NetCGlyc 1.0: Prediction of mammalian C-mannosylation sites

Running Title: Prediction of C-mannosylation

Key words: machine learning / neural networks / C-mannosylation / prediction

Karin Julenius

Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-171 77
Stockholm, Sweden

Stockholm Bioinformatics Center, Albanova, Stockholm University, SE-106 91
Stockholm, Sweden

1 To whom correspondence should be addressed; email: karin.julenius@sbc.su.se; fax:
+46 8 313 445; tel: +46 8 524 86 976

© The Author 2007. Published by Oxford University Press. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org
Abstract

C-mannosylation is the attachment of an α-mannopyranose to a tryptophan via a C-C link. The sequence WXXW, in which the first Trp becomes mannosylated, has been suggested as a consensus motif for the modification, but only 2/3 of known sites follow this rule. We have gathered a data set of 69 experimentally verified C-mannosylation sites from literature. We analyzed these for sequence context and found that apart from Trp in position +3, Cys is accepted in the same position. We also find a clear preference in position +1, where a small and/or polar residue (Ser, Ala, Gly, and Thr) is preferred and a Phe or Leu discriminated against. The Protein Data Bank was searched for structural information and five structures of C-mannosylated proteins were obtained. We showed that modified tryptophan residues are at least partly solvent-exposed. A method predicting the location of C-mannosylation sites in proteins was developed using a neural network approach. The best overall network used a 21-residue sequence input window plus information on the presence/absence of the WXXW motif. NetCGlyc 1.0 correctly predicts 93% of both positive and negative C-mannosylation sites. This is a significant improvement over the WXXW consensus motif itself, which only identifies 67% of positive sites. NetCGlyc 1.0 is available at

http://www.cbs.dtu.dk/services/NetCGlyc/. Using NetCGlyc 1.0, we scanned the human genome and found 2573 exported or transmembrane transcripts with at least one predicted C-mannosylation site.
Introduction

Among post-translational modifications, protein glycosylation is more abundant and structurally diverse than all the other types combined (Hart, G.W. 1992, Seitz, O. 2000). Glycosylation is known to affect protein folding, localization and trafficking, protein solubility, antigenicity, biological activity and half-life, as well as cell-cell interactions (Varki, A. 1993). 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 (Spiro, R.G. 2002). In mammals, seven different monosaccharides and six amino acid types participate in these bonds so that at least eleven sugar-amino acid combinations exist (Ohtsubo, K. and Marth, J.D. 2006).

C-mannosylation is the attachment of an α-mannopyranosyl residue to the indole C2 of tryptophan via a C-C link (de Beer, T., Vliegenthart, J.F., et al. 1995, Hofsteenge, J., Muller, D.R., et al. 1994). First example of glycosylation of a tryptophan residue (with a hexose of unknown type) was discovered in a neuropeptide from a stick insect (Gade, G., Kellner, R., et al. 1992). Since then, numerous C-mannosylation sites have been found in mammalian proteins, of which the first was RNase 2 (Furmanek, A. and Hofsteenge, J. 2000, Hofsteenge, J., Muller, D.R., et al. 1994). In all mammalian cases, the glycan has been found to be a single α-mannopyranose. The transfer of mannose to the protein is catalyzed by the enzyme C-mannosyltransferase and this probably occurs in the endoplasmic reticulum (ER) (Doucey, M.A., Hess, D., et al. 1998, Perez-Vilar, J., Randell, S.H., et al. 2004). C-mannosyltransferase activity towards peptides derived from human RNase has been found in C. elegans, amphibians, birds and mammals, but not in E. coli, insects or yeast (Doucey, M.A., Hess, D., et al. 1998, Furmanek, A. and Hofsteenge, J. 2000, Krieg, J., Glasner, W., et al. 1997). At
present, little is known about the function of C-mannosylation, but two recent studies indicate that it is likely required for proper folding of Cys subdomains in two mucins (Perez-Vilar, J., Randell, S.H., et al. 2004) and that it may have a pathological role in diabetic complications under hypoglycemic conditions (Ihara, Y., Manabe, S., et al. 2005).

A study involving site-directed mutagenesis of RNase 2 showed that the sequence WXXW, in which the first Trp becomes mannosylated, is the specificity determinant for C-mannosylation (Krieg, J., Hartmann, S., et al. 1998). In thrombospondin repeats, containing the motif WXXWXXWXXC (in some cases with one or two of the tryptophan residues substituted by other amino acids), C-mannosylation was found on one, two or all three tryptophans (Hofsteenge, J., Blommers, M., et al. 1999). The shortest peptide still valid as a substrate for C-mannosyltransferase found so far is WAKW (Hartmann, S. and Hofsteenge, J. 2000). However, in two particular thrombospondin repeats (from Complement component C6 and C7) the first tryptophan is mutated to phenylalanine or tyrosine respectively (Hofsteenge, J., Blommers, M., et al. 1999) and two recently discovered C-mannosylation sites in bovine lens fiber membrane intrinsic protein show no relationship at all to the WXXW motif (Ervin, L.A., Ball, L.E., et al. 2005). This indicates that although the WXXW motif seems to be a sufficient requirement for C-mannosylation, it does not seem to be a necessary one.

