NOTE TO USERS

Page(s) not included in the original manuscript and are unavailable from the author or university. The manuscript was scanned as received.

73,135-148 ,150

This reproduction is the best copy available.

UMI®

DISSERTATION

MODELING OF RECOMBINANT ENZYME INACTIVATION AND PREDICTION OF *N*-LINKED GLYCOSYLATION SITE-OCCUPANCY AND MICROHETEROGENEITY

Submitted by

Ryan S. Senger

Department of Chemical Engineering

In partial fulfillment of the requirements

for the degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2005

COLORADO STATE UNIVERSITY

January 7, 2005

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY RYAN S. SENGER ENTITLED MODELING OF RECOMBINANT ENZYME INACTIVATION AND PREDICTION OF *N*-LINKED GLYCOSYLATION SITE-OCCUPANCY AND MICROHETEROGENEITY BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Kristina D. Rinker

Robert W. Woody

M. N. Karim

Advisor

Department Head

ABSTRACT OF DISSERTATION

MODELING OF RECOMBINANT ENZYME INACTIVATION AND PREDICTION OF *N*-LINKED GLYCOSYLATION SITE-OCCUPANCY AND MICROHETEROGENEITY

The inactivation of the tissue-type plasminogen activator protein (tPA) by a glycation reaction with glucose was identified as another possible mechanism between hyperglycemia and cardiovascular disease. Kinetic modeling revealed identical rates of glycation for glycosylation variants of r-tPA. Glycation at a single residue was found necessary to result in enzymatic activity loss. Computational techniques identified possible residues in the protease domain at which inactivating glycation may occur. A glucose-independent inactivation mechanism, as a result of protein-protein interactions, was also observed and found dependent on r-tPA glycosylation at N184. Simulations were performed for the optimization of fed batch feeding parameters for the production of r-tPA in a stirred-tank reactor in the presence of these inactivation mechanisms. The optimal harvest period was identified as the total r-tPA activity of the culture approached a maximum value, which served as the objective function of the optimization. Feeding profiles in the presence and absence of specified metabolite control were examined.

Novel neural network-based models were developed for the prediction of *N*-linked glycosylation characteristics. Variable site-occupancy and microheterogeneity classification were found to be predictable quantities of polypeptide glycosylation.

Intracellular oligosaccharide transfer to a polypeptide is known to be either robust or dependent on cell culture conditions during pharmaceutical production. Model predictions were optimized when based on an input of a portion of the polypeptide primary sequence. Further intracellular enzymatic processing of the oligosaccharide results in complex-type, high mannose or hybrid branching of the glycan structure. A neural network model was created for the prediction of the major fraction of a heterogeneous mixture of glycoforms. Predicted values of secondary structure elements and residue solvent accessibility were found to best predict neural network testing data sets. The primary structure was effectively eliminated from the neural network input vector space. These results further emphasized the notion that site-occupancy remains dependent upon the primary sequence of the polypeptide and glycosylation microheterogeneity remains governed by secondary structure elements and three-dimensional properties of the folded glycoprotein.

Ryan S. Senger
Department of Chemical Engineering
Colorado State University
Fort Collins, CO 80523
Spring, 2005

ACKNOWLEDGEMENTS

I give special thanks to my family for their continued support and encouragement throughout my graduate career. I thank my advisor, Dr. M. Nazmul Karim, for the guidance, education and special opportunities he presented me with as a graduate student. I also thank my committee members for the interest and insightful suggestions to this research. Finally, I acknowledge the fine work of my undergraduate summer research students.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Motivation for research

The implications of protein glycosylation on recombinant protein biological activity have long been an important topic in pharmaceutical research. In addition, protein glycosylation has been found in much medical research to play significant roles in disease as well. Not only does protein glycosylation play an important role in biological activity, it is known to affect other protein properties such as folding, solubility, stability, intracellular location and specific targeting (Parekh *et al.*, 1987; Cumming, 1991; Stanley, 1992). Knowing the importance and specific roles of glycosylation with respect to recombinant protein products, process models describing product production and inactivation should be written with respect to glycoforms of such recombinant products. Furthermore, as glycosylation by eukaryotic cell cultures generally results in a heterogeneous mixture of glycoform products (Parekh, 1994), multiple product models should be required to effectively model the glycoform compositon of a recombinant glycosylated product. However, little has been done to date with respect to product model generation with respect to glycosylation and especially with respect to glycosylation prediction upon production by a genetically engineered cell line in bioprocessing applications. Two such topics of glycosylation-based modeling were addressed in this research. First, kinetic product models were developed to describe the

glycosylation-dependent rates of inactivation of the recombinant tissue-type plasminogen activator (r-tPA) protein when produced in a stirred tank fed batch reactor. These were coupled with glycosylation-dependent models of recombinant product production derived in previous research (Senger and Karim, 2003a). Feeding parameters were then optimized based on an objective function of the total r-tPA activity of the bioreactor. Consequently, the glycosylation-dependent mechanisms of independently identified inactivation reactions, including a glucose-dependent mechanism, were effectively modeled for a heterogeneous glycoform population in this research. Furthermore, neural network-based algorithms were developed for the prediction of variable site-occupancy and microheterogeneity classification of *N*-linked glycosylation for recombinant proteins by CHO cell cultures. The motivation of this particular part of the research remained to evaluate the effects of genetic mutations on the resulting glycosylation profiles of recombinant proteins when produced by CHO cell cultures under normal growth conditions. Prediction methods of this type allow for the theoretical prediction of glycosylation characteristics of a mutated polypeptide sequences. These types of developments aid in the conservation of laboratory resources in mutant protein experimental validation. These modeling developments of this research are novel in that glycosylation-based modeling in bioprocessing applications has been hindered, despite their apparent usefulness, due to the complex mechanisms associated with the *N*-linked glycosylation pathway.

1.2 Identification of r-tPA inactivation mechanisms

The r-tPA protein is an important pharmaceutical product based on the importance of the wild-type protein in the fibrinolytic system *in vivo*. The wild-type tPA protein exists as an important component of this system *in vivo* as a component of the plasminogen activator family of serine protease enzymes (Grossbard, 1987; Lijnen and Collen, 1987). These proteins are responsible for the activation of the zymogen, plasminogen, to the active form of plasmin, which has fibrinolytic properties. Thus, study of inactivation mechanisms of the r-tPA protein *in vitro* may apply to mechanisms *in vivo* that result in impaired fibrinolysis, which is a strong precursor for a thrombolytic event. The study of free glucose interactions with the r-tPA protein was investigated, as much medical research has recently linked the hyperglycemic condition to impaired fibrinolysis. However, as noted in much medical research, not all mechanisms have been identified to date that fully explains this increased risk (Kannel and McGee, 1979; Panahloo and Yudkin, 1996). In addition, a free glucose interaction with r-tPA is also of great interest in the bioprocessing industry as glucose is a very common and essential nutrient of CHO cell cultivation. A known mechanism in the food processing industry and in medical research with hemoglobin, glycation, involves the covalent interaction of free glucose and specific lysine residues in certain proteins and may result in protein structural rearrangements in some cases (Shapiro *et al.*, 1980; Bunn and Forget, 1993). In addition, other researchers have reported glucose-independent recombinant protein inactivation due to protein-protein interactions resulting in either autolysis or product aggregation (Saido *et al.*, 1994; Baki *et al.*, 1996; Lang and Schleef, 1996; Stevens *et al.*, 1996; Kendrick *et al.*, 1998; Roberts *et al.*, 2003; Wang and Kelner, 2003; Shiraki *et al.*,

2004). These two mechanisms were investigated with the Type I and Type II glycoforms of the r-tPA protein separately. The two glycoforms of r-tPA differ in the presence of glycosylation at N184 (Grossbard, 1987; Lijnen and Collen, 1987; Parekh *et al.*, 1989; Spellman *et al.*, 1989; Wittwer and Howard, 1990). Kinetic models were written for inactivation of the separate glycoforms of the r-tPA based on these two investigated mechanisms of possible inactivation. Comparison of the resulting rate constants alluded to the stabilization effect of the glycoform at N184 of the r-tPA protein with respect to the specific inactivation mechanism.

1.3 Fed batch optimization in the presence of product inactivation

In the presence of product inactivation, careful consideration must be given to the harvest period for a bioreactor in order to maximize recombinant product activity in a pharmaceutical production application. This becomes a much more sensitive problem should a fed nutrient increase the rate of product inactivation. This problem is further compounded by the presence of heterogeneous glycosylation of the recombinant product should glycoforms of the product have differing rates of production, inactivation by multiple mechanisms, as well as specific activity. A simulation study was performed with separate product models for the fully-glycosylated Type I r-tPA and the partially-glycosylated Type II r-tPA glycoform. Product models were written that included mechanisms of inactivation corresponding to glycation and glucose-independent protein-protein interactions identified in the previous part of this research with rate constants that corresponded to incubation in CHO cell culture media. Feeding profiles of D-glucose, L-glutamine and L-asparagine were optimized, through the use of dynamic programming

methods, for maximum r-tPA activity at an optimum harvest period. This problem displayed high levels of complexity in that certain feeding profiles driving high rates of culture growth and product production may also drive product inactivation. In addition, minimization of product inactivation may sacrifice cell culture growth rates and viability, resulting in low rates of productivity.

1.4 Prediction of *N*-linked glycosylation

As the use of glycosylation-dependent modeling was shown effective in modeling inactivation mechanisms and optimizing recombinant glycosylated product, prediction of glycosylation is an important development in the area of glycosylation-dependent modeling. Factors such as the cell culture phenotype, environmental conditions and the primary sequence and structure elements of a protein have been noted as vital to *N*-linked glycosylation site-occupancy and glycan microheterogeneity (Parekh, 1994). Much research has been performed on these first two factors of glycosylation determination. In addition, statistical analyses have been performed in recent research with respect to primary sequence factors affecting site-occupancy of *N*-linked glycosylation sites (Petrescu *et al.*, 2004) and experimental research has been dedicated toward probing the effects of mutations on glycosylation characteristics (Wilhelm *et al.*, 1990; Shakin-Eshleman, 1996; Kasturi *et al.*, 1997; Mellquist *et al.*, 1998). But, neither of these approaches has led to the development of a predictive model for glycosylation characteristics for production by a defined organism. In this research, artificial neural networks were investigated as data-based models for prediction of glycosylation variable site-occupancy and microheterogeneity classification. Variable site-occupancy exists for

glycosylation sites such as N184 of r-tPA, in which the presence of a glycan structure at a specified site is dictated by environmental conditions. In contrast, robust glycosylation, as in the case of N117 and N448 of r-tPA, is defined as the presence or absence of a glycan structure at a defined glycosylation site independent of cell culture environmental conditions. Commonly, variable site-occupancy results in a heterogeneous mixture of glycoforms. The Type I and Type II glycoforms of r-tPA are an example of such a heterogeneous mixture. On the other hand, robust glycosylation results in a near homogeneous glycoform population. Following initial glycan attachment in the endoplasmic reticulum (ER), many glycosyltransferase enzymatic reactions are responsible for defining the specific oligosaccharide branching patterns of glycan structures. These reactions also result in heterogeneous mixtures of glycoforms varying in glycosylation microheterogeneity. Common categories of branching patterns include: complex-type, high mannose and hybrid structures (Hubbard and Ivatt, 1981; Kornfeld and Kornfeld, 1985; Roth, 1987; Parekh, 1994). Neural networks were used as data-based models to map inputs of primary sequence, predicted secondary structure elements and predicted residue solvent accessibility values to glycosylation characteristics related to site-occupancy and microheterogeneity classification. This methodology enabled the investigation of relevant inputs for particular glycosylation characteristic prediction. Incorporation of various residues in the neighborhood of the glycosylation site allowed for the determination of the effective glycosylation window in each case. The glycosylation window was defined as the number of neighboring amino acid residues, on both the *N*-terminal and *C*-terminal sides of the glycosylation site, that exert influence over glycosylation characteristics in each case.

1.5 Outline of contributions

This research has been divided into four separate sections, with each consisting of an independent document with specific introduction and background, methods, results and discussion and conclusions sections. Glycosylation-dependent kinetic modeling applications are contained in the first two such sections, and neural network-based modeling for glycosylation prediction is contained in the following two chapters. The investigation of r-tPA inactivation mechanisms and kinetic modeling of inactivation is included in Chapter 2. This chapter also contains a study using computational methods and circular dichroism (CD) spectroscopy to further probe the structural rearrangements leading to enzyme inactivation and possible sites of the non-enzymatic glycation reaction. Possible correlations of *in vitro* experiments with r-tPA to *in vivo* mechanisms of impaired fibrinolysis are also included in this chapter. The modeling and optimization of a fed batch process, with experimental validation, is included in Chapter 3. This chapter also contains kinetic studies for the determination of inactivation rate constants in a changing CHO cell culture media environment. Glycosylation characteristics predictions were separated into two separate chapters. Chapter 4 contains the development of neural network-based models for variable site-occupancy predictions. The prediction model for microheterogeneity classification is contained exclusively in Chapter 5. Statistical analyses accompany both of the glycosylation prediction models. In addition, in both cases, model validation consists of prediction of previously published experimentally validated data in which site-directed mutations effectively altered the studied glycosylation characteristics. All references are contained in a single section following Chapter 5. In addition, Appendix A includes a reference list of all notation used

throughout chapters 2 through 5. A detailed description of fed batch equations used in the dynamic programming methods of Chapter 2 are contained in Appendix B. Comprehensive results of selected simulations of Chapter 2 are included in Appendix C. The complete data sets used for construction of the neural network training and testing data sets for the predictive models of Chapter 4 and Chapter 5 are contained in Appendix D and Appendix E, respectively. All computer programs are included in Appendix F along with descriptive commentaries about variables and specific program executions.

Chapter 2

EXAMINING THE RELATIONSHIP BETWEEN HYPERGLYCEMIA AND CARDIOVASCULAR DISEASE: GLYCATION KINETICS OF TYPE I AND TYPE II GLYCOFORMS OF THE TISSUE-TYPE PLASMINOGEN ACTIVATOR PROTEIN

2.1 Introduction and background

2.1.1 Connection between hyperglycemia and cardiovascular disease

Thrombosis is a significant problem in diabetics. It has been found that diabetic patients have a higher morbidity and mortality rate from acute myocardial infarction than non-diabetic patients (Yudkin and Hendra, 1992; Hurst and Lee, 2003). Accumulating evidence points to the circumstance that impaired fibrinolysis plays a direct role in the increased incidence of atherosclerosis and thrombosis in diabetic patients (Panahloo and Yudkin, 1996). Diabetic vessel walls have higher levels of plasminogen activator inhibitor type 1 (PAI-1), which may significantly contribute to the formation of vulnerable plaques that are prone to rupture (Pandolfi *et al.*, 2001). Inhibitors of fibrinolysis are elevated in atherosclerotic plaques (Robbie *et al.*, 1996). A fine balance exists between plasminogen activators and inhibitors *in vivo*. This balance makes up the fibrinolytic system of the body, and prevention of thrombosis is among the most researched functions of this system. Since patients suffering from diabetes mellitus possess a much higher risk of premature atherosclerosis, especially coronary artery disease (CAD), this is suggestive of an imbalance (or less than optimal balance) due to decreased production or inactivity of the plasminogen activator (PA) proteins and/or increased production of the plasminogen activator inhibitor proteins (PAI-1). This

imbalance in the fibrinolytic system (or impaired fibrinolysis) not only increases the risk of a thrombolytic event, but it plays a major role in the formation and progression of atherosclerotic plaques (Panahloo and Yudkin, 1996). In addition to impaired fibrinolysis, endothelial dysfunction, diabetic dyslipidemia, oxidative stress, and autonomic neuropathy have been identified as possible mechanisms of increased atherosclerosis in diabetic patients (Hurst and Lee, 2003). The governing mechanism behind the increased risk of cardiovascular disease experienced by diabetic patients is of great concern to medical research as conventional risk factors, such as smoking, obesity, blood pressure, and serum lipids fail to fully explain the excess risk in diabetic patients (Kannel and McGee, 1979).

2.1.2 Glycation reaction details

Interactions between glucose and proteins *in vivo* and *in vitro* have been known for some time as the *browning* or Malliard reaction. In particular, this is a non-enzymatic reaction also known as *glycation*. It involves the condensation of aldehydes, ketones, or reducing sugars with susceptible amino groups of a polypeptide and is heterogeneous by nature. Specifically, the glycation reaction mechanism is the initial formation of a Schiff base between the carbonyl group on glucose with the amino group of an amino acid. This initial step is reversible and the rate-limiting step of glycation. This step is followed by an Amadori rearrangement, which is internally catalyzed by neighboring carboxyl groups (Shapiro *et al.*, 1980, Bunn and Forget, 1993). The glycation reaction has been observed most readily at the NH_2 terminal amino groups of polypeptide chains. The relatively low pK_a of these groups make them effective nucleophiles at physiological pH.

However, this reaction has also been observed at the ε-amino group of lysine residues, which normally have pK_a values in the vicinity of 10 to 11. Both types of glycation reactions have been found to occur under physiological conditions, as is the case with hemoglobin. For example, hemoglobin has been found to be glycated twofold to threefold in the case of diabetic subjects relative to hemoglobin of non-diabetic subjects. Glycation of hemoglobin *in vivo* has been found to take place slowly and continuously over the entire life-span of a red blood cell (Bunn and Forget, 1993). In fact, about 1% of total lysine residues of hemoglobin become glycated over a 120-day span under hyperglycemic conditions *in vivo*, and those ε-amino groups of lysine residues that showed preference for glycation were typically on the surface of the protein (Shapiro *et al.*, 1980). Alteration of protein three-dimensional structure is expected upon glycation as the formation of a ketoamine lowers the pK_a value of the glycated residue (Bunn *et al*, 1979). More importantly, glycated hemoglobin demonstrated altered physical and functional properties as compared to the non-glycated protein. In particular, it was observed that the oxygen binding properties of hemoglobin were compromised upon glycation (Bunn and Cahil, 1981).

2.1.3 The tissue-type plasminogen activator (tPA) protein

The tPA protein is a component of the fibrinolytic enzyme system, which has the main goal of activating the proenzyme plasminogen to the active form, plasmin. Thus, the role of plasmin remains, subsequently, to break down and remove fibrin from the vascular bed (Grossbard, 1987; Lijnen and Collen, 1987). Recombinant tPA (r-tPA) is an important pharmaceutical product. The tPA protein is a serine protease comprised of a

polypeptide chain of 527 amino acid residues. Robust *N*-linked glycosylation, from a site-occupancy standpoint, is only observed at two sites in the protein (N117 and N448), and a site of variable occupancy exists at N184 (Grossbard, 1987; Lijnen and Collen, 1987; Parekh *et al.*, 1989; Spellman *et al.*, 1989; Wittwer and Howard; 1990). The fully-glycosylated protein is commonly referred to as the *Type I* glycoform, and *Type II* describes the partially-glycosylated (glycan absent at N184) glycoform. The glycosylation of the tPA protein has also been shown to be of great importance to the performance of the tPA protein in the activation of plasminogen, as specific activity was shown to decrease 6-10 fold for the non-glycosylated protein, as compared to glycosylated tPA, in fibrinolytic assays (Little *et al.*, 1984). The tPA protein primary structure also contains 35 cysteine residues, leading to potentially 17 disulfide bonds in the secondary structure. Common regions of the protein resulting from the secondary structure conformation include two kringle regions, a growth factor, a finger region, and the protease catalytic site, which is formed by the H322, R371 and S478 residues. Also characteristic of the tPA protein is its enzymatic conversion to a two-chain molecule through cleavage of the R275-I276 peptide bond, which leaves the molecule linked by a single disulfide bond (Grossbard, 1987; Lijnen and Collen, 1987; Wittwer and Howard, 1990).

2.1.4 Identifying mechanisms of r-tPA inactivation

A further examination of the effect of hyperglycemia on the fibrinolytic system was investigated. The significance of this research is that a new mechanism in the area of *impaired fibrinolysis* was explored. Since enzyme activity of plasminogen activator

proteins is determined by three-dimensional protein structure, it was hypothesized that any glycation of the tPA protein near the active site may impair enzymatic activity through structural rearrangement. Should this occur *in vivo*, an imbalance in the fibrinolytic system would result and increase the likelihood of cardiovascular disease. As glycation is known to be a slow non-enzymatic reaction *in vivo*, a sodium phosphate buffer was used. This buffer system is known to increase the rate of protein glycation if present in the system (Watkins *et al.*, 1987; Lapola *et al.*, 1996). In addition, the question was addressed as to whether inactivation by this mechanism would differ between the Type I and Type II glycoforms of the protein since the presence of a glycan at N184 is responsible for differences in specific activity (Parekh *et al.*, 1989; Wittwer *et al.*, 1989; Wittwer and Howard, 1990). This research has implications to both clinical studies and to the production of r-tPA in the pharmaceutical industry. A loss of enzymatic activity due to glycation during production using mammalian cell cultures would suggest that optimal cultivation conditions may exist with lower glucose concentrations than that typically used during mammalian cell cultivation. Finally, the extent of Type I and Type II tPA glycation was examined through an analysis of kinetic data and bioinformatics techniques were utilized to postulate possible sites of glycation resulting in inactivation.

2.2 Theoretical development

2.2.1 Development of inactivation mechanisms

A reaction mechanism was developed for both the inactivation of Type I and Type II r-tPA. The reaction mechanism accounted for two modes of inactivation. First, since the r-tPA protein is a serine protease of high molecular weight, a glucose-independent mechanism of low-order kinetics was investigated. In particular, the glucose-independent inactivation mechanism modeled in this research may be due to an autolysis mechanism of the serine protease or due to protein aggregation since large concentrations of r-tPA were used (Saido, *et al.*, 1994; Baki *et al.*, 1996; Lang and Schleef, 1996; Stevens *et al.*, 1996; Kendrick *et al.*, 1998; Roberts *et al.*, 2003; Wang and Kelner, 2003; Shiraki *et al.*, 2004). Secondly, inactivation as a result of glycation was examined. For both reactions, the assumption was made that upon inactivation, all (not partial) fibrinolytic activity was compromised. The rate of the glucose-independent inactivation mechanism of the active r-tPA protein, A, to an inactive form, I, was described by the rate constant, k_0. The overall rate of the reversible covalent attachment of glucose, G, to the active protein, A, to the formation of a Schiff base intermediate, B, to the final ketoamine, C, was approximated by the irreversible rate constant, k_1. In addition, the possible extent of inactivating glycation was unknown for r-tPA and was described by the stoichiometric coefficient, z. These reaction constants were determined independently for Type I and Type II r-tPA glycoforms. This overall reaction mechanism is shown in Figure 2.1, and the following rate equation (2.1) was derived.

