ADCHEM 2006

International Symposium on Advanced Control of Chemical Processes
Gramado, Brazil – April 2-5, 2006

PREDICTION OF GLYCOSYLATION SITE-OCCUPANCY USING ARTIFICIAL NEURAL NETWORKS

Ryan S. Senger and M. Nazmul Karim

Department of Chemical Engineering
Texas Tech University
Lubbock, TX 79409 USA

Abstract: A novel neural network-based model was developed to predict *N*-linked glycosylation site-occupancy characteristics. The model classified potential glycosylation sites as displaying *variable* site-occupancy or *robust* glycosylation when produced by CHO cell cultures under normal growth conditions. The term *variable* site-occupancy describes heterogeneous glycan attachment to a specified protein site. This phenomenon results in a heterogeneous mixture of glycosylated and unglycosylated proteins when produced in mammalian cell culture. The model input consists of amino acid residues around the site of glycosylation. Simulation of the model strongly correlated with previously published experimental results by Kasturi *et al*. (1997) and Mellquist *et al*. (1998).

Keywords: Artificial intelligence, Biotechnology, Computer-aided design, Mathematical models, Neural networks

1. INTRODUCTION

Glycosylated Pharmaceutical Proteins. Protein glycosylation is a vital post-translational modification of many proteins with therapeutic properties. The glycosylation pathway begins in the endoplasmic reticulum of a cell with the attachment of an oligosaccharide to a N-*X*-S/T (where *X* is not praline) polypeptide sequence. The attachment of a glycan structure is then followed by enzymatic trimming and processing of the attached oligosaccharide (glycan) structure (Kornfeld and Kornfeld, 1985; Roth, 1987; Silberstein and Gilmore, 1996). Glycan attachment and remodeling processes occur in the endoplasmic reticulum (ER) and Golgi apparatus of many cell types; however, only a select number of cell types produce glycosylation variants compatible in humans. Glycosylation is of great importance in bioprocessing because of its large influence on *in vivo* properties of therapeutic proteins, including specific activity. In addition, intramolecular influences of glycosylation on protein structure include: proper folding, intracellular location, biological activity, solubility, antigenicity, biological half-life and protease sensitivity. Similarly, intermolecular characteristics affected by protein glycosylation include: targeting to lysosomes, tissue targeting, cell-cell adhesion and binding of pathogens (Stanley, 1992). Cell types commonly selected by the pharmaceutical industry, for expression of glycosylated proteins, include human melanoma cells, baby hamster kidney (BHK) cells, and Chinese hamster ovary (CHO) cells.

Optimization with Respect to Glycosylation. The heterogeneity observed with glycosylation during

bioprocessing and its relevance on the biological activity of therapeutic proteins has led to a new area of optimization in the bioprocessing industry. This new area of glycosylation optimization is currently divided into two parts: (1) the initial attachment of the oligosaccharide to the protein and (2) the processing of glycan branches. For some proteins, the initial glycan attachment process has been found to be *robust*, resulting in a homogenously glycosylated or unglycosylated polypeptide sequence. However, for others, such as the recombinant tissue-type plasminogen activator (r-tPA) protein, this process is *variable*, resulting in a mixture of heterogeneous isoforms (or glycoforms) of fully, partially and unglycosylated species (Kornfeld and Kornfeld, 1985; Grossbard, 1987; Wittwer and Howard, 1990; Andersen *et al*., 2000; Senger and Karim, 2003a). Manipulations of process variables and culture medium conditions have been found to largely impact the degree to which r-tPA is glycosylated at site N184. However, culture conditions resulting in homogenously glycosylated r-tPA have not been found (Andersen *et al*., 2000; Senger and Karim, 2003a,b). Given that other glycosylation sites of r-tPA experience homogenous glycosylation (Grossbard, 1987; Wittwer and Howard, 1990), a better understanding of the mechanisms that cause a particular glycosylation site to display *variable* site-occupancy is desired. Glycosylation optimization is of great benefit to pharmaceutical manufacturing in terms of production costs and would result in tighter control of product specific activity.