According to estimates based on the Swiss-Prot database, more than half of all proteins are glycosylated (Apweiler, R., Hermjakob, H., et al. 1999). However, despite the fact that human proteins are the most studied of all and that only proteins with some experimental verification are present in Swiss-Prot, only about 1.7% of human Swiss-Prot entries have experimentally verified glycosylation site information. To bridge an enormous gap between an exponentially increase in gene sequences in databases and a linear increase in proteins investigated for post-translational modifications, prediction methods are needed. Prediction of glycosylation sites is
a valuable tool when trying to characterize a new protein, e.g. for the interpretation of mass spectrometry results. Further, prediction of glycosylation sites is one of the important features when predicting orphan protein function (Jensen, L.J., Gupta, R., et al. 2003). Since glycosylation may affect the structure of the protein and occurs primarily in surface-exposed regions, predicted glycosylation sites may be used to improve protein structural prediction as well. Prediction can also be useful in protein engineering to incorporate or abolish glycosylation sites and to design competitive inhibitors of glycosyltransferases (Hansen, J.E., Lund, O., et al. 1998).

We have analyzed experimentally verified C-mannosylation sites with respect to sequence and structure. We have trained a predictor method, NetCGlyc 1.0, which correctly predicts 93% of both positive and negative C-mannosylation sites. This is a significant improvement over the WXXW consensus motif, which identifies only 67% of the positive sites. NetCGlyc 1.0 is publicly available at [http://www.cbs.dtu.dk/services/NetCGlyc/](http://www.cbs.dtu.dk/services/NetCGlyc/). Using NetCGlyc 1.0, we scanned the human genome for predicted C-mannosylation sites.

**Results**

**Sequence analysis**

From literature, we gathered a dataset of 12 native proteins plus 27 naturally occurring or engineered mutants/peptides that contain a total of 69 experimentally verified C-mannosylation sites and 88 non-modified sites. The sequence neighbourhood around the sites can be illustrated using sequence logos based on Shannon information content (Schneider, T.D. and Stephens, R.M. 1990), shown in Figure 1A, or Kullback-Leibler information content (Kesmir, C., van Noort, V., et al. 2003), shown in Figure 1B. The Shannon information logo
(1A) is based only on the residue occurrence in different positions in positive sites and show that the strongest discrimination is clearly at position +3, where mostly tryptophan and cysteine are accepted. Our analysis also indicates a strong preference for small and/or polar residues like serine, alanine, glycine and threonine at position +1 not previously reported. A repetition pattern where a tryptophan adjacent to a serine/glycine is repeated every 3 residues on either side of the glycosylation site is also evident. This is probably arising from the C-mannosylation sites located in thrombospondin repeats which contain WXXWXXWXXC motifs where one, two or all three of the tryptophans are glycosylated.

The Kullback-Leibler information logo (1B) is based on both positive and negative sites. Residues over-represented in positive sites are shown as normal letters and those that are over-represented in negative sites are shown as upside-down letters. Note that the modified tryptophan residue in the middle is cancelled out entirely since both positive and negative sites all have a tryptophan at that position. Not surprisingly, the strongest preference is again found at position +3, where tryptophan and to some extent cysteine is preferred and most other residues are discriminated against. We find that phenylalanine and leucine, both large and hydrophobic, are not tolerated at position +1 of the positive sites. We also find a number of residues at different positions, even surprisingly far away from the attachment site, that seem to be inconsistent with C-mannosylation: arginine/lysine at position -9, glutamine at positions -6 and 4, phenylalanine at position -5, histidine at position 5, aspartic acid at position 9 and alanine at position 10. Whether these are true reflections on the requirements for C-mannosylation or a result of insufficient sequence sampling in the data set is hard to say at this point.
**Structural analysis**

Using FeatureMap3D (Wernersson, R., Rapacki, K., et al. 2006), we were able to identify five NMR or X-ray structures in the worldwide Protein Data Bank (Berman, H., Henrick, K., et al. 2006) showing the structure of C-mannosylated proteins. Two of the structures (1SZL and 1LSL) show the structure of thrombospondin repeats. The fold of a thrombospondin repeat contains two \( \beta \) strands along with a third, fairly extended, but not hydrogen-bonded stretch running parallel to the \( \beta \) sheet (Figure 2B) (Paakkonen, K., Tossavainen, H., et al. 2006, Tan, K., Duquette, M., et al. 2002). The three, potentially glycosylated, tryptophans are situated in the non-\( \beta \) stretch. The aromatic rings of the three tryptophans are parallel to each other at a \( C_\alpha - C_\alpha \) distance of 8.3-8.5 Å, which is too long to allow aromatic stacking (\( \pi - \pi \) interactions).