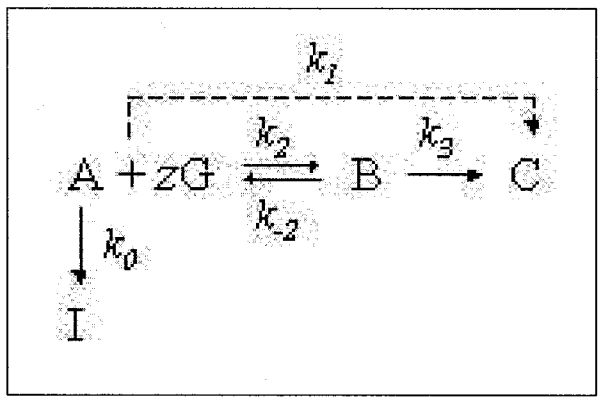

Figure 2.1. Reaction mechanism for active r-tPA, *A*, inactivation by a glucose-independent mechanism, k_0, and by glycation by glucose, k_1. The following species were included in the mechanism: active r-tPA protein, *A*; glucose, *G*; Schiff base intermediate, *B*; ketoamine, *C*, following an Amadori rearrangement and inactive r-tPA, *I*, following a low-order inactivation mechanism.

$$\frac{dA}{dt} = -k_0 A - k_1 A G^z \tag{2.1}$$

2.3 Materials and methods

2.3.1 Glycation of the r-tPA protein

Purified r-tPA standards consisting of 86% Type I glycoform (with 14% Type II) and 80% Type II glycoform (with 20% Type I) were used in this research. For simplicity, these are referred to as *Type I* and *Type II standards* in all figures. Corrections have been made to account for the mixture of glycoforms and have been labeled accordingly. Protein standards were incubated for a period of 24 hours under physiological temperature (37°C) and pH (7.2), in 20 mM sodium phosphate buffer, and with specified amounts of free D-glucose. A sodium phosphate buffer was chosen as it

has been shown to promote the rate of protein glycation (Watkins *et al.*, 1987; Lapola *et al.*, 1996). Free glucose concentrations of 0, 0.9, 1.8, and 5 g L^{-1} were investigated. A glucose concentration of 1.8 g L^{-1} is typical for a hyperglycemic condition *in vivo*, and a glucose concentration of 5 g L^{-1} is typical for mammalian cell culture for recombinant protein production. For all experiments, the specific r-tPA protein glycoform concentration was held constant at 2.3 $x10^{-6}$ mol L^{-1}. The reaction was continuously stirred and sampled frequently. Gas exchange was allowed with the surrounding environment, and the relative humidity was controlled at 100% to prevent evaporation.

2.3.2 Activity analysis

An assay utilizing a chromogenic substrate, specific for tPA, was utilized for r-tPA activity determination (Spectrozyme; American Diagnostica). Absorbance measurements were obtained using a Beckman DU640 spectrophotometer operated at 405 nm.

2.4 Results and discussion

2.4.1 Evidence of natural inactivation and glycation

The results of r-tPA enzymatic activity loss in 20 mM sodium phosphate buffer and specified glucose concentrations are shown in Figure 2.2. The dimensionless activity was reported for normalization of Type I and Type II specific activities. It is well known that the specific activity of the Type II glycoform is greater than that of the Type I glycoform (Parekh *et al.*, 1989; Wittwer *et al.*, 1989). In 20 mM sodium phosphate buffer solution and in tris buffer (data not shown), it was observed that the control

experiments incurred a loss of activity over the duration of the 24-hour period. As the pH values of the incubated solutions were found to have remained constant in all cases, it was determined the normal inactivation of r-tPA in a control environment was significant, yet occurred with glycosylation-dependent rate constants. This is suggestive of a buffer-independent, glucose-independent degradation mechanism probably resulting from alterations in the three-dimensional protein structure in response to contact with other protein molecules. Serine protease autolysis and protein aggregation are the most common mechanisms for explaining these observations. Since this glucose-independent inactivation mechanism was observed at significantly different rates for the Type I and Type II glycoforms, this suggests that the oligosaccharide at N184 plays a role in inhibiting this inactivating mechanism of the tPA protein. Thus, the bulky oligosaccharide at this position may provide steric hindrance preventing an inactivating three-dimensional conformation change that originates in the kringle 2 region of the tPA protein. This may also suggest that the oligosaccharide at N184 partially shields amino acid residues involved in this inactivation mechanism. The Type II glycoform, which maintains a higher specific activity, showed a shorter half-life in both control experiments as well as during incubation with glucose. Thus, the overall magnitude of the glycation and glucose-independent (glycosylation-dependent) inactivation rate constants were found to both be significant in describing the overall rate of r-tPA inactivation in the presence of free glucose.

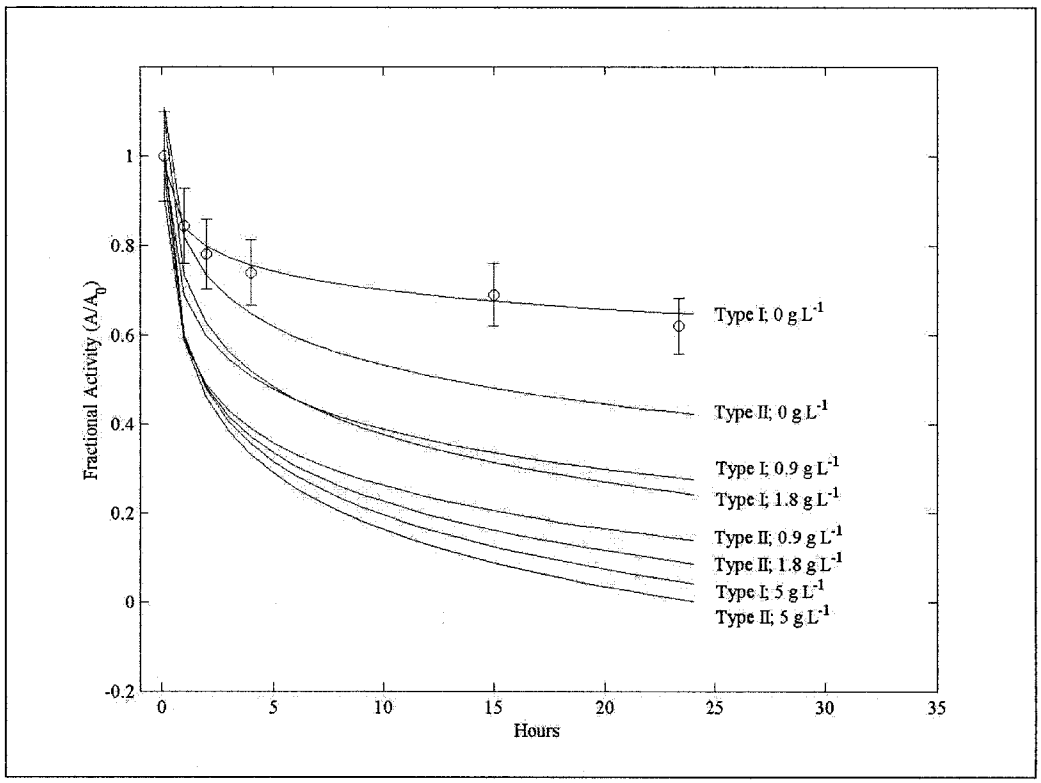

Figure 2.2. Fractional activity of Type I and Type II r-tPA as a function of time for specified glucose concentrations in 20 mM sodium phosphate buffer. Fractional activity was defined as the ratio of time-dependent active r-tPA glycoform concentration, A, to the initial active concentration, A_0. Error bars represent one standard deviation.

The glucose-independent inactivation mechanism rate constant, k_0, due to possible protein-protein interactions, was investigated for the Type I and Type II glycoforms independently using the control standards. Although this rate constant may not be representative of characteristics *in vivo*, this parameter is necessary at the protein concentration of this work to accurately measure the overall glycation rate constant, k_1. The measurement of the inactivation rate constant, k_0, was done through the application of numerous kinetic models and minimization of the mean-square error (MSE) between predicted and experimental values. Low-order rate constants have been noted in the

literature for mechanisms of inactivating autolysis and protein aggregation (Saido, *et al.*, 1994; Baki *et al.*, 1996; Lang and Schleef, 1996; Stevens *et al.*, 1996; Kendrick *et al.*, 1998; Roberts *et al.*, 2003; Wang and Kelner, 2003; Shiraki *et al.*, 2004). So, for both the Type I and Type II glycoforms, a second-order model was found to effectively model the experimental data; although, the argument may be made that this inactivation may actually be of even higher-order for the Type I glycoform. Applications of low and higher-order models with optimized rate constants are displayed in Figure 2.3. The optimized k_0 coefficients of the second-order model and those corresponding pure glycoform coefficients are displayed in Table 2.1. The rate of the glucose-independent inactivation of the Type II glycoform of r-tPA exceeded that of the Type I glycoform by a factor of 2.9. This result further suggests that the bulky oligosaccharide at N184 of the Type I glycoform may contribute to the stability of the protein in response to the natural deactivating mechanism originating in the kringle 2 domain that ultimately compromises the enzymatic activity of r-tPA.

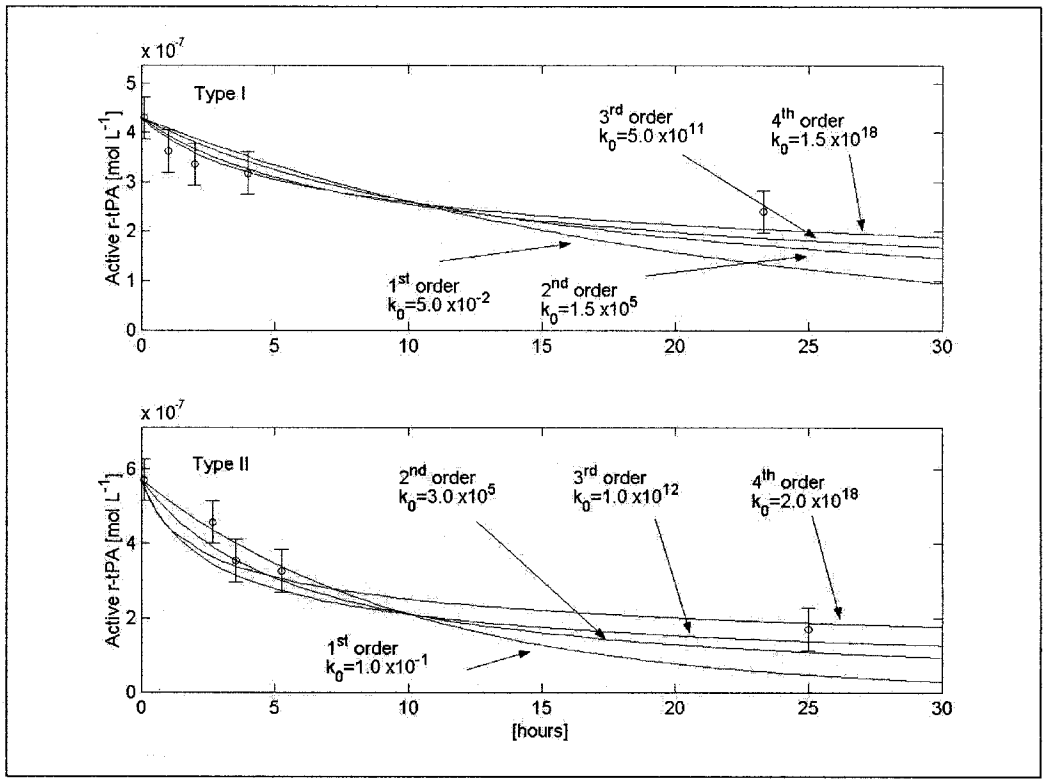

Figure 2.3. Natural inactivation of Type I and Type II r-tPA modeled by 1st, 2nd, 3rd and 4th order kinetic models. Rate constants, k_0, were optimized for each model by minimization of MSE between experimental and model data. Error bars represent one standard deviation.

	Experimental Glycoform Mixtures[a]	Pure Glycoform[b]
Type I	1.5×10^5	1.2×10^5
Type II	3.0×10^5	3.5×10^5

Table 2.1. Optimized values of the rate constant, k_0 [mL μg^{-1} h^{-1}], for second-order kinetics of the glucose-independent inactivation mechanism of Type I and Type II r-tPA glycoforms.

[a] Experimental glycoform mixtures consist of 86% Type I/15% Type II and 20% Type I/80% Type II.

[b] A calculated quantity.

2.4.2 Determination of the inactivation rate constants and degree of glycation

Calculation of the overall glycation rate constant, k_1, may only be achieved if the degree of glycation (or stoichiometric coefficient, z) is known. Although experimental methods exist that could provide additional insight on the number of glucose molecules that non-enzymatically bind to the r-tPA protein, this parameter may be calculated through the evaluation of simple kinetic models. This approach also reduced the propagation of error that an added experimental technique would introduce. The kinetic analysis approach is preferred in this case because information regarding the stoichiometric coefficient, z, is known. For example, it is a safe assumption that the number of glucose binding molecules per r-tPA molecule is a whole number. In addition, previous research with hemoglobin has yielded values between *one* and *four* based on experimental techniques (Bunn *et al.*, 1979; Shapiro *et al.*, 1980; Zhang *et al.*, 2001). Thus, the approach in this case was to assign whole number values to the stoichiometric

coefficient of glucose, z, and optimize the glycation rate constant, k_1, for three independent glucose concentrations for both the Type I and Type II experimental glycoform mixtures. Successful optimization was realized when identical values of the glycation rate constant were calculated for all three glucose concentrations with a given glucose stoichiometric coefficient. The initial assumption was made that the glucose stoichiometric coefficient, z, may be different for the Type I and Type II glycoforms.

Due to experimental error, generating identical values of the glycation rate constant, k_1, for three independent glucose concentrations for a particular glycoform, was not reasonable. A method for rate constant comparison was applied in which rate constants were divided by the average rate constant at a fixed value of the glucose stoichiometric coefficient, z. Optimization occurred by minimization of the standard deviation between these three resulting values. These results for the Type I and Type II r-tPA glycoforms, for various values of the stoichiometric coefficient, z, are presented in Tables 2.2 and 2.3, respectively. The inactivation model was applied to all data with optimized rate constants and glucose stoichiometric coefficient. Results are shown in Figures 2.4 and 2.5.

z	0.9 g L^{-1} glucose	1.8 g L^{-1} glucose	5.0 g L^{-1} glucose	Standard Deviation[b]	Weighted Mean[c]	Pure Glycoform Mean[d]
1	**1.1×10^1**	**1.3×10^1**	**1.7×10^1**	**0.24**	**1.3×10^1**	**1.3×10^1**
2	3.8×10^2	1.2×10^3	3.4×10^3	1.19	1.3×10^3	1.1×10^3
3	1.2×10^4	1.6×10^5	8.0×10^5	1.83	2.3×10^5	1.9×10^5
5	1.5×10^7	1.5×10^9	3.0×10^{10}	2.56	6.6×10^9	5.5×10^9
10	1.0×10^{15}	5.0×10^{18}	8.0×10^{21}	2.88	1.6×10^{21}	1.3×10^{21}

Table 2.2. Optimized values of the glycation rate constant, k_1 [mL μg^{-1} h^{-1}], for Type I r-tPA experimental standard[a] with specified glucose concentrations and values of the glucose stoichiometric coefficient, z. The optimum solution is highlighted in bold.

[a] Experimental glycoform mixtures consist of 86% Type I/15% Type II and 20% Type I/80% Type II.

[b] Standard deviation of normalized values.

[c] The values corresponding to glucose concentrations 1.8 g L^{-1} and 5.0 g L^{-1} were given slightly more weight because a larger number of data points existed in comparison to the 0.9 g L^{-1} data set.

[d] A calculated value.

z	0.9 g L^{-1} glucose	1.8 g L^{-1} glucose	5.0 g L^{-1} glucose	Standard Deviation[b]	Weighted Mean[c]	Pure Glycoform Mean[d]
1	**9.0**	**3.2**	**2.7×10^{1}**	**0.83**	**1.5×10^{1}**	**1.6×10^{1}**
2	3.3×10^{2}	3.2×10^{2}	5.4×10^{3}	1.24	2.4×10^{3}	2.7×10^{3}
3	1.0×10^{4}	9.0×10^{4}	1.0×10^{6}	1.30	4.2×10^{5}	4.8×10^{5}
5	1.5×10^{7}	1.0×10^{9}	3.0×10^{10}	1.40	1.2×10^{10}	1.4×10^{10}
10	8.0×10^{14}	1.0×10^{19}	8.0×10^{21}	1.44	3.2×10^{21}	3.7×10^{21}

Table 2.3. Optimized values of the glycation rate constant, k_1 [mL μg^{-1} h^{-1}], for Type II r-tPA experimental standard[a] with specified glucose concentrations and values of the glucose stoichiometric coefficient, z. The optimum solution is highlighted in bold.

[a] Experimental glycoform mixtures consist of 86% Type I/15% Type II and 20% Type I/80% Type II.

[b] Standard deviation of normalized values.

[c] The values corresponding to glucose concentrations 0.9 g L^{-1} and 5.0 g L^{-1} were given slightly more weight because a larger number of data points existed in comparison to the 1.8 g L^{-1} data set.

[d] A calculated value.

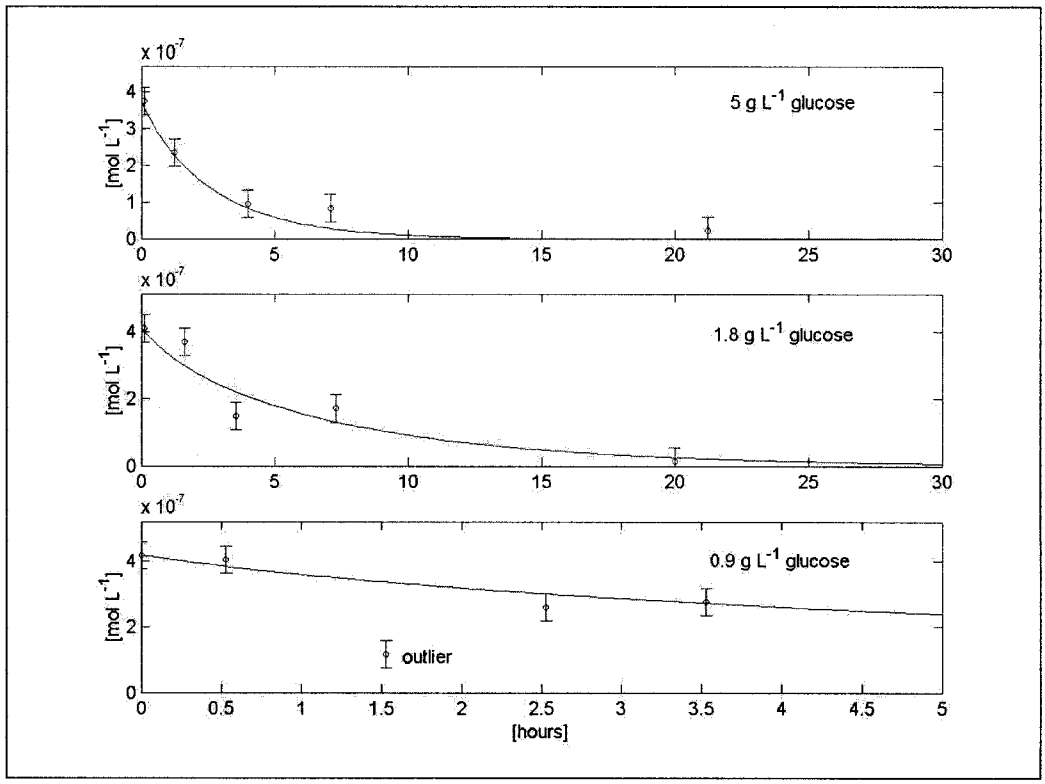

Figure 2.4. Raw data (circles) and optimized model (lines) for Type I glycoform experimental standard[a] with specified glucose concentrations. Model parameters used were 1.5 x10^5 [mL μg^{-1} h^{-1}] for the natural inactivation mechanism, k_0, and 1.3 x10^1 [mL μg^{-1} h^{-1}] for the glycation rate constant, k_1.

[a] Experimental glycoform mixtures consist of 86% Type I/15% Type II and 20% Type I/80% Type II.

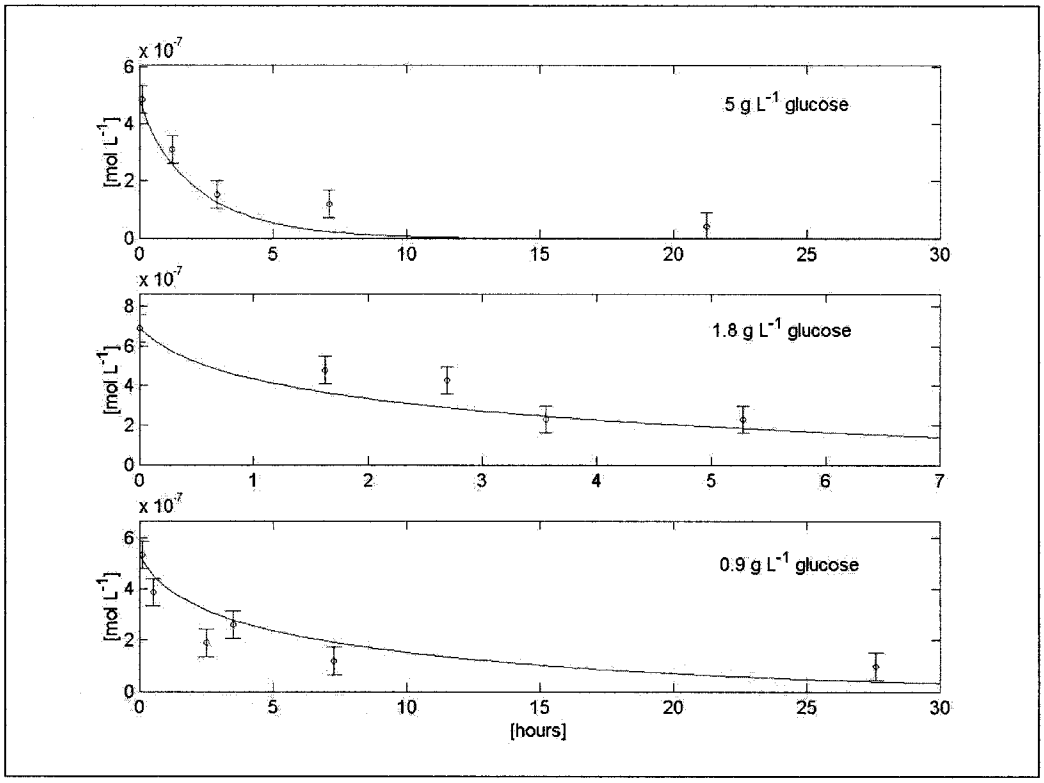

Figure 2.5. Raw data (circles) and optimized model (lines) for Type II glycoform experimental standard[a] with specified glucose concentrations. Model parameters used were 3.0 x10^5 [mL μg^{-1} h^{-1}] for the natural inactivation mechanism, k_0, and 1.5 x10^1 [mL μg^{-1} h^{-1}] for the glycation rate constant, k_1.

[a] Experimental glycoform mixtures consist of 86% Type I/15% Type II and 20% Type I/80% Type II.

The optimization related to kinetic modeling has identified that both major glycoforms of the r-tPA protein are subject to enzymatic activity loss due to glycation at a single amino acid residue. The optimized rate coefficients, k_0 and k_1, adequately model the experimental data for multiple glucose concentrations and for both glycoforms. In addition, the rate constant of glycation of Type II r-tPA was calculated to be slightly

larger than that of the Type I glycoform, but this claim cannot be validated due to experimental error. However, since the significant difference between the natural inactivation rate constants, k_0, for the two glycoforms was not observed for the glycation rate constant, k_1, it may be postulated that the bulky oligosaccharide located at N184 in the Type I glycoform does not influence the glycation reaction. Since the N184 residue is located in the kringle 2 region of the protein, separate from the protease region, this result may seem expected. However, the lysine-binding site, located in the kringle 2 region of r-tPA, has been documented to be responsible for initial substrate binding in the activation of plasminogen (Hoylaerts *et al.*, 1982). Thus, it may be postulated that the glycation reaction leading to enzymatic activity loss occurs in the protease domain of the r-tPA protein.