Neural Networks in Structural Bioinformatics. Neural network-based models have been developed for the prediction of many structural characteristics of proteins, based on the protein amino acid sequence. In particular, this area of structural bioinformatics has expanded to predict secondary and some three-dimensional structures (Rost and Sander, 1993; Jones, 1999; Kelley *et al*., 2000; Pollastri *et al*., 2002a,b; Baldi and Pollastri, 2003). In general, neural networks have been an intricate part of these model-developments in that their capability for structure prediction has far exceeded that of first-principle (deterministic) models. This is due in large part to the expanding data bank of protein structure and genomic research (Rost, 2001).

Using Neural Networks to Predict Glycosylation Site-Occupancy Characteristics. A novel neural-network model has been developed in this research for predictions of glycosylation site-occupancy as homogeneous (*robust*) or heterogeneous (*variable*). The development of this model has allowed insight into many questions concerning protein glycosylation. The phenomena of variable site-occupancy, in the absence of substrate limitation, was found related to primary sequence characteristics. The number of amino acid residues around the site of glycosylation with influence on glycosylation characteristics was found much larger than what has been cited by previous research. Thus, the goal of this research is to develop a model of glycosylation site-occupancy so the optimization problem of glycosylation site-occupancy may be addressed through site-directed mutations of the protein sequence rather than manipulation of cell culture variables.

2. SYSTEMS AND METHODS

Acquired Data and Amino Acid Residue Quantification. Data for the construction of a glycosylation site-occupancy prediction model consisted of glycosylation sites, the amino acid sequence around this site and whether a particular site promoted homogeneous (*robust*) or heterogeneous (*variable*) glycosylation site-occupancy when produced by mammalian cell cultures. All data was acquired from a literature search interfaced with protein sequence databases. The entire data set consisted of 48 glycosylation sites. Five sequences (~10%) were reserved as a neural network testing data set. Greater than 40% of sequences of the data set were classified as displaying *variable* site-occupancy in the literature. For the input of particular amino acid residues into a neural network model, the identities of all amino acids were first converted to numerical values. Individual amino acid residues were grouped into eleven classes based on similar characteristics, such as charge, size, and hydrophobicity and assigned numerical values based on research by Kasturi *et al*. (1997) and Mellquist *et al*. (1998). Quantification of the target (site-occupancy classification) was also required. Glycosylation sites displaying *variable* site-occupancy were assigned the value *1*, and *robust* sites were assigned *0*. It is noted that a *robust* site may be homogeneously glycosylated or homogenously unglycosylated. Statistical models have been developed to discern between these two types of *robust* site-occupancy (Petrescu *et al*., 2004).

Table 1 Primary Sequence Quantification

Amino Acid Classes	Amino Acids	Assigned Value
Hydroxy	T	1
Hydroxy	S	2
Basic	K R H	3
Thioether	M	4
Alkyl	A V L I	5
Carboxamide	N Q	6
Unsubstituted	G	7
Acidic	D E	8
Mercapto	C	9
Aromatic	F Y W	10
Cyclic	P	11

Neural Network Architecture. Elman recurrent neural networks were used for the construction of the neural network-based model. Recurrent neural networks are renowned for their ability to learn non-causal data sets. The neural network inputs consisted of quantified amino acid residues around the site of glycosylation. Targets consisted of glycosylation site-occupancy assigned values. All neural networks consisted of a single hidden layer with hyperbolic sigmoid transfer functions. A single output neuron was used with a log sigmoid transfer function. A single perceptron neuron was used following the output neuron for two-dimensional classification. Thus, this neuron acted as a rounding function of the recurrent neural network output value. The number of hidden layer neurons was adjusted so that the number of adjustable network parameters (weight and bias values) always remained less than the number of data points used in the training procedure. Initial (prior to training) weight and bias values were assigned random values. Network training was performed using gradient decent with momentum and adaptive learning rate back-propagation. All neural networks were trained for 2000 epochs. Each neural network was independently initiated and trained 100 times. Results were averaged. The entire data set was cross-correlated using a testing set size of approximately 10% of the training set size.

Optimization of the Amino Acid Input Sequence. The number of amino acids on the *N*-terminus and *C*-terminus sites of the glycosylation site was varied, and neural network training and testing set analysis was performed in each case. This goal of this study was to identify relevant amino acids in prediction of glycosylation site-occupancy classification. In this study, the objective function of optimization problem was the mean-square error between the target values of glycosylation site-occupancy and the neural network-predicted values.