In two particular thrombospondin repeats (from Complement component C6 and C7), C-mannosylation is found in this structural context without the presence of a true WXXW motif. Instead, the first tryptophan is mutated to phenylalanine or tyrosine respectively.

Two structures show similar local structures around the C-mannosylation site as compared to the thrombospondin repeats, 1EER (figure 2A) and 1F42 (not shown). Again, the glycosylated tryptophan is situated in a fairly extended, non-hydrogen bonded stretch running parallel to a \( \beta \) strand (Syed, R.S., Reid, S.W., et al. 1998, Yoon, C., Johnston, S.C., et al. 2000). The aromatic rings are parallel to each other at a \( C_\alpha - C_\alpha \) distance of 8.6 and 8.7 Å respectively.

One structure shows an entirely different local structure, 2BZZ (Figure 2C). The two tryptophans are located in an \( \alpha \)-helix and rotated so that the aromatic rings are face to edge at a \( C_\alpha - C_\alpha \) distance of 5.1 Å, indicating aromatic stacking between the rings (Baker, M.D., Holloway, D.E., et al. 2006). The protein has been co-crystallized with a ligand (not shown), but a ligand-free structure not available in the Protein Data Bank show very similar orientations of the tryptophan rings (Mosimann, S.C., Newton, D.L., et al. 1996).
Unfortunately, no structure was found for the only protein where the C-mannosylation sites are completely unrelated to the WXXW motif, lens fiber membrane intrinsic protein.

Based on the available structures, we find that the accessible surface according to DSSP is 30-147 Å² (mean 71 Å²) for glycosylated tryptophans and 0-85 Å² (mean 39 Å²) for non-glycosylated tryptophans, showing that modified tryptophans are, on average, more solvent-exposed and all of them are solvent exposed to a certain extent.

**Prediction of C-mannosylation sites**

Before developing a predictor using machine learning, we investigated what prediction performance is obtained when searching for the simple consensus pattern suggested: WXXW, where the first tryptophan would be glycosylated (Krieg, J., Hartmann, S., et al. 1998). This is the approach used so far and must ultimately be out-performed for a more complex machine learning approach to be worthwhile. In our dataset consisting of 69 positive and 88 negative sites, the consensus pattern predictor correctly identifies 67% of the positive sites and 93% of the negative sites (see Table 2). This means that the consensus rule does not apply for as much as one third of the positive sites in our data set. Since most experimental studies have so far been directed towards sites which follow the WXXW rule, our data set is, if anything, biased towards sites that do follow it. The number of true sites missed when using the consensus pattern predictor could therefore be much higher. As a test we trained neural networks based only on the information whether the WXXW pattern was present or not. Not surprisingly, these networks all had predictive performances identical to the consensus predictor itself.
In order to develop a more complex predictor, we used a neural network strategy developed for the prediction of mucin-type glycosylation sites (Julenius, K., Molgaard, A., et al. 2005). We transformed the sequence information (letters) in various ways into numbers that the neural network predictor can understand, to learn what type of encoding would work best for this particular predictor problem. We used sparse encoding (the standard way), profile encoding (the corresponding row in the BLOSUM62 matrix), PSI-BLAST profile encoding (the corresponding row in the profile computed from PSI-BLAST) and amino-acid composition. We also trained networks based only on sequence-derived features: predicted secondary structure, predicted surface accessibility and predicted disorder (three different definitions). The window size presented to the network varied up to 21 residues, with the possibly glycosylated tryptophan in the middle. Initially, we trained neural network predictors with five hidden neurons for all possible networks involving single features. The complexity of the neural network architecture, and therefore the number of parameters that needs to be learned, increases with the window size and the number of hidden neurons used. For these predictors, the Matthews correlation coefficient was calculated using a cross-validation scheme (see methods section) and the results are shown in Figure 3. The consensus pattern search performance (0.63) is shown as a thin black line. Of the four different ways to present the sequence, profile encoding was the most successful, with correlation coefficients above 0.80 for window sizes 7 and 11. Of the sequence-derived features, disorder prediction according to the DSSP loop/coil definition, and surface accessibility were the most successful with correlation coefficients above 0.63 for many window sizes.

