, 2002; Jentoft, 1990).

The biosynthesis of mucin-type glycosylation takes place in the rough endoplasmatic reticulum and the Golgi complex after N-glycosylation, folding, and oligomerization (Asker *et al.*, 1995; Peters *et al.*, 1989). As opposed to the *en bloc* transfer of the high-mannose oligosaccharide involved in N-glycosylation, O-glycosylation is a stepwise process including one monosaccharide at a time. The addition of GalNAc to serine and threonine residues is what governs the site specificity, and this process is mediated by at least 14 different UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferases (Wang *et al.*, 2003). From sequence similarity, it is estimated that there are up to 24 unique GalNAc-transferases genes; see Ten Hagen *et al.* (2003) for a recent review. The different transferases have

[1]To whom correspondence should be addressed; e-mail: karin.julenius@sbc.su.se

overlapping, but different specificities and are differentially expressed (Sørensen *et al*., 1995; Ten Hagen *et al*., 2003; Van den Steen *et al*., 1998). Although no consensus sequence has been formulated, many studies have noted the skew in amino acid composition around mucin-type O-glycosylation sites (Christlet and Veluraja, 2001; Elhammer *et al*., 1993; Hansen *et al*., 1998; Wilson *et al*., 1991, for example) with a higher frequency of prolines, serines, threonines, and alanines than expected. A number of studies have investigated the effect of flanking residues in *in vitro* experiments on synthetic peptides (Nishimori *et al*., 1994; O'Connell *et al*., 1992; Yoshida *et al*., 1997; Young *et al*., 1979) and especially the importance of prolines at certain positions has been confirmed. There is now strong support for the theory that mucin-type glycosylation of multisite substrates proceed in a hierarchical manner, because some of the characterized UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferases seem to only glycosylate peptides, which are already partly glycosylated (Bennett *et al*., 1999; Ten Hagen *et al*., 1999, 2001). This could partly be explained by a recent nuclear magnetic resonance (NMR) study that showed that the preferred substrates of different transferases had different secondary structure in terms of slightly different dihedral angles and that previous glycosylation of a nearby residue affected these structural propensities (Kinarsky *et al*., 2003).

Prediction of glycosylation sites is a valuable tool when trying to characterize a new protein, for example, to help interpret mass spectrometry results. Predicted mucin-type O-glycosylation is one of the important features when predicting orphan protein function (Jensen *et al*., 2002, 2003), and because O-glycosylation affects the structure of the protein and occurs primarily in surface-exposed regions, predicted glycosylation sites may be used to improve protein structure prediction as well. Prediction can also be useful in protein engineering to engineer or abolish O-glycosylation sites and to design competetive inhibitors of glycosyltransferases (Hansen *et al*., 1998).

The most well-known and tested prediction methods for mucin-type O-glycosylation sites are a matrix statistics method (Elhammer *et al*., 1993), a vector projection method (Chou *et al*., 1995; Chou, 1995), and a neural network method (Hansen *et al*., 1995, 1998). All these methods have been based on quite limited data, and when compared in independent experimental studies, none have shown convincing predictive performance (Gerken *et al*., 1997; Neumann *et al*., 1998). Gerken *et al*. (1997) failed to find any correlation between the outputs of the predictor methods and the experimentally determined degree of glycosylation for individual serines and threonines in a highly glycosylated mucin peptide, something neither of the methods were intended for. There exists also three other predictors developed using different neural network methods (Cai and Chou, 1996; Cai *et al*., 1997, 2002). The main problem with these predictors is that although modern machine learning approaches have been used, the data sets have not been updated. The training set consists of 195 positive and 110 negative sites and the test set only of 26 positive and 4 negative sites. In two of the articles (Cai and Chou, 1996; Cai *et al*., 2002) the only performance reported is the number of correct predictions: 26 and 23 out of 30, respectively. Note that a prediction method that predicts all sites to be positive will be correct for 26 out of 30 sites, but not very useful.

The neural network method developed by Hansen *et al*. (1998) is available online (www.cbs.dtu.dk/services/netoglyc-2.0) and had ~5000 queries/month during 2003. It was trained on data available at that time, in total 299 O-GalNAc sites from mammalian proteins. Through continuous updates of our glycosylation database OGlycBase (www.cbs.dtu.dk/databases/oglycbase), we now have access to 421 experimentally verified sites, an increase of more than 40%. When working with small data sets like this, the increase in available data motivates an update, and we also wanted to try predicting not only from sequence but from sequence derived features such as predicted structure. Elhammer *et al*. (1993) and Hansen *et al*. (1998) showed that glycosylation correlates with predicted secondary structure and a number of experimental studies show that UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase substrates adopt an extended β-like or turn-like conformation (Coltart *et al*., 2002; Kinarsky *et al*., 2003; Kirnarsky *et al*., 1998; O'Connell *et al*., 1991; Schuman *et al*., 2003) and that mucin-type glycosylation induces a more rigid extended structure (Schuman *et al*., 2000, 2003; Tagashira *et al*., 2002).

We have searched the Protein Data Bank (Westbrook *et al*., 2003) for structural information on 86 mammalian proteins containing a total number of 421 experimentally verified mucin-type glycosylation sites. Twelve structures were obtained. We found that all sites were found in coil or turn regions either located near the N- or C-termini of the proteins, in linker regions between domains, or in coil regions connecting secondary structure elements. We found that a glycosylated serine and threonine are less likely to be precisely conserved between mammalian protein homologs and more likely to be surface exposed than a nonglycosylated serine or threonine. We have trained a new predictor method, NetOGlyc 3.1, which correctly predicts 76% of the positive sites and 93% of the negative sites. We show that NetOGlyc 3.1 can predict sites for completely new proteins with no loss in performance.

Results

Structural context of O-glycosylation sites

The Protein Data Bank (Westbrook *et al*., 2003) was scanned for structural information about the protein sequences in our data set. Of the 86 sequences in the data set, 14 were represented by structures in the PDB. Of these 14 structures, 2 were represented twice—in all 12 nonredundant structures. Sequence identities between query protein and corresponding protein structure were above 94% in all cases. Two of the structures contained more than one mucin-type O-glycosylation site (2GMF: three Ser and one Thr site and 1AUQ: one Ser and one Thr site). Of the 12 structures, 8 were of recombinant proteins expressed in *Escherichia coli*, and they can therefore not carry any mucin-type glycans (Spiro, 2002). Because of this they may represent the native protein structure involved in the recognition event between the glycosyltransferase and its

Table I. Structural context of mucin-type glycosylation sites

Protein		Source	Resol.	Site present[a]	Site[b]	S.S.[c]
C terminal site						
EOTA_HUMAN	2EOT._	*E. coli*	NMR	no	94(Thr)71	C
N terminal site						
LITA_HUMAN	1QDD.A	Human[d]	1.30 Å	yes	27(Thr)5	C
IL2_HUMAN	1IRL._	*E. coli*	NMR	no	23(Thr)3	C
CSF2_HUMAN	2GMF.A	*E. coli*	2.40 Å	no	22(Ser)5	C
				no	24(Ser)7	C
				no	26(Ser)9	T
				no	27(Thr)10	T
Interdomain linker region site						
PLMN_HUMAN	1KI0.A	*P. pastoris*	1.75 Å	no	268(Ser)249	C
VWF_HUMAN	1AUQ._	*E. coli*	2.30 Å	no	1263(Ser)500	C
				no	1468(Thr)705	C
ALC_MOUSE	2FBJ.H	Mouse[d]	1.95 Å	no	101(Ser)219	C
Intradomain connecting two α-helices						
INA2_HUMAN	1ITF._	*E. coli*	NMR	no	129(Thr)106	C
IL6_HUMAN	1ALU._	*E. coli*	1.90 Å	no	166(Thr)138	C
CSF3_HUMAN	1CD9.A	*E. coli*	2.80 Å	no	166(Thr)134	C
EPO_HUMAN	1EER.A	*E. coli*	1.90 Å	no	153(Ser)126	C
Intradomain, in (very) disordered coil region						
GLHA_HUMAN	1E9J.A	Human[d]	NMR	possibly	63(Thr)39	C

[a] O-glycosylation site detectable in electron density map or not.
[b] The location of the O-glycosylation site. The number before the parentheses refers to the numbering in the query sequence and the number after the parentheses refers to the number of the residue in the PDB entry.
[c] DSSP secondary structure. "T" is hydrogen bonded turn and "C" is random coil.
[d] Not recombinant.

protein substrate. Only 1 of the 12 structures is annotated in the PDB as being glycosylated, namely, the crystal structure of human lithostathine, 1QDD (Gerbaud *et al.*, 2000), where Thr5 is annotated as being O-glycosylated (see Table I). In two structures (1KI0 and 2FBJ), it has not been possible to detect any O-glycosylation in the electron density maps, but this does not necessarily mean that it is not there. It may be so flexible that it becomes invisible. It is unclear whether the O-glycosylation site in 1E9 J is glycosylated or not.