Comprehensive Model Construction. Once successful neural networks were identified, with an optimum input sequence length, a comprehensive model was constructed and further tested on published experimental data. The comprehensive model consisted of 20 neural networks that were found to correctly classify all elements of the neural network testing data sets. Networks from all cross-correlation iterations were used to construct the model. This composition of the comprehensive predictive model enabled the model to return an overall prediction value as well as a confidence interval. In particular, all neural networks were simulated, and results were averaged before perceptron classification. This method returned an overall model prediction. The confidence interval represents the fraction of neural networks returning the dominant classification. Thus, the overall model prediction consisted of a value of either *1* or *0*, and the confidence level of prediction was a value between 0.5 (low confidence) and 1 (high confidence).

Further Simulations. Published experimental data by Kasturi *et al.*, (1997) and Mellquist *et al.*, (1998) was used to further test the comprehensive neural network-based model. The published data focused on the effects of site-directed mutations around *variable* site-occupancy glycosylation site N39 of the rabies virus glycoprotein (rgp). Results of this work found that specific site-directed mutations resulted in the transformation of this glycosylation site from *variable* to *robust* site-occupancy. All of these sequences were simulated using the predictive model. In addition, further simulations were performed on simple theoretical sequences to examine the effects of charged amino acid residues around the site of glycosylation. Alanine (uncharged), lysine (negatively-charged) and aspartate (positively-charged) residues were used in this study.

3. RESULTS AND DISCUSSION

Optimization of the Glycosylation Window Length. The number of amino acids surrounding the site of glycosylation was termed the *glycosylation window*. Optimization of the glycosylation window length was performed by full neural network analysis with various input sequence lengths. Previous research has identified amino acid residues of the N-*X*-S/T-*Y*

glycosylation sequence as having significant influence over glycosylation site-occupancy characteristics (Shakin-Eshleman, 1996; Kasturi *et al*., 1997; Mellquist *et al*., 1998; Petrescu *et al*., 2004). The data-based method of analysis of this research allowed for the impact of 20 amino acid residue sites to be analyzed for influences on glycosylation characteristics. For each input sequence, the mean-square error was calculated between averaged neural network predictions and target values (glycosylation classification). Results are displayed as Figure 1. The starting residue of the input sequence is displayed on the abscissa as (*n*-x). The ending residue of a glycosylation window is displayed on the ordinate axis as (*n*+y), where *n* is the site of glycosylation. For example, a glycosylation window originating at (*n*-5) and extending to (*n*+4) contains a total of 10 amino acids: 5 residues on the *N*-terminus side of the glycosylation site, the glycosylation site itself (*n*) and 4 residues on the *C*-terminus side of the glycosylation site. The average standard deviation of all data points of Figure 1 was calculated as approximately 5% of the given data point value. Results showed a minimum mean-square error value of 0.0767 for the glycosylation window originating at (*n*-5) and extending to (*n*+4). The size of this glycosylation window is larger than others determined by experimental methods, and these are the first results of our knowledge to suggest influence of residues on the *N*-terminus side of a glycosylation site on glycosylation site-occupancy. For comparison, the data set was predicted by random values (in the absence of neural network training). These results were classified by the perceptron to yield and average mean-square error value of 0.7.

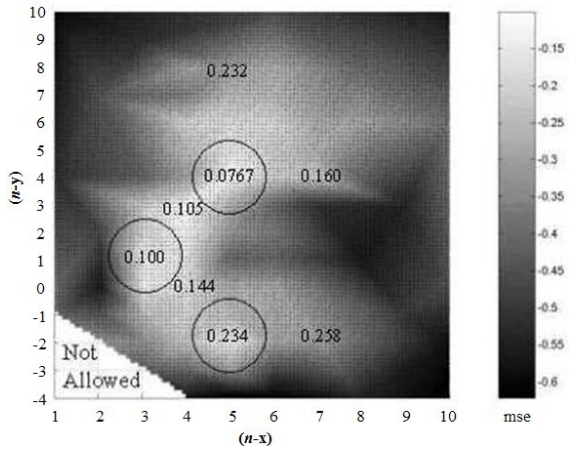

Fig. 1. Glycosylation window length optimization. Starting residue (abscissa). Ending (ordinate).