To find the best possible combination of features, we used a greedy strategy, trying to combine what appeared to be good input information when training the single feature networks. We also combined with the information on the presence/absence of the WXXW
motif. For feature combinations that seemed promising, networks with varying number of hidden neurons (different network complexity) were trained. The very best combination was sparse encoding in a 21-residue window, plus information on the presence/absence of the WXXW motif, using eight hidden neurons. This network correctly identifies 93% of both the positive and the negative sites (see Table 2). Figure 4 shows the trade-off between making many positive predictions, of which some are false and making fewer 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. For comparison, the performance of the consensus pattern search is marked with an X.

**Scanning the human genome**

All human transcripts with signal peptides and/or transmembrane helices were scanned with NetCGlyc 1.0 for predicted C-mannosylation sites. Since C-mannosylation occurs in the ER, only tryptophans either in extracellular proteins or on the extracellular side of membrane proteins can be mannosylated. Of the 14554 downloaded transcripts, 2573 (18%) were predicted to contain at least one C-mannosylation site. These proteins were investigated for Gene Ontology (GO) annotation and the results are shown in Table 3. An enrichment factor $>1$ indicates that the term is over-represented for the C-mannosylated proteins. Of the 3713 predicted sites, 1366 were located at the first tryptophan in a WXXW motif, 214 were located at the second tryptophan in a WXXW motif and 2133 were found in different sequence contexts.

Investigating proteins with more than five predicted sites, we find that proteins with thrombospondin repeats are highly over-represented (e.g. semaphorins, brain-specific
angiogenesis inhibitors, ADAMTS’s) as would be expected. More surprisingly, we also find many proteins related to low-density lipoprotein receptor. Looking more closely at this class of proteins, we find that a substantial number of LDL-receptor class B repeats, also called YWTD repeats, have an additional tryptophan, making the repeated sequence YWTDW. According to PROSITE (http://expasy.org/prosite/), there are 47 such YWTDW repeats in the human proteome and our predictor predicts most of these to be positive for C-mannosylation. There are three available crystal structures (PDB ID 1IJQ, 1NPE and 1N7D) of LDL-receptor class B repeats from two different proteins (human LDL receptor and mouse nidogen 1). In both proteins, six repeats are packed very closely together in a six-bladed β-propeller (Jeon, H., Meng, W., et al. 2001, Rudenko, G., Henry, L., et al. 2002, Takagi, J., Yang, Y., et al. 2003). Because of a close contact with a hydrophobic residue on the preceding repeat (phenylalanine in both structures), the first tryptophan of the YWTDW sequence is inaccessible to the solvent. If all LDL-receptor class B repeats fold into six-bladed β-propellers, we find C-mannosylation at these sites highly unlikely. However, in the case of two of the YWTDW repeats, an additional tryptophan precedes the YWTDW repeat, making the sequence WMYWTDW. Judging from the available structures in which the corresponding position is occupied by a phenylalanine and an asparagine, respectively, the first tryptophan is more solvent-accessible and this residue is therefore a more likely a C-mannosylation site.

One of the characteristic structural features of type I cytokine receptors is a WSXWS motif in the C-terminal domain (Bazan, J.F. 1990). This has, at least in the case of erythropoietin receptor, been shown to be C-mannosylated (Furmanek, A., Hess, D., et al. 2003). We extracted 29 human protein sequences with annotated WSXWS motifs from Swiss-Prot and performed prediction of C-mannosylation sites using NetCGlyc 1.0. 27 of 29 proteins have at least one predicted site. The two exceptions were Growth hormone receptor (P10912) and Interleukin-3 receptor alpha chain (P26951), both with degenerated motifs (YGEFS and
Interestingly, several receptors contain more than one predicted C-mannosylation site. Interleukin-6 receptor subunit beta (P40189), Leptin receptor (P48357) and Leukemia inhibitory factor receptor (P42702) each contain as much as four predicted sites and what seems to be two WSXWS motifs. Type I cytokine receptors are classified as GO:0004896 (hematopoietin/interferon-class cytokine receptor activity), which explains the high enrichment factor (4.09) of this GO-term among the human transcripts predicted to be C-mannosylated (Table 3).

## Discussion

The structural analysis indicates that aromatic stacking may play a role in the substrate recognition of C-mannosyltransferase, at least in the case of substrates that contain the WXXW motif. Modified tryptophan residues are typically at least partly solvent-exposed, while non-modified tryptophans may be completely buried in the interior of the protein. Previous studies have shown C-mannosylation to take place very early, probably even before the folding of the protein (Doucey, M.A., Hess, D., et al. 1998, Perez-Vilar, J., Randell, S.H., et al. 2004). To explain the differences in solvent-accessibility of different tryptophans, we suggest that the modification probably, at least in some proteins, affects the folding of the protein. It would be interesting to investigate what prevents C-mannosylation of the YWTDW motifs of LDL-receptor class B repeats before folding, since this would then prevent the correct folding into six-bladed β-propellers.