A summary of where in the structures the mucin-type glycosylation sites are located can be found in Table I. All sites, both the glycosylated and the unglycosylated, were found in coil or turn regions. Seven were found near the N- or the C-termini of the polypeptide chains, four of these in the same structure (2GMF). Four sites were located in linker regions between two domains. There were four intradomain sites, all located in coil regions connecting two α-helices. These coil regions were loosely associated with the globular domains. All sites were localized in or close to mainly α-helical domains, except one found in the α-subunit of human chorionic gonadotropin. This preference for coil regions could potentially be used in a mucin-type O-glycosylation site predictor by providing it with predicted structural information.

Sequence conservation and surface accessibility

We investigated whether glycosylated serine and threonine residues are more likely to be conserved between close protein homologs than nonglycosylated serine and threonine residues. Because there are not enough examples of proteins where more than one homolog have been investigated for glycosylation sites, we aligned each proteins in our data set against all its mammalian homologs. A conservation of a threonine or serine residue does not guarantee that the glycosylation site is in fact conserved, but a mutation to anything other than serine or threonine proves that it is not. We were interested to see if there is any additional selective pressure on the glycosylated residues, presumably from the need to conserve the glycan itself, so we investigated both conservation, allowing for no mutations, and what we call semi-conservation, allowing for mutation between serine and threonine only. The results can be seen in Table II and indicate that there is no extra selective pressure on the glycosylated residues in terms of precise site conservation.

On the contrary, glycosylation makes serine and threonine less likely to be conserved. Although the difference in sequence conservation is opposite to what we expected, the fact that there is a difference at all could potentially be used for improving a glycosylation site predictor.

To rule out the possibility that the glycosylation sites are selectively conserved compared to other residues in the disordered and surface-exposed regions of the proteins where glycosylation sites are typically found, we specifically investigated sequence conservation for residues in close proximity (distance <5 amino acids) to glycosylation sites. The sequence conservation varies widely depending on the type of amino acid residue investigated (from 25.9% for methionine to 100% for cysteine), so we choose to restrict our comparisons to serines and threonines (Table II). The sequence conservation for residues in close proximity of glycosylation sites is lower than for other nonglycosylated residues, but not as low as for the glycosylated residues.

Surface accessibility prediction was performed on the 86 proteins in our data set and the result can be seen in Table III. Glycosylated serine and threonines are more surface exposed, and this information is hidden in the sequence and detected by the surface accessibility predictor. Although in principle a neural network trained on mucin-type O-glycosylation sites should be able to pick up this on its own if enough training examples are supplied, providing the network with this information could help when the data are limited, as in our case. The surface accessibility predictions were used already in NetOGlyc 2.0 (Hansen *et al.*, 1998) by letting it control the threshold for positive assignment at the output. This time we want to incorporate the surface accessibility prediction data in the input information to the network instead.

Predictive performance

The concept of using sequence derived features for mucin-type glycosylation prediction is illustrated in Figure 1. The sequence itself need to be translated from letters to numbers before it is presented to the network, and this can be done in various ways: as sparse encoding (the standard way), BLOSUM62 profile encoding (the corresponding row in the

Table II. Conservation of glycosylated serines and threonines

Residue	Type of site	% Conserved	% Semi-conserved[a]	Number of aligned residues[b]
T	Glycosylated	57.5	61.6	1415
	Prox. nonglyc.[c]	65.3	69.5	475
	All nonglyc.	69.8	75.4	4753
S	Glycosylated	39.3	43.2	506
	Prox. nonglyc.[c]	54.7	61.5	605
	All nonglyc.	68.5	72.9	5740

[a] T or S in the aligned sequence so that there is a possibility for conservation of the glycosylation site even though the residue itself is not.

[b] The number of aligned positions investigated. For each glycosylated protein this equals the number of serines/threonines in each category times the number of aligned homologuous sequences. The product is summed over all the proteins in our data set.

[c] Nonglycosylated residues within close proximity of a glycosylation site (<5 aa distance).

Table III. Surface accessibility of Ser and Thr residues

Residue	Type of site	% Surface exposed	Number of sites
T	Glycosylated	69.8	258
	Nonglycosylated	32.4	1202
S	Glycosylated	70.9	148
	Nonglycosylated	37.1	1486

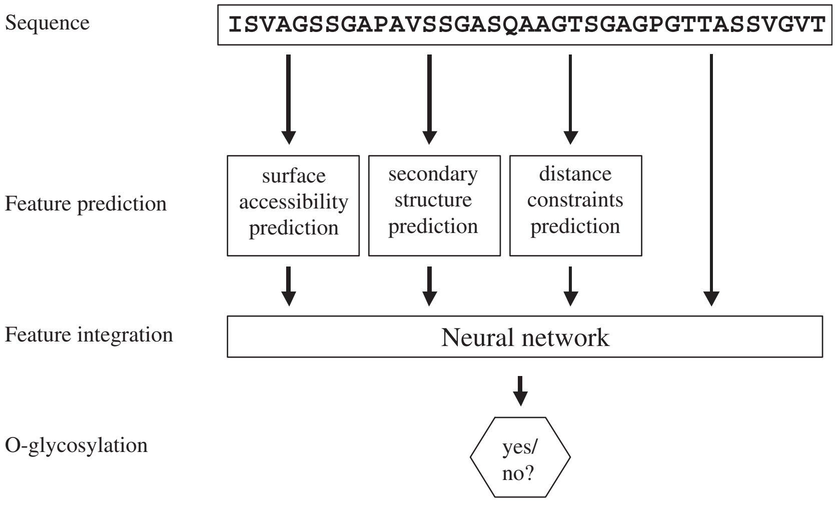

Sequence
ISVAGSSGAPAVSSGASQAAGTSGAGPGTTASSVGVT
Feature prediction
surface accessibility prediction
secondary structure prediction
distance constraints prediction
Feature integration
Neural network
O-glycosylation
yes/ no?

Fig. 1. The concept of using a combination of sequence and sequence derived features to predict glycosylation sites.

BLOSUM62 matrix), PSI-BLAST profile encoding (the corresponding row in the profile computed from PSI-BLAST), reduced alphabet (sparse encoding with fewer letters), or as amino acid composition. Cross-validation was used, so that the 421 positive and 2063 negative sites were divided into three groups of 828 sites each with a minimum of sequence similarity between the three groups. These were used so that every network was trained three times, using two sets as training set and one set as test set. As performance measure we used the joint Matthews correlation coefficient (Matthews, 1975) of the three resulting networks on their respective test sets and this is what we aimed to maximize.