Simulations of rgp Wild-Type and Mutants Using Comprehensive Predictive Model. The comprehensive neural network-based predictive model was constructed using the optimized glycosylation window input length. The model consisted of 20 independent neural networks, and represented all iterations of the cross-correlation analysis. All neural networks of the comprehensive model classified corresponding testing data sets with 100% accuracy following perceptron classification of recurrent network output values. To further verify the predictive model, wild-type and site-directed mutations of the rgp protein glycosylation site N37 were simulated and compared to published experimental observations by Kasturi *et al*. (1997) and Mellquist *et al*. (1998). These experimental studies examined the influence of amino acid residues at positions *X* and *Y* of the N-*X*-S/T-*Y* glycosylation sequence. These sites correspond to (*n*+1) and (*n*+4) using the terminology developed for the glycosylation window length optimization. Results reported by Kasturi *et al*. and Mellquist *et al*. (1998) reported *glycosylation efficiency*. In short, the glycosylation efficiency is defined as the fraction of fully glycosylated rgp (N37). Thus, reported glycosylation efficiency between 0 and 1 corresponds to *variable* site-occupancy glycosylation. A glycosylation efficiency value of 1 corresponds to homogeneous (*robust*) glycosylation, and a glycosylation efficiency of 0 corresponds to a glycosylation site that is homogeneously (*robust*) unglycosylated. Values of glycosylation efficiency were interpolated from the published experimental studies, taking into account experimental error, and these are listed in Table 2. A total of 19 rgp mutants, in addition to the wild-type protein, were evaluated by the comprehensive predictive model. The sequence identity (details of site-directed mutations), the overall predictive model classification and the confidence level are also reported in Table 3.

Discussion of Prediction Results. Overall, the predictive model showed 95% accuracy in predicting glycosylation site-occupancy characteristics for this set of published experimental data. One rgp mutant (simulation 15a; S39T G40W) was incorrectly classified by the predictive model. However, in this case, the confidence level of the prediction was low (0.65). Although, the unsuccessful model prediction contained an aromatic residue (tryptophan), other predictions involving aromatic residues were classified correctly (simulation 8a). Sequences displaying *variable* site-occupancy, to a glycosylation efficiency of 0.9 (simulations 2a-4a, 8a, 10a-13a, 14a-15a) were correctly classified by the model as promoting *variable* site-occupancy. For

sequences that were constructed that resulted in glycosylation efficiency of or exceeding 0.95, most model predictions (simulations 6a, 7a, 20a) classified these sequences as having *robust* glycosylation. The exception in this case was for the L38N S39T mutant (simulation 6a), which displayed a glycosylation efficiency of approximately 0.95. *Variable* site-occupancy was predicted in this case, but a lower confidence level in this case (0.7) suggested that many neural networks of the predictive model recognized this sequence as promoting *robust* glycosylation. In addition, given the experimental error in the case of simulation 6a of roughly 5%, *variable* site-occupancy may accurately describe this system. The other sequences of simulations 7a and 20a displayed glycosylation efficiencies that exceeded 0.95. Of further importance is that the sensitivity of the predictive model was evaluated by this set of simulations. In short, this set of model predictions correctly classified a polypeptide sequence as having *variable* site-occupancy glycosylation characteristics for glycosylation efficiencies ranging between 0.1 to >0.95.

Simulations of Theoretical Sequences. Due to the success of the comprehensive predictive model in classification of rgp wild-type and variant sequences, the same simulation technique was applied to theoretical sequences. Glycosylation site-occupancy classification in the presence of charged amino acid residues was studied in the following simulations. For consistency, only alanine (A) was used as the uncharged residue outside of the required N-*X*-S/T glycosylation sequence. In addition, aspartate (D) and lysine (K) were used as the negatively and positively-charged residues, respectively. Simulations suggested that a sequence consisting of alanine or aspartate residues throughout the glycosylation window (except for the glycosylation sequence at (*n*) and (*n*+2) would result in robust glycosylation with a high confidence level. With positively-charged lysine residues occupying the glycosylation window and serine or threonine at position (*n*+2), variable site-occupancy glycosylation was predicted with a confidence level of 0.95. Further simulations examined the influence of particular locations within the glycosylation window. For example, a glycosylation window consisting of lysine residues was substituted with aspartate and alanine residues until a robust glycosylation site-occupancy prediction was achieved. A summary of these simulation results is presented in Table 3, and the actual simulation results are given in Table 4. These types of simulation experiments were performed for both serine and threonine in the (*n*+2) position. It was found through simulation that replacement of serine in the glycosylation sequence with threonine increases the robustness of glycan attachment. This evidence further supports this idea, as it was suggested by previous research (Kasturi *et al*., 1997; Mellquist *et al*., 1998; Petrescu *et al*., 2004).