The results of the training on predicted features (Figure 3) are in agreement with the results of the structural analysis. The fact that predicted surface accessibility proved to be good input information for the network method can be explained by the fact that glycosylated
tryptophans are more solvent accessible than the tryptophans which are not modified. Predicted disorder according to DSSP loop/coil definition was much better input information than any of the two other predicted disorder measures. In four of the five available structures, the glycosylated tryptophan is located in a fairly extended, non-hydrogen bonded stretch. These stretches are classified as loop or coil according to the DSSP definition, but are not particularly disordered according to the two other definitions, which require the loop/coil to have elevated temperature factor, “hot loops”, or atom coordinates to be missing in the structure. It is hardly surprising that prediction of a disorder definition that seems to apply to a large part of glycosylated tryptophans is good input information to the predictor network.

We were able to develop a predictor that predicts more sites than the WXXW consensus rule (higher sensitivity) without making any additional false predictions. Getting higher sensitivity without loss of specificity is usually very difficult, but can probably be explained by the fact that there is a lot of sequence information in various positions of the aligned sites (Figure 1) apart from the tryptophan in position +3. Our method is able to use these additional sites in an optimal way. We would like to point out that although this is the case, NetCGlyc 1.0 will work best on WXXW-related sites since most of the sites in the training examples were of this type. If future experiments show that C-mannosylation is common in other sequence contexts as well, NetCGlyc will be retrained to accommodate this.

By training a predictor, NetCGlyc 1.0, and making it publicly available among our other predictors for different types of glycosylation sites at our web page, www.cbs.dtu.dk/services, we hope to bring attention to this newly discovered type of glycosylation. The glycan is very small, only one hexose, which is probably why the modification was left undiscovered for so long. One hexose would not change the migration rate on a SDS PAGE gel enough to attract
attention to its presence, as compared to the large glycans of N-glycosylation and proteoglycans, or the numerous glycans of a mucin. The two newly discovered sites in Lens fiber membrane intrinsic protein (Ervin, L.A., Ball, L.E., et al. 2005) indicate that although tryptophan is the rarest of the amino acid residues, its modification with α-mannopyranose does not require the presence of a WXXW motif and may actually be more common than we think.

Materials and Methods

Data set

Experimentally verified mammalian C-mannosylation sites were extracted from O-GlycBase v6.00 (www.cbs.dtu.dk/databases/OGLYCBASE/) (Gupta, R., Birch, H., et al. 1999), Swiss-Prot (Boeckmann, B., Bairoch, A., et al. 2003) and through literature searches. We also found one protein reported to have no C-mannosylation sites. 12 native proteins plus 27 naturally occurring or engineered mutants/peptides were gathered in this way. The original articles were checked for the protein region investigated for glycosylation sites in each case. Tryptophans located in investigated regions and not reported as positive or partial sites were used as negative sites. Partly glycosylated tryptophans were used as positive sites. No tryptophans located in signal peptides were used. The dataset consisted of 69 positive and 88 negative sites.

Neural network training

For readability, this section was shortened to suit the average reader of Glycobiology. For details on sequence encoding, feature encoding and neural networks, see the supplementary material.
A neural network does not understand letters, so the amino acid sequence and different features must be translated into numbers. This is called the encoding and can be done in a number of ways. Each number that is presented to the neural network constitutes 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, J., Krogh, A., et al. 1991, Qian, N. and Sejnowski, T.J. 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, S. and Henikoff, J.G. 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, S.F., Madden, T.L., et al. 1997).
- **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, J.E., Lund, O., et al. 1998).
- **Protein disorder** was predicted using DisEMBL (Linding, R., Jensen, L.J., et al. 2003). DisEMBL predicts disorder according to three different definitions: 1) loops/coils as defined by DSSP (Kabsch, W. and Sander, C. 1983); 2) hot loops, being loops according to DSSP with a high degree of mobility as determined from $C_a$ temperature factors; 3) missing coordinates in X-ray structures.
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, since 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 loose the ability to make predictions for
examples that were not in the training set (the ability to generalize). This second problem is
sometimes called over-training 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 over-
training occurs and if the problem can be detected, it can also be avoided. The risk of over-
training increases with decreasing data set size.