Predictors were trained using window sizes between 1 and 35 amino acids and different in-data information (one feature at the time): sparse encoding, BLOSUM62 profile encoding, PSI-BLAST profile encoding, 5-letter reduced alphabet, 8-letter reduced alphabet, amino acid composition, secondary structure, average secondary structure, protein distance constraints, surface accessibility, and average surface accessibility. The performance of these predictors can be seen in Figure 2 and show that each of these features clearly have predictive potential because all correlation coefficients are greater than zero. This means that there is in fact a preference of mucin-type O-glycosylation for a certain secondary structure, certain protein distance constraints, a certain value of surface accessibility, and so on (otherwise the correlation coefficients would be close to zero). The comparably low predictive performance of the networks trained only on secondary structure information show that it is not likely to be the most discriminating condition that needs to be fulfilled for a serine/threonine to be glycosylated, and this is probably the reason for the bad performance (a network trained only on secondary structure would predict all sites with the correct secondary structure to be positive and this would lead to a large number of false positives). Averaged information was as powerful as position specific information. This can be seen from comparing the curve for surface accessibility with the one for average surface accessibility, from comparing secondary structure with average secondary structure, and from comparing amino acid composition with the different sequence encoding methods. The only exception from this rule is that the networks trained on PSI-BLAST encoded sequence information perform better than amino acid composition for window sizes up to 15 amino acid residues. A PSI-BLAST encoded sequence contains information about sequence conservation between related proteins, and this additional information gives even a network trained only on a three-residue window surprisingly high performance. But because the number of input neurons increases linearly with increasing window size for sequence information, the high network complexity causes problems with overtraining for large windows, and this is probably the reason why BLAST encoding does not outperform amino acid composition for larger windows.

Overall, a network trained on amino acid composition in a 31-residue window (with only 21 input neurons) outperforms all other single-feature networks. We analyzed a linear network (no hidden neurons) trained on amino acid composition in a 31-residue window to see directly the effect of the different amino acids on the prediction (correlation coefficient = 0.54). The residues that makes a glycosylation more likely are (in decreasing order of their tendency to promote glycosylation): Pro, Thr, Ser, end of sequence, and Ala. One residue is essentially

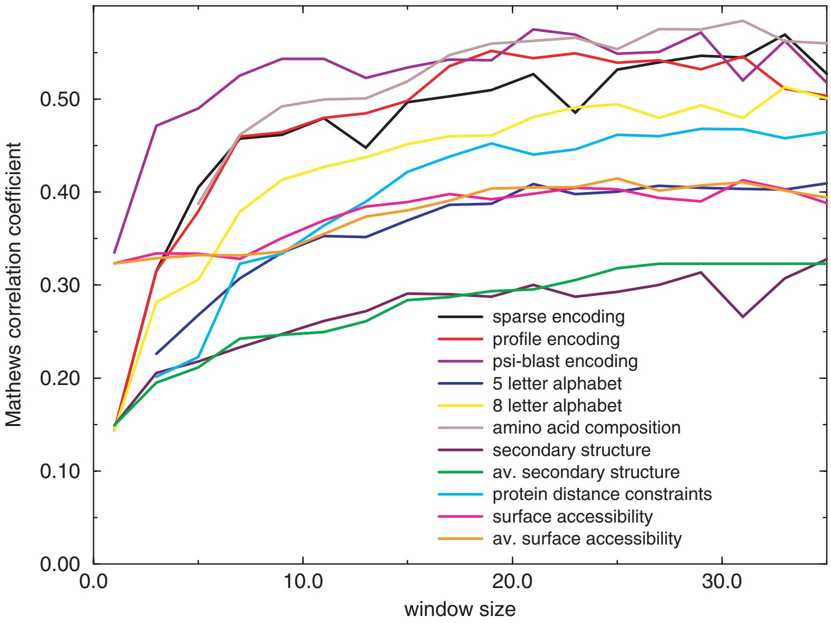


Fig. 2. Glycosylation predictor performance of networks trained on different in-data information using seven hidden neurons. Window size is the number of amino acids for which the information is provided, with the potentially glycosylated Ser/Thr residue in the middle. The Matthews correlation coefficient is a measure of the prediction performance. A perfect predictor would have a correlation coefficient of 1 and a predictor making random guesses would have a correlation coefficient of 0.

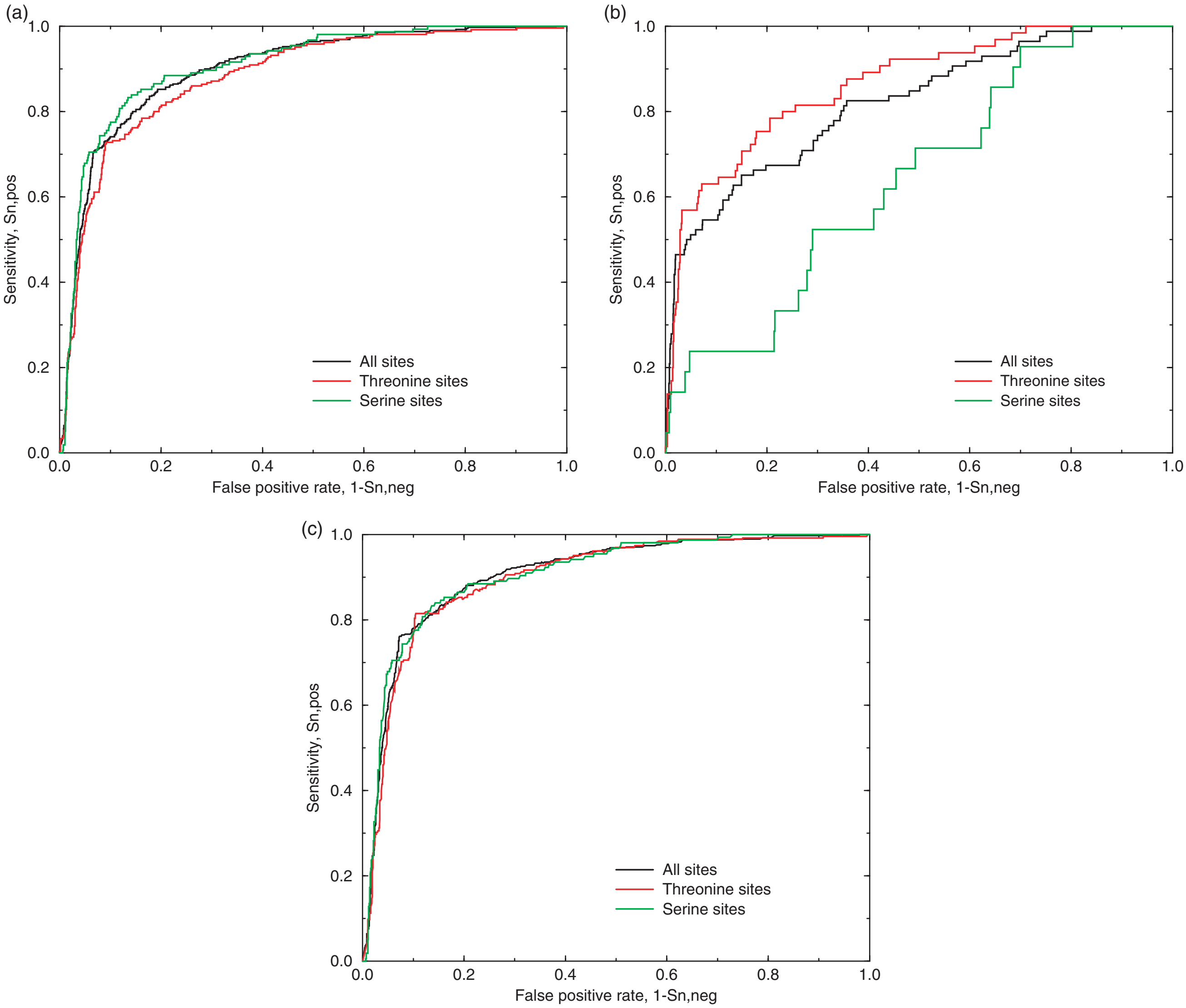


Fig. 3. ROC curves showing predictor performances. The sensitivity is the fraction of positive sites correctly predicted. The false positive rate is the fraction of negative sites wrongly predicted to be positive. A predictor making random guesses would perform along the diagonal and a perfect predictor along the *y*-axis. (**a**) Performance of NetOGlyc 3.0 (best overall network) on all sites. (**b**) Performance of best isolated site network on isolated sites only. (**c**) Performance of NetOGlyc 3.1 (combination of the best general network and the best single site network) on all sites.

glycosylation neutral, Glu, and the rest makes glycosylation less likely (in decreasing order of their tendency to promote glycosylation): Val, Gly, Met, Ile, His, Gln, Trp, Asp, Arg, Phe, Tyr, Lys, Cys, Asn, and Leu. Note that these rankings are based on single amino acids and not correlated pairs or other combinations.