2.4.3 Identifying possible sites of glycation responsible for inactivation

Computational methods were employed in an attempt to locate possible sites of r-tPA glycation in the protease region of the protein. It is known that, aside from the NH_2 terminus, ε-amino groups of lysine residues are most susceptible to the non-enzymatic attachment of glucose. In addition, it is reasonable to assume that regions of the protein most favorable to this reaction must be highly exposed to solvent conditions. This is a problem that may be initially addressed through computational methods in bioinformatics. In particular, the residue solvent accessibility may be predicted through the use of the ACCpro algorithm (http://www.igb.uci.edu/tools/scratch). This algorithm was constructed using bi-directional recurrent neural networks and has achieved better than 70% accuracy (Pollastri *et al.*, 2002). Initial predictions were performed on the

alpha and beta chains of hemoglobin since glycation sites have been well characterized. It was observed that lysine residues subject to glycation were located in regions in which neighboring residues were highly exposed to solvent. This may explain, in part, why neighboring cationic residues have been observed to enhance the glycation reaction. Individual lysine residues of hemoglobin were generally predicted as highly solvent exposed. This was due to the high pK_a of the ε-amino group. In addition, most sites of glycation (but not all) occurred in areas of helical secondary structure. Solvent exposure predictions of the r-tPA molecule revealed that four lysine residues reside in regions of the active site with possible high solvent exposure of surrounding residues. Lysine residues located within the active site of r-tPA with high solvent exposure of neighboring residues include K361, K378, K429, and K505. It is noted that these residues reside far from the region of variable glycosylation (N184), which may further explain why Type I and Type II r-tPA glycoforms experienced similar inactivation rate constants, k_I, due to the glycation mechanism.

A three-dimensional representation of the catalytic domain was constructed using PyMOL™ (Delano Scientific) from existing crystallographic data (1A5H) (Renatus *et al.*, 1997). This structure is displayed in Figure 2.6. Secondary structures were represented in the polypeptide backbone, and residues corresponding to the active site were highlighted with a mesh surface representation. The lysine residues suspected as possible sites of inactivating glycation (K361, K378, K429, and K505) were highlighted as well. Further conclusions were drawn from visual analysis of this three-dimensional representation. For example, it is apparent that K361 resides closest (relative to the other

highlighted lysine residues) to the active site of r-tPA. In addition K378 resides in a coiled region of the subunit that is shielded from the active site by rigid β-sheet structures. The residues K505 and K429 reside in highly unordered regions of the proteolytic domain; however, these structures reside a further distance from the active site than K361, which also resides on an unordered coil structure. From this analysis, it is postulated that glycation of K361 is most likely to result in a rearrangement of the three-dimensional protein structure that results in shielding of the r-tPA active site, due to its close proximity to the protease active site.

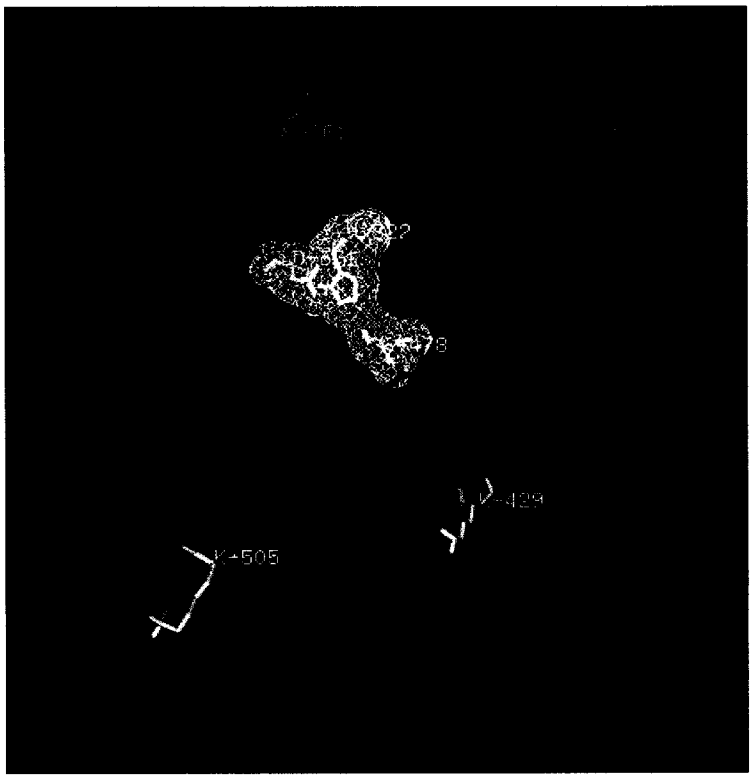

Figure 2.6. Three-dimensional representation of the proteolytic domain of r-tPA from crystallographic data (1A5H) (Renatus *et al.*, 1997). The polypeptide backbone, including secondary structure, is represented in green. The active site residues (with side-chains) have been encased in white mesh, and relevant lysine residues have been displayed (with side-chains) in yellow.

2.5 Conclusions

Many mechanisms exist that contribute to the link between hyperglycemia and cardiovascular disease. This research is novel in that it has identified the inactivation of the tissue plasminogen activator protein by a glycation reaction in the protease domain, within the active site. This in vitro study of the recombinant tPA protein represents a new mechanism under the category of *impaired fibrinolysis* should it be found to apply *in vivo*. Although the kinetic rate constants derived in this research are not directly

applicable to *in vivo* studies or glycation of the r-tPA protein in pharmaceutical production, it has shown that the glycation rate constants of the Type I and Type II glycoforms that result in inactivation are nearly identical. Further, kinetic analyses suggested that glycation at a single residue is responsible for the inactivation of the tPA enzyme. This enabled the use of bioinformatics tools to identify four possible sites in the region of the active site of r-tPA as possibly susceptible to the glycation reaction (K361, K378, K429, and K505). Three-dimensional structure representation of the protease domain revealed that K361 resides nearest the active site, but it resides in a structurally unordered region. This led to the postulation that K361 is the most likely lysine residue to disrupt the active site as a result of glycation. Kinetic analyses also revealed that the glycan structure at N184 of tPA contributes to the relative stability of the protein from a glucose-independent inactivation mechanism that is suspected to be either autolysis or aggregation. In addition, the glycan at N184 does not contribute to inactivation by glycation although it was found to inhibit inactivation by protein-protein interactions. The two mechanisms of r-tPA inactivation explored by this research were present in multiple buffer solutions. The implications of this research are great in both clinical science and biotechnology. A more complete understanding of the mechanisms *in vivo* that contribute to cardiovascular disease as a result of elevated blood glucose levels is essential to developing effective treatments. In the production of r-tPA as a pharmaceutical product, the loss of product activity due to a glycation mechanism is an important addition to governing biochemical models. Normally, high glucose concentrations in mammalian cell bioreactors leads to higher cell densities and concentration of recombinant protein product. In the case of loss of activity due to

glycation and natural inactivation, conservation of recombinant enzyme activity may possibly be optimized under lower glucose concentrations and shorter fed batch durations. The identification of glycosylation-dependent inactivation rate constants in the case of inactivation due to protein-protein interactions is significant to bioprocess modeling in that the notion is further emphasized that separate product models should be written to describe the dynamic behavior of differently glycosylated isoforms of a recombinant product.

Chapter 3

OPTIMIZATION OF FED BATCH PARAMETERS AND HARVEST TIME OF CHO CELL CULTURES FOR A GLYCOSYLATED PRODUCT WITH MULTIPLE MECHANISMS OF INACTIVATION

3.1 Introduction and Background

3.1.1 Recent developments in bioreactor feeding strategies

The use of fed batch bioreactors for recombinant protein production is common in the pharmaceutical industry. The implementation of intelligent feeding strategy has led to significant increases in productivity by enabling a cell culture to overcome such obstacles as substrate inhibition during early stages of growth and both substrate limitations and product inhibition during later stages of growth. In general, the overall growth state and biomass viability of the culture has been used as a benchmark for feeding decisions, as productivity is directly tied to this state of the culture. Significant research in this area was performed by Simon and Karim (2002) in which CHO cell apoptosis was effectively controlled through the use of variable feed rates of D-glucose, L-asparagine and L-glutamine. This research also identified the relevant free amino acids for effective feeding based on a neural network-based sensitivity analysis. By implementing this feed strategy, the authors of this work effectively extended the stationary phase of the CHO cell culture growth cycle, dramatically increasing the productivity of the culture. In addition, many other feeding strategies make use of one or multiple metabolites as set points for control algorithms, which then dictate variable

feeding flow rates. The design of these types of control strategies that optimize a desired culture state are classified under the category of optimal control (Rani and Raw, 1999). Many approaches to these types of problems, for which feed-flow rates appear linearly in process models, have been addressed in recent literature (Mahadevan and Doyle III, 2003; Kapadi and Gudi, 2004; Kookos, 2004; Sarkar and Modak, 2004; Skolpap *et al.*, 2004).

3.1.2 Product inactivation

Of particular interest to recombinant protein production is conservation of recombinant protein activity. Much effort is included in pre-production decisions such as host selection, glycosylation considerations and medium optimization to optimize recombinant protein activity. However, recombinant protein activity may be compromised in a fed batch culture environment following protein production by the culture. For instance, as proteins are dynamically-driven molecules, specific interactions with the surrounding environment may lead to structural rearrangements that compromise the active site of the recombinant protein. And, as is the case with batch and fed batch bioreactors, the culture environment consistently changes. Thus, the residence time of the protein in the changing cell culture environment itself, following cellular secretion, is vital if such structural changes leading to inactivation are observed for a particular protein. A common mechanism of inactivation for serine proteases is autolysis. Low-order kinetics have been reported for this type of inactivation (Saido *et al.*, 1994; Baki *et al.*, 1996; Stevens *et al.*, 1996). Another mechanism of recombinant protein inactivation occurs via aggregation. Many researchers have reported incidence of recombinant

protein inactivation by this mechanism for a wide variety of recombinant proteins. From a kinetic standpoint, this type of inactivation has also been found to occur with low-order kinetics (Lang and Schleef, 1996; Kendrick *et al.*, 1998; Roberts *et al.*, 2003; Wang and Kelner, 2003; Shiraki *et al.*, 2004). In turn, the activity loss due to recombinant protein aggregation in a fed batch bioreactor can be significant and should be taken into account in process models when applicable. A second mode of recombinant protein inactivation has been known in the food industry for many years as the *browning* or Malliard reaction. More specifically, this reaction is known as *glycation* and involves the irreversible condensation of aldehydes, ketones or reducing sugars with susceptible amino acid residues of a polypeptide chain (Shapiro *et al.*, 1980; Bunn and Forget, 1993). This mechanism was identified in Chapter 2 as a mode of inactivation for the recombinant tissue-type plasminogen activator (r-tPA) protein in the presence of free glucose in solution. It was hypothesized that the glycation mechanism led to r-tPA inactivation as the site of this covalent reaction occurred near the active site of the protein and induced a structural rearrangement. Specific lysine residues in the catalytic domain of the r-tPA protein have been postulated as possible glycation sites of the protein.

3.1.3 Modeling and optimization

Given various modes of recombinant protein inactivation, these effects must be incorporated into the process models in order to design feeding profiles that will not only optimize recombinant protein production but, more importantly, will maximize recombinant protein activity at the reactor harvest. Thus, time dependence is introduced into these problems and is defined as the *harvest period* for the reactor. Ideally, a reactor

is harvested at a maximum value of recombinant protein activity. In the cases of recombinant protein inactivation due to glycation, autolysis or aggregation, it is assumed that activity cannot be recovered through methods of protein refolding. These inactivation mechanisms are directly linked to the design of fed batch feeding strategies. For example, the design of a feeding strategy to extend the stationary phase of a mammalian cell culture may prove detrimental if the rate of recombinant protein inactivation exceeds that of recombinant protein production during the extended stationary phase. In addition, a high glucose set point in a variable feed-flow control algorithm may result in an excess of inactive recombinant protein due to glycation. This research examines feed strategies for a CHO cell culture producing r-tPA, which has been shown to have two modes of protein inactivation. Feed-flow rates of glucose and amino acids (an asparagine and glutamine mixture) were varied to optimize r-tPA activity at the optimum reactor harvest period. In turn, the harvest period was determined by the maximum r-tPA activity. Two types of feeding strategies were investigated by simulation studies. First, set points of glucose and fed amino acids were chosen as control variables and variable feed-flow was simulated. Second, constant feed-flow rates were investigated, in the absence of metabolite concentration control. During this set of simulations, two cases of feed-flow initiation were also investigated. In the first case, feed-flow was initiated as the first derivative of r-tPA activity with respect to time approached zero. In the second case, feed-flow was initiated as the second derivative of r-tPA activity with respect to time approached zero. The first case of feed-flow initiation followed exponential culture growth, while the second case initiated feed-flow during the exponential growth phase at the maximum value of r-tPA activity increase with respect to

time. The optimization problem was solved in all cases with respect to maximizing r-tPA activity at the harvest period. This provided for a novel comparison of feeding strategies for CHO cell cultures in the presence of multiple mechanisms of product inactivation. Feed-flow initiation was also investigated at culture inoculation (initiation at time equal to 0 hours). Results were found inferior to the other two cases of feed-flow initiation due to substrate inhibition by glucose (data not shown).

3.2 Theoretical Developments

3.2.1 Consideration of the intrinsic culture state

The intrinsic culture state was defined by Senger and Karim (2003a) and used in neural network-based culture state predictions (Senger and Karim, 2003b). In general, the intrinsic culture state takes into account the contributions of lysed cells to the overall culture state. This is of importance primarily to mammalian cell cultures in which significant lysed cell densities may be observed as a function of the hydrodynamic environment of the bioreactor. The intrinsic total cell density, X_t, is defined as the sum of the apparent total cell density, $X_{t,app}$, and the lysed cell density, X_l. In this case, the assumption was made that the lysed cell density cannot be determined by direct cell counting methods. On the other hand, the apparent total cell density is the quantity returned from a direct count of both viable and dead cells. Further, it is assumed that all lysed cells are dead, so the intrinsic dead cell density, X_d, is defined as the sum of the (countable) apparent dead cell density, $X_{d,app}$, and the lysed cell density. In terms of kinetic growth models, the Monod-based intrinsic growth rate, μ_{int}, was used as the kinetic rate constant to describe the rate of growth of the intrinsic total cell density. The

intrinsic cell death rate constant, k_d, was developed from competitive enzyme kinetics correlations to effectively model cell death. All other kinetic models were written in terms of the apparent viable cell density, $X_{v,app}$, of the culture. In addition, with an indirect measure of cell lysis, further described in Senger and Karim (2003a), a cell lysis rate, k_l, constant was formulated as a function of the hydrodynamic environment. The theoretical development for a batch culture is given in equations (3.1) through (3.6).

$$X_t = X_{v,app} + X_{d,app} + X_l \tag{3.1}$$

$$X_d = X_{d,app} + X_l \tag{3.2}$$

$$\frac{dX_t}{dt} = \mu_{int} X_{v,app} \tag{3.3}$$

$$\frac{dX_d}{dt} = k_d X_{v,app} \tag{3.4}$$

$$\frac{dX_{v,app}}{dt} = (\mu_{int} - k_d) X_{v,app} \tag{3.5}$$

$$\frac{dX_l}{dt} = k_l X_{t,app} \tag{3.6}$$

3.2.2 Cell growth and death models

Monod-type cell growth models, derived from enzyme kinetics, that encompass multiple substrates for growth as well as multiple inhibitors are certainly no longer new developments in microbial growth modeling. However, these enzyme kinetics concepts have been extended in this research for a novel development of a macroscopic intrinsic cell death model, from which the intrinsic cell death rate, k_d, was calculated. In

mammalian cell culture simulations, a cell death model is of utmost importance due to the extensive degree of cell death commonly observed throughout the culture. For example, CHO cell culture viability has been observed to approach values less than 50% during the latter stages of culture growth (Senger and Karim, 2003a). As with enzyme kinetics-derived cell growth models involving inhibitors, an interaction with a cell growth inhibitor leads to a reversible dead-end complex from which the cell cannot multiply. In the same respect, in the developed cell death model, cellular interaction with substrate leads to a reversible dead-end complex from which the cell cannot die. In addition, with the cell death model, the cell death pathway occurs through a pathway in which the cell has only interacted with inhibitor, *I*. This cell death pathway may also be considered as the absence of an interaction with substrate, *S*. Multiple complex macroscopic cell death mechanisms, based on non-competitive enzyme kinetics, were proposed and evaluated. However, the cell death mechanism based on competitive enzyme kinetics converged much faster in parameter optimization to experimental data despite having the same number of adjustable parameters as more complex macroscopic mechanisms (data not shown). The competitive enzyme kinetics-based macroscopic intrinsic cell death mechanism is displayed in equations (3.7) and (3.8). The resulting model for the intrinsic death rate, k_d, is presented in equation (3.9). The intrinsic growth rate, μ_{int}, was also approximated by a Monod-type model utilizing multiple substrates of glucose, asparagine and glutamine and is presented in equation (3.10). Since asparagine and glutamine were fed in a constant ratio and utilize the same metabolic pathways, these amino acids were grouped together as the quantity, *AA*. In addition, inhibition terms were included for glucose, *glc*, lactate, *lac*, and total (protonated and unprotonated) ammonia, *Am*. In the

cell death model, substrates were grouped as S and consisted of glucose, glutamine and asparagine. Inhibitors in the cell death model, I, included lactate and total ammonia.

$$X_{v,app} + I \xleftrightarrow{K_I} X_{v,app} I \xrightarrow{k} X_d \tag{3.7}$$

$$X_{v,app} + S \xleftrightarrow{K_S} X_{v,app} S \ \{\text{dead-end complex for cell death}\} \tag{3.8}$$

$$k_d = \frac{k_{d,max}(I)}{K_I + I + \frac{K_I}{K_S} S} \tag{3.9}$$

$$\mu_{int} = \frac{\mu_{max}(glc)(AA)}{(K_{glc} + glc)(K_{AA} + AA)\left(\frac{glc}{K_{d,glc}} + 1\right)\left(\frac{lac}{K_{lac}} + 1\right)\left(\frac{Am}{K_{Am}} + 1\right)} \tag{3.10}$$

3.2.3 Product model development

As discussed in Senger and Karim (2003a), production of the r-tPA protein by CHO cell cultures may be modeled as a growth associated product when the intrinsic growth rate is considered. The r-tPA protein displays heterogeneous glycosylation from a site-occupancy standpoint when produced in CHO cell cultures. The fully glycosylated protein is commonly termed the *Type I* glycoform; whereas, the *Type II* glycoform is unglycosylated at site N184 (Grossbard, 1987; Wittwer and Howard, 1990). Consideration of site-occupancy glycosylation is important in the case of r-tPA as the two major glycoforms have displayed significantly different thrombolytic activity. In particular, specific clot lysis activity values of 363 IU μg^{-1} and 459 IU μg^{-1} are reported in the literature for Type I and Type II glycoforms, respectively, of the tPA protein isolated from human colon fibroblast cells (Wittwer *et al.*, 1989). Activity as a function

of glycosylation has been further investigated for the r-tPA protein from other cell types, including CHO cell expression (Parekh *et al.*, 1989). In addition, the presence of the glycan structure at N184, which lies in the kringle 2 domain of the protein, has shown influence on the inactivation rate kinetics of the protein in Chapter 1. In particular, inactivation rates between the Type I and Type II glycoforms were observed with respect to differences in the glucose-independent inactivation mechanism, which is postulated to be either autolysis or protein aggregation in the case of r-tPA. However, rate differences were not observed for inactivation due to the glycation mechanism. This provided further evidence for the site of inactivating glycation in the catalytic domain as opposed to the kringle 2 domain. The assumption was made in the proposal of product models that only active r-tPA protein was produced by the CHO cell culture, and different product models could be written for total Type I and Type II r-tPA protein and active Type I and Type II r-tPA glycoforms. Product models developed in batch operation are listed in equations (3.11) through (3.14). All kinetic models of the entire CHO cell culture process model are listed in Appendix B.

$$\frac{d(TypeI_{total})}{dt} = \alpha\mu_{int} X_{v,app} \tag{3.11}$$

$$\frac{d(TypeII_{total})}{dt} = \beta\mu_{int} X_{v,app} \tag{3.12}$$

$$\frac{d(TypeI_{active})}{dt} = \alpha\mu_{int} X_{v,app} - k_{I,0}(TypeI)^2 - k_{Gly}(TypeI)(glc) \tag{3.13}$$

$$\frac{d(TypeII_{active})}{dt} = \beta\mu_{int} X_{v,app} - k_{II,0}(TypeII)^2 - k_{Gly}(TypeII)(glc) \tag{3.14}$$

3.3 Materials and methods

3.3.1 CHO cell line, media and bioreactor description

The CHO cell line CRL-9606 was obtained from ATCC. This particular cell line was designed to produce r-tPA using a $DHFR^-$ expression system with methotrexate (MTX) for selection. The culture was adapted to CD-CHO protein-free media (Invitrogen) deficient in hypoxanthine and thymidine. The culture medium was supplemented with L-glutamine to a final concentration of 3.4 mM and MTX to a final concentration of 500 nM. Batch and fed batch CHO cultivations were performed using a 3-L stirred bioreactor (Applicon) with pH, temperature, and dissolved oxygen controls monitored by an ADI 1030 Bio Controller (Applicon). The bioreactor vessel had a diameter of 13 cm, and a 3-blade pitched impeller (blades pitched 45° to vertical) with a diameter of 4.5 cm was used. The impeller was suspended 3 cm above the reactor base. Stirring speed was held constant throughout all experiments at 80 RPM. A working volume of 1.5-L was used in batch experiments and the maximum volume used in fed batch experiments was 2.0-L. Temperature was controlled at 37 °C using an external heating jacket. Gentle sparging of CO_2 and the addition of 1.0 N NaOH were used to control pH at a value of 7.2. In addition, dissolved oxygen was controlled at 40% of air saturation by gentle sparging of O_2 gas through the base of the reactor.

3.3.2 Fed batch experiments

Two fed batch experiments were performed in this study to validate simulation results. CHO cell-seeding density was held constant at 1.0×10^5 viable cells per mL. The culture viability during seeding was close to 100%. Fed batch feeding initiation occurred

at 170 hours, and was based on simulation results, which determined this time period as the point where the first derivative of total r-tPA activity approached zero. The concentration of D-glucose feed to the reactor, S_{glc}, was held constant in both experiments at 50 g L^{-1}. The amino acids L-glutamine and L-asparagine were fed from the same reservoir. The total concentration of amino acids was 5 g L^{-1}. In particular, the concentration of glutamine, S_{Gln}, in the feed reservoir was 1.11 g L^{-1}, and the concentration of asparagine, S_{Asn}, was 3.89 g L^{-1}. The ratio of glutamine to asparagine in the amino acids feed reservoir was held constant with the original composition of the CD-CHO media. It is noted that the amino acids feed reservoir was kept at 5 °C in order to retard glutamine degradation. In the first fed batch experiment, fixed-mass feed-flow rates of glucose and amino acids were 5.0 g h^{-1} and 0.5 g h^{-1}, respectively. In the second experiment, these fixed-mass feed-flow-rates were 60 g h^{-1} and 18.9 g h^{-1}, respectively. In both cases, the initial batch volume of the reactor was 1.5-L. Feeding was implemented until the reactor volume reached 2.0-L. Following the conclusion of reactor feeding, batch operation was resumed at the final reactor volume. The cultures were monitored until the viable cell density approached zero.