The predictive performance was monitored using the Matthews correlation coefficient
\[C = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fn)(tn + fp)(tp + fn)(tp + fp)}}\]

where \(tp\) is the number of correctly predicted positive sites (true positives), \(tn\) the number of
correctly predicted negative sites (true negatives), \(fn\) the number of sites falsely predicted to
be negative (false negatives) and \(fp\) 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, $S_{n,\text{pos}}$, was computed as:

$$S_{n,\text{pos}} = \frac{t_p}{t_p + f_n}$$

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

$$S_p = \frac{t_p}{t_p + f_p}$$

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

$$S_{n,\text{neg}} = \frac{t_n}{t_n + f_p}$$

A region of 21 residues around each (positive or negative) site was extracted (10 amino acids on each side of the tryptophan). The sites were aligned according to the central tryptophan and an un-rooted neighbour-joining tree was constructed using CLUSTAL X (Thompson, J.D., Gibson, T.J., et al. 1997). From this tree, groups of closely related sites were identified. One or more of these groups were collected into larger sets, in total six, each containing both positive and negative sites and of roughly equal size. Between sites belonging to different sets, sequence identity did not exceed 50%. The six sets were used so that every network was trained six times, using five sets as training set and one set as test set. The reported cross-
validation performance is the joint performance of the six resulting networks on their respective test sets.

**Scanning the humane genome**

Sequences and their GO annotations for all human protein transcripts (build NCBI36) with either signal peptide and/or transmembrane helices were downloaded from [http://www.ensembl.org](http://www.ensembl.org) using the EnsMart system. Looking at GO annotations, “cellular component” terms were ignored. We compared the occurrences of different GO terms of the proteins predicted to be C-mannosylated to the occurrence of the different GO terms of all the protein transcripts, since some GO terms are more frequently occurring than others. The enrichment factor was calculated as the ratio between the occurrence of the term for the C-mannosylated sequences and the occurrence of the term for a random sample of the same size. An enrichment factor >1 indicates that the term is over-represented for the C-mannosylated proteins.

Proteins with annotated WSXWS motifs were extracted from Swiss-Prot by searching for the term “WSXWS motif.” in the “features” section of all entries. Human proteins were identified using the last part of the entry name, which is “_HUMAN” for human proteins. In total, 29 human type I cytokine receptors were identified in this way.

**Acknowledgments**

Kristoffer Rapacki is thanked for technical assistance in making the web predictor operational. Anne Mølgaard is thanked for help with the analysis of protein structures. Timo Pikkarainen is thanked for critical reading of the manuscript. This work was supported by Knut and Alice Wallenberg foundation.
References


**Figure legends**

**Figure 1.** Sequence logos for C-mannosylation sites. Position zero denotes the location of the glycosylated tryptophan residue. Amino acids are represented by their one-letter code and the letters are coloured according to the following scheme: hydrophobic residues in black, polar residues in green, acidic residues in red and basic residues in blue. A) The Shannon logo shows the frequencies of amino acid residues at each position in positive sites, as the relative heights of letters, along with the degree of sequence conservation as the total height of a stack of letters. B) The Kullback-Leibler logo shows the differences in frequencies of amino acid residues at each position in positive sites as compared to negative sites. Amino acids over-represented in positive sites are shown as regular letters; those over-represented in negative sites are shown as upside-down letters. Histidines over-represented in negative sites (upside-down H’s) are shown in light blue for clarity. The larger the skew is, the larger the letter will be.

**Figure 2.** Protein structures of C-mannosylated proteins. Figures were prepared using PyMol (http://pymol.sourceforge.net/). Glycosylated tryptophans are shown in yellow, WXXW non-glycosylated tryptophans in purple. A) Human erythropoietin receptor (1EER) (Syed, R.S., Reid, S.W., et al. 1998) B) Thrombospondin repeats (1LSL) (Tan, K., Duquette, M., et al. 2002) C) Eosinophil-derived neurotoxin (2BZZ) (Baker, M.D., Holloway, D.E., et al. 2006).

**Figure 3.** Cross-validation performance of neural networks trained on different features using five hidden neurons. The window size is the number of amino acids for which the information in question is presented to the network, with the tryptophan in question located in the center.
of the window. See Method section and the supplementary material for a detailed description of the different features. Sparse, profile and blast profile encoding are three different ways of representing the amino acid sequence. Amino acid composition is the frequency of amino acid residues within the window. Secondary structure, surface accessibility and disorder (three different definitions) are predicted from the amino acid sequence. The thin black line denotes the performance of a WXXW consensus pattern search.