To find the best possible combination of features, we used a greedy strategy, trying to combine what appeared to be good input information from the results of the single-feature networks. For feature combinations that seemed promising, networks with varying number of hidden neurons (different network complexity) were trained. We also tried linear combinations of different networks and trained networks where the input was the output from a number of single-feature networks. The very best combination was profile encoding in a 1-residue window, plus amino acid composition in a 31-residue window, plus average surface accessibility in a 25-residue window using seven hidden neurons. The performance of this network can be seen in Figure 3a and in Table IV. The figure shows the trade-off between making many positive predictions, of which some are false, and making few predictions and thereby missing some. A curve reaching far up into the upper left corner is to be preferred, and completely random designation would perform along the diagonal. ROC curves are widely used in describing the quality of a classification method such as a predictor or a medical diagnostic tool. When you want to make a classification like sick/healthy or glycosylated/nonglycosylated you typically have to set a threshold. If you set a high threshold you will get few positives, but a higher percentage of the predictions you make will in fact be true (in our example, 40% of the positives can be found with only about 3% of the

Table IV. Comparison between the predictive performance of NetOGlyc 2.0, NetOGlyc 3.0, and NetOGlyc 3.1

Residue	Method	C[a]	$S_{n,pos}$ (%)[b]	S_p (%)[c]	$S_{n,neg}$ (%)[d]	Test set[e]
T	NetOGlyc 2.0	0.60	87.6	51.1	86.2	cross
	NetOGlyc 2.0	0.36	63.6	30.8	85.5	new
	NetOGlyc 3.0	0.63	72.5	70.1	90.9	cross
	NetOGlyc 3.0-old[f]	0.56	68.0	63.0	89.4	new
	NetOGlyc 3.1[g]	0.67	81.5	69.5	89.5	cross
S	NetOGlyc 2.0	0.54	75.2	41.5	92.6	cross
	NetOGlyc 2.0	0.20	50.0	13.2	86.1	new
	NetOGlyc 3.0	0.62	66.7	65.8	95.3	cross
	NetOGlyc 3.0-old[f]	0.77	70.0	90.3	99.0	new
	NetOGlyc 3.1[g]	0.62	66.7	65.8	95.3	cross
S+T	NetOGlyc 2.0	0.58	82.9	48.9	89.7	cross
	NetOGlyc 2.0	0.28	58.3	21.3	85.8	new
	NetOGlyc 3.0	0.63	70.3	68.5	93.4	cross
	NetOGlyc 3.0-old[f]	0.66	68.9	72.9	95.3	new
	NetOGlyc 3.1[g]	0.66	76.0	68.2	92.8	cross

[a] Matthews correlation coefficient.
[b] Positive site sensitivity (the fraction of positive sites correctly predicted).
[c] Specificity (the fraction of all positive predictions that are correct).
[d] Negative site sensitivity (the fraction of negative sites correctly predicted).
[e] Indicates whether the performance is the cross-validation performance or the performance on the independent new set.
[f] A version of NetOGlyc 3.0 trained on the same data as NetOGlyc 2.0 for the purpose of fair comparison.
[g] The combined performance of NetOGlyc 3.0 (the best general network) with the best network trained only on isolated glycosylation sites.

negatives being wrongly predicted to be positive). If a low threshold is used, you will find more of the true positives, but you will also get more false positives (80% of the positives found will give about 15% wrong predictions of the negative sites). Because nonglycosylated serines and threonines typically are much more common than glycosylated ones, it is normally preferred to keep the false positive rate as low as possible, because otherwise the specificity (the fraction of predicted sites that are in fact glycosylated) becomes very low. The maximum Matthews correlation coefficient is obtained when a threshold of 0.5 is used and the resulting detailed performance can be seen in Table IV. This is also the default threshold of the Web server of NetOGlyc 3.0, but ultimately the choice is up to the user.

In Table IV the performances of NetOGlyc 3.0 and NetOGlyc 2.0 are compared. Looking only at the reported cross-validation performance, the differences are not that dramatic. NetOGlyc 3.0 has a higher correlation coefficient and a considerably higher specificity, but also a lower sensitivity for the positive sites. (If desired, the balance between sensitivity and specificity can be changed by changing the prediction threshold, so this is not a problem.) It appears that the reported performance of NetOGlyc 2.0 has been somewhat overestimated when tested on completely new proteins (Gerken *et al.*, 1997; Neumann *et al.*, 1998). To compare the two networks on equal terms and to ensure that we do not make the same mistake of overestimating the performance on unknown proteins, we therefore trained a version of NetOGlyc 3.0 on an old set, corresponding to the only site information available when NetOGlyc 2.0 was developed. The performances of this version of NetOGlyc 3.0 (NetOGlyc 3.0-old) and NetOGlyc 2.0 on the proteins in the new set were then compared. Although the performance of NetOGlyc 3.0-old on the new set is comparable to the cross-validation performance, the correlation coefficient of NetOGlyc 2.0 plunges from 0.58 to 0.28. This is largely due to a specificity of 21%, meaning that almost four out of five positive predictions are false.

Mucin-type O-glycosylation sites seem to fall within two different categories. The majorities of the sites occur in highly glycosylated regions where the distance to the closest neighboring glycosylation site is short. NetOGlyc 3.0 performs well on these sites. There are, however, a smaller group of isolated (single) sites in our data set. A previous database study suggests that single and multiple sites may be slightly different from each other (Christlet and Veluraja, 2001). When we examine the performance on isolated sites only, it is much lower than for multiple sites. To improve the prediction on isolated sites, we trained networks only on these (distance to closest neighboring mucin-type glycosylation site >10 amino acids), in total 65 threonine sites and 21 serine sites. The best network uses substitution matrix profile encoding (BLOSUM62) in a 9-residue window and averaged surface accessibility in a 17-residue window. The Matthew correlation coefficient is 0.46, which is to be compared to 0.24 for NetOGlyc 3.0 on these sites. The ROC curve in Figure 3b show that the perfomance on threonine sites is much better than for serine sites. This is due to the small number of isolated serine sites compared to isolated threonine sites. We have tried to improve the performance on serine sites by various means but believe that nothing short of more known sites can solve this problem.

To provide an easy-to-use all-around predictor, we devised an algorithm for combining NetOGlyc 3.0 and the single-site predictor:

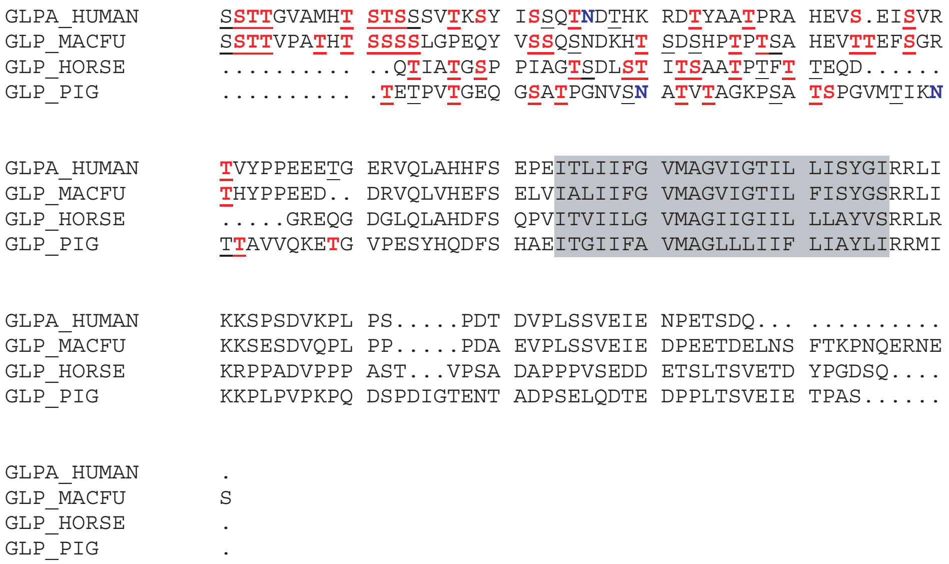
1. The sequence is run through both predictors.
2. All NetOGlyc 3.0 predictions above a certain threshold are accepted.
3. For serines/threonines where there are no predicted sites within 10 aa on either side, accept single-site predictions above a certain threshold.

The thresholds where optimized independently and found to be 0.5 in both cases for threonine sites, which makes sense because that is the threshold that gives the best performance in each individual case. For serine sites, adding sites predicted by the single-site predictor adds too many false positive sites, and the optimum is actually to stick with the NetOGlyc 3.0 prediction only. The new, combined predictor is called NetOGlyc 3.1, and its performance can be seen in Figure 3c and Table IV. As you can see, the performance on serine sites is identical between NetOGlyc 3.0 and NetOGlyc 3.1, but for threonine sites NetOGlyc 3.1 is outstanding.

Discussion

The fact that there is no extra evolutionary pressure to conserve site-specific mucin-type glycosylated serines and threonines compared to nonglycosylated serines and threonines was a surprise. To understand why, several points have to be made. One is that nonglycosylated serine and threonine residues often occur in the well-conserved core of a protein, whereas glycosylated serines and threonines occur in disordered and surface-exposed regions with little overall sequence conservation. Our predictions show that only about 35% of nonglycosylated serines and threonines are surface exposed, whereas 70% of glycosylated are. A priori, the buried core residues are more likely to be subjected to a high evolutionary pressure, leading to sequence conservation. This does not seem to be the whole explanation, though. Taking only serines and threonines found close to glycosylation sites into account, nonglycosylated residues are still more likely to be conserved than glycosylated residues. The same is true when comparing only serines and threonines predicted to be surface exposed (data not shown). The second point is that the loops where glycosylation occur often vary in length and the problem of aligning two loops of different length and weaker sequence similarity (than for the protein core) is not trivial. Although a low linear conservation of the glycosylated residues can be detected, there is a quite high structural one, as described in the previous section.

Another point is that the function of mucin-type glycosylation in the highly glycosylated mucin proteins, which are responsible for a large number of glycosylation sites in our data set, is believed mainly to be to change the biophysical properties of the protein: to protect it from cleavage, change the size and charge distribution of the protein, make the protein bind more water, and change the structure to be stiffer and more extended. In neither of these functions the exact number or positions of glycosylated residues would be important. Rather, the glycosylation would be conserved more as a bulk property. In fact, this can be observed for highly glycosylated homologs within our data set, see Figure 4. The mucin-type glycosylation is clearly conserved, but only on an overall, bulk level. This does not exclude the



```
GLPA_HUMAN      SSTTGVAMHT STSSSVTKSY ISSQTNDTHK RDTYAATPRA HEVS.EISVR
GLP_MACFU       SSTTVPATHT SSSSLGPEQY VSSQSNDKHT SDSHPTPTSA HEVTTEFSGR
GLP_HORSE       .......... ..QTIATGSP PIAGTSDLST ITSAATPTFT TEQD......
GLP_PIG         .......... .TETPVTGEQ GSATPGNVSN ATVTAGKPSA TSPGVMTIKN

GLPA_HUMAN      TVYPPEEETG ERVQLAHHFS EPEITLIIFG VMAGVIGTIL LISYGIRRLI
GLP_MACFU       THYPPEED.. DRVQLVHEFS ELVIALIIFG VMAGVIGTIL FISYGSRRLI
GLP_HORSE       .....GREQG DGLQLAHDFS QPVITVIILG VMAGIIGIIL LLAYVSRRLR
GLP_PIG         TTAVVQKETG VPESYHQDFS HAEITGIIFA VMAGLLLIIF LIAYLIRRMI

GLPA_HUMAN      KKSPSDVKPL PS.....PDT DVPLSSVEIE NPETSDQ... ..........
GLP_MACFU       KKSESDVQPL PP.....PDA EVPLSSVEIE DPEETDELNS FTKPNQERNE
GLP_HORSE       KRPPADVPPP AST...VPSA DAPPPVSEDD ETSLTSVETD YPGDSQ....
GLP_PIG         KKPLPVPKPQ DSPDIGTENT ADPSELQDTE DPPLTSVEIE TPAS......

GLPA_HUMAN      .
GLP_MACFU       S
GLP_HORSE       .
GLP_PIG         .
```

Fig. 4. Multiple alignment of glycophorin A from four different organisms performed with CLUSTAL W. Experimentally verified mucin-type glycosylation sites are colored red and N-glycosylation sites are colored blue. (The experimentally investigated parts of the proteins are residues 1 to 61 for GLPA_HUMAN and the entire sequence for the other three proteins.) Glycophorin A is a membrane protein, where the N-terminal part is extracellular and the C-terminal part is cytosolic (mucin-type glycosylation is only possible in the extracellular domain). The location of the transmembrane helix is shown in gray. Underscored residues are predicted to be glycosylated by NetOGlyc 3.1 (predictions in the cytosolic domain are ignored).

possibility that individual mucin-type glycosylation sites may be highly specific and therefore highly conserved between species; one example may be human and bovine corticotropin, COLI_HUMAN and COLI_BOVIN, which have identical sequences from position −10 to +20 relative to their only mucin-type glycosylation site, respectively.

The third point is that a large part of the endothelial glycocalyx consists of mucin-type glycosylated proteins (Jentoft, 1990). One of the functions of the glycocalyx is to protect the cell surface from bacteria and viruses. Therefore we speculate that a higher mutation rate in the mucin-type glycosylated protein regions might be an evolutionary advantage, because this could render an infectious agent harmless to other organisms than the one it coevolved with. This is essentially the same strategy used by viruses themselves for their surface proteins, like HIV1 gp120 (Hansen *et al.*, 1998). The fact that PSI-BLAST encoding works so much better than the other types of sequence encoding for a small sequence window shows that this difference in sequence conservation between glycosylated and nonglycosylated residues has prediction potential. With a data set large enough so that overtraining could be avoided, it should have been possible to incorporate this into the final method as well. The fact that averaged information like amino acid composition was as powerful as position specific information like sequence for prediction purposes also support the theory that mucin-type glycosylation is a bulk property.

The action of the different UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferases on a Ser/Thr/Pro-rich domain is highly complex. In a hierarchical manner a number of enzymes glycosylate the serine and threonine residues in the surface accessible loops that have the right amino acid composition and adopts the right extended conformation. The glycosylation of the different sites takes place in a specific order, depending on the transferases present in the tissue, and due to steric hindrance from the flanking glycosylation sites, some sites may be only partially glycosylated or not at all (Gerken, 2004; Hanisch *et al.*, 2001; Kato *et al.*, 2001; Takeuchi *et al.*, 2002). Unfortunately, NetOGlyc 3.1 does not hold the key to understanding all of this complexity. It is based on *in vivo* data, which is neither tissue- nor transferase-specific. In a highly glycosylated Ser/Thr/Pro-rich domain, it is likely to predict all the threonines and serines as glycosylation sites, even the ones that are not glycosylated or only to a lesser extent. Nevertheless, it is a powerful tool when it comes to identifying the glycosylated regions in a protein and for finding isolated threonine sites.

NetOGlyc 3.1 is only intended for extracellular protein sequences. Intracellular proteins or the cytosolic domains of membrane proteins will never encounter the UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferases performing the mucin-type O-glycosylation, because these are located in the Golgi complex. Therefore, all sequences submitted to NetOGlyc 3.0 are routinely checked for signal peptide using the SignalP prediction server (Bendtsen *et al.*, forthcoming; Nielsen *et al.*, 1997). For membrane proteins, the responsibility to only consider predictions in the potentially extracellular domains is left up to the user.

In several studies, threonine has been proven to be a better substrate for mucin-type glycosylation than serine (for example, Kinarsky *et al.*, 2003; O'Connell *et al.*, 1992; Yoshida *et al.*, 1997). At the same time, serine is a more common amino acid residue overall. The fact that we would normally expect a smaller percentage of serines to be glycosylated as compared to threonines makes the correct prediction of serine sites harder. In Table IV we can see that we will normally find fewer of the positive sites (the positive site sensitivity) and a fewer percentage of the predicted sites will be correct (the specificity) for serines than for threonines. The fact that we were able to specifically improve the performance on isolated sites for threonines and not for serines when developing NetOGlyc 3.1 indicates that the recognition sequence are sligthly different between isolated serine and threonine sites. With only 21 isolated serine sites, we have every reason to believe that a sufficient increase in the number of known isolated serine sites would make it possible to make a similar improvement in the prediction of serine sites using the method described here for threonine.

Materials and methods

Data set

Eighty-six mammalian protein sequences with one or more experimentally verified mucin-type glycosylation site were extracted from O-GlycBase v6.00 (www.cbs.dtu.dk/databases/oglycbase) (Gupta *et al.*, 1999). One protein, mouse interleukin-3, was added for negative site information because it has been shown to have no O-glycosylation (Knepper *et al.*, 1992). All original articles on the site assignments of the glycosylated proteins were investigated. Signal peptides and parts of the protein not investigated for glycosylation were masked out along with everything but the positive sites of nonexhaustive studies. All serine and threonine residues in nonmasked regions having neither experimental nor predicted glycosylation were used as negative sites. Serine and threonine residues reported to be partially glycosylated were used as positive sites.

Structural context

The program GetStruct (www.cbs.dtu.dk/services/getstruct) was used with default parameters to extract structural information about the glycosylation sites in our data set from the PDB database (Westbrook *et al.*, 2003). GetStruct performed BLAST (Altschul *et al.*, 1997) alignments of the sequences in our data set versus the sequences in the PDB with the aim of obtaining one hit structure for each query (input) sequence. Only structures with at least 90% sequence identity to the query (input) sequences were considered. With a few notable exceptions (Dalal *et al.*, 1997; Gerstein and Levitt, 1998; Riesner, 2003), a clear amino acid sequence relationship between two proteins implies that they have similar structure (Chothia and Lesk, 1986). Therefore, at the required levels of sequence similarity (90% or more), the found structures can be expected to be good representatives of the structures of the glycoproteins.

The reported localization of the O-glycosylation sites are indicated relative to their position in the query sequence. Thus, a site that is close to the N-terminal in a structure but in the middle of the query sequence, is classified as being in an interdomain region (the assumption being that the structurally determined unit is a full domain).

Sequence conservation and surface accessibility

For each of the 86 proteins in our data set, close protein homologs were identified by searching SWISS-PROT (Boeckmann *et al.*, 2003) for mammalian proteins with entry names with identical prefix. Example: As homologs to bovine fibronectin (SWISS-PROT entry name FINC_BOVIN), FINC_HUMAN, FINC_MOUSE, and FINC_RAT were identified. To avoid fragment proteins in the study, proteins with less than half the length of the query protein were discarded. Multiple alignment of the sequences was performed using CLUSTAL W (Thompson *et al.*, 1994). The sequence conservation was estimated on a residue for residue basis.

Surface accessibility was predicted using a neural network method called surfg (Hansen *et al.*, 1998). Surfg gives both a direct output score and a smoothed score. Both are between 0 and 1, with a score above 0.5 if the amino acid residue is predicted to be buried and a score below 0.5 if it is predicted to be surface exposed. The serine and threonine residues for which the smoothed score is below 0.5 were considered to be predicted surface exposed.

Neural network training

For readability, this section was shortened to suit the average readers of *Glycobiology*. For details on sequence encoding, feature encoding, and neural networks, see the supplementary material online.

A neural network does not understand letters, so the amino acid sequence and different features must be translated into numbers. This is called encoding and can be done in a number of ways. Each number that is presented to the neural network make up what is called an input neuron. The goal is to provide the network with as much information as possible while still keeping the number of input neurons as low as possible.

- *Sparse encoding* (Hertz et al., 1991; Qian and Sejnowski, 1988) is the conventional way to convert the amino acid sequence into numerical form.
- With *profile encoding*, the input for each amino acid consisted of the corresponding row in the BLOSUM62 matrix (Henikoff and Henikoff, 1992).
- With *PSI-BLAST encoding*, the input for each amino acid consisted of the corresponding row in the position-specific scoring matrix computed from three cycles of PSI-BLAST (Altschul *et al.*, 1997).
- The *5-letter alphabet* encoding was conventional sparse encoding, but with a reduced alphabet (Soumpasis, personal communications).
- The *8-letter alphabet* is another reduced alphabet.
- *Amino acid composition* was calculated for a sequence window around each particular site.
- *Surface accessibility* was predicted using a neural network method called surfg (Hansen *et al.*, 1998).
- *Secondary structure* was predicted using PSIPRED (Jones, 1999; McGuffin *et al.*, 2000) using position-specific scoring matrices computed from three cycles of PSI-BLAST (Altschul *et al.*, 1997).
- *Protein distance constraints* were predicted using DistanceP (Gorodkin *et al.*, 1999).

The neural networks were of the two-layer feed-forward type, trained by standard back-propagation. Network complexity was varied by changing the number of neurons in the input layer as well as in the hidden layer to find the optimal complexity for this particular prediction problem. This is important, because a network with too little complexity (too few neurons) will lack the ability to learn the training examples, and a network with too much complexity (too many neurons) will learn the examples too well and lose the ability to make predictions for examples that were not in the training set (the ability to generalize). This second problem is sometimes called overtraining and is one of the reasons why it is so important to make sure that the examples in the test set are different and unrelated to the examples in the training set. If the sets are unrelated to each other, the performance on the test set will decrease when overtraining occurs, and if the problem can be detected, it can also be avoided. The risk of overtraining is greater the smaller the data set is.

The predictive performance was monitored using the Matthews correlation coefficient (Matthews, 1975) during training and test of the networks:

$$C = \frac{t_p t_n - f_p f_n}{\sqrt{(t_n + f_n)(t_n + f_p)(t_p + f_n)(t_p + f_p)}} \quad (1)$$

where t_p is the number of correctly predicted positive sites (true postitives), t_n the number of correctly predicted negative sites (true negatives), f_n the number of sites falsely predicted to be negative (false negatives), and f_p the number of sites falsely predicted to be positive (false positives). The Matthews correlation coefficient will always be a value between −1 and 1 where a predictor that always is wrong will have a correlation coefficient of −1, one that is always right will have a correlation coefficient of 1, and one that makes random guesses will have a correlation coefficient of 0. It takes into account the performance on both the positive and the negative sites and is widely used for classification problems such as this one.

The fraction of positive sites correctly predicted, the positive site sensitivity, $S_{n,pos}$, was computed as

$$S_{n,pos} = \frac{t_p}{t_p + f_n} \quad (2)$$

The fraction of all positive classifications that are correct, the specificity S_p, was computed as

$$S_p = \frac{t_p}{t_p + f_p} \quad (3)$$

The fraction of negative sites correctly predicted, the negative site sensitivity, $S_{n,neg}$, was computed as

$$S_{n,neg} = \frac{t_n}{t_n + f_p} \quad (4)$$

In the data set, proteins were identified as closely related if at least two of the following criteria were fulfilled: (1) similar protein names, (2) SWISS-PROT entry name with identical prefix, and (3) high sequence identity. Examples: Human lithostathine 1α, LITA_HUMAN, and human lithostathine

1β, LITB_HUMAN (86% sequence identity); human and bovine corticotropin, COLI_HUMAN, and COLI_BOVIN (80% sequence identity). Of these groups of related proteins, only the most well-studied in each group was used for negative site information. The positive sites were scanned for similarities within the group and those with identical residues from −5 to +5 were excluded. This resulted in one protein (COLI_BOVIN) being altogether masked out, so our data set consisted of 85 proteins. Using only the most well-studied protein from each group, the proteins were divided into three sets of equal size with minimal sequence overlap between the sets using a heuristic described in Jensen *et al*. (2003). After this division was performed, the closely related proteins were manually placed in the same partition as their representative. For computational reasons, we needed to have the same number of sites in each partition. To achieve this, some negative sites were randomly omitted. The result was a total of 421 positive (265 Thr and 156 Ser) and 2063 negative sites (903 Thr and 1160 Ser) divided into three sets of 828 sites each. These were used so that every network was trained three times, using two sets as training set and one set as test set. The reported cross-validation performance is the joint performance of the three resulting networks on their respective test sets.

To be able to truly compare our performance to the performance of NetOGlyc 2.0 (Hansen *et al*., 1998), we also trained on a reduced set, consisting only of proteins entered into O-GLYCBASE (Gupta *et al*., 1999) before 20 January 1997. These were the 65 proteins available for training of NetOGlyc 2.0 and is referred to as the old set. The same division of sets were used, and the result was 331 positive and 1190 negative divided into three sets of 507 sites each. The best window and feature combination as for the whole set was used, but the number of hidden neurons was varied (0–15), and the best number was chosen based on the cross-validation performance. The 20 proteins entered into the database after NetOGlyc 2.0 was trained could then be used to compare the performance of NetOGlyc 2.0 and NetOGlyc 3.0 directly. This is referred to as the new set and consists of 90 positive sites (50 Thr and 40 Ser) and 489 negative sites (188 Thr and 301 Ser). The reported performance of NetOGlyc 3.0-old on this set is the performance of the average output from the three cross-validation networks trained on the old set.

Acknowledgments

Kristoffer Rapacki, Hans Henrik Stærfeldt, and Lars Juhl Jensen are thanked for technical assistance. This work was supported by the Danish National Research Foundation, the Danish Center for Scientific Computing and Knut and Alice Wallenbergs Foundation.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Apweiler, R., Hermjakob, H., and Sharon, N. (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta*, **1473**, 4–8.

Asker, N., Baeckstrom, D., Axelsson, M.A.B., Carlstedt, I., and Hansson, G.C. (1995) The human MUC2 apoprotein appears to dimerize before O-glycosylation and shares epitopes with the “insoluble” mucin of rat small intestine. *Biochem. J.*, **308**, 873–880.

Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. (forthcoming) Improved prediction of signal peptides—signalp 3.0. *J. Mol. Biol.*

Bennett, E.P., Hassan, H., Hollingsworht, M.A., and Clausen, H. (1999) A novel human UDP-N-acetyl-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase, GalNAc-T7, with specificity for partial GalNAc-glycosylated acceptor substrates. *FEBS Lett.*, **460**, 226–230.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., J., M.M., Michoud, K., O’Donovan, C., Phan, I., and others. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Cai, Y.D. and Chou, K.C. (1996) Artificial neural network model for predicting the specificity of GalNAc-transferase. *Anal. Biochem.*, **243**, 284–285.

Cai, Y.D., Liu, X.J., Xu, X.B. and Chou, K.C. (2002) Support vector machines for predicting the specificity of GalNAc-transferase. *Peptides*, **23**, 205–208.

Cai, Y.D., Yu, H. and Chou, K.C. (1997) Artificial neural network method for predicting the specificity of GalNAc-transferase. *J. Protein Chem.*, **16**, 689–700.

Carraway, K.L. and Hull, S.R. (1991) Cell surface mucin-type glycoproteins and mucin-like domains. *Glycobiology*, **1**, 131–138.

Chothia, C. and Lesk, A.M. (1986) Relationship between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–827.

Chou, K.C. (1995) A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. *Protein Sci.*, **4**, 1365–1383.

Chou, K.C., Zhang, C.T., Kezdy, F.J. and Poorman, R.A. (1995) A vector projection method for predicting the specificity of GalNAc-transferase. *Proteins*, **21**, 118–126.

Christlet, T.H.T. and Veluraja, K. (2001) Database analysis of O-glycosylation sites in proteins. *Biophys. J.*, **80**, 952–960.

Coltart, D.M., Royyuru, A.K., Williams, L.J., Glunz, P.W., Sames, D., Kuduk, S.D., Schwarz, J.B., Chen, X.T., Danishefsky, S.J., and Live, D.H. (2002) Principles of mucin architecture: Structural studies on synthetic glycopeptides bearing clustered mono-, di-, tri-, and hexasaccharide glycodomains. *J. Am. Chem. Soc.*, **124**, 9833–9844.

Dalal, S., Balasubramanian, S. and Regan, L. (1997) Protein alchemy: changing beta-sheet into alpha-helix. *Nat. Struct. Biol.*, **4**, 548–552.

Elhammer, Å.P., Poorman, R.A., Brown, E., Maggiora, L.L., Hoogerheide, J.G. and Kézdy, F.J. (1993) The specificty of UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferase as inferred from a database of *in vivo* substrates and from the *in vitro* glycosylation of proteins and peptides. *J. Biol. Chem.*, **268**, 10029–10038.

Gerbaud, V., Pignol, D., Loret, E., Bertrand, J.A., Berland, Y., Fontecilla-Camps, J.C., Canselier, J.P., Gabas, N., and Verdier, J.M. (2000) Mechanism of calcite crystal growth inhibition by the N-terminal undecapeptide of lithostathine. *J. Biol. Chem.*, **275**, 1057–1064.

Gerken, T.A. (2004) Kinetic modeling confirms the biosynthesis of mucin core 1 (beta-Gal(1-3)alpha-GalNAc-O-ser/thr) O-glycan structures are modulated by neighboring glycosylation effects. *Biochemistry*, **43**, 4137–4142.

Gerken, T.A., Owen, C.L., and Pasumarthy, M. (1997) Determination of the site-specific O-glycosylation pattern of the porcine submaxillary mucin tandem repeat glycopeptide. *J. Biol. Chem.*, **272**, 9709–9719.

Gerstein, M. and Levitt, M. (1998) Comprehensive assessment of automatic structural alignment against a manual standard; the scop classification of proteins. *Protein Sci.*, **7**, 445–456.

Gorodkin, J., Lund, O., Andersen, C.A., and Brunak, S. (1999) Using sequence motifs for enhanced neural network prediction of protein distance constraints. In Lengauer, T., Schneider, R., Bork, P., Brutlag, D., Glasgow, J., Mewes, H.W., and Zimmer, R. (Eds.), *Proceedings of the Seventh International Conference for Molecular Biology*. pp. 95–105.

Gupta, R., Birch, H., Rapacki, K., Brunak, S., and Hansen, J.E. (1999) O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res.*, **27**, 370–372.

Hanisch, F.G., Reis, C.A., Clausen, H., and Paulsen, H. (2001) Evidence for glycosylation-dependent activities of polypeptide N-acetylgalactosaminyltransferases rGalNAc-T2 and -T4 on mucin glycopeptides. *Glycobiology*, **11**, 731–740.

Hansen, J.E., Lund, O., Engelbrecht, J., Bohr, H., Nielsen, J.O., Hansen, J.E., and Brunak, S. (1995) Prediction of O-glycosylation of mammalian proteins: specificity patterns of UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase. *Biochem. J.*, **308**, 801–813.

Hansen, J.E., Lund, O., Gooley, A.A., Williams, K.L., and Brunak, S. (1998) NetOGlyc. Prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconj. J.*, **15**, 115–130.

Hart, G.W. (1992) Glycosylation. *Curr. Opin. Cell Biol.*, **4**, 1017–1023.

Henikoff, S. and Henikoff, J.G. (1992) Amino acid subsitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919.

Hertz, J., Krogh, A., and Palmer, R. (1991) *Introduction to the theory of neural computation.* Redwood City, CA: Addison-Wesley.

Jensen, L.J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Stærfeldt, H.H., Rapacki, K., Workman, C., and others. (2002) Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.*, **319**, 1257–1265.

Jensen, L.J., Gupta, R., Stærfeldt, H.H., and Brunak, S. (2003) Prediction of human protein function according to Gene Ontology categories. *Bioinformatics*, **19**, 635–642.

Jentoft, N. (1990) Why are proteins O-glycosylated? *Trends Biochem. Sci.*, **15**, 291–294.

Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Kato, K., Takeuchi, H., Miyahara, N., Kanoh, A., Hassan, H., Clausen, H., and Irimura,T. (2001) Distinct orders of GalNAc incorporation into a peptide with consecutive threonines. *Biochem. Biophys. Res. Commun.*, **287**, 110–115.

Kinarsky, L., Suryanarayanan, G., Prakash, O., Paulsen, H., Clausen, H., Hanisch, F.G., Hollingsworth, M.A., and Sherman, S. (2003) Conformational studies on the MUC1 tandem repeat glycopeptides: implication for the enzymatic O-glycosylation of the mucin protein core. *Glycobiology*, **13**, 929–939.

Kirnarsky, L., Nomoto, M., Ikematsu, Y., Hassan, H., Bennet, E.P., Cerny, R.L., Clausen, H., Hollingsworth, M.A., and Sherman, S. (1998) Structural analysis of peptide substrates for mucin-type O-glycosylation. *Biochemistry*, **37**, 12811–12817.

Knepper, T.P., Arbogast, B., Schreurs, J., and Deinzer, M.L. (1992) Determination of the glycosylation patterns, disulphide linkages, and protein heterogeneities of baculovirus- expressed mouse interleukin-3 by mass spectrometry. *Biochemistry*, **31**, 11651–11659.

Matthews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

McGuffin, L.J., Bryson, K., and Jones, D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.

Neumann, G.M., Marinaro, J.A., and Bach, L.A. (1998) Identification of O-glycosylation sites and partial characterization of carbohydrate structure and disulfide linkages of human insulin-like growth factor binding protein 6. *Biochemistry*, **37**, 6572–6585.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.

Nishimori, I., Johnson, N.R., Sanderson, S.D., Perini, F., Mountjoy, K., Cerny, R.L., Gross, M.L., and Hollingsworth, M.A. (1994) Influence of acceptor substrate primary amino acid sequence on the activity of human UDP-N-acetylgalactosamine:polypeptide N-Acetylgalactosaminyltransferase. *J. Biol. Chem.,* **269**, 16123–16130.

O'Connell, B., Tabak, L.A., and Ramasubbu, N. (1991) The influence of flanking sequences on O-glycosylation. *Biochem. Biophys. Res. Commun.*, **180**, 1024–1030.

O'Connell, B.C., Hagen, F.K., and Tabak, L.A. (1992) The influence of flanking sequence on the O-glycosylation of threonine *in vitro*. *J. Biol. Chem.*, **267**, 25010–25018.

Peters, B.P., Krzesicki, R.F., Perini, F., and Ruddon, R.W. (1989) O-glycosylation of the α-subunit does not limit the assembly of chorionic gonadotropin αβ dimer in human malignant and nonmalignant trophoblast cells. *Endocrinology*, **124**, 1602–1612.

Qian, N. and Sejnowski, T.J. (1988) Prediction the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, **202**, 865–884.

Riesner, D. (2003). Biochemistry and structure of prp(c) and prp(sc). *Br. Med. Bull.*, **66**, 21–33.

Schuman, J., Qiu, D., Koganty, R.R., Longenecker, B.M., and Campbell, A.P. (2000) Glycosylations versus conformational preferences of cancer associated mucin core. *Glycoconj. J.*, **17**, 835–848.

Schuman, J., Campbell, A.P., Koganty, R.R., and Longenecker, B.M. (2003) Probing the conformational and dynamical effects of O-glycosylation within the immunodominant region of a MUC1 peptide tumor antigen. *J. Peptide Res.*, **61**, 91–108.

Seitz, O. (2000) Glycopeptide synthesis and the effects of glycosylation on protein structure and activity. *Chembiochem.*, **1**, 214–246.

Sørensen, T., White, T., Wandall, H.H., Kristensen, A.K., Roepstorff, P., and Clausen, H. (1995) UDP-N-acetyl-α-D-galactosamine: polypeptide N-acetylgalactosaminyltransferase. *J. Biol. Chem.*, **270**, 24166–24173.

Spiro, R.G. (2002) Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology*, **12**, 43R–56R.

Strous, G.J. and Dekker, J. (1992) Mucin-type glycoproteins. *Crit. Rev. Biochem. Mol. Biol.*, **27**, 57–92.

Tagashira, M., Iijimia, H., and Toma, K. (2002) An NMR study of O-glycosylation induced structural changes in the α-helix of calcitonin. *Glycoconj. J.*, **19**, 43–52.

Takeuchi, H., Kato, K., Hassan, H., Clausen, H., and Irimura, T. (2002) O-GalNAc incorporation into a cluster acceptor site of three consecutive threonines. Distinct specificity of GalNAc-transferase isoforms. *Eur. J. Biochem.*, **269**, 6173–6183.

Ten Hagen, K.G., Tetaert, D., Hagen, F.K., Richet, C., Beres, T.M., Gagnon, J., Balys, M.M., VanWuyckhuyse, B., Bedi, G.S., Degand, P., and Tabak, L.A. (1999) Characterization of a UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferase that displays glycopeptide N-acetylgalactosaminyltransferase activity. *J. Biol. Chem.*, **274**, 27867–27874.

Ten Hagen, K.G., Bedi, G.S., Tetaert, D., Kingsley, P.D., Hagen, F.K., Balys, M.M., Beres, T.M., Degand, P., and Tabak, L.A. (2001) Cloning and characterization of a ninth membre of the UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase family, ppGaNTase-T9. *J. Biol. Chem.*, **276**, 17395–17404.

Ten Hagen, K.G., Fritz, T.A., and Tabak, L.A. (2003) All in the family: the UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferases. *Glycobiology*, **13**, 1R–16R.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Van den Steen, P., Rudd, P.M., Dwek, R.A., and Opdenakker, G. (1998) Concepts and principles of O-linked glycosylation. *Crit. Rev. Biochem. Mol. Biol.*, **33**, 151–208.

Varki, A. (1993) Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology*, **3**, 97–130.

Wang, H., Tachibana, K., Zhang, Y., Iwasaki, H., Kameyama, A., Cheng, L., Guo, J., Hiruma, T., Togayachi, A., Kudo, T., and others. (2003) Cloning and characterization of a novel UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase, pp-GalNAc-T14. *Biochem. Biophys. Res. Commun.*, **300**, 738–744.

Westbrook, J., Feng, Z., Chen, L., Yang, H., and Berman, H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.

Wilson, I.B., Gavel, Y., and von Heijne, G. (1991) Amino acid distributions around O-linked glycosylation sites. *Biochem. J.*, **275**, 529–534.

Yoshida, A., Suzuki, M., Ikenaga, H., and Takeuchi, M. (1997) Discovery of the shortest sequence motif for high level mucin-type O-glycosylation. *J. Biol. Chem.*, **272**, 16884–16888.

Young, J.D., Tsuchiya, D., Sandlin, D.E., and Holroyde, M.J. (1979) Enzymic O-glycosylation of synthetic peptides from sequences in basic myelin protein. *Biochemistry*, **18**, 4444–4448.