3.3.3 Culture state analyses

Both apparent total and viable cell densities were determined through a direct counting technique using a hemacytometer (Hausser Scientific) and the trypan blue (Invitrogen) staining procedure. Lysed cell density was determined through the measurement of free DNA in the culture media. Fluorescent measurements were made of the complex between free DNA and the fluorescent dye 4',6-diamidino-2-phenylidole

(DAPI) as described by Brunk *et al.* (1979). The fluorescence buffer used in the assay was composed of 100 mM NaCl, 10 mM EDTA and 10 mM Tris buffer, with a DAPI content of 100 ng mL^{-1}. The cell culture supernatant was added to the fluorescence buffer for a final ratio of 1:10. CHO DNA was used as an analytical standard. Relations are presented in Senger and Karim (2003a) that relate free DNA to lysed cell density based on the specific growth rate of the culture. Metabolite concentrations of glucose, glutamine, asparagine and lactate were determined using a single HPLC assay developed by Stoll *et al.* (1994). An Amminex Carbohydrate HPX-87C (250 x 4 mm) chromatography column (Biorad), operated at 85 °C, was used with a Waters HPLC system and refractive index detector. The eluent used was 2 mM $Ca(NO_3)_2$ with an eluent flow rate of 0.6 mL min^{-1}. Glucose and lactate concentrations were further verified by use of a YSI analyzer model 2700 according to the manufacturer's protocol. Total ammonia concentration (both NH_3 and NH_4^+) was determined using a spectrophotometric ammonia assay number 171-C (Sigma) with a Beckman DU640 spectrophotometer, operated at 340 nm. Total r-tPA protein concentration and total Type I and Type II glycoform concentrations were determined by an HPLC assay originally developed by Xu and Cacia (2000). Two separate assays were conducted, both utilizing a Symmetry 300 C18, 5-μm particle size (4.6 x 250 mm) chromatography column (Waters). Glycoform analysis required initial denaturation and digestion of the protein. The sample size used in the analysis was 250 μL of cell-free culture supernatant, and the analysis was performed at 30 °C. A linear eluent gradient from 70/30/0.1 water/acetonitrile/trifluoroacetic acid (TFA) to a ratio of 50/50/0.1 water/acetonitrile/TFA was executed over 20 minutes after an initial hold for 5 minutes.

An eluent flow rate of 1.0 mL min^{-1} was used. Fluorescence detection of the r-tPA protein and glycoforms was necessary using a Waters 474 programmable fluorescence detector, operated with an excitation wavelength of 275 nm and an emission wavelength of 340 nm. Total r-tPA activity was determined by direct analysis of CHO cell culture supernatant using a Spectrozyme® assay that contains a chromogenic substrate specific for tPA (American Diagnostica). Absorbance measurements were obtained at 405 nm, and an r-tPA activity standard (American Diagnostica) was obtained for specific quantitation.

3.3.4 r-tPA inactivation

Purified r-tPA samples consisting of 86% Type I r-tPA (with 14% Type II r-tPA) and 80% Type II r-tPA (with 20% Type I r-tPA) were incubated in CD-CHO media at 37 °C. Total r-tPA activity measurements were taken on regular intervals. The samples were incubated with fresh medium (5 g L^{-1} glucose) as well as used, cell-free medium. The used medium experiments were performed to analyze glucose concentrations of 3 g L^{-1} and 0 g L^{-1}. It is noted that the composition of the used culture medium differed significantly from fresh medium in regards to medium components other than glucose concentration. These experiments also helped determine whether extracellular byproducts or other consumed components of fresh medium played an important role in the r-tPA inactivation mechanism.

3.3.5 Simulations and optimization methods

Kinetic equations comprising the process model (as shown in Appendix B) were solved simultaneously using a fourth-order Runge-Kutta numerical method with a time step of 0.01 hours. All mathematical simulations were performed using the MATLAB® 7.0.1 software package. Constrained parameter optimization was performed simultaneously for the adjustable parameters of the intrinsic growth and intrinsic cell death models. Yield coefficients were determined from numerical analysis of experimental data. The objective function of the parameter optimization numerical techniques was the sum of the mean-square error values between experimental data and model predictions for the intrinsic total cell density, X_t, the apparent viable cell density, $X_{v,app}$, and the intrinsic dead cell density, X_d. In this case, the design vector consisted of all adjustable parameters. The design vector was updated according to a steepest gradient ascension technique with a step size equal to one one-hundredth of the original value. The numerical technique was verified through the use of multiple starting points. Initially, constraints were placed within plus or minus 20% of the particular parameter estimates from numerical analysis of the experimental data set. Parameters related to maximum intrinsic growth and death rates, μ_{max} and $k_{d,max}$, were left unconstrained. Steepest gradient ascension optimization was also utilized for the objective function of total r-tPA activity. In this case, the amino acids feed-flow rate (or set point) and the glucose feed-flow rate (or set point) were used as the design vector for optimization. Total r-tPA activity was used as an objective function, as opposed to active r-tPA per volume, due to the dilution factor in fed batch feeding. In fed batch simulations with metabolite control, the following simple variable volumetric feed-flow equations (3.15)

and (3.16) were implemented as metabolite values reached specified set point values. The relatively small order of magnitude of the time step in the simulation enabled effective metabolite control in simulation studies. Activity values for the Type I and Type II r-tPA glycoforms used in simulation studies were obtained from Parekh *et al.* (1989) and Wittwer *et al.* (1989).

$$F_{glc} = \frac{F_{AA}(glc) + \frac{1}{Y_{X_{V,app}/glc}} V\mu_{int} X_{V,app}}{(S_{glc} - glc)} \tag{3.15}$$

$$F_{AA} = \frac{F_{glc}(AA) + Vk_{deg}(Gln) + \left(\frac{1}{Y_{X_{v,app}/Gln}} + \frac{1}{Y_{X_{V,app}/Asn}} \right) V\mu_{int} X_{V,app}}{(S_{AA} - AA)} \tag{3.16}$$

3.4 Results and discussion

3.4.1 Parameter optimization

Multiple batch CHO cell cultivations were executed and analyzed for the purpose of kinetic parameter optimization. Values for all optimized parameters, including yield coefficients and inactivation terms, are displayed in Table 3.1. In addition, the process model, with optimized parameters, was simulated for 200 hours, and results are shown along with experimental data in Figures 3.1-3.4. The parameter optimization techniques used in this research provided exceptional fit to experimental data obtained from separate batch experiments. Error bars in Figures 3.1-3.6 represent one standard deviation of experimentally measured data points from repeated experiments. In addition, the macroscopic cell death model derived from enzyme kinetics accurately predicted the

intrinsic dead cell density of the culture as shown in Figure 3.1c. Results of r-tPA inactivation experiments are shown in Figure 3.5 for the Type I glycoform and Figure 3.6 for the Type II glycoform. As found in previous experiments, the kinetic rate constant of glycation, k_{Gly}, remained constant for the two glycoforms in fresh and nutrient-depleted culture media. In addition, the rate constant of glucose-independent inactivation of the Type I glycoform, $k_{I,0}$, possibly due to autolysis or protein aggregation, remained equal to roughly half of the value of the rate constant for the Type II glycoform, $k_{II,0}$. The influence of changing medium conditions due to nutrient depletion and cellular byproduct addition to the culture medium did not appear to have a significant effect on these kinetic rate constants. Specific values of the glycation and natural inactivation rate constants are also listed in Table 3.1.

μ_{max}	0.23 [h^{-1}]	$Y_{Xv,app/glc}$	11.1 [x10^8 cells g^{-1}]	k_l	9.0 x10^{-4} [h^{-1}]
K_{glc}	2.5 [g L^{-1}]	$Y_{Xv,app/gln}$	190 [x10^8 cells g^{-1}]	k_{deg}	3.0 x10^{-4} [h^{-1}]
$K_{d,glc}$	10.0 [g L^{-1}]	$Y_{Xv,app/asn}$	94.8 [x10^8 cells g^{-1}]	$k_{I,0}$	1.74 x10^5 [mL μg^{-1} h^{-1}]
K_{AA}	1.5 [g L^{-1}]	$Y_{Xv,app/lac}$	24.1 [x10^8 cells g^{-1}]	$k_{II,0}$	3.48 x10^5 [mL μg^{-1} h^{-1}]
K_{lac}	14.5 [g L^{-1}]	$Y_{Xv,app/Am}$	1300 [x10^8 cells g^{-1}]	k_{Gly}	2.0 x10^{-3} [L g^{-1} h^{-1}]
K_{Am}	3.5 [g L^{-1}]	α	0.115 [μg x10^5 $cells^{-1}$]		
$k_{d,max}$	0.008 [h^{-1}]	β	0.130 [μg x10^5 $cells^{-1}$]		
K_I	0.625 [g L^{-1}]				
K_S	6.5 [g L^{-1}]				

Table 3.1. Optimized parameter values, yield coefficients and kinetic rate constants for the process model listed in Appendix A.

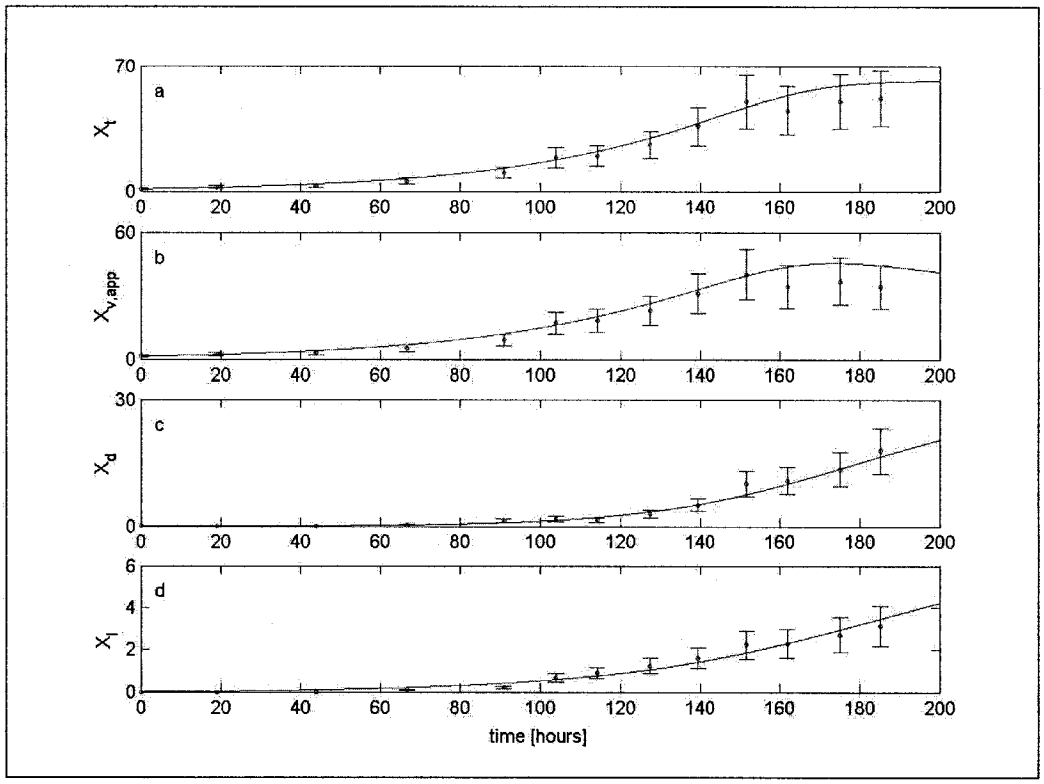

Figure 3.1. Experimental data (circles) and model predictions (lines) of: (a) intrinsic total cell density, X_t, (b) apparent viable cell density, $X_{v,app}$, (c) intrinsic dead cell density, X_d, and (d) lysed cell density. All cell density values have units of [x10^5 cells mL^{-1}].

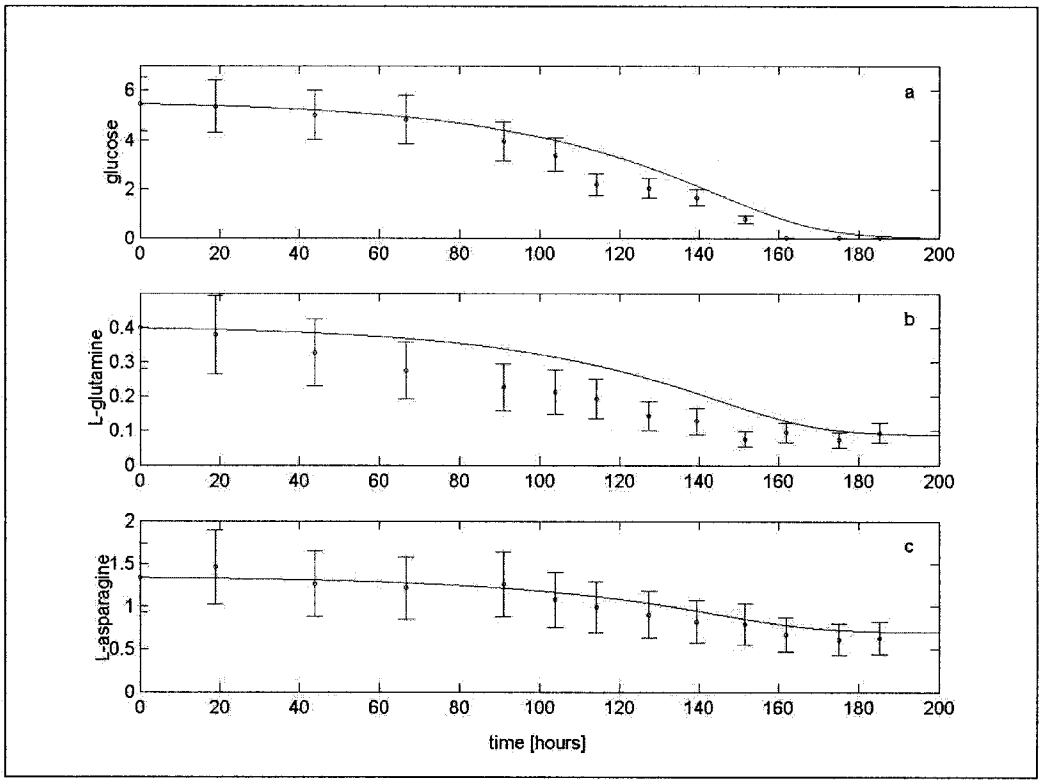

Figure 3.2. Experimental data (circles) and model predictions (lines) of: (a) free glucose concentration, (b) free glutamine concentration, and (c) free asparagine concentration in cell culture supernatant. All concentrations have units of [g L^{-1}].

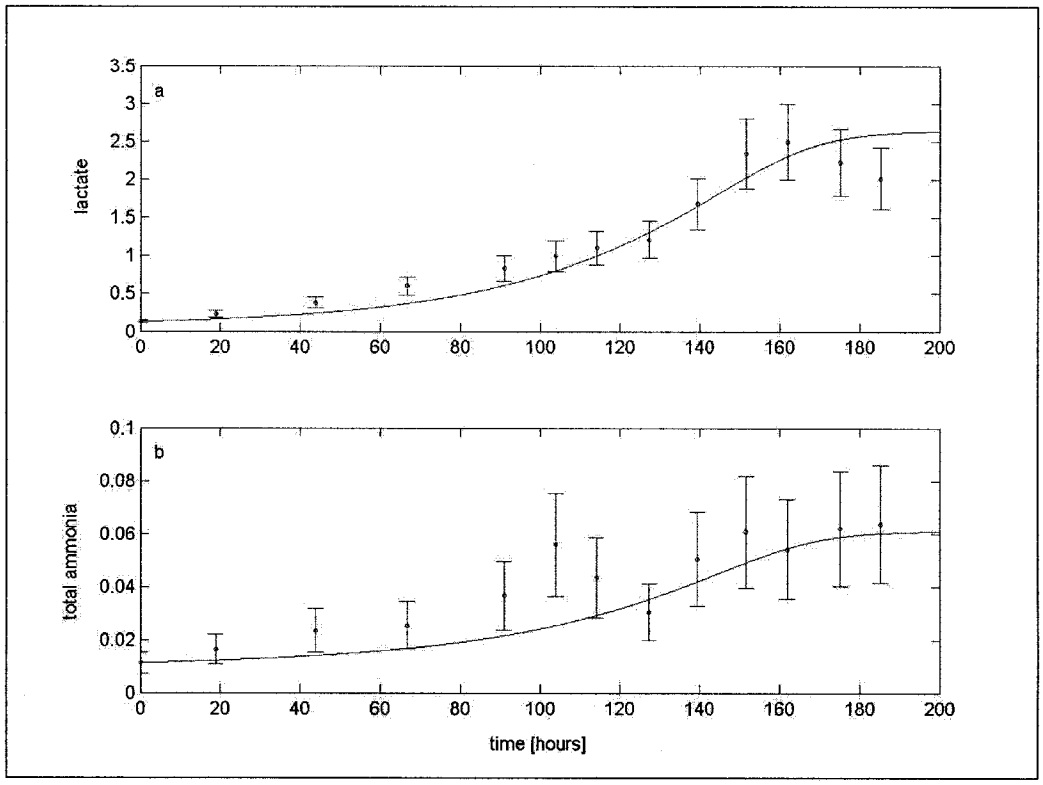

Figure 3.3. Experimental data (circles) and model predictions (lines) of: (a) lactate concentration and (b) total ammonia concentration in cell culture supernatant. All concentrations have units of [g L^{-1}].

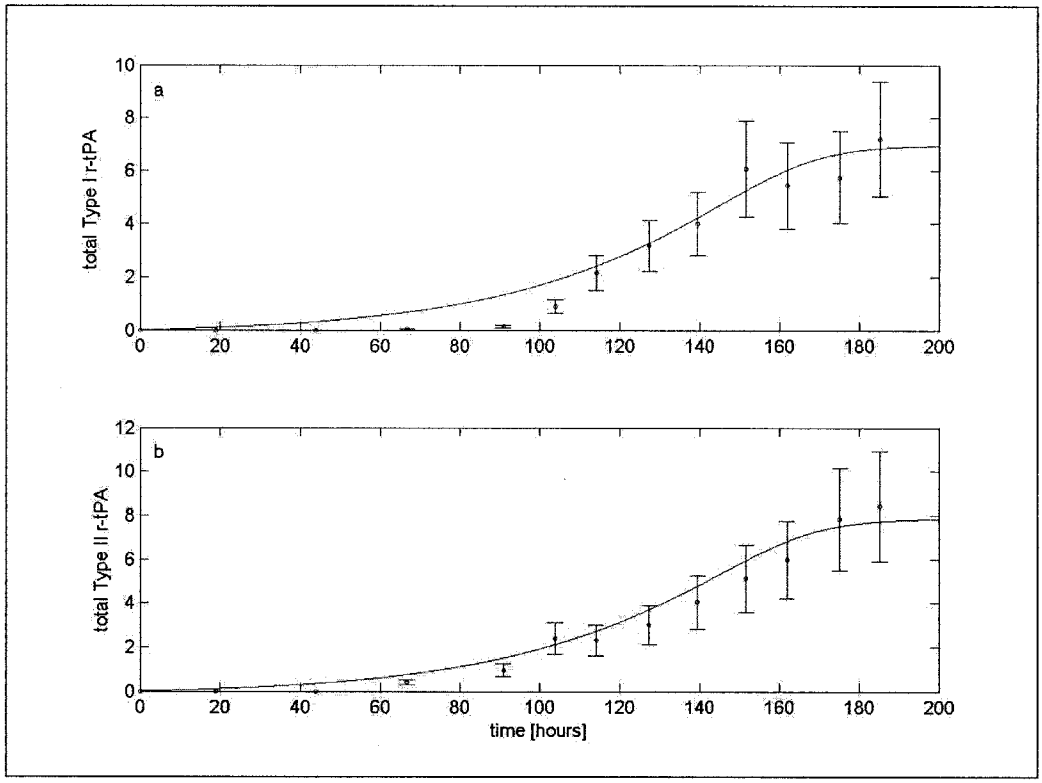

Figure 3.4. Experimental data (circles) and model predictions (lines) of: (a) total Type I r-tPA concentration and (b) total Type II r-tPA concentration in cell culture supernatant. All concentrations have units of [$\mu g\ mL^{-1}$].

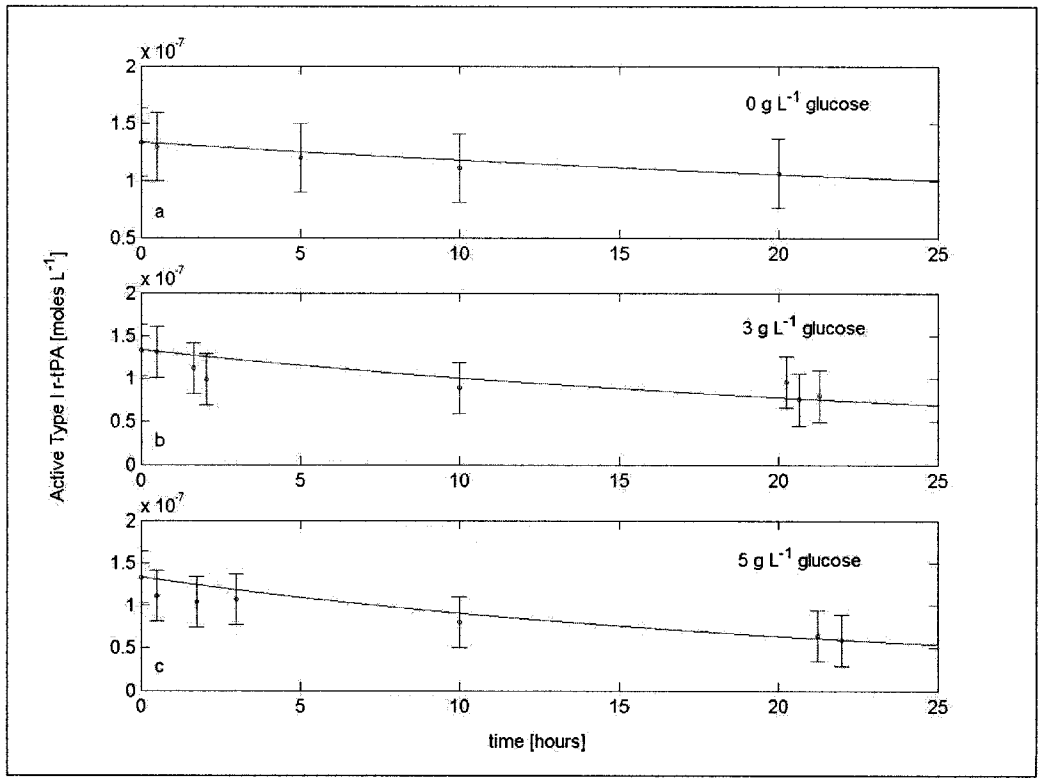

Figure 3.5. Experimental data (circles) and model predictions (lines) for Type I r-tPA inactivation in the presence of (a) 0 g L^{-1} glucose (control), (b) 3 g L^{-1} glucose, and (c) 5 g L^{-1} glucose. Fresh and depleted CHO cell culture supernatants were used as the buffer medium in the following glycoform inactivation experiments.

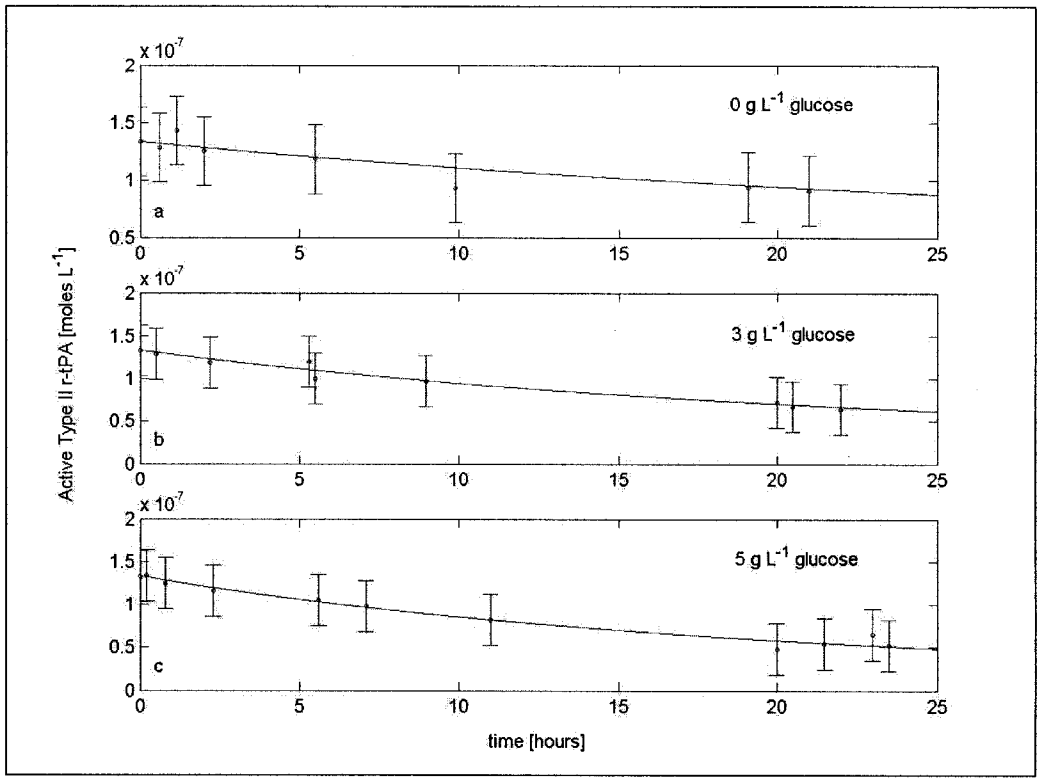

Figure 3.6. Experimental data (circles) and model predictions (lines) for Type II r-tPA inactivation in the presence of (a) 0 g L^{-1} glucose (control), (b) 3 g L^{-1} glucose, and (c) 5 g L^{-1} glucose. Fresh and depleted CHO cell culture supernatants were used as the buffer medium in the following glycoform inactivation experiments.

3.4.2 Fed batch optimization with controlled metabolite concentrations

Simulations of the fed batch CHO cell culture were performed with set points on metabolite concentrations of glucose and fed amino acids (glutamine and asparagine). The simulations were conducted in that once a particular metabolite concentration reached the set point in the initial batch process, feeding of that particular component was initiated. Since glutamine and asparagine were fed together, these quantities were considered as a lumped free amino acids parameter. Set points of amino acids concentration were evaluated from 0.001 g L^{-1} to 3 g L^{-1}. In the same way, glucose set

points were evaluated from 0.001 g L^{-1} to 5 g L^{-1}, and the objective function of total r-tPA activity at the optimum harvest period was maximized. Results are shown as a contour plot in Figure 3.7. The color-coded z-axis represents values of total r-tPA activity [IU]. It is noted that in batch cultivation, simulation results projected a maximum of 3.24 $x10^6$ [IU] at optimum batch harvest (170 hours) for a 1.5-L batch reactor. Results in Figure 3.7 suggest that fed batch operation can actually result in decreased maximum productivity at optimum harvest for the fed batch if metabolite set points are not chosen properly. Optimization methods also revealed a maximum of 5.64 $x10^6$ [IU] at optimum harvest (207 hours) for a glucose set point of 1.51 g L^{-1} and an amino acids set point of 1.18 g L^{-1}. This shows an increase in productivity of the fed batch process of nearly 75% for correctly chosen metabolite set points. It is further noted that regions of decreased productivity occurred in regions of insufficient amino acids feed with high glucose set points. This further suggests that excess glucose fed to the culture, but not effectively utilized by the CHO cell culture, results in increased rates of r-tPA inactivation due to the glycation mechanism.

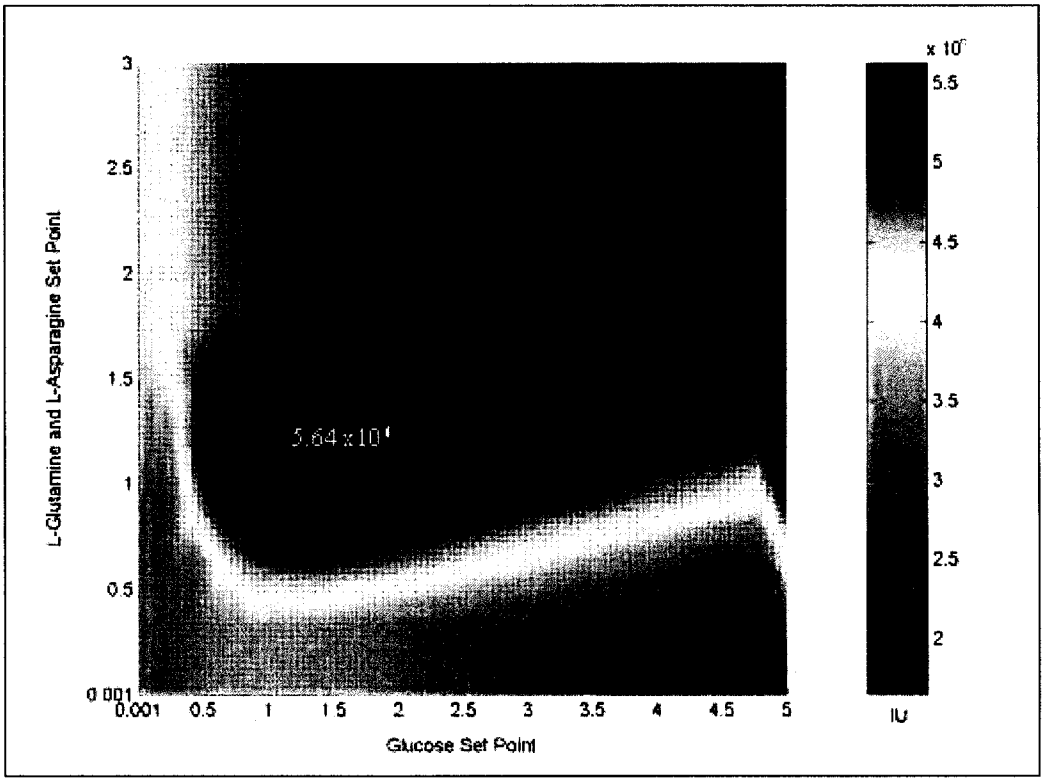

Figure 3.7. Total r-tPA activity (IU) as a function of glucose and combined free amino acids (glutamine and asparagine) set points. Set points are represented with units of [g L^{-1}].

3.4.3 Fed batch optimization with fixed feed-flow rates

The alternative to controlling metabolite concentrations with variable feed-flows is design of a fed batch process that implements fixed feed-flows at a particular time period. In this case, the period of feed implementation was dictated according to the objective function, total r-tPA activity. Two time periods were chosen for feed implementation. First, feed-flow was initiated in simulation when the first derivative of total r-tPA activity with respect to time approached zero. In batch simulations, this occurred at 170 hours. As an alternative, feed-flow initiation was also investigated at the time period at which the second derivative of total r-tPA activity with respect to time

approached zero, which occurred at 140 hours in batch culture simulations. In both cases, the mass feed-flow rate of glucose and combined amino acids (glutamine and asparagine) were used as the design vector to maximize total r-tPA activity [IU]. In both cases, the feed concentrations of glucose and amino acids were kept near the solubility limits at values of 50 g L^{-1} and 5 g L^{-1}, respectively. Contour plots of total r-tPA activity [IU] with various fixed mass feed-flow rates are shown in Figures 3.8 and 3.9. Results suggested an optimum ratio of feeding mass feed-flow rates of glucose and amino acids existed in both cases. This ratio was investigated in depth for both cases, and results are displayed in Figure 3.10. In both cases, a ratio of the glucose mass feed-flow rate, $M_{Glucose}$, to the amino acids mass feed-flow rate, $M_{Amino\ Acids}$, of 3.15 led to an optimum solution for maximizing total r-tPA activity. Of course, total feed-flow rate along this ratio presented another optimization problem. It was observed that increasing the total feed-flow along the optimum mass feed-flow rate ratio led to increased overall r-tPA activity at optimum harvest. Limiting values in these cases for instantaneous feed addition in optimal ratio was found to be above 6.0 x10^6 [IU] for both cases of feed-flow initiation. Thus, in this case involving multiple mechanisms of product inactivation and substrate inhibition, rapid feeding of an optimal ratio of glucose, glutamine and asparagine maximized total r-tPA activity in a fed batch reactor. Feed initiation as the second derivative approached zero provided slightly higher values of total r-tPA activity; however, feed ratios heavy in glucose resulted in total r-tPA activity values at optimum harvest lower than that observed for the batch culture, itself. In addition, in practice, the first derivative of a culture state is much easier to identify as opposed to the second derivative of the culture state.

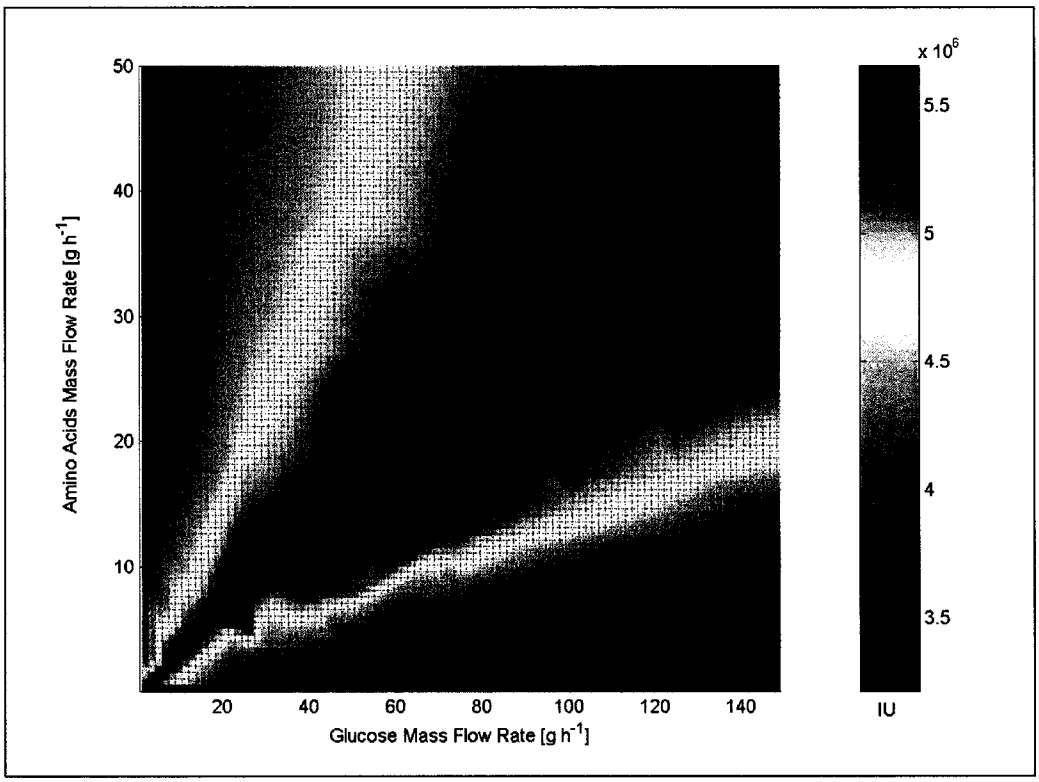

Figure 3.8. Simulation results of total r-tPA activity [IU] as a function of fixed glucose and amino acids feed-flow rates. Flow was initiated as the first derivative if total r-tPA activity with respect to time approached zero. Mass feed-flow rates have units of [g h^{-1}].

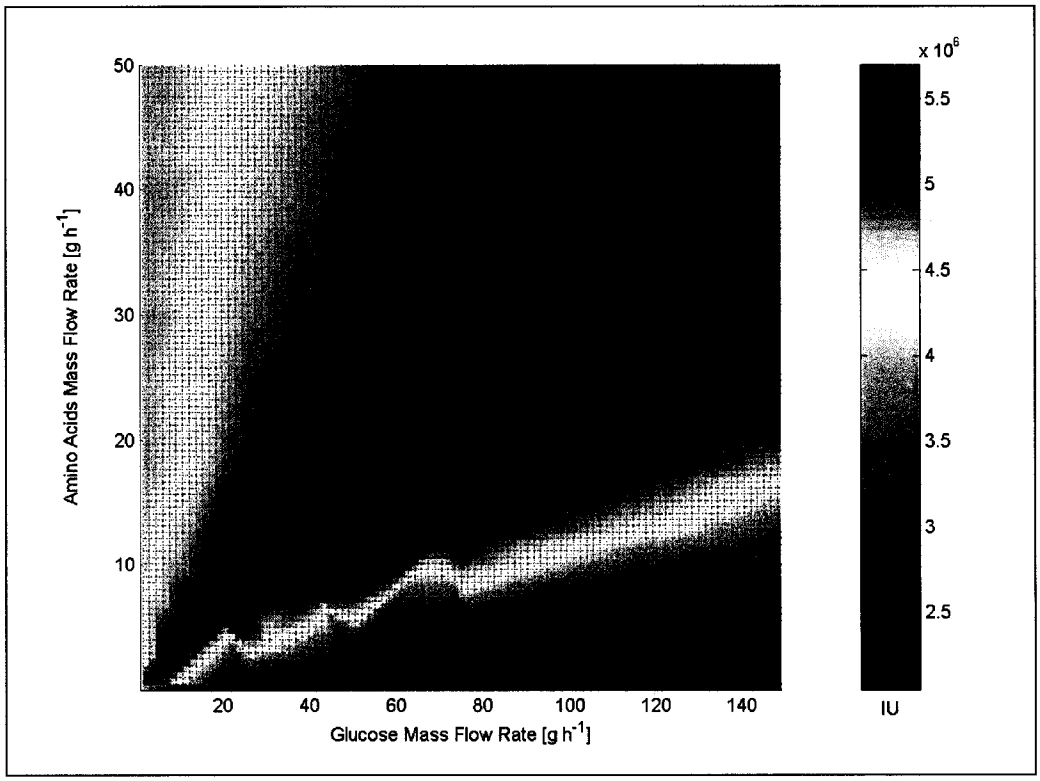

Figure 3.9. Simulation results of total r-tPA activity [IU] as a function of fixed glucose and amino acids feed-flow rates. Flow was initiated as the second derivative if total r-tPA activity with respect to time approached zero. Mass feed-flow rates have units of [g h^{-1}].

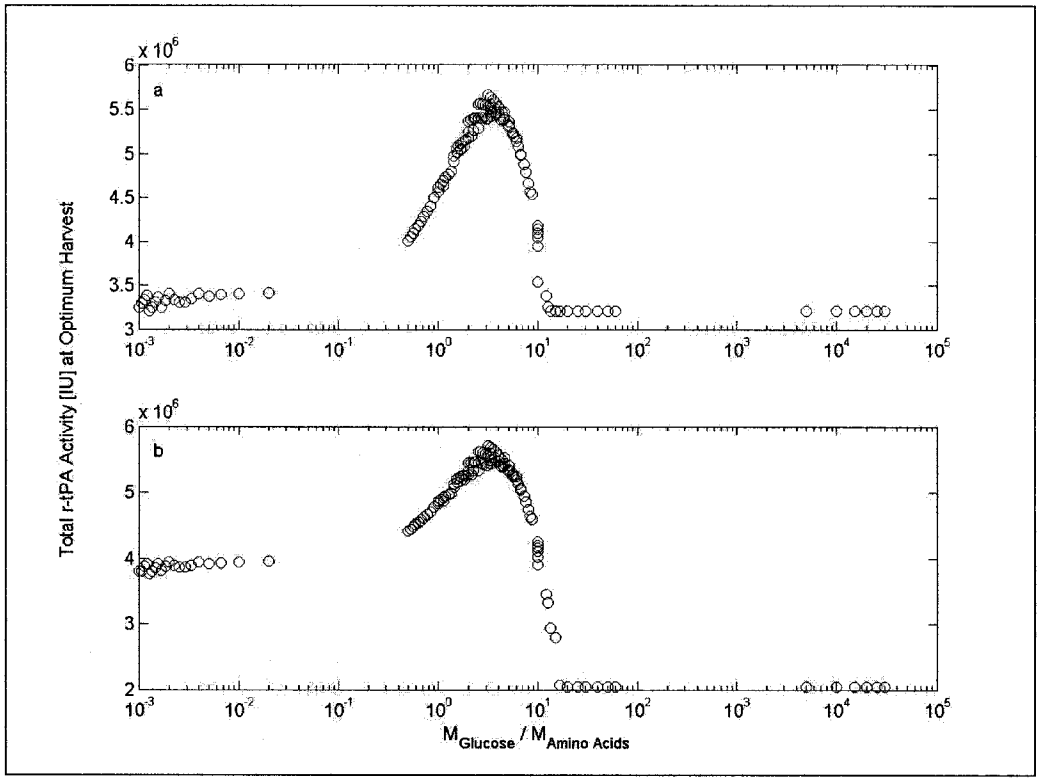

Figure 3.10. Total r-tPA activity [IU] as functions of the ratio of glucose mass feed-flow rate, $M_{Glucose}$, to amino acids mass feed-flow rate, $M_{Amino\ Acids}$. Feed-flow initiation as the first derivative of total r-tPA activity with respect to time approaches zero is shown in figure (a), and feed-flow initiation as the second derivative approaches zero is shown in figure (b). Both graphs show an optimum ratio at 3.15.

3.4.4 Further analysis of culture states around local and global optima

Due to the complex nature of the system, the overall feed rate and composition play an important role in determining the driving mechanisms to the final overall culture state at optimum harvest. For example, a system fed excess glucose displayed growth rate limitations based on the concentration of free amino acids in the system; however, excess glucose drove the glycation inactivation mechanism. So, although such a system may display high concentrations of viable cell density and total r-tPA concentrations, the

amount of active r-tPA at optimal harvest may be lower than cultures displaying lower maximum concentrations of viable cells. In addition, cultures limited by glucose certainly displayed a lesser amount of glycated r-tPA; however, growth limitations limited overall r-tPA production by the culture. It was found through simulation studies that systems in which all glucose and almost all free amino acids of glutamine and asparagine were consumed led to the optimum solution of maximized total r-tPA activity at harvest of the reactor. In addition, as the mass feed-flow rates and metabolite control set points were varied, the optimum harvest time of the reactor fluctuated widely. In addition, since the Type I and Type II glycoforms of r-tPA have different rates of production by the culture as well as rates of inactivation due to possibly autolysis or aggregation (but not glycation), the ratio of active Type I to Type II glycoforms was found to vary at optimum harvest times as well. Table 3.2 contains results of selected simulations for variable feed-flow rate experiments with metabolite concentration control. Graphical representations of these simulations are contained in Appendix C. In addition to set point, optimal harvest time and total r-tPA activity at harvest, the duration of feeding, the active Type I / Type II ratio at harvest are reported. In addition, limits of total r-tPA, glucose and free amino acids are defined as the steady state values of these culture states. It is noted that these culture states reach steady state values as the viable cell density of the culture reaches zero. Tables 3.3 and 3.4 contain values for selected simulations for fixed feed rate simulations in the case of feed-flow implementation as the first derivative of total r-tPA activity with respect to time approaches zero and implementation as the second derivative with respect to time approaches zero, respectively. Graphical results of these simulations are also contained in Appendix C.

Glucose Set Point [g L^{-1}]	Amino Acids Set Point [g L^{-1}]	Optimal Harvest Time [h]	Total r-tPA Activity at Harvest [x10^6 IU]	Duration of Feeding [hours]	Active Type I / Type II Ratio at Harvest	Maximum $X_{v,app}$ [x10^5 cells mL^{-1}]	Limit of Total r-tPA [µg mL^{-1}]	Limit of Glucose [g L^{-1}]	Limit of Amino Acids [g L^{-1}]
0.50	0.50	205	3.47	242	1.13	47.3	25.8	0	0.44
1.51	1.18	207	5.64	54	1.13	58.5	20.1	0	0.90
1.00	2.70	169	4.90	155	1.04	43.9	13.7	0	1.50
4.50	0.50	304	2.28	163	0.96	59.3	31.1	0.45	0.02

Table 3.2. Results of selected variable feed-flow simulations.

Glucose Mass-flow Rate [g h^{-1}]	Amino Acids Mass-flow Rate [g L^{-1}]	Optimal Harvest Time [h]	Total r-tPA Activity at Harvest [x10^6 IU]	Active Type I / Type II Ratio at Harvest	Maximum $X_{v,app}$ [x10^5 cells mL^{-1}]	Limit of Total r-tPA [µg mL^{-1}]	Limit of Glucose [g L^{-1}]	Limit of Amino Acids [g L^{-1}]
120.0	5.0	170	3.21	1.02	55.5	29.1	2.7	0.03
120.0	40.0	212	5.52	1.06	58.2	19.7	0	1.06
20.0	5.0	215	5.40	1.06	60.2	20.9	0	0.83
20.0	40.0	188	3.87	1.05	44.9	13.6	0	1.82
80.0	25.0	210	5.42	1.06	57.4	19.6	0	0.97

Table 3.3. Results of selected constant feed-flow simulations with feed initiation as the first derivative of total r-tPA activity with respect to time approached zero. This point occurred at 170 hours in simulation.

Glucose Mass-flow Rate [g L^{-1}]	Amino Acids Mass-flow Rate [g L^{-1}]	Optimal Harvest Time [h]	Total r-tPA Activity at Harvest [x10^6 IU]	Active Type I / Type II Ratio at Harvest	Maximum $X_{v,app}$ [x10^5 cells mL^{-1}]	Limit of Total r-tPA [$\mu g\ mL^{-1}$]	Limit of Glucose [g L^{-1}]	Limit of Amino Acids [g L^{-1}]
120.0	5.0	140	2.05	0.94	56.8	29.2	2.7	0.03
120.0	40.0	197	5.55	1.04	60.7	19.7	0	1.06
20.0	5.0	203	5.41	1.04	62.7	20.9	0	0.84
20.0	40.0	172	4.31	1.01	39.0	12.0	0	1.81
80.0	25.0	197	5.44	1.04	59.9	19.6	0	0.98

Table 3.4. Results of selected constant feed-flow simulations with feed initiation as the second derivative of total r-tPA activity with respect to time approached zero. This point occurred at 140 hours in simulation.

3.4.5 Experimental validation of simulation results

Two fed batch experiments were performed with constant feed-flow rates initiated as the first derivative of total r-tPA activity with respect to time approached zero. Experiments were performed to validate simulation studies. In the first experiment, shown in Figure 3.11, low values of glucose and amino acids mass feed-flow rates were chosen. In particular, these values were 5.0 g h^{-1} and 0.5 g h^{-1}, respectively. In the second experiment, shown in Figure 3.12, the mass feed-flow rates were chosen along the optimum feed-flow rate ratio, as shown in Figure 3.8. In this case, the glucose mass feed-flow rate, $M_{Glucose}$, was chosen as 60 g h^{-1}, and the amino acids (glutamine and asparagine) mass feed-flow rate, $M_{Amino\ Acids}$, was chosen as 18.9 g h^{-1}. In both cases, very good agreement was observed between simulations and experimental results. In Figure 3.11 and Figure 3.12 simulation and experimental values of the apparent viable cell density, $X_{v,app}$, and the total r-tPA activity are displayed.

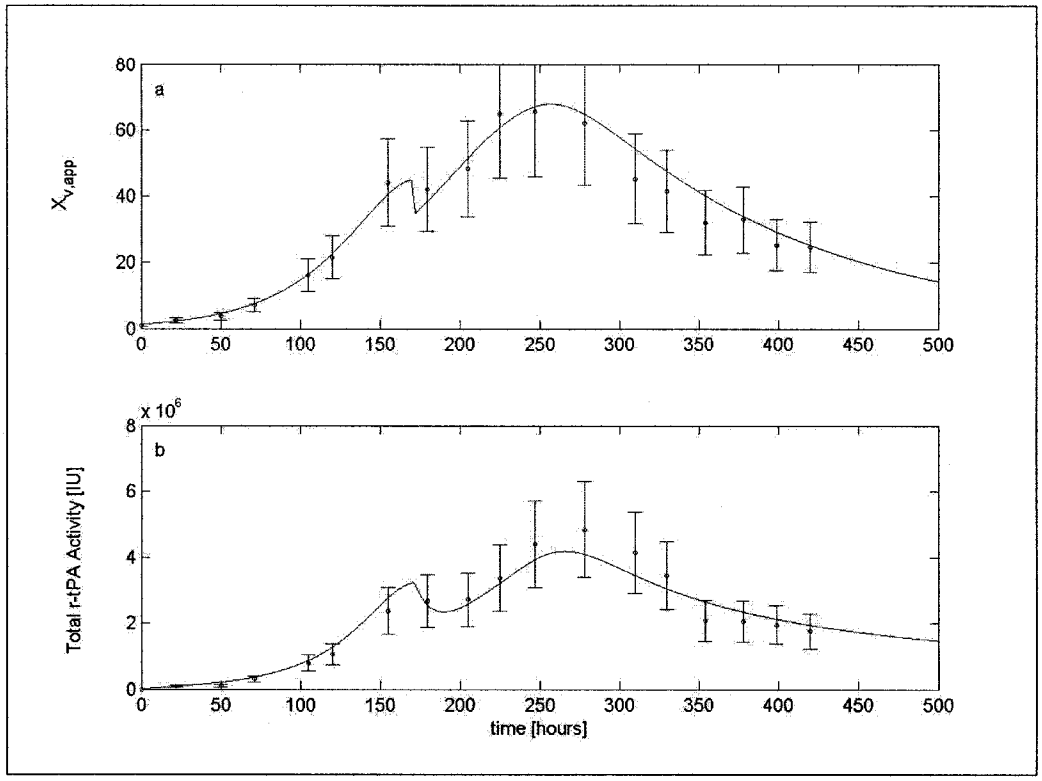

Figure 3.11. Simulation results (line) and experimental results (circles) of (a) apparent viable cell density, $X_{v,app}$ [x10^5 cells mL^{-1}], and (b) total r-tPA activity for a fixed feed-flow fed batch CHO cell cultivation. Glucose mass feed-flow rate, $M_{Glucose}$, was 5 g h^{-1}, and the amino acids mass feed-flow rate, $M_{Amino\ Acids}$, was 0.5 g h^{-1}. Feed was initiated at 170 hours.

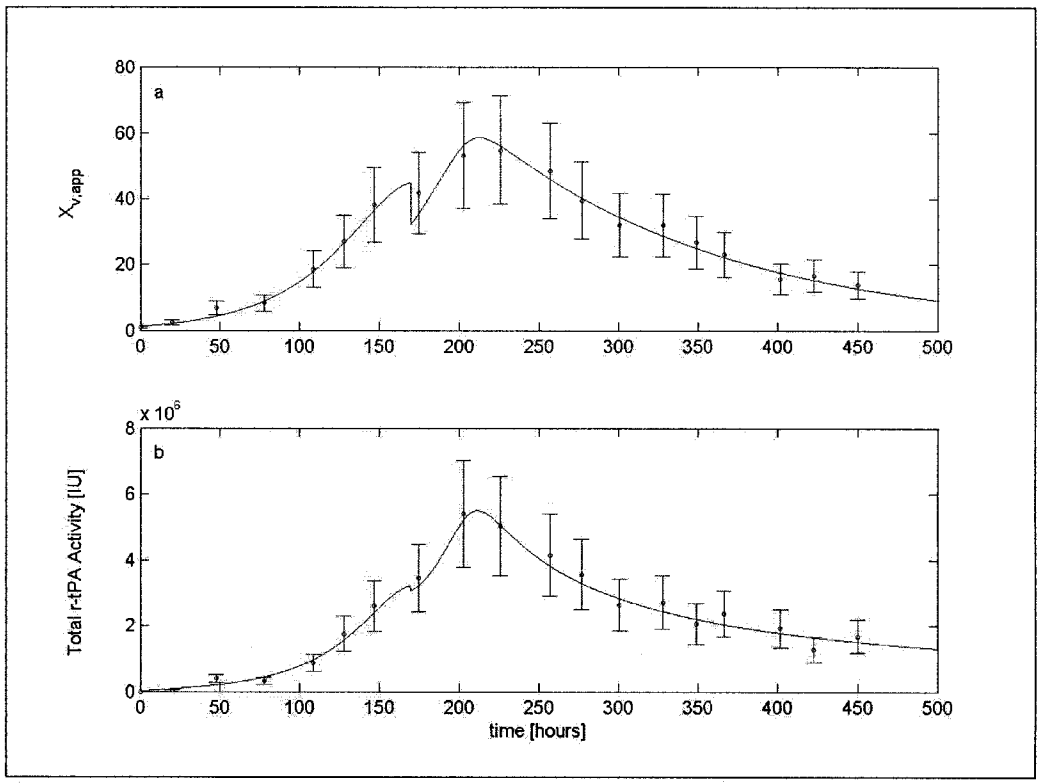

Figure 3.12. Simulation results (line) and experimental results (circles) of (a) apparent viable cell density, $X_{v,app}$ [x10^5 cells mL^{-1}], and (b) total r-tPA activity for a fixed feed-flow fed batch CHO cell cultivation. Glucose mass feed-flow rate, $M_{Glucose}$, was 60 g h^{-1}, and the amino acids mass feed-flow rate, $M_{Amino\ Acids}$, was 18.9 g h^{-1}. Feed was initiated at 170 hours.

3.5 Conclusions

Optimization of fed batch feeding variables with an active product objective function is important when dealing with a product with inactivating mechanisms. In the absence of inactivating mechanisms, product concentration may be directly linked to the viable biomass culture state. With knowledge of intracellular mechanisms, many possibilities exist for optimization using viable biomass as an objective function or control variable. This was the case in the work of Simon and Karim (2002) as variable

feeding profiles of glucose, glutamine and asparagine were used to control CHO cell apoptosis in a stirred bioreactor. However, with product inactivation mechanisms that show a time-dependence, such as with inactivating autolysis and protein aggregation, the overall duration of the fed batch culture must be viewed as a variable to be minimized while maximizing the overall product activity in the culture. Inactivation mechanisms were identified in this research and in Chapter 1 following the work by Simon and Karim (2002), so this work in no way discounts their efforts as apoptosis control would certainly lead to optimum production of a product not displaying inactivation mechanisms. In the case of the r-tPA protein, two separate inactivation mechanisms were identified and one rate constant was dependent upon the presence of a glycan moiety at N184 of the protein. This mechanism, in particular, is believed to be due to serine protease autolysis or protein aggregation, which are both normally associated with low-order rate constants. In the case of both glycoforms of r-tPA, second-order rate constants best fit experimental data. The second inactivation mechanism explored was found due to a covalent reaction with free glucose, known as glycation. The glycation reaction rate constant was found to be independent of the presence of glycosylation at N184. Thus, product models were proposed including these inactivation mechanisms with optimized rate constants. Fed batch simulations were performed for two types of feeding schemes. First, metabolite control was assumed, and glucose and the free amino acids, glutamine and asparagine, were fed into the reactor. Optimization of the fed batch simulation was performed with metabolite set points serving as the design vector and total active r-tPA of the reactor serving as the objective function. A glucose set point of 1.51 g L^{-1} and a combined amino acids set point of 1.18 g L^{-1} maximized the objective function. A second set of

simulations was performed in which constant feeding rates were explored. Feed-flow was initiated in simulations for the cases as the first derivative of the objective function with respect to time approached zero and for the case in which the second derivative of the objective function with respect to time approached zero. These simulations identified an optimum mass feed-flow rate ratio of glucose to amino acids of 3.15. In this family of optimum solutions, rapid feeding further maximized the objective function. In addition, comparison of the two feeding strategies identified the constant feed-flow rate feeding procedure as yielding slightly higher values of the maximized objective function than variable feeding strategies in the presence of metabolite control. Experimental results were used to verify simulation results of constant feed-flow simulations. This system provided the opportunity to examine a unique problem in optimization as competing mechanisms led to possible inhibition of culture growth or product inactivation. For example, excess amino-acid feed led to free glutamine accumulation in the reactor. The degradation of glutamine produced ammonia, which, in turn, inhibits culture growth. On the other hand, excess free glucose in the reactor drove the glycation product inactivation mechanism. Low feeding rates, resulting in a prolonged fed batch operation duration, drove the glucose-independent protein inactivation mechanism, which is probably due to autolysis or aggregation. Optimal solutions in fed batch operation were found to increase the yield of total r-tPA activity by greater than 75% as compared to batch cultivation.

Chapter 4

VARIABLE SITE-OCCUPANCY CLASSIFICATION OF *N*-LINKED GLYCOSYLATION USING ARTIFICIAL NEURAL NETWORKS

4.1 Introduction and background

4.1.1 Glycosylated pharmaceutical products

Protein glycosylation is an important post-translational modification that involves the covalent attachment of an oligosaccharide to a polypeptide chain and its further enzymatic processing. Glycan attachment to an N-*X*-S/T polypeptide sequence (where *X* is not proline) is accomplished by the membrane-bound oligosaccharyl transferase enzyme. Completion of this reaction determines glycosylation macroheterogeneity (Kornfeld and Kornfeld, 1985; Roth, 1987; Silberstein and Gilmore, 1996; Aebi and Hennet, 2001; Spiro, 2002). Glycan attachment and remodeling processes occur in the endoplasmic reticulum (ER) and Golgi apparatus of almost all cell types; however, only a select number of cell types produce glycosylation variants compatible in humans. Glycosylation plays an important role in the determination of biological activity of many pharmaceutical products and proteins with biological activity *in vivo*. 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 selected by the pharmaceutical industry, for expression of glycosylated proteins, commonly include human melanoma cells, baby hamster kidney (BHK) cells, and Chinese hamster ovary (CHO) cells.

The importance of protein glycosylation has developed a new area for optimization in the bioprocessing industry. An even more specific application of such optimization has progressed with respect to variable site-occupancy glycosylation (Andersen *et al.*, 2000; Senger and Karim, 2003a). For some proteins, the 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; Wittwer and Howard, 1990). For the case study of r-tPA protein, three sites of *N*-linked protein glycosylation are available. The r-tPA protein is a multi-domain serine protease and is a natural component of the human fibrinolytic system. Two of the glycosylated sites of the r-tPA protein (N117 and N448) experience robust glycosylation and are fully glycosylated under normal cultivation conditions. However, a variable site of glycosylation site-occupancy exists at N184 (Grossbard, 1987). Manipulation of culture conditions such as temperature, culture butyrate levels, and shear stress applied to the culture have been shown to impact the degree to which this variable glycosylation site may become occupied (Andersen *et al.*, 2000; Senger and Karim, 2003a,b).

4.1.2 Challenges facing protein glycosylation and possible solutions

In many cases, such as with r-tPA, industrial fed-batch production results in a heterogeneous mixture of glycoforms. Of significance to the pharmaceutical industry in the case of r-tPA is that glycosylation site-occupancy at N184 of r-tPA is a strong indicator of specific enzymatic activity (Einarsson *et al.*, 1985; Wittwer and Howard, 1990). Furthermore, the separation of glycoforms from cell culture supernatant is difficult from a process standpoint, and usually results in abnormally high production costs. Downstream production costs account for roughly 80% of pharmaceutical production costs; high production costs are known to limit the use and availability of effective pharmaceutical products (Blanch and Clark, 1997). Thus, considerable attention has been dedicated to metabolic engineering of cell cultures in order to manipulate the glycoform production ratios of the culture during the cultivation process (Goochee and Monica, 1990; Andersen *et al.*, 2000; Senger and Karim, 2003a,b). In addition, many researchers have focused on genetic manipulations in attempts to understand and address the problem of protein glycosylation heterogeneity (Paques *et al.*, 1992; Shakin-Eshleman, 1996; Nishikawa and Mizuno, 2001). In addition, statistical analyses have been performed by Petrescu *et al.* (2004) and have identified sequence and structural characteristics common to robustly occupied and unoccupied glycosylation sites. Glycosylation optimization is of great benefit to pharmaceutical manufacturing in terms of production costs and would result in tighter control of product specific activity. Thus, a better understanding of variable site-occupancy glycosylation of polypeptide sequences may lead to advances resulting in homogenous glycosylation of pharmaceutical products during production.

4.1.3 Importance of computational techniques

Computational methods are used frequently for the prediction of structural characteristics of proteins. A major advantage of combining computational predictions with experimental observation is that simulations may replace limited laboratory resources in specific cases. In the area of pharmaceutical product development and protein activity enhancement from site-directed mutations, combinatorial laboratory methods have proven successful. But, a combination of these experimental techniques with computational methods can greatly reduce the number of site-directed mutations that must be performed and analyzed in a laboratory. Thus, it is noted that computational methods cannot eliminate the need for experimental validation of theoretical predictions. Currently, computational methods exist for the structural prediction of: protein secondary structures (Kneller *et al.*, 1990; Rost and Sander, 1993; Jones, 1999; Baldi *et al.* 2000; Pollastri *et al.*, 2002a), three-dimensional folding and backbone structure (Kelley *et al.*, 2000; Combet *et al.*, 2002; Mallick *et al.*, 2002; Baldi and Pollastri, 2003), solvent accessibility (Pollastri *et al.*, 2002b), residue contacts (Pollastri *et al.*, 2001; Pollastri *et al.*, 2002b), disulfide bonding (Fariselli *et al.*, 1999), cofactor binding (Kleiger and Eisenberg, 2002), long-range surface accessibility (Yeates, 1995), transmembrane properties of helices (Eisenberg *et al.*, 1982) as well as many other characteristics. However, computational methods do not exist for the prediction of glycosylation characteristics of a polypeptide sequence when produced in cell types common for pharmaceutical production.

4.1.4 The role of neural networks in bioinformatics

The area of structural bioinformatics has expanded to predict two- and three-dimensional structures and characteristics of polypeptide structures through the use of neural network-based models. In general, neural networks have been an integral part of this process 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). Among the most predicted protein characteristics is secondary structure of alpha helix, beta sheet, and unordered structures. Recent advances have resulted in the accuracy of these predictions to approach 76% out of a theoretical maximum accuracy of about 88%, which is determined from the accuracy of experimental methods that compose the reference set of these neural models (Rost, 2001; Rost and Eyrich, 2001).

4.1.5 Using neural networks to predict glycosylation characteristics

A novel neural-network model has been developed in this research for predictions of glycosylation characteristics with respect to variable site-occupancy. Whereas the recent statistical analysis by Petrescu *et al.* (2004) discerns relationships resulting in occupied and unoccupied glycosylation sites, this research identifies characteristics resulting in variable site-occupancy. The development of this model has allowed insight to many questions concerning protein glycosylation. In addition to determining that the phenomena of variable site-occupancy in the absence of substrate limitation is related to primary sequence characteristics, it has also been determined that the number of amino acid residues around the site of glycosylation with influence on glycosylation

network training was more successful with the clustered data set of Table 4.1. Quantification was also required of the variable site-occupancy classification (target data sets) as well. A glycosylation site in which variable site-occupancy was observed was represented as *1*; whereas, robust glycosylation was quantified as *0*. It is duly noted that variable glycosylation is defined in this model as that in which a heterogeneous mixture of glycosylated and unglycosylated polypeptide is produced by a culture in which substrate limitation is not a factor. In addition, a robust glycosylation process results in a homogenous species that may either be glycosylated or unglycosylated.

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

Table 4.1. Primary sequence quantification based on amino acid residue clustering and site-occupancy affinity as described by Kasturi *et al.* (1997) and Mellquist *et al.* (1998).

4.2.2 Neural network architecture

Elman recurrent neural networks were investigated as possible artificial neural network architectures since recurrent networks are renowned for the ability to learn sequenced data. Of the components of this model data set, five data points (glycosylation

sites) were reserved for testing of the recurrent network. The testing set consisted of two variable sites and three robust sites of glycosylation. Thus, 43 sequences, including 18 sequences with variable-site occupancy, were designated for network training. Data points consisted of a single amino acid and glycosylation classification. Thus, all unglycosylated amino acids received a classification of *0*. Only asparagine residues with variable glycosylation received a classification of *1*. The hidden layers of the recurrent neural network utilized hyperbolic tangent sigmoid transfer functions. The output layer of the recurrent network used a log sigmoid transfer function, which returns a value between 0 and 1 by definition. A single perceptron was employed following training of the recurrent neural network for two-dimensional data classification. The perceptron was not trained, but it was manually configured to mimic a simple rounding function of recurrent neural network output values. Network training was performed using gradient descent with momentum and adaptive learning rate backpropagation for duration of 2000 epochs. Network architecture was optimized for the given data set by adjusting the overall number of neurons given a single hidden layer. Multiple hidden layers were investigated but were not found to improve results (data not presented). The details of performance optimizations are described in the following section. Simulation results of the testing set were then compared to testing set data used for network training by calculation of the mean-square error.

4.2.3 Optimization of input sequence length and network architecture

Optimization of the length of input sequence around the site of glycosylation (or *glycosylation window*) as well as optimization of neural network architecture (number of

hidden layer neurons) was performed. Thus, the resulting optimization problem contained three dimensions for optimization in which the mean-square error of the testing data set was minimized. The problem of parameterization was followed closely with the optimization of input sequence length. As the input length of a sequence is modified, so must the input and output layers of the neural network. This, in turn, changes the number of adjustable parameters (weights and biases) associated with the neural network model, given that the number of hidden layer neurons remains constant. Since an output value is associated with each residue, input sequence length also directly determines the size of the output layer. For consistency, the number of hidden layer neurons was adjusted so that the number of data points used for training per adjustable parameter of the neural network remained constant. The number of hidden layer neurons for each input sequence length, given a ratio of data points per adjustable parameter of roughly 1, is shown in Figure 4.1. The glycosylation window was evaluated for a total of 21 amino acid residues beginning 10 residues toward the *N*-terminus (upstream) (*n*-10) of the glycosylation site and extending ten residues beyond (*n*+10) the glycosylation site (*n*). As neural networks convergence and prediction of the testing data set is dependent on the initialized weight and bias values of the network, 100 separate iterations of each data point were evaluated using networks initialized with random weight and bias values. Mean-square error results of testing data set predictions for a given sequence length were then averaged to minimize bias. The two-dimensional sequence length optimization problem was solved using a steepest gradient evaluation until the mean square error of the testing data set was minimized. Multiple starting points were utilized in the optimization technique in order to identify global and local minima values. Following

sequence optimization, the optimized sequence length was used for architecture optimization. The optimum sequence was trained with separate neural networks differing in the number of hidden layer neurons. The solution to this one-dimensional optimization problem straight-forward, and the previous two-dimensional problem was re-solved using optimized network architecture structures. This iterative procedure converged in few iterations.

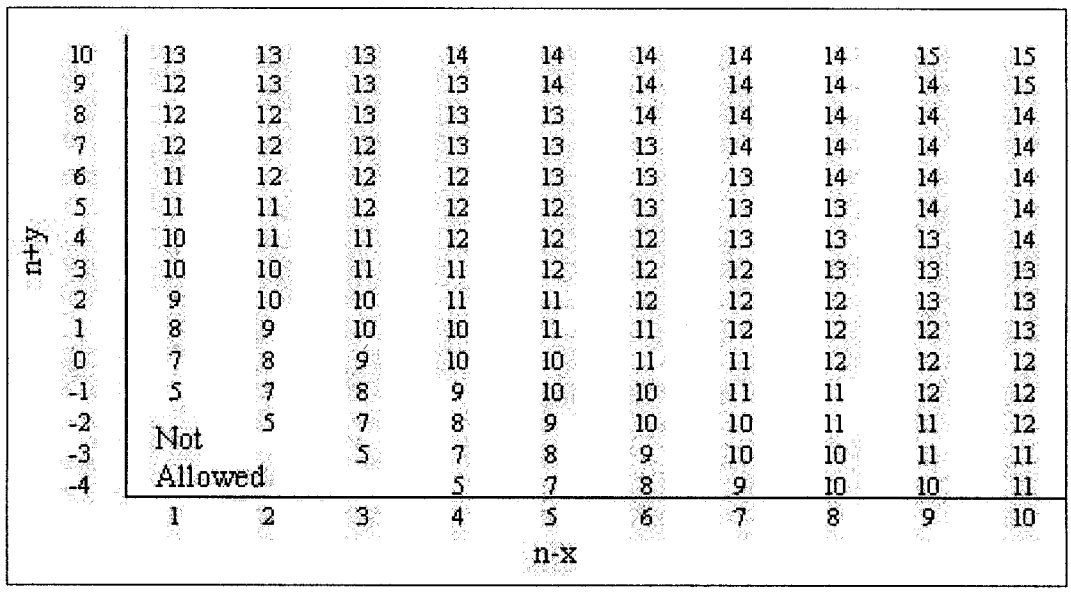

Figure 4.1. Number of neurons in the single hidden layer of recurrent neural networks for input sequence lengths defined by a starting residue (x) (along the abscissa) upstream of the glycosylation site (n) and extending to the terminating residue (y), described by the ordinate axis.