**Figure 4.** ROC curve showing predictor performance of NetCGlyc. 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. The performance of the consensus pattern search (WXXW) is marked with an X.
### Table 1. Structural context of C-mannosylation sites

<table>
<thead>
<tr>
<th>Swiss-Prot</th>
<th>Protein name</th>
<th>PDB entry</th>
<th>Identity(^a)</th>
<th>Resol.</th>
<th>TSR(^b)</th>
<th>Site(^c)</th>
<th>S.S.(^d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P35446</td>
<td>Spondin-1 (F-spondin)</td>
<td>1SZL</td>
<td>100 %</td>
<td>NMR</td>
<td>yes</td>
<td>420(Trp)448</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>423(Trp)451</td>
<td>C</td>
</tr>
<tr>
<td>P29460</td>
<td>Interleukin-12 subunit (\beta)</td>
<td>1F42</td>
<td>97 %</td>
<td>2.50</td>
<td>no</td>
<td>297(Trp)297</td>
<td>C</td>
</tr>
<tr>
<td>P13671</td>
<td>Complement component C6</td>
<td>1LSL</td>
<td>44 %</td>
<td>1.90</td>
<td>yes</td>
<td>547(Trp)420</td>
<td>C</td>
</tr>
<tr>
<td>P10643</td>
<td>Complement component C7</td>
<td>1LSL</td>
<td>43 %</td>
<td>1.90</td>
<td>yes</td>
<td>481(Trp)477</td>
<td>C</td>
</tr>
<tr>
<td>P07357</td>
<td>Complement component C8 (\alpha) chain</td>
<td>1LSL</td>
<td>44 %</td>
<td>1.90</td>
<td>yes</td>
<td>522(Trp)420</td>
<td>C</td>
</tr>
<tr>
<td>P07358</td>
<td>Complement component C8 (\beta) chain</td>
<td>1LSL</td>
<td>52 %</td>
<td>1.90</td>
<td>yes</td>
<td>519(Trp)423</td>
<td>C</td>
</tr>
<tr>
<td>P14753</td>
<td>Erythropoietin receptor</td>
<td>1EER</td>
<td>82 %</td>
<td>1.90</td>
<td>no</td>
<td>208(Trp)209</td>
<td>C</td>
</tr>
<tr>
<td>P10153</td>
<td>Ribonuclease 2</td>
<td>2BZZ</td>
<td>99 %</td>
<td>0.98</td>
<td>no</td>
<td>11(Trp)1007</td>
<td>H</td>
</tr>
</tbody>
</table>

\(^a\) Sequence identity between query protein and PDB sequence.
\(^b\) C-mannosylation is located in a thrombospondin repeat or not.
\(^c\) The location of the C-mannosylation site. The number before the parenthesis refers to the numbering in the mature query protein and the number after the parenthesis refers to the numbering in the PDB entry.
\(^d\) DSSP secondary structure. “H” is \(\alpha\)-helix, “C” is random coil.
### Table 2. Performance of the NetCGlyc predictor.

<table>
<thead>
<tr>
<th>Method</th>
<th>C&lt;sup&gt;a&lt;/sup&gt;</th>
<th>S&lt;sub&gt;n,pos&lt;/sub&gt;&lt;sup&gt;b&lt;/sup&gt; (%)</th>
<th>S&lt;sub&gt;p&lt;/sub&gt;&lt;sup&gt;c&lt;/sup&gt; (%)</th>
<th>S&lt;sub&gt;n,neg&lt;/sub&gt;&lt;sup&gt;d&lt;/sup&gt; (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WXXW pattern search</td>
<td>0.63</td>
<td>66.7</td>
<td>88.5</td>
<td>93.2</td>
</tr>
<tr>
<td>NetCGlyc</td>
<td>0.86</td>
<td>92.8</td>
<td>91.4</td>
<td>93.2</td>
</tr>
</tbody>
</table>

<sup>a</sup> Matthews correlation coefficient

<sup>b</sup> Positive site sensitivity (the fraction of positive sites correctly predicted)

<sup>c</sup> Specificity (the fraction of all positive predictions that are correct)

<sup>d</sup> Negative site sensitivity (the fraction of negative sites correctly predicted)
Table 3. GO annotations for human proteins predicted to be C-mannosylated.

<table>
<thead>
<tr>
<th>Occurrence</th>
<th>Enrichment factor</th>
<th>GO term</th>
<th>GO annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>442</td>
<td>1.29</td>
<td>GO:0004872</td>
<td>receptor activity</td>
</tr>
<tr>
<td>257</td>
<td>1.00</td>
<td>GO:0005515</td>
<td>protein binding</td>
</tr>
<tr>
<td>227</td>
<td>0.92</td>
<td>GO:0007165</td>
<td>signal transduction</td>
</tr>
<tr>
<td>195</td>
<td>1.34</td>
<td>GO:0005509</td>
<td>calcium ion binding</td>
</tr>
<tr>
<td>152</td>
<td>1.32</td>
<td>GO:0006810</td>
<td>transport</td>
</tr>
<tr>
<td>148</td>
<td>0.84</td>
<td>GO:0007186</td>
<td>G-protein coupled receptor protein signaling pathway</td>
</tr>
<tr>
<td>129</td>
<td>1.25</td>
<td>GO:0016740</td>
<td>transferase activity</td>
</tr>
<tr>
<td>127</td>
<td>1.15</td>
<td>GO:0006811</td>
<td>ion transport</td>
</tr>
<tr>
<td>119</td>
<td>1.51</td>
<td>GO:0005524</td>
<td>ATP binding</td>
</tr>
<tr>
<td>110</td>
<td>1.01</td>
<td>GO:0007155</td>
<td>cell adhesion</td>
</tr>
<tr>
<td>109</td>
<td>1.28</td>
<td>GO:0005215</td>
<td>transporter activity</td>
</tr>
<tr>
<td>108</td>
<td>1.09</td>
<td>GO:0006508</td>
<td>proteolysis</td>
</tr>
<tr>
<td>102</td>
<td>1.34</td>
<td>GO:0008152</td>
<td>metabolism</td>
</tr>
<tr>
<td>101</td>
<td>1.48</td>
<td>GO:0000166</td>
<td>nucleotide binding</td>
</tr>
<tr>
<td>88</td>
<td>1.28</td>
<td>GO:0016787</td>
<td>hydrolase activity</td>
</tr>
<tr>
<td>83</td>
<td>1.10</td>
<td>GO:0001584</td>
<td>rhodopsin-like receptor activity</td>
</tr>
<tr>
<td>77</td>
<td>1.18</td>
<td>GO:0008270</td>
<td>zinc ion binding</td>
</tr>
<tr>
<td>75</td>
<td>0.94</td>
<td>GO:0046872</td>
<td>metal ion binding</td>
</tr>
<tr>
<td>74</td>
<td>0.98</td>
<td>GO:0006955</td>
<td>immune response</td>
</tr>
<tr>
<td>71</td>
<td>1.15</td>
<td>GO:0005554</td>
<td>molecular function unknown</td>
</tr>
<tr>
<td>65</td>
<td>1.35</td>
<td>GO:0007275</td>
<td>development</td>
</tr>
<tr>
<td>65</td>
<td>1.19</td>
<td>GO:0005216</td>
<td>ion channel activity</td>
</tr>
<tr>
<td>62</td>
<td>1.23</td>
<td>GO:0000004</td>
<td>biological process unknown</td>
</tr>
<tr>
<td>57</td>
<td>1.41</td>
<td>GO:0005529</td>
<td>sugar binding</td>
</tr>
<tr>
<td>52</td>
<td>1.08</td>
<td>GO:0006118</td>
<td>electron transport</td>
</tr>
<tr>
<td>49</td>
<td>2.29</td>
<td>GO:0016887</td>
<td>ATPase activity</td>
</tr>
<tr>
<td>49</td>
<td>0.63</td>
<td>GO:0050896</td>
<td>response to stimulus</td>
</tr>
<tr>
<td>48</td>
<td>1.90</td>
<td>GO:0004930</td>
<td>G-protein coupled receptor activity</td>
</tr>
<tr>
<td>47</td>
<td>4.09</td>
<td>GO:0004896</td>
<td>hematopoietin/interferon-class (D200-domain) cytokine receptor activity</td>
</tr>
<tr>
<td>47</td>
<td>1.48</td>
<td>GO:0006468</td>
<td>protein amino acid phosphorylation</td>
</tr>
<tr>
<td>47</td>
<td>1.47</td>
<td>GO:0006814</td>
<td>sodium ion transport</td>
</tr>
<tr>
<td>46</td>
<td>1.21</td>
<td>GO:0007399</td>
<td>nervous system development</td>
</tr>
<tr>
<td>45</td>
<td>1.72</td>
<td>GO:0006812</td>
<td>cation transport</td>
</tr>
<tr>
<td>43</td>
<td>1.78</td>
<td>GO:0006816</td>
<td>calcium ion transport</td>
</tr>
<tr>
<td>43</td>
<td>1.39</td>
<td>GO:0016757</td>
<td>transferase activity, transferring glycosyl groups</td>
</tr>
<tr>
<td>42</td>
<td>1.26</td>
<td>GO:0005975</td>
<td>carbohydrate metabolism</td>
</tr>
<tr>
<td>42</td>
<td>1.30</td>
<td>GO:0006629</td>
<td>lipid metabolism</td>
</tr>
<tr>
<td>42</td>
<td>1.26</td>
<td>GO:0030154</td>
<td>cell differentiation</td>
</tr>
<tr>
<td>41</td>
<td>1.84</td>
<td>GO:0004222</td>
<td>metalloendopeptidase activity</td>
</tr>
</tbody>
</table>
Figure 1
Figure 3

Matthews correlation coefficient vs. window size for different encoding methods and disorder indicators:
- Sparse encoding
- Profile encoding
- Blast profile encoding
- Amino acid composition
- Secondary structure
- Surface accessibility
- Disorder, DSSP loop/coil
- Disorder, hot loops
- Disorder, missing coordinates
Figure 4