Table 2 rgp variant and wild-type predictions with confidence level and published experimental results

Sim.	Sequence	Overall Model Classification and Confidence Level	Published Glycosylation Efficiency
1a	Wild-type	1 (1.00)	0.35
2a	S39T	1 (0.95)	0.80
3a	L38N	1 (1.00)	0.70
4a	L38S	1 (1.00)	0.90
5a	L38W	0 (0.60)	0.10
6a	L38N S39T	1 (0.70)	0.95
7a	L38G S39T	0 (0.60)	>0.95
8a	G40F	1 (0.75)	0.40
9a	G40P	0 (0.75)	0.05
10a	G40H	1 (1.00)	0.55
11a	G40M	1 (1.00)	0.70
12a	G40N	1 (1.00)	0.80
13a	G40S	1 (1.00)	0.80
14a	G40C	1 (0.90)	0.80
15a	S39T G40W	0 (0.65)	0.80
16a	S39T G40H	1 (1.00)	0.85
17a	S39T G40M	1 (1.00)	0.90
18a	S39T G40N	1 (1.00)	0.90
19a	S39T G40T	1 (1.00)	0.90
20a	S39T G40C	0 (0.55)	>0.95

Sim. is an abbreviation for "Corresponding Simulation."
The Confidence Level is listed in parentheses.
Published glycosylation efficiency values were interpolated from Kasturi *et al*. (1997) and Mellquist *et al*. (1998).

Table 4 Generalizations from simulations

Residues:	Type of influence on glycosylation site-occupancy:
Lysine (positive charge)	Promotes *variable*
Aspartate (negative charge)	Promotes *robust*
Alanine (no charge)	Promotes *robust* with less influence than aspartate
Location in glycosylation window:	**Level of influence on glycosylation site-occupancy:**
(*n*+1)	Highest influence
(*n*+3), (*n*+4), (*n*-1)	Moderate influence
(*n*-5)…(*n*-2)	Lowest influence, but significance was observed
(*n*+2)	Threonine results in more *robust* glycosylation

Table 5 Amino acid residues and simulation results of theoretical sequences

-5	-4	-3	-2	-1	*n*	+1	+2	+3	+4	O.C. (C.L.)
A	A	A	A	A	N	A	S	A	A	0 (0.75)
A	A	A	A	A	N	A	T	A	A	0 (1.00)
A	A	A	A	A	N	P	S	A	A	0 (1.00)
A	A	A	A	A	N	P	T	A	A	0 (1.00)
K	K	K	K	K	N	K	S	K	K	1 (1.00)
K	K	K	K	K	N	K	T	K	K	1 (0.95)
D	D	D	D	D	N	D	S	D	D	0 (1.00)
D	D	D	D	D	N	D	T	D	D	0 (1.00)
A	A	A	A	A	N	K	S	A	A	1 (0.95)
A	A	A	A	A	N	A	S	K	A	1 (0.70)
A	A	A	A	K	N	A	S	A	A	0 (0.60)
A	A	A	A	A	N	K	S	K	A	1 (1.00)
A	A	A	A	A	N	D	S	D	A	0 (1.00)
A	A	A	K	K	N	A	S	A	A	0 (0.60)
A	A	K	K	K	N	A	S	A	A	1 (0.65)
K	K	K	K	K	N	A	S	A	A	1 (0.80)
A	A	K	K	K	N	A	T	A	A	0 (0.60)
A	K	K	K	K	N	A	T	A	A	1 (0.70)
D	D	K	K	K	N	A	S	A	A	0 (0.80)
D	D	D	D	D	N	K	S	D	D	1 (0.90)
D	D	D	D	D	N	A	S	D	D	0 (0.90)
D	D	D	D	D	N	K	T	D	D	1 (0.50)
K	K	K	K	K	N	D	S	K	K	0 (0.85)

O.C. is an abbreviation for "Overall Model Classification."
C.L. is an abbreviation for "Confidence Level."
The Confidence Level is listed in parentheses.

4. CONCLUSIONS

A novel neural network-based predictive model has been developed for the classification of *N*-linked glycosylation as heterogeneous (*variable*) or homogeneous (*robust*) for proteins produced by mammalian cell culture. Amino acid residues around the site of glycosylation were found to impact site-occupancy characteristics. In particular, an optimization study found that 5 residues on the *N*-terminus side and 4 residues on the *C*-terminus side of the glycosylation site directly influence these characteristics. The neural network-based predictive model classified published experimental findings regarding the impact of amino acid residues on site-occupancy characteristics with 95% accuracy. Further simulations with theoretical amino acid sequences revealed negatively-charged promote *robust* glycosylation and that the (*n*+1) position of the glycosylation window had the most influence over glycosylation site-occupancy characteristics. Elimination of *variable* site-occupancy will have significant impact in the pharmaceutical industry. *Robust* glycosylation results in homogenous product production. This is of utmost importance to quality control and optimization of biological activity of glycosylated recombinant proteins with therapeutic properties.

REFERENCES

Andersen, D.C., T. Bridges, M. Gawleitzek and C. Hoy (2000). Multiple cell culture factors can affect the glycosylation of Asn-184 in CHO-produced tissue-type plasminogen activator, *Biotechnol. Bioeng*., **70**, 25-31.

Baldi P. and G. Pollastri (2003). The principled design of large-scale recursive neural network architectures-DAG-RNNs and the protein structure prediction problem, *Machine Learning*, **4**, 575-602.

Grossbard, E.B. (1987). Recombinant tissue plasminogen activator: A brief review, *Pharm. Res*., **4**, 375-378.

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

Kasturi, L., C. Hegang and S.H. Shakin-Eshleman (1997). Regulation of *N*-linked core glycosylation: use of a site-directed mutagenesis approach to identify Asn-Xaa-Ser/Thr sequons that are poor oligosaccharide acceptors. *Biochem. J*., **323**, 415-419.

Kelley, L.A., R.M. MacCallum and M. J. E. Sternberg (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM, *J. Mol. Biol*., **299**, 499-520.

Kornfeld, R. and S. Kornfeld (1985). Assembly of asparagine-linked oligosaccharides, *Ann. Rev. Biochem*., **54**, 631-664.

Mellquist, J.L., L. Kasturi, S.L. Spitalnik and S.H. Shakin-Eshleman (1998). The amino acid following an Asn-X-Ser/Thr sequon is an important determinant of *N*-linked core glycosylation efficiency. *Biochemistry*, **37**, 6833-6837.

Petrescu, A.J., A.L. Milac, S.M. Petrescu, R.A. Dwek and M.R. Wormald (2004). Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding, *Glycobiology*, **14**, 103-114.

Pollastri G., D. Przybylski, B. Rost and P. Baldi (2002a). Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles, *Proteins*, **47**, 228-235.

Pollastri, G., P. Baldi, P. Fariselli and R. Casadio (2002b). Prediction of coordination number and relative solvent accessibility in proteins, *Proteins*, **47**, 142-153.

Rost B. (2001). Review: Protein secondary structure prediction continues to rise, *J. Struct. Biol*., **134**, 204-218.

Rost, B. and C. Sander (1993). Prediction of protein secondary structure at better than 70% accuracy, *J. Mol. Biol*, **232**, 584-599.

Roth, J. (1987). Subcellular organization of glycosylation in mammalian cells, *Biochim. Biophys. Acta*, **906**, 405-436.

Senger, R.S. and M.N. Karim (2003a). Effect of shear stress on intrinsic CHO culture state and glycosylation of recombinant tissue-type plasminogen activator protein, *Biotechnol. Prog*., **19**, 1199-1209.

Senger, R.S. and M. N. Karim (2003b). Neural-network-identification of tissue-type plasminogen activator protein production and glycosylation in CHO cell culture under shear environment, *Biotechnol. Prog*., **19**, 1828-1836.

Shakin-Eshleman, S.H. (1996). Regulation of N-linked core-glycosylation, *Trends Glycosci. Glyc*., **8**, 115-130.

Silberstein, S. and R. Gilmore (1996). Biochemistry, molecular biology, and genetics of the oligosaccharyltransferase, *FASEB J*., **10**, 849-858.

Stanley, P. (1992). Glycosylation engineering, *Glycobiology*, **2**, 99-107.

Wittwer, J. and S.C. Howard (1990). Glycosylation at Asn184 inhibits the conversion of single-chain to two-chain tissue-type plasminogen activator by plasmin, *Biochemistry*, **29**, 4175-4180.