4.2.4 Simulations

Following identification of optimum sequence input length and recurrent neural network architecture, a separate set of recurrent neural networks was identified for making predictions of polypeptide sequences. In order for a recurrent neural network to

be considered for making glycosylation predictions, the mean-square error for prediction of the testing data set (following perceptron classification) was required to be zero. Only networks consisting of optimized input sequence length and architecture were considered for sequence predictions outside of the reserved testing data set. Twenty such networks were located, and each contained different weight and bias values upon the completion of training. Recurrent network simulation results were individually classified by the same perceptron classifier and results were averaged. This averaged result was reclassified but also used as a confidence level for the final classification. For example, if 18 of 20 networks predicted variable site-occupancy glycosylation for a particular sequence, results were returned as *1* (variable) with a confidence level of 0.9. Simulations were performed on N39 of wild-type and variants of the rabies virus glycoprotein (rgp). Only proteins experimentally evaluated in Kasturi *et al.* (1997) and Mellquist *et al.* (1998) were simulated. Further simulations were performed on simple theoretical sequences containing the N-*X*-S/T glycosylation marker to evaluate the effects of charged residues. Alanine, lysine and aspartate residues, along with the necessary N-*X*-S/T sequence, were used to construct hypothetical polypeptide sequences for simulation.

4.3 Results and discussion

4.3.1 Optimization of the glycosylation window length

Using multiple trainings of the neural network model, the number of residues surrounding the site of glycosylation (or *glycosylation window*) with influence on glycosylation characteristics was examined. Extensive laboratory research has identified primary structure at the *X* and *Y* positions of the N-*X*-S/T-*Y* *N*-linked glycosylation

sequence as vital to determining the site-occupancy of the glycosylation site (Shakin-Eshleman, 1996; Kasturi *et al.*, 1997; Melquist *et al.*, 1998; Petrescu *et al.*, 2004). However, the neural network model approach of this research allowed for the consideration of a much larger optimum sequence length in determining site-occupancy characteristics. As a starting point, a twenty-one amino acid residue glycosylation window was examined. This sequence consisted of 10 residues on either side (n-10 and n+10) of the glycosylated asparagine residue (n). Results of this two-dimensional optimization problem for recurrent neural network architecture with a single hidden layer are shown in Figure 4.2. For each sequence the number of adjustable parameters per data point was held constant at just less than 1, which was determined from convergence of the three-dimensional optimization problem including neural network architecture. Figure 4.2 reports the average mean-square error prediction of the testing data set for 100 independent iterations of the input sequence. The starting residue of the input sequence is displayed on the abscissa as (n-x) in Figure 4.2; whereas, the ordinate (n+y) describes the terminating residue of the sequence relative to the site of glycosylation (n). As described previously, for each iteration, prior to network training, the recurrent neural network was initiated with random weight and bias values. Multiple evaluations of data points were performed to determine an error estimate of the given procedure. Local and global optima points were independently evaluated at least 5 times of 100 separate iterations. The average standard deviation of the data points in Figure 4.2 was calculated as 0.02, or approximately ± 5% of the given data point value. 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 an average mean-square error value of

0.7. As shown by Figure 4.2, the optimal sequence for the determination of variable site-occupancy was identified as the sequence beginning five residues prior (n-5) to the glycosylation site (n) and extending four residues beyond (n+4) the site of glycan attachment. This is a much larger sequence than has been previously reported for the determination of variable site-occupancy glycosylation. Also evident in Figure 4.2 is that multiple local optima exist as well. One such local optimum occurs at a sequence length of four residues and defined by five residues prior (n-5) to the glycosylation site through two residues prior (n-2) to the glycosylation site. This further suggests that residues preceding (upstream of) the site of glycosylation may influence glycosylation characteristics in terms of variable site-occupancy. A local optimum observed at a sequence starting three residues prior (n-3) to the glycosylation site and extending one residue beyond (n+1) this site suggests close-range interactions (1-3 residues) may exist separately of long-range interactions (~4-5 residues) in determining glycosylation site-occupancy. Prediction of site-occupancy results in high values of the mean square error as the glycosylation window extends beyond the limits of (n-5) to (n+4) even though this same information is contained within larger sequences. At larger than optimum glycosylation window lengths, superfluous information is input into the training recurrent neural network, and with a glycosylation window smaller than the optimum, relevant information regarding glycosylation site-occupancy is not learned by the neural network.

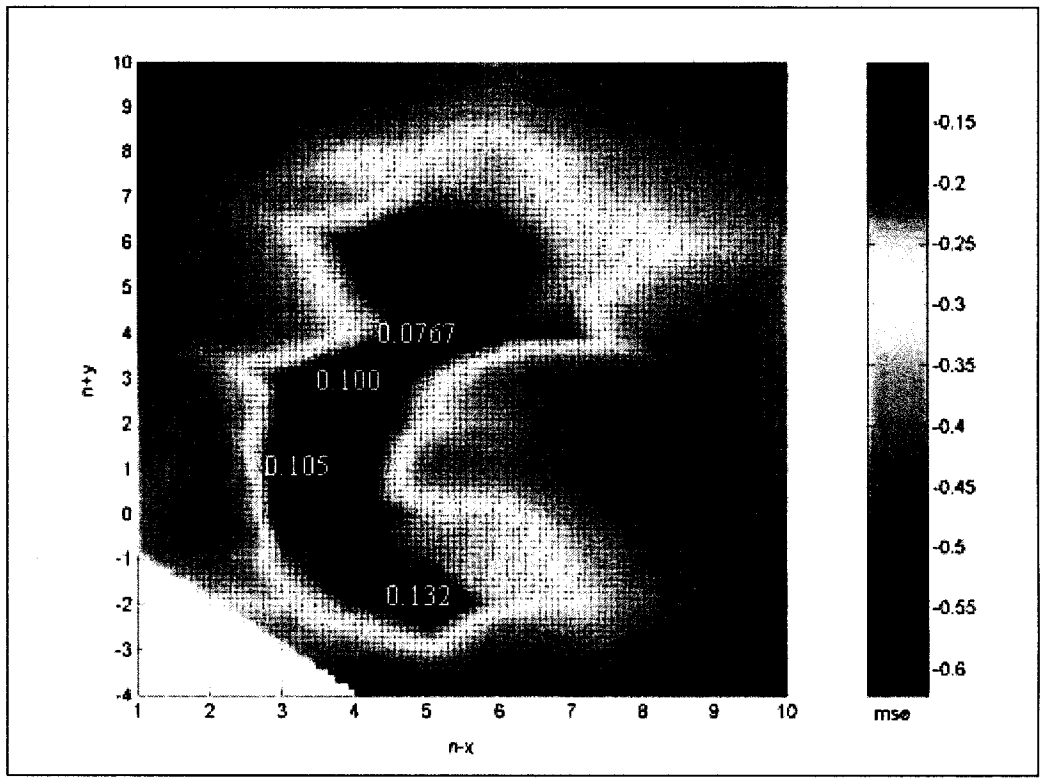

Figure 4.2. Sequence input length optimization for glycosylation site-occupancy prediction. Averaged mean-square error values are displayed for input sequence lengths initiating upstream of the site of glycosylation (*n*). Starting residues are displayed on the abscissa axis, and terminating residues are displayed on the ordinate.

4.3.2 Optimization of neural network architecture

Results of neural network architecture optimization for a glycosylation window spanning (*n*-5) to (*n*+4) are shown in Figure 4.3. Data points represent the mean of at least 3 separate training trials of 100 iterations each and evaluations of the testing data set. Error analysis was generated from the standard deviations of the means of 100 iterations. Optimal neural network architecture was recognized at the point where the number of adjustable parameters was roughly equal to the number of data points used in the training procedure. Few evaluations of the system were made for a situation of

overparameterization, which is defined as the number of adjustable parameters exceeds the number of data points used in network training. A steady decrease in the mean-square error for prediction of the testing data set was observed until the number of adjustable parameters of the neural network approached the number of data points used for training. As the system became overparameterized by up to 200 adjustable parameters, no significant change in the mean-square error was detected.

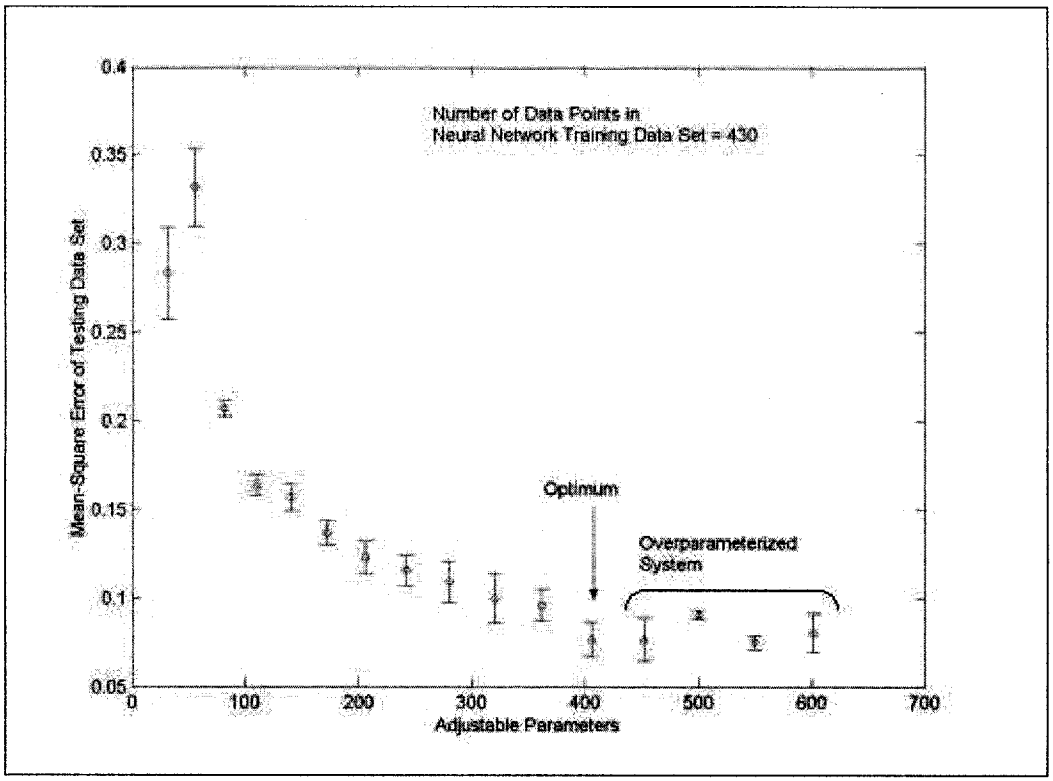

Figure 4.3. Recurrent neural network architecture optimization for an optimized input glycosylation window consisting of 5 residues prior (n-5) to 4 residues downstream (n+4) of the glycosylation site (n). Error bars represent one standard deviation.

4.3.3 Selected neural networks for further simulations

Following sequence input length and neural network architecture optimizations, neural networks were located for the purpose of making simulations of sequences not

included as part of the neural network training or testing data sets. Twenty independent neural networks were identified, with optimized sequence input length and architecture, that classified the neural network testing set with a mean-square error of zero. All neural networks consisted of different weight and bias values and individual recurrent network values were different for the prediction of each protein sequence (data not shown). The mean recurrent network output, classification and confidence level are presented in Table 4.2 for each of the components of the neural network testing data set. All classifications were consistent with published experimental observations, and the confidence level for each prediction represented the highest possible value. For networks with optimized input sequence length and architecture, networks initiated with random weight and bias values prior to training were found to converge to predict the neural network testing data set with a mean-square error value of zero for better than 67% of all networks trained. Therefore, locating such networks to meet requirements to perform further simulations was not a difficult task.

Input Sequence	Average Recurrent Network Result	Overall Model Classification	Confidence Interval	Published Classification	Reference
Transferrin Receptor (G724S) (site 722)	9.84×10^{-1}	1	1	1	(Williams and Enns, 1993)
Transferrin Receptor (G724S) (site 251)	1.21×10^{-4}	0	1	0	(Williams and Enns, 1993)
r-tPA (site 117)	1.39×10^{-2}	0	1	0	(Grossbard, 1987)
r-tPA (site 184)	8.75×10^{-1}	1	1	1	(Grossbard, 1987)
r-tPA (site 448)	3.65×10^{-2}	0	1	0	(Grossbard, 1987)

Table 4.2. Neural network testing data set components with predicted and published experimental classifications. Classification of *1* corresponds to variable site-occupancy, and a classification of *0* corresponds to robust glycosylation.

4.3.4 Simulations of rabies virus glycoprotein wild-type and mutants

To further verify model predictions beyond the neural network testing data set, simulations were made of wild-type and mutant variants of the rabies virus glycoprotein (rgp). Particular variants were chosen, as comprehensive experimental research has been previously published analyzing the effects of site-directed mutations around the site of glycosylation (*n*) on glycosylation efficiency (Kasturi *et al.*, 1997; Mellquist *et al.*, 1998). Specifically, this research examines mutations at the *X* and *Y* positions of the N-*X*-S/T-*Y* sequence of N37 of rgb on the degree of glycosylation variable site-occupancy. Glycosylation efficiency of mutant species of the rgb protein is reported in these publications. Glycosylation efficiency between values of 0 and 1 corresponds to variable site-occupancy glycosylation, and a glycosylation efficiency of 0 or 1 describes robust glycosylation of the polypeptide sequence in question. Values of glycosylation efficiency presented in these publications were interpolated, taking into account experimental error, and are listed in Table 4.3 for a total of 19 mutant species and the wild-type sequence. Also contained in Table 4.3 are the results of simulations of these sequences. The overall model classification is reported with a confidence level. Mutations of the rgb protein resulting in a loss of glycosylation (L38W and G40P; simulations 5a and 9a) were correctly predicted as displaying robust glycosylation. In addition, sequences displaying variable site-occupancy, to a glycosylation efficiency of 0.9 (simulations 2a-4a, 8a, 10a-13a, 14a-19a), were correctly classified by the model. The model failed to predict variable site-occupancy glycosylation for the S39T G40W sequence (simulation 15a). However, it is noted that this robust prediction was made with a low confidence level (0.65). The model did correctly predict variable site-occupancy glycosylation for the

G40F aromatic substitution (simulation 8a). This suggests future additions to the neural network training data set should include additional sequences with aromatic residues in which variable site-occupancy was observed. As sequences were constructed that resulted in glycosylation efficiency of or exceeding 95% (L38N S39T, L38G S39T, S39T G40C; simulations 6a, 7a, 20a), most model predictions classified these sequences as having robust glycosylation. The exception in this case was for the L38N S39T mutant (simulation 6a). Variable site-occupancy glycosylation was predicted in this case, but a lower confidence level in this case (0.7) suggests many networks recognized this sequence as robust. In addition, given the experimental error in this case of roughly ±5%, variable glycosylation may accurately describe this system. Of further importance is that the sensitivity of model predictions was analyzed in this set of simulations. In short, this set of model predictions will classify a polypeptide sequence as variable if glycosylation efficiency is between 10% and 95% for a given sequence.

Corresponding Simulation	Sequence	Overall Model Classification	Confidence Level	Published Glycosylation Efficiency	Experimental and Simulation Correlation?
1a	Wild-type	1	1	0.35	Yes
2a	S39T	1	0.95	0.8	Yes
3a	L38N	1	1	0.7	Yes
4a	L38S	1	1	0.9	Yes
5a	L38W	0	0.6	0.1	Yes
6a	L38N S39T	1	0.7	0.95	Yes
7a	L38G S39T	0	0.6	>0.95	Yes
8a	G40F	1	0.75	0.4	Yes
9a	G40P	0	0.75	0.05	Yes
10a	G40H	1	1	0.55	Yes
11a	G40M	1	1	0.7	Yes
12a	G40N	1	1	0.8	Yes
13a	G40S	1	1	0.8	Yes
14a	G40C	1	0.9	0.8	Yes
15a	S39T G40W	0	0.65	0.8	No
16a	S39T G40H	1	1	0.85	Yes
17a	S39T G40M	1	1	0.9	Yes
18a	S39T G40N	1	1	0.9	Yes
19a	S39T G40T	1	1	0.9	Yes
20a	S39T G40C	0	0.55	>0.95	Yes

Table 4.3. Wild type and variant rgb sequences with predicted classification and confidence levels and published glycosylation efficiency (Kasturi *et al.*, 1997; Mellquist *et al.*, 1998). Classification of *1* corresponds to variable site-occupancy, and a classification of *0* corresponds to robust glycosylation.

4.3.5 Statistical analysis of the glycosylation window

Polypeptide sequences of the neural network training and testing data sets were divided into sequences of robust and variable site-occupancy. For each of these data sets, the occurrence of specific amino acid residues for each site of the glycosylation window was analyzed and reported as a percentage value. Results for variable site-occupancy glycosylation are shown in Figure 4.4, and results for robust glycosylation are shown in

Figure 4.5. Results suggested relative importance of primary sequence characteristics at different positions of the glycosylation window. For example, the ratio of threonine to serine at (*n*+2) of the N-*X*-S/T glycosylation sequence is much larger in robust sequences (1.55) than in variable site-occupancy sequences (0.82). This further provides evidence that threonine at the (*n*+2) position results in more complete glycosylation. However, it also suggests other factors are involved in variable site-occupancy glycosylation classification. One such difference detected between these two figures is that sequences of variable site-occupancy contained a much larger fraction of positively charged residues (lysine, arginine and histidine) in the sites (*n*-1) through (*n*-3) of the glycosylation window. In addition, at site (*n*-1) it was observed that most robust sequences contained hydrophobic residues; whereas, a larger fraction of sequences with variable site-occupancy contain hydrophilic residues. The site directly following glycosylation (*n*+1) was found to have some major differences between sequences promoting robust and variable site-occupancy. In particular, charged residues of variable site-occupancy sequences were positively charged. This is in contrast to robust glycosylated sequences, which contained negatively charged residues at this position. The other incidence of a large fraction of negatively charged residues in robust glycosylated sequences occurred at the (*n*+3) site. Finally, a high incidence of phenylalanine residues was observed for sequences of variable site-occupancy glycosylation at the (*n*+4) site of the glycosylation window. Finally, no single amino acid residue or position of the glycosylation window was found to completely dictate variable site-occupancy glycosylation. Instead, results indicate that many possible

mechanisms exist for variable site-occupancy, and these mechanisms may show a high degree of sequential dependence throughout the glycosylation window.

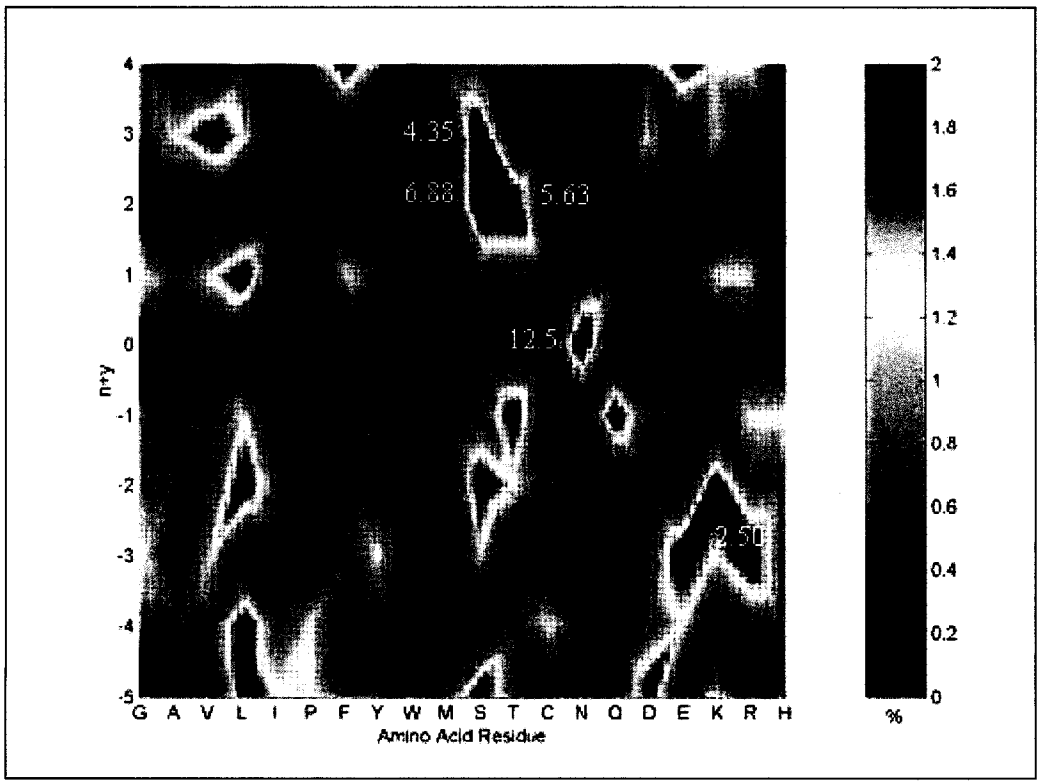

Figure 4.4. Statistical analysis of variable site-occupancy glycosylation sequences. The percentage of residue occurrence is plotted for each position of the glycosylation window. For occurrences greater than 2%, the value has been manually listed.

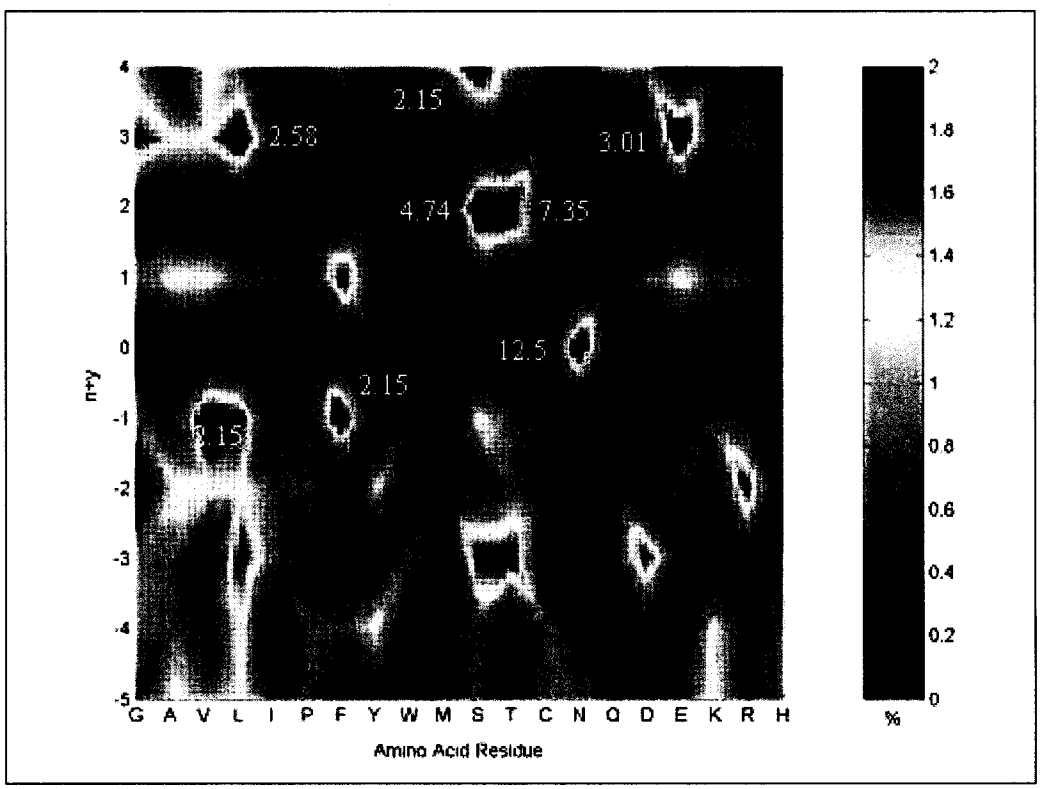

Figure 4.5. Statistical analysis of robust glycosylation sequences. The percentage of residue occurrence is plotted for each position of the glycosylation window. For occurrences greater than 2%, the value has been manually listed.

4.3.6 Simulations of theoretical sequences

Due to the high level of agreement between simulation results and variable site-occupancy classification of rgp wild-type and variant sequences, the same simulation technique was applied to theoretical sequences listed in Table 4.4. Variable site-occupancy glycosylation in the presence of charged residues was studied in the following simulations. For consistency, only alanine was used as the uncharged residue outside of the required N-*X*-S/T glycosylation sequence. In addition, aspartate and lysine were used as charged residues in this study. Simulations suggested that a sequence consisting of

alanine residues throughout the glycosylation window (except for the glycosylation sequence at (*n*) and (*n*+2)) would result in robust glycosylation with confidence intervals of 0.75 and 1 (simulations 1b and 2b). Robust glycosylation was predicted with a confidence interval of 1 for a N-P-S/T sequence, which is traditionally known as a robust, unglycosylated sequence. With positively charged lysine residues occupying the glycosylation window and serine at position (*n*+2), variable site-occupancy glycosylation was predicted with a confidence level of 0.95. This was also the case for threonine at position (*n*+2) with a confidence level of 0.95 (simulations 5b and 6b). Robust glycosylation with a confidence level of 1 was observed for a glycosylation window consisting of negatively charged aspartate residues (simulations 7b and 8b). Simulations 9b through 12b in Table 4.4 use the alanine glycosylation window sequence of simulation 1b as a starting point. Lysine residues were substituted in spaces close to the glycosylation site until variable site-occupancy was achieved similar to simulation 5b. Substitutions of lysine residues at either (*n*+1) or (*n*+3) were required to achieve a variable site-occupancy prediction (simulation 9b, 10b, 12b). However, these same substitutions with aspartate residues (simulation 13b) resulted in a robust glycosylation prediction. Substitution of lysine upstream of the glycosylation site at position (*n*-1) resulted in a robust glycosylation prediction, however (simulation 11b). In fact, lysine substitutions at (*n*-2) and (*n*-3) were also required to generate a variable site-occupancy prediction (simulations 14b-16b). As shown in simulations 17b and 18b, the number of lysine residue substitutions required to promote variable site-occupancy was greater for the N-*X*-T sequence than the N-*X*-S sequence. This further supports the idea presented in Kasturi *et al.* (1997); Mellquist *et al.* (1998) and Petrescu *et al.* (2004) that replacement

of serine in the glycosylation sequence with threonine increases the robustness of glycan attachment. The influence of upstream positions (*n*-4) and (*n*-5) were demonstrated in simulation 19b as the sequence of simulation 15b was transformed to a prediction of robust glycosylation through substitutions in these positions. Simulations 20b through 23b suggest the importance of the position (*n*+1) in glycosylation site-occupancy determination in the presence of charged residues.

Corresponding Simulation	*n*-5	*n*-4	*n*-3	*n*-2	*n*-1	*n*	*n*+1	*n*+2	*n*+3	*n*+4	Overall Model Classification	Confidence Level
1b	A	A	A	A	A	N	A	S	A	A	0	0.75
2b	A	A	A	A	A	N	A	T	A	A	0	1
3b	A	A	A	A	A	N	P	S	A	A	0	1
4b	A	A	A	A	A	N	P	T	A	A	0	1
5b	K	K	K	K	K	N	K	S	K	K	1	1
6b	K	K	K	K	K	N	K	T	K	K	1	0.95
7b	D	D	D	D	D	N	D	S	D	D	0	1
8b	D	D	D	D	D	N	D	T	D	D	0	1
9b	A	A	A	A	A	N	K	S	A	A	1	0.95
10b	A	A	A	A	A	N	A	S	K	A	1	0.7
11b	A	A	A	A	K	N	A	S	A	A	0	0.6
12b	A	A	A	A	A	N	K	S	K	A	1	1
13b	A	A	A	A	A	N	D	S	D	A	0	1
14b	A	A	A	K	K	N	A	S	A	A	0	0.6
15b	A	A	K	K	K	N	A	S	A	A	1	0.65
16b	K	K	K	K	K	N	A	S	A	A	1	0.8
17b	A	A	K	K	K	N	A	T	A	A	0	0.6
18b	A	K	K	K	K	N	A	T	A	A	1	0.7
19b	D	D	K	K	K	N	A	S	A	A	0	0.8
20b	D	D	D	D	D	N	K	S	D	D	1	0.9
21b	D	D	D	D	D	N	A	S	D	D	0	0.9
22b	D	D	D	D	D	N	K	T	D	D	1	0.5
23b	K	K	K	K	K	N	D	S	K	K	0	0.85

Table 4.4. Theoretical polypeptide sequences with glycosylation site-occupancy prediction and confidence level. Classification of *1* corresponds to variable site-occupancy, and a classification of *0* corresponds to robust glycosylation.

4.4 Conclusions

A novel neural network-based model has been developed, verified and used to make predictions regarding the nature of variable site-occupancy glycosylation. Specifically, the model was developed to perform a two-dimensional classification of *N*-linked glycosylation macroheterogeneity. Glycosylation site-occupancy dependent upon cell culture conditions in the absence of substrate limitation was classified as *variable site-occupancy*; whereas, *robust* was used to describe glycosylation not showing this dependence. Possible implications of this model include the specific engineering (by site-directed mutation) of protein sequences to eliminate variable sites of glycosylation site-occupancy. Elimination of variable site-occupancy glycosylation will have significant impact in the pharmaceutical industry in that robust glycosylation results in homogenous product production with respect to recombinant product glycosylation. Further application of this model may also lie in the identification of glycosylation characteristics, from a macroheterogeneity standpoint, of newly discovered proteins in which titers are low. Neural network training algorithms defined an optimal sequence length to enable variable site-occupancy classification from primary sequence information. This sequence was defined as extending 5 residues from the glycosylation site toward the *N*-terminus and 4 residues toward the *C*-terminus. In addition, correct classification of a neural network testing data set led to the identification of networks suitable for construction of a model for further predictions. In turn, variable site-occupancy classification of published experimental data was correctly predicted. However, the need for an expanded neural network training data set to include more sequences containing aromatic residues with variable site-occupancy glycosylation has

been noted. Further simulations with charged and uncharged residues of lysine, aspartate, and alanine demonstrated the usefulness of this model. Simulation results suggested a decrease in glycosylation site-occupancy robustness in the presence of positively charged residues and demonstrated the influence of residues upstream of the glycosylation site.

Chapter 5

MICROHETEROGENEITY CLASSIFICATION OF *N*-LINKED GLYCOSYLATION USING ARTIFICIAL NEURAL NETWORKS

5.1 Introduction and background

5.1.1 Functions of glycosylated proteins

Most secreted and membrane-associated proteins of eukaryotic cells are glycosylated. Different types of glycosylation linkages are common in eukaryotic systems: *N*-linked and *O*-linked glycosidic linkages as well as incorporation into the glycosylphosphatidylinositol (GPI) membrane anchor. Enzymatic processing commonly results in complex-type, high mannose and hybrid glycan structures. In addition, various terminal glycan modifications include sialylation, fucosylation, galactosylation, sulfation, and phosphorylation (Hubbard and Ivatt, 1981; Kornfeld and Kornfeld, 1985; Roth, 1987; Parekh, 1994). Many combinations of glycan structures with varying terminal linkages are known and have been found to influence both intermolecular and intramolecular properties of proteins. Examples of intramolecular influences of protein glycosylation include: proper protein folding, intracellular location, biological activity, solubility, antigenicity, biological half-life and protease sensitivity. Intermolecular properties found altered due to glycosylation include: targeting to lysosomes, tissue targeting, cell-cell adhesion and binding of pathogens (Parekh *et al.*, 1987; Takeuchi *et al.*, 1989; Barton *et al.*, 1991; Cumming, 1991; Stanley, 1992; Qasba *et al.*, 1997; Almond *et al.*, 2004). Thus, protein glycosylation has been found to influence isoelectric properties as well as

protein three-dimensional structure and surface topology (Parekh, 1991). Optimization of protein glycosylation of pharmaceutical products has led to interesting developments in pharmaceutical research. Approaches to this problem have included investigation of cell types with differing glycosyltransferase activities, manipulation of environmental conditions to influence glycosylation characteristics and manipulation of the product polypeptide sequence (Goochee and Monica, 1990). An important consideration of this optimization process has persisted in that glycosylation site engineering continues to result in the production of a population of glycoforms. Many levels of glycosylation heterogeneity are associated with eukaryotic expression (Parekh, 1994).

5.1.2 Mechanisms behind *N*-linked glycan microheterogeneity classification

As glycosylation is the most complicated of all post-translational modifications, numerous glycosyltransferase enzymes, with high degrees of substrate specificity, in the endoplasmic reticulum (ER) and Golgi apparatus are responsible for determining the final structure of protein *N*-linked protein glycans. It is noted that cytosolic glycosylation process have been observed for nuclear proteins as well as for those proteins remaining in the cytosol (Shoup and Touster, 1976; Daniel *et al.*, 1994; Spiro, 2002), but these processes are not discussed further here. The focus of this research remains those proteins in the secretory pathway with *N*-linked glycosidic linkages, which is common for pharmaceutical proteins produced by mammalian cell systems. As a result of varying glycosyltransferase activity, the topography of *N*-linked glycans can take on complex-type, high mannose or hybrid conformations, and these conformations are dictated by the level of glycosyltransferase processing. Glycosylation events in the ER have been found

to occur cotranslationally (Robbins *et al.*, 1977; Tabas *et al.*, 1978; Chen and Lennarz, 1978; Attkinson and Lee, 1984). These processes, in particular, first include the initial attachment of a lipid-linked oligosaccharide, $Glc_3Man_9GlcNAc_2$, to a polypeptide sequence of N-*X*-S/T, where *X* is not proline, by the membrane-bound oligosaccharyl transferase enzyme in the rough ER (RER) (Marshall, 1972; Hubbard and Ivatt, 1981; Kornfeld and Kornfeld, 1985; Roth, 1987; Parekh, 1994). Factors affecting the robustness of this attachment reaction have been the focus of much previous research for pharmaceutical proteins (Goochee and Monica, 1990; Andersen et al., 2000; Senger and Karim, 2003a,b). This was also the topic for model development in Chapter 4. Following glycan attachment, the terminal glucose of the bulk oligosaccharide glycan is enzymatically removed by a membrane-bound specific α-1,2 glucosidase I enzyme. The following glucose residues are removed by a membrane-bound α-1,3 glucosidase II enzyme. The removal of at least one branched α-1,2-linked mannose residue has also been cited as occurring cotranslationally in the RER. It is noted that these enzymes, generally, do not contact many glycans exhibiting high mannose branching patterns. Major differences in glycan processing are observed upon transfer to the cis-Golgi cisternae after significant protein folding events have occurred of the primary sequence. Upon entering the Golgi apparatus, many high mannose residues do not experience further glycan trimming by glycosyltransferase enzymes in the cis and medial cisternae. Evidence of high mannose glycan interaction with mannosidase enzymes has been noted by the observation of high mannose glycans containing as little as five mannose residues. Many hypotheses have been proposed for this phenomenon. Most suggest that glycans retaining high mannose microheterogeneity classification reside in regions of low solvent

exposure following initial protein folding rearrangements, as glycan microheterogeneity determination does not occur cotranslationally (Hubbard and Ivatt, 1981; Kornfeld and Kornfeld, 1985; Roth, 1987; Parekh, 1994). Glycans destined for complex-type or hybrid microheterogeneity classifications interact with the Golgi α-1,2 mannosidase enzyme, which further cleaves all α-1,2 linked mannose residues to yield a $Man_5GlcNAc_2$ glycan structure. Thus, only high mannose glycan structures contain terminal α-1,2 linked mannose residues. An *N*-acetylglucosamine residue is added to a terminal α-1,3 linked terminal mannose residue by the *N*-acetylglucosaminyltransferase I enzyme in the medial-Golgi cisternae. At this point, glycan structures destined to become of complex-type microheterogeneity classification interact with the Golgi α-mannosidase II enzyme, and the removal of remaining terminal α-1,3 and α-1,6 linked mannose residues is catalyzed. Glycan structures resulting in hybrid glycosylation classification have been found to not interact with this enzyme, resulting in a branch consisting of terminal mannose residues that cannot be processed further by other glycosyltransferase enzymes in the medial and trans cisternae of the Golgi. However, following interaction with the Golgi α-mannosidase II, another *N*-acetylglucosamine residue is added to this site by the *N*-acetylglucosaminyltransferase II enzyme. Glycan structures experiencing this level of enzymatic processing are classified as complex-type and undergo further processing in the trans Golgi cisternae. Further processing includes the addition of terminal galactose, fucose and sialic acid residues. Covalent additions such as phosphorylation, acetylation, sulfation and sialylation may also occur at this point to glycans of all types of microheterogeneity classification. The comprehensive reviews by Hubbard and Ivatt

(1981), Kornfeld and Kornfeld (1985), Roth (1987) and Parekh (1994) were used to summarize the *N*-linked glycosylation pathway.

5.1.3 Factors influencing glycosylation microheterogeneity and control mechanisms

Control studies of glycan structures have revealed that the primary polypeptide sequence as well as the expression system are vital to glycan oligosaccharide structures. In particular, two modes of research have shown the influence of the primary sequence on glycosylation characteristics (Kornfeld and Kornfeld, 1985; Parekh, 1994; Roth *et al.*, 2002). First, different polypeptides, produced by the same expression system, have been shown to be glycosylated differently. Second, for a particular polypeptide sequence, a wide variety of glycan structures is not observed depending on cell-type expression. Rather, an expression system variation has been found to result in a family of closely related glycan structures (Anderson and Grimes, 1982; Sweidler *et al.*, 1985; Lee *et al.*, 1990; Wilhelm *et al.*, 1990; Davidson *et al.*, 1991). In addition, the conformation of the polypeptide backbone has been shown important to oligosaccharide processing of proteins exiting the ER (Hunt *et al.*, 1983; Carver and Cumming, 1987; Hubbard, 1988). However, simply because a certain expression system yields a complex-type microheterogeneity glycan classification for a particular glycosylated polypeptide does not imply that this same microheterogeneity classification should be observed in all expression systems. In fact, the degree of primary sequence influence, given different expression systems, is site-dependent. Thus, the polypeptide sequence, itself, is hypothesized to influence interaction with glycosyltransferase enzymes for

microheterogeneity classification determination through a combination of the primary sequence with secondary and tertiary protein structures (Parekh, 1994).

5.1.4 The role of neural networks in protein structure predictions

Neural networks have played an important role in the prediction of secondary structure elements of polypeptide structures since the development of the NNPREDICT (Kneller *et al.*, 1990) and PHD (Rost and Sander, 1993) algorithms. Since these developments, the prediction accuracy of neural network-based secondary structure algorithms has approached 78% by the SSpro and SSpro8 algorithms (Baldi *et al.*, 1999) of a theoretical limit of 88% (Rost, 2001; Rost and Eyrich, 2001). Despite the fact that secondary structure prediction has remained a bench-mark problem for the development of neural network-based protein structure predictions, the scope of this prediction methodology has expanded 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 (Pollastri *et al.*, 2001; Pollastri *et al.*, 2002; Pollastri and Baldi, 2002). The binding states of specific cysteine residues were predicted through a further development by Fariselli *et al.* (1999). The polypeptide sequence was coupled with structural characteristics of enzymes for a neural network-based identification of the catalytic residues in recent research by Gutteridge *et al.* (2003). In addition, the interface of protein-protein interactions (Ofran and Rost, 2003) as well as protein subcellular location (Cai *et al.*, 2002) have become predictable entities using neural network-based algorithms with primary sequence inputs. Advances have

also been noted in the area of post-translational modifications predictions by Blom *et al*, (2004) in prediction of phosphorylation sites. Progress has also been made in the use of neural networks for the prediction of *N*-linked glycosylation characteristics. The model developed in Chapter 4 identified glycosylation site-occupancy characteristics were predictable by primary sequence inputs. In addition, the neural network model was applied to identify the optimum window of residues for glycosylation site-occupancy prediction as well as the effects of neighboring charged residues on glycosylation site-occupancy characteristics.

5.1.5 Neural networks for glycosylation microheterogeneity classification

This research is closely related to the research reported in Chapter 4; however, the concepts explored in this previous research were further extended to predict glycosylation microheterogeneity classification. In general, microheterogeneity classification was divided into two categories: *complex-type* and *high mannose*. Hybrid glycan structures were not considered in this study, as hybrid glycan structures were not found to be the major glycoform component of any proteins comprising the reference data set. In addition, it is noted that glycosylation microheterogeneity is heterogeneous by nature for eukaryotic expression (Goochee and Monica, 1990; Parekh, 1994). Thus, the scope of this research remains to identify the classification of the major glycoform component of a possibly heterogeneous mixture of glycoforms. Due to the fact that microheterogeneity classification occurs from glycosyltransferase activity in the Gogi apparatus, another degree of complexity was added to this problem as opposed to the previous problem of site-occupancy prediction. In the presence of completely or partially-folded polypeptide

structures, glycosyltransferase activity was expected to be governed by not only the primary sequence but also by secondary structure elements as well as solvent accessibility. Therefore, these predictable elements were examined separately and in addition to primary sequence inputs for the prediction of microheterogeneity classification. Predicted secondary structure elements and solvent accessibility were examined as possible input quantities despite findings that glycan structures may influence polypeptide backbone conformations (Bosques *et al.*, 2004). Data complied for reference set construction consisted of proteins and glycosylation characteristics from CHO cell culture expression.

5.2 Systems and methods

5.2.1 Data acquisition and reference set construction

Data of glycosylation characteristics for particular *N*-linked glycosylated polypeptide sequences was obtained from a massive literature search. In particular, data was collected for CHO expression of proteins in which comprehensive glycosylation analysis had been performed. As stated previously, eukaryotic protein expression of glycosylated proteins commonly results in a heterogeneous mixture of glycoforms. In this case, only the major glycoform was considered. Of course, care was taken in selecting proteins for the reference set in which the major glycoform of a reported heterogeneous mixture comprised a rather large fraction of the mixture. As much heterogeneity is commonly observed among complex-type and high mannose glycans, these glycoforms were clustered into two categories for the purpose of this research. In addition, hybrid glycosylation microheterogeneity classification was not considered in

this research as hybrid glycan structures failed to comprise a major glycoform fraction in any of the case studies reviewed. Thus, predictions were performed to classify the major glycoform from CHO culture expression as *complex-type* or of *high mannose* glycosylation microheterogeneity classification. In all, 120 protein sequences and glycosylation characteristics made up the entire reference set, which is included in Appendix E. The primary sequence, predicted secondary structures and predicted solvent accessibility were mapped to glycosylation microheterogeneity classification by the use of recurrent neural networks. Data sets were constructed for neural network training procedures and consisted of data for 110 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 of the complete reference data set. 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 readily available on the world-wide-web: Jpred (http://www.compbio.dundee.ac.uk/~www-jpred/) (Cuff and Barton, 1999; Cuff and Barton, 2000); NNPREDICT (http://www.cmpharm.ucsf.edu/%7Enomi/nnpredict.html) (Kneller *et al.*, 1990);

PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/) (Jones, 1999); PROFsec (http://cubic.ioc.columbia.edu/predictprotein) (Rost and Sander, 1993; Rost, 2001); SSpro and SSpro8 (http://www.igb.uci.edu/tools/scratch) (Baldi *et al.*, 1999; Pollastri *et al.*, 2002b). Multiple predictions of secondary structure were quantified and averaged. Predictions of solvent accessibility were obtained, using a 25% threshold, from Jnet (http://www.compbio.dundee.ac.uk/~www-jpred/jnet/) (Cuff and Barton, 1999) and from the SCRATCH server by the ACCpro algorithm (http://www.igb.uci.edu/tools/scratch) (Pollastri *et al.*, 2002a).

5.2.2 Quantification of sequences, structures and glycosylation characteristics

Of course, in order to perform mapping operations using neural networks, the conversion of the primary sequence, predicted secondary structure, predicted solvent accessibility and glycosylation characteristics to numerical values was necessary. The primary structure was quantified by amino acid clusters, as described in Chapter 4. Residues predicted to be incorporated in helical secondary structures were quantified as *1*, and residues predicted in extended structures, such as β-sheets, were quantified as *2*. All other secondary structure elements and unordered regions were assigned the value *3*. Residues predicted to be exposed to solvent at a 25% threshold were assigned the value *0*, and predicted buried residues were quantified as *1*. Finally, high mannose glycosylation microheterogeneity classification was labeled *0.25*, and complex-type microheterogeneity was given the value *0.75*. Residues not displaying glycosylation were quantified as *0*.

5.2.3 Neural network architecture and glycosylation window optimization

Elman recurrent neural networks with a single hidden layer were used for construction of all neural network models. Neural networks were initiated with random weight and bias values prior to training procedures. 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 overparameterization of the system but ensured optimum neural network performance, as verified in Chapter 4 for these systems. The optimum glycosylation window was also defined in Chapter 4 as the number of residues, initiating on the *N*-terminal side of the glycosylation site (*n*) and extending a specified number of residues on the *C*-terminal side of the glycosylation site. Various lengths of the glycosylation window were evaluated on the effectiveness of neural network training. For each glycosylation window data point, and for each cross-validation iteration, 100 independent neural networks were initiated with random weight and bias values and trained for a total of 2000 epochs apiece. Data points reported for glycosylation windows represent an average of all testing data set predictions. Neural network training effectiveness was monitored by calculation of the mean-square error (MSE) for prediction of the testing data set. The glycosylation windows for primary sequence, predicted secondary structure and predicted solvent accessibility neural network inputs were initially evaluated independently.

5.2.4 Predictive model construction

Following the establishment of optimized glycosylation windows for primary sequence, predicted secondary structure and predicted solvent accessibility inputs, these independent inputs were grouped in an effort to improve the overall prediction of the testing data sets. For example, the input vector space was expanded to include primary sequence, predicted secondary structure and predicted solvent accessibility data, all with optimized glycosylation windows, for glycosylation microheterogeneity prediction. Multiple arrangements of inputs were examined to investigate input relevance in microheterogeneity prediction, 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. These networks were used to make further predictions of mutated sequences found in the literature. The use of multiple networks for this purpose allowed for the calculation of a mean network prediction with a confidence level based on the fraction of neural networks returning the dominant output value. All neural networks comprising the final predictive model contained different values of weights and biases.

5.2.5 Further simulations

The effectiveness of the model was demonstrated by simulation of mutants of the recombinant tissue-type plasminogen activator (r-tPA) protein in research performed by Wilhelm *et al.* (1990). The construction of r-tPA variants through deletion mutations and domain insertions was found to alter the glycosylation of r-tPA at N117 from high

mannose microheterogeneity to complex-type in some cases. Prediction of these observations was found to be a suitable test for the constructed glycosylation microheterogeneity classification model.

5.3 Results and discussion

5.3.1 Probing the suitability of model inputs: primary sequence

Prior to neural network training, a statistical analysis was performed to examine the input variables of primary structure, predicted secondary structure and predicted solvent accessibility, over a wide range of the glycosylation window, for glycosylation predictions. The total reