Prediction of $N$-linked glycan branching patterns using artificial neural networks

Ryan S. Senger a,*, M. Nazmul Karim b

a Delaware Biotechnology Institute, University of Delaware, Newark, DE 19711, USA
b Department of Chemical Engineering, Texas Tech University, Lubbock, TX 79409, USA

Received 27 January 2007; received in revised form 11 October 2007; accepted 11 October 2007
Available online 30 October 2007

Abstract

A model was developed for novel prediction of $N$-linked glycan branching pattern classification for CHO-derived $N$-linked glycoproteins. The model consists of 30 independent recurrent neural networks and uses predicted quantities of secondary structure elements and residue solvent accessibility as an input vector. The model was designed to predict the major component of a heterogeneous mixture of CHO-derived glycoforms of a recombinant protein under normal growth conditions. Resulting glycosylation prediction is classified as either complex-type or high mannose. The incorporation of predicted quantities in the input vector allowed for theoretical mutant $N$-linked glycan branching predictions without initial experimental analysis of protein structures. Primary amino acid sequence data were effectively eliminated from the input vector space based on neural network prediction analyses. This provided further evidence that localized protein secondary structure elements and conformational structure may play more important roles in determining glycan branching patterns than does the primary sequence of a polypeptide. A confidence interval parameter was incorporated into the model to enable identification of false predictions. The model was further tested using published experimental results for mutants of the tissue-type plasminogen activator protein [J. Wilhelm, S.G. Lee, N.K. Kalyan, S.M. Cheng, F. Wiener, W. Pierzchala, P.P. Hung, Alterations in the domain structure of tissue-type plasminogen activator change the nature of asparagine glycosylation, Biotechnology (N.Y.) 8 (1990) 321–325].

© 2007 Elsevier Inc. All rights reserved.
1. Introduction and background

1.1. N-linked glycan branching

Most secreted and membrane-associated proteins of eukaryotic cells are subjected to N-linked glycosylation, among other post-translational modifications. N-linked glycosylation encompasses a large degree of heterogeneity and enables variations in terminal linkages such as sialylation, fucosylation and sulfation [1–3]. Many combinations of glycan structures with varying terminal linkages have been found to influence both intermolecular and intramolecular properties of polypeptides including biological activity, intracellular location and tissue-targeting [4–9]. It is noted that cytosolic glycosylation has been observed for nuclear proteins as well as for those proteins remaining in the cytosol [10–12], but these processes are not discussed further here. The focus of this research remains to those proteins in the secretory pathway. As a result of varying glycosyltransferase activity in the rough endoplasmic reticulum (RER) and Golgi apparatus, the topography of N-linked glycan structures can take on broadly-classified high mannose, hybrid or complex-type configurations, and these are dictated by the level of substrate-limited glycosyltransferase processing. N-linked glycosylation events in the RER include the initial attachment of a lipid-linked oligosaccharide, Glc3Man9GlcNAc2, to a N-X-S/T polypeptide sequence, where X is not proline [1–3,13]. Factors affecting the robustness of this attachment reaction in mammalian expression systems have been the focus of much previous research for pharmaceutical proteins [14–18]. Following N-linked glycan attachment, the terminal glucose is enzymatically removed by a membrane-bound specific α-1,2 glucosidase I. Remaining glucose residues are removed by the membrane-bound α-1,3 glucosidase II enzyme, leaving a high mannose glycan structure. The enzymatic removal of at least one branched α-1,2-linked mannose residue also occurs in the RER by the membrane-bound α-1,2-mannosidase [1–3]. Glycan structures not contacted by this enzyme generally remain destined to exhibit high mannose branching patterns (although exceptions exist); where as, glycan structures processed by α-1,2-mannosidase receive the potential for further processing in the cis and medial-Golgi compartments to produce complex-type or hybrid glycan structures. Mechanistic explanations as to why N-linked glycan structures of certain glycosylated polypeptides are more highly processed to complex-type branching than those N-linked glycan structures of other glycosylated polypeptides, given the same organism and substrate limitations, remains unclear. However, it has been suggested that glycan structures retaining high mannose branching reside in regions of low solvent exposure following initial protein folding rearrangements in the RER [1–3].

1.2. Regulation of N-linked glycan branching

Studies to identify the contributing factors that impact N-linked glycosylation have revealed that the amino acid sequence of the glycosylated polypeptide, itself, the expression system and the extracellular environment are all vital to determining glycan branching patterns and terminal
The significance of the amino acid sequence of the glycoprotein is obvious in that different glycoproteins, produced by the same mammalian expression system, often exhibit different glycosylation. The importance of the particular mammalian expression system is apparent as a conserved glycoprotein, when produced by multiple expression systems, rarely shows identical glycosylation in all cases. However, interplay of the glycoprotein primary sequence given multiple mammalian expression systems exists as families of closely related glycan structures have been found to result for several glycoproteins [20–24]. Given the observation that a particular mammalian expression system yields a particular N-linked glycan branching pattern for a particular glycoprotein, this does not imply that this same branching pattern should be anticipated for this glycoprotein from all mammalian expression systems. This is especially true as non-mammalian expression systems are considered. In fact, given different expression systems, the degree to which glycosylation is influenced by the primary sequence of the glycoprotein, itself, and the expression system is believed to be glycosylation site-dependent. In addition, the configuration of the polypeptide backbone has been shown important to oligosaccharide processing of proteins exiting the ER [25–27]. As a further example, the modification potential of hydroxy-amino acids of a polypeptide sequence has been linked to conformational structures and independent of the primary sequence [28]. Thus, the interaction of a glycoprotein with membrane-bound glycosyltransferase enzymes (of a mammalian expression system) is hypothesized to be regulated by defined secondary structure elements and three-dimensional polypeptide conformations. These structural elements may hinder or promote interactions between the glycan structure and membrane-bound glycosyltransferases, but the details of these interactions remain largely unknown.

1.3. Neural networks in predictions of protein structure elements and post-translational modifications

The development of high-throughput technologies has not eluded post-translational modifications and glycosylation, leading to the development of specialized databases and predictive models [29,30]. In particular, artificial neural networks have served as a useful model structure to discern complicated relationships and develop predictive models based on non-linear mappings between primary sequence data and the state of a post-translational modification. In particular, neural networks have been applied for identification of phosphorylation sites [29,31], O-GalNAc glycosylation [32] and N-linked glycosylation sites via sequon identification. Neural networks have also played an important role in the prediction of secondary structure elements of polypeptides from primary sequence data [33,34]. Prediction accuracy of neural network-based secondary structure algorithms has approached 78% [35] of a theoretical limit of 88% [36,37]. Despite the fact that secondary structure prediction has remained a benchmark problem for the development of neural network-based protein structure predictions, the scope of this prediction methodology has extended to many other areas of protein structure. For example, the ACCpro and CONpro prediction algorithms were developed with the use of bi-directional neural networks for the prediction of residue solvent exposure and the relative number of residue contacts at a specified distance, respectively [38–41]. The binding states of specific cysteine residues were predicted through a further development [42]. The polypeptide sequence was also coupled with structural characteristics of enzymes for a neural network-based identification of the catalytic residues [43]. In addition, the interface of protein–protein interactions [44] as well as intracellular glycoprotein location [45] have become predictable entities using neural network-based algorithms with primary
sequence inputs. Progress has also been made in the use of neural networks for the prediction of N-linked glycosylation site-occupancy in which the condition of variable site-occupancy was found predictable by primary sequence inputs [18].

1.4. Neural networks for N-linked glycan branching pattern prediction

A neural network-based model was developed in this research for the prediction of N-linked glycan branching characteristics of a glycosylated polypeptide. The model predicts glycan branching for the major fraction of a heterogeneous mixture of glycan structures synthesized by eukaryotic expression. Glycan branching classification was divided into two categories: high mannose and complex-type. Hybrid glycan structures were not considered in this study, as hybrid structures were not found to be a major glycoform component of any proteins comprising the reference data set. In addition, it is noted that N-linked glycan branching is highly heterogeneous by nature for eukaryotic expression [15]. Thus, the scope of this research remains to identify the classification of the major glycoform component of a possibly heterogeneous mixture of glycoforms. In the presence of completely or partially-folded polypeptide structures in the RER, the determination of glycan branching by glycosyltransferase activity was expected to become governed by not only the primary sequence but also by secondary structure elements as well as solvent accessibility. Therefore, these structural inputs were examined separately and in addition to primary sequence inputs for the prediction of glycan branching classification. It is noted that predicted secondary structure elements and solvent accessibility were examined as possible input quantities despite findings that glycan structures may influence polypeptide backbone configurations [46]. Data compiled for reference set construction consisted of proteins and glycosylation characteristics from Chinese hamster ovary (CHO) cell culture expression.

2. Systems and methods

2.1. Data acquisition and reference set construction

Data of N-linked glycan branching patterns for particular glycosylated polypeptide sequences were obtained from a massive literature search. In particular, data were collected for CHO expression of proteins in which comprehensive glycosylation analysis had been performed. As eukaryotic protein expression of glycosylated proteins commonly results in a heterogeneous mixture of glycoforms, only the major glycoform was considered. Also, only conditions in which a particular glycoform comprised a large fraction of the heterogeneous mixture were considered. Both classes of N-linked glycans (complex-type and high mannose) exhibit considerable heterogeneity; however, these glycoforms were clustered into two categories for the purpose of this research. In addition, hybrid glycan structures were not considered in this research as these structures failed to comprise the major glycoform fraction of any of the literature cases reviewed. Thus, predictions were performed to classify the major glycoform from CHO culture expression as complex-type or of high mannose branching classification. In all, 158 protein sequences and their corresponding glycosylation branching classifications made up the entire reference set. The primary sequence, predicted secondary structures and predicted solvent accessibility were mapped to corresponding
N-linked glycan branching classification by the use of recurrent neural networks. Data sets were constructed for neural network training procedures and consisted of data for 148 proteins. Data of the remaining 10 proteins were included in a neural network testing data set. It is noted that components of the neural network testing data sets were not included in any neural network training procedures. A cross-validation training procedure was used in neural network evaluation. A cross-validation procedure requires multiple neural network training procedures in which new components of the total reference set are selected for the neural network testing set and the remaining components of the total reference set are used as the training data set. Of course, in this procedure, each neural network was independent as the cross-validation procedure progressed. The technique made use of predicted values of secondary structure and solvent exposure values to enable easy theoretical mutant evaluation. Secondary structure predictions were obtained from the following algorithms and servers: Jpred [47]; NNPREDICT [33]; PSIPRED [48]; PROFsec [34,36]; SSpro and SSpro8 [35,41]. Multiple predictions of secondary structure were quantified and averaged. Predictions of solvent accessibility were obtained, using a 25% threshold, from Jnet [49] and from the SCRATCH server by ACCpro [40].

2.2. Numerical representation of sequences, structures and glycan branching

To train and test neural network models, the conversion of the primary sequence (amino acid residues), predicted secondary structure (helix, extended or unordered), predicted solvent accessibility (buried or exposed) and glycosylation characteristics (complex-type or high mannose) to numerical values was necessary. The primary sequence was quantified by amino acid clusters, as previously described [18]. To our knowledge, the conversion of secondary structures, solvent accessibility and glycan branching into numerical representation for use in neural network training has not been optimized in published research. Thus, arbitrary values were assigned in this research. Residues predicted to be incorporated in helical secondary structures were assigned values of 1, and residues predicted in extended structures, such as β-sheets, were assigned values of 2. All other secondary structure elements were given the value 3. Residues predicted to be exposed to solvent at a 25% threshold were assigned the value 0, and predicted buried residues were assigned a value of 1. Finally, high mannose glycan branching classification was given the value 0.25, and complex-type was given the value 0.75.

2.3. Neural network architecture and glycosylation window optimization

Elman recurrent neural networks with a single hidden layer were constructed in MATLAB® (The Mathworks, Inc., Natick, MA) using the Neural Network Toolbox. Neural networks were initiated with random weight and bias values prior to training procedures. In addition, 100 independent neural networks were initiated and trained for 2000 epochs. As the number of hidden layer neurons is a common adjustable parameter in neural network model construction, careful attention was given to the number of total adjustable parameters (weight and bias values) associated with the neural network. The number of hidden layer neurons was selected so that the total number of adjustable parameters associated with the neural network approached, but was less than, the total number of data points used in neural network training procedures. This method avoided the problem of over-parameterization of the system but ensured optimum neural network
The optimum glycosylation window was defined as the number of residues, initiating on the N-terminus site of the glycosylation site and extending a specified number of residues on the C-terminus side of the glycosylation site. Various lengths of the glycosylation window were evaluated on the effectiveness of neural network training. Neural network training effectiveness was monitored by calculation of the mean-square error (MSE) for prediction of the testing data set. The MSE was based on the correct and incorrect predictions of the data set. The overall MSE value was calculated as the mean of the 100 independent initiation and training procedures. For the case of cross-validation, these MSE values were averaged for each of the independent testing data set predictions. The glycosylation windows for primary sequence, predicted secondary structure and predicted solvent accessibility neural network inputs were evaluated independently.

2.4. Predictive model construction

Following the establishment of optimized glycosylation windows for: (i) primary sequence, (ii) predicted secondary structure, and (iii) predicted solvent accessibility inputs independently, these inputs were grouped in effort to improve the overall prediction of the testing data sets. For example, the input vector space was expanded to simultaneously include all independent inputs above to predict glycosylation branching patterns. Multiple arrangements of inputs were examined to investigate input relevance in predictions, as the overall goal was minimization of the MSE of testing data set prediction. Once the relevant inputs and optimized glycosylation windows were identified, 30 independent neural networks were located in which the testing data sets were predicted with a minimized MSE value. The use of multiple networks for this purpose allowed for the calculation of a mean network prediction with a confidence interval calculated as the fraction of all neural networks returning the dominant output value. The confidence interval has a range of 0.5 (lowest confidence) to 1 (highest confidence). All neural networks included in the final predictive model contained different values of weights and biases. The effectiveness of the model was demonstrated by simulation of mutants of the tissue-type plasminogen activator (tPA) protein previously published research. The construction of tPA variants through deletion mutations and domain insertions was found to alter the glycosylation of tPA at N117 from high mannose N-linked glycan branching to complex-type classification in some cases.

3. Results and discussion

3.1. Statistical analysis of the data set

Discernable relationships were sought between a polypeptide sequence and resulting N-linked glycosylation branching patterns to determine a model structure for predicting glycosylation branching classification. Relative occurrences of: (i) amino acid residues, (ii) predicted secondary structures, and (iii) predicted solvent accessibility were calculated for residues surrounding the site of N-linked glycosylation. All polypeptide sequences of the data set were included in the analysis, and for each sequence, 20 residues on either side of the glycosylation site were analyzed. The results were clustered based on experimentally observed glycan branching (complex-type or high mannose) and yielded significant differences (see Supplementary Figs. 1–6).
From these data, it was observed that the serine to threonine ratio of the N-X-S/T glycosylation sequence was 0.77 for polypeptides with complex-type glycosylation and 0.39 for sequences with high mannose glycosylation. Also apparent were high occurrences of unsubstituted, hydrophobic residues on the N-terminus side of the glycosylation site for sequences with complex-type glycosylation. Larger occurrences of hydrophilic, including charged, residues were observed on the C-terminus side of the glycosylation site for sequences with high mannose glycosylation. Sequences with high mannose glycosylation showed a larger occurrence of predicted helical secondary structures than those with complex-type glycosylation. Also, for sequences of high mannose glycosylation a higher incidence of residues with predicted low solvent accessibility (buried) residues were observed on either side of the glycosylation site. These results suggest a combination of electrostatic and structural influences of the polypeptide chain are involved in governing interactions between N-linked glycan structures and membrane-bound glycosyltransferase enzymes. Also, the analysis suggests a complex interplay exists between these influences, and these can not yet be modeled deterministically. Thus, artificial neural networks may be one model platform capable of identifying governing relationships and their connections given representative experimental data.

3.2. Glycosylation window optimization and reference set cross-validation

The primary sequence, predicted secondary structures and predicted relative solvent accessibility, and combinations of these, were further evaluated by recurrent neural network models for prediction of N-linked glycan branching classification. In all cases, the number of residues around the glycosylation site (called the glycosylation window) was varied in attempt to minimize the average MSE of testing data set predictions. The glycosylation window leading to the best overall prediction of the testing data sets in cross-validation experiments was defined as the optimized glycosylation window for all three cases. Results of glycan branching classification prediction using primary sequence data is shown in Fig. 1. The abscissa specifies the start of the glycosylation window and the ordinate specifies the termination of the glycosylation window. Thus, the coordinates at each point locate the edges of a window relative to the glycosylation site (at coordinates 0,0). Negative coordinates denote residue positions to the N-terminus side of the glycosylation site, and residues to the C-terminus side are labeled with positive coordinates. It is noted that the mirror-image portions of Figs. 1–3 have been omitted. Three notable regions led to better predictions of the testing data sets in cross-validation experiments when using primary sequence data as the sole input vector (Fig. 1). One such region is defined by glycosylation windows starting between residues -16 and -9 (starting on the N-terminus side of the glycosylation site) and extending to between residues 15 and 20 (terminating on the C-terminus side). Another significant set of glycosylation windows is located between residues numbered 2 (starting) through 6 (terminating). This optimum solution is possibly an artifact of the data set due to the relatively small and position of the resulting glycosylation window. In the data set, the relative occurrence of serine at residue number 2 is higher in the data set for complex-type glycans than for high mannose glycan branching (discussed previously). Regardless of whether this observation holds true given larger data sets, it is recognized from the relative occurrence ratios of serine to threonine (presented previously) that further sequence and/or structural information (larger glycosylation window) is needed to make accurate predictions. Thus, the improved predictions for these short glycosylation
windows represent a family of local optima. Recurrent neural network testing MSE results with predicted secondary structure inputs are shown in Fig. 2. Near-optimum glycosylation windows exist with starting residues numbered −10 to 10 and extending to terminating residues numbered 16 to 20. These results indicate the importance of secondary structure elements to the C-terminus side of the glycosylation site in determining glycan branching. At the same time, the precedence of the glycosylation window terminating residues demonstrates long-range (~20 residues) influence of secondary structure elements. Fig. 2 also shows the priority of the secondary structure to the C-terminus side of the glycosylation site. Glycosylation windows incorporating residues largely to the N-terminus side of the glycosylation site produced sub-optimal results. Finally, results with predicted solvent accessibility inputs are shown in Fig. 3. Several regions of glycosylation windows appeared optimized for the prediction of the testing set data. A large optimized region is located between starting residues −20 and −10 and extending to between terminating residues 6 through 20. An additional optimum family of solutions starts at residues 4 through 9 and extends to residues 8 through 20. The notable region of poor predictions was observed for short glycosylation windows around the glycosylation site. This is an expected result as glycosylation sites (for both complex-type and high mannose glycans) are often located in highly exposed regions.
Other results of Fig. 3 also suggest the dominant influence of the residues located to the C-terminus side of the glycosylation window.

3.3. Combination of input vectors to improve predictions

Using optimized glycosylation windows from single input cross-validation experiments, the combination of input vectors was examined in effort to improve prediction of the testing data sets. The input vectors examined consisted of the following data (discussed previously): (i) primary sequence data, (ii) secondary structure predictions, and (iii) solvent exposure predictions. Two optimized glycosylation windows were chosen for each input vector (see Figs. 1–3). One input vector was chosen that spanned the N-terminus and C-terminus sides of the glycosylation site, and one input vector was chosen from optimized windows solely on the C-terminus side of the glycosylation window. Separate input vectors were combined in all combinations, and cross-validation experiments were repeated with the extended input vectors. This case allowed for a further increase in the number of neurons of the neural network, since the number of data points was greatly expanded with multiple input vector incorporation. Results of averaged testing set predic-
tions for each case are shown in Table 1. Input vector combinations resulted in decreased prediction accuracy in all cases except for one. This exception consisted of input vectors spanning residues of the N-terminus side and C-terminus side of the glycosylation site with predicted secondary structure and predicted solvent accessibility data. Consequently, this case also dramatically improved prediction accuracy relative to input vectors of a single entity. Thus, the optimized input vector for prediction of N-linked glycan branching classification contained (i) the predicted secondary structure data with a glycosylation window starting at residue 9 and extending to residue 16 as well as (ii) the predicted solvent accessibility with a glycosylation window starting at residue 16 and terminating at residue 17. Importantly, primary sequence data were effectively eliminated from the optimal input vector space.

The neural network-based model correctly predicted over 79% of the glycosylated polypeptides in all testing data sets. This model performance was compared to that of a completely random model that generates either complex-type or high mannose output values. The completely random model is defined as having a probability of success, in all predictions, of 0.5, and results conform to a binomial probability distribution. Further defining the distribution is the number of identical
trials by the model, which was equal to the number of polypeptide sequences in all testing data sets. The (binomial) probability of obtaining 79% (or greater) correct predictions of the testing data sets using the completely random model is $4.2 \times 10^{-14}$. Thus the neural network-based model adds significant order to the system. As additional experimental data are published and included in the training data set, the performance of the neural network-based model is expected to improve.

### 3.4. The prediction model

Using the optimized input vector given in the previous section, 30 independent recurrent neural networks were independently trained and comprised the overall predictive model. Individual neural networks of the overall model were selected from cross-validation experiments, and all 30 trained neural networks contained different sets of weight and bias values. Further, networks selected for the predictive model were found to predict the specific testing data set to a minimum MSE value, and in some cases, this value was found to be zero. Overall, individual neural networks comprising the predictive model had a rate of success greater than 90% for correctly classifying branching characteristics of their respective testing data sets. Final predictions were made by the model by averaging the results of the 30 independent recurrent neural networks. The average recurrent neural network result was then mapped using a simple non-trained perceptron classifier. In general, average recurrent neural network results less than 0.5 were mapped to complex-type branching. On the other hand, values greater than, or equal to 0.5, were mapped as high mannose branching classification. Consequently, the use of multiple neural networks enabled the calculation of the confidence interval of the prediction. The confidence interval is defined as

| Primary sequence starting residue
d | Primary sequence terminating residue
d | Predicted secondary structure starting residue
d | Predicted secondary structure terminating residue
d | Predicted solvent exposure starting residue
d | Predicted solvent exposure terminating residue
d | Resulting average MSE of testing data sets
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>6</td>
<td>6</td>
<td>17</td>
<td>5</td>
<td>15</td>
<td>0.0845</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>None</td>
<td>6</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>−10</td>
<td>18</td>
<td>−9</td>
<td>17</td>
<td>−16</td>
<td>13</td>
<td>0.0955</td>
</tr>
<tr>
<td>−10</td>
<td>18</td>
<td>−9</td>
<td>17</td>
<td>None</td>
<td>None</td>
<td>0.0800</td>
</tr>
<tr>
<td>None</td>
<td>18</td>
<td>None</td>
<td>None</td>
<td>−16</td>
<td>13</td>
<td>0.0845</td>
</tr>
<tr>
<td>None</td>
<td>−9</td>
<td>17</td>
<td>−16</td>
<td>13</td>
<td>0.0465</td>
<td></td>
</tr>
</tbody>
</table>

a Glycosylation windows are defined as coordinates of starting and terminating residues relative to the glycosylation site at (0,0). Negative coordinates are located on the N-terminus side, and positive coordinates denote the C-terminus side of the glycosylation site.

b The MSE value of 0.0465 results from 79% correct predictions of all testing data sets. The probability that a completely random (2 possible outcome) model would achieve this level of accuracy (or better) is $4.2 \times 10^{-14}$. 

---

**Table 1**

MSE results of specified glycosylation windows combining primary sequence, predicted secondary structure and predicted solvent exposure (25% threshold) for $N$-linked glycan branching classification prediction

<table>
<thead>
<tr>
<th>Primary sequence starting residue</th>
<th>Predicted secondary structure starting residue</th>
<th>Predicted secondary structure terminating residue</th>
<th>Predicted solvent exposure starting residue</th>
<th>Predicted solvent exposure terminating residue</th>
<th>Resulting average MSE of testing data sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>6</td>
<td>6</td>
<td>17</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>None</td>
<td>17</td>
<td>None</td>
<td>5</td>
</tr>
<tr>
<td>None</td>
<td>6</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>5</td>
</tr>
<tr>
<td>None</td>
<td>18</td>
<td>−9</td>
<td>17</td>
<td>−16</td>
<td>13</td>
</tr>
<tr>
<td>−10</td>
<td>18</td>
<td>−9</td>
<td>17</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>−9</td>
<td>None</td>
<td>None</td>
<td>−16</td>
<td>13</td>
</tr>
<tr>
<td>None</td>
<td>−9</td>
<td>17</td>
<td>−16</td>
<td>13</td>
<td>0.0465</td>
</tr>
</tbody>
</table>
the fraction of individual neural networks returning the overall classified result. Thus, the confidence interval maintains a range between 0.5 and 1. A confidence interval value close to 0.5 represents a split two-dimensional classification decision by the predictive model.

3.5. Further model simulations

The predictive model was applied to a set of mutation studies performed by Wilhelm et al. [20] with the tissue-type plasminogen activator (tPA) protein. In particular, N-linked glycan branching at N117 was studied [20] with respect to specified domain addition and sequence deletion mutations. The wild-type tPA protein contains a large fraction of high mannose glycan structures at N117 when produced by CHO cell cultures [50]. In addition to evaluation of the wild-type protein, N-linked glycan branching was evaluated at N117 for the following deletion mutants by Wilhelm et al. [20]: residues 2–89 (Δ2–89); residues 44–48 (Δ44–48); and residues 55–62 (Δ55–62). In each case, the effect of the specified sequence deletion resulted in complex-type glycan branching at N117. Glycosylation of tPA at N117 occurs in the kringle I domain of the protein, and the kringle I domain is preceded, in the N-terminus direction, by a growth factor region and a finger domain [51]. Four hybrid structures of tPA were prepared in the research by Wilhelm et al. [20] by rearrangement of these domains and through the addition of the growth factor and kringle domains of urokinase. Hybrid A was composed of (in order) the urokinase epidermal growth factor, the urokinase kringle domain and the tPA kringle I domain followed by the kringle II and protease domains of tPA. Hybrid B consisted of the tPA finger domain followed by the epidermal growth factor. The urokinase kringle domain was inserted before the kringle I domain of tPA. In hybrid C, the urokinase kringle domain was inserted following the kringle I domain of tPA. Finally, in hybrid D, an extra copy of the tPA kringle II domain was inserted preceding the tPA kringle I domain. Glycosylation in the extra kringle II domain was blocked in this case. Hybrids A, B and D were found to have complex-type N-linked glycan branching at N117 (wild-type residue numbering) by Wilhelm et al. [20] through experimental methods, and hybrid C was found to conserve high mannose glycan branching. The decision to exclude primary structure data from the predictive model input vector was further reinforced by this case study. The deletion and addition mutations in these cases were performed well-beyond 20 residues away from the N117 glycosylation site in the N-terminus direction. Thus, primary sequence data of the glycosylation window for all mutant sequences was conserved in all cases, further justifying the removal of this data from the predictive model. However, this was found to not be the case for predicted secondary structure and predicted solvent exposure data (data not shown). As shown in Table 2, all hybrid sequences and the Δ44–48 deletion mutant were classified correctly by the predictive model with a relatively high confidence interval in most cases. The deletion mutants Δ2–89 and Δ55–62 were classified incorrectly by the model. However, the somewhat low confidence interval of 0.63 was observed for these incorrect predictions. These cases expose limitations of the predictive model, but more importantly, these predictions demonstrate the usefulness of the confidence interval in making judgment of model predictions. Further, this case study illustrates the usefulness of predicted quantities as model inputs as these model predictions were possible without experimental determination of secondary structure elements and residue solvent accessibility.

A comparison of the results in Table 2 to the completely random model was performed to further assess the usefulness of the neural network-based predictive model. As discussed previously,
the random model: (i) produces one of two outputs in independent trials, (ii) has a probability of success of 0.5, and (iii) results conform to a binomial probability distribution. In this comparison, to effectively show the usefulness of the model, the probability of success must be shown to be greater than 0.5. As shown in Table 2, six of eight sequences were correctly predicted by the model. The mean and standard deviation of the binomial distribution (number of trials, \( n = 8 \); probability of success per trial, \( p = 0.5 \)) are 4.0 and \( \sqrt{2} \), respectively. Thus, correct predictions of six of eight sequences in Table 2 results in a performance of approximately 1.4 standard deviations greater than the mean of the binomial distribution of the random model. In addition, the binomial probability of the random model correctly predicting six or more classifications out of eight possible is 0.14. However, this analysis does not make use of the confidence interval or the fact that the neural network-based predictive model consists of 30 independent neural networks. In Table 2, the eight model predictions require a total of 240 independent neural networks, all of which return one of two possible outcomes after classification (complex-type or high mannose). From a back-calculation involving the confidence interval, 187 of the 240 neural networks of Table 2 made correct predictions. The mean and standard deviation of this binomial distribution (\( n = 240; \ p = 0.5 \)) from the random model are 120 and 7.8, respectively. Thus, the performance of all neural networks included in the predictive model is 8.7 standard deviations above the mean value returned by the binomial distribution of the random model. In addition, the probability that the random model returns 187 or more correct predictions (out of 240 total) is \( 5.2 \times 10^{-19} \). Thus, the neural networks of the predictive model offer a probability of correct prediction that is significantly greater than 0.5. This illustrates the effectiveness of the model for making glycosylation branching classification predictions.

3.6. Model limitations

As demonstrated in the previous example, although the neural network-based predictive model adds significant order to the system, model limitations exist for the prediction of \( N \)-linked glycan branching classification. But, these limitations may be revealed in certain cases by examination of the confidence interval. Also, incorrect predictions of a neural network-based model can be used

<table>
<thead>
<tr>
<th>Input sequence</th>
<th>Average recurrent network result</th>
<th>Overall model classification(^{a})</th>
<th>Confidence interval(^{b})</th>
<th>Published classification(^{a})</th>
<th>In agreement?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tissue plasminogen activator (N117)</td>
<td>0.3245</td>
<td>0.25</td>
<td>0.87</td>
<td>0.25</td>
<td>Yes</td>
</tr>
<tr>
<td>tPA (N117) Δ2–89</td>
<td>0.4123</td>
<td>0.25</td>
<td>0.63</td>
<td>0.75</td>
<td>No</td>
</tr>
<tr>
<td>tPA (N117) Δ44–48</td>
<td>0.6175</td>
<td>0.75</td>
<td>0.80</td>
<td>0.75</td>
<td>Yes</td>
</tr>
<tr>
<td>tPA (N117) Δ55–62</td>
<td>0.3855</td>
<td>0.25</td>
<td>0.63</td>
<td>0.75</td>
<td>No</td>
</tr>
<tr>
<td>tPA (N117) hybrid A</td>
<td>0.7915</td>
<td>0.75</td>
<td>0.93</td>
<td>0.75</td>
<td>Yes</td>
</tr>
<tr>
<td>tPA (N117) hybrid B</td>
<td>0.7915</td>
<td>0.75</td>
<td>0.93</td>
<td>0.75</td>
<td>Yes</td>
</tr>
<tr>
<td>tPA (N117) hybrid C</td>
<td>0.4080</td>
<td>0.25</td>
<td>0.63</td>
<td>0.25</td>
<td>Yes</td>
</tr>
<tr>
<td>tPA (N117) hybrid D</td>
<td>0.6175</td>
<td>0.75</td>
<td>0.80</td>
<td>0.75</td>
<td>Yes</td>
</tr>
</tbody>
</table>

\(^{a}\) Model classification of 0.75 is complex-type glycan branching and 0.25 is high mannose.

\(^{b}\) Confidence interval values range from 0.5 (lowest confidence) to 1.0 (highest confidence).
to identify possible discrepancies and under-represented areas of the data set. As the data set of this model expands (as additional data is published), incorrect model predictions will be used to identify new useful data, and future data sets will be built using this iterative process. Although, from a theoretical standpoint, other limitations are expected until they can be addressed by computational predictions. For example, the effect of neighboring glycan structures on secondary structure and residue solvent accessibility is not a predictable quantity at this time. Also, predictions of secondary structures and relative solvent accessibility assume a fully-folded polypeptide sequence. Whether this case is true upon glycan interaction (or non-interaction) with the α-1,2-mannosidase (a main determinant in high mannose versus complex-type branching potential) in the RER remains to be seen. Until these effects can be effectively modeled, these limitations will exist within the neural network-based prediction model.

4. Conclusions

N-Linked glycan branching patterns resulting from eukaryotic expression of a particular glycosylated polypeptide has been found to be highly heterogeneous in many applications. However, a large fraction of these glycan structures usually exist as complex-type or high mannose structures. This major fraction was found to be a predictable characteristic using neural network-based models. The combination of multiple predictions of secondary structure elements and residue solvent accessibility were found to be an optimum input vector for prediction of branching classification. In addition, primary structure was investigated and later eliminated from the model input vector space. In all, 30 independent recurrent neural networks were used to construct the predictive model. In addition, further predictions of mutant tPA sequences illustrated the usefulness of this model. The incorporation of a confidence interval aided the identification of false predictions in the case of the tPA variant case study. The effective elimination of the primary sequence data from the model input space further reinforced the premise that N-linked glycan branching is governed not only by the glycosyltransferase enzymes native to the microorganism and culture environmental conditions, but also by the secondary structure elements and three-dimensional structure of the glycosylated polypeptide itself. To our knowledge, this is the first computational model for the prediction of N-linked glycan branching patterns that allows for the evaluation of mutant polypeptide sequences.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.mbs.2007.10.005.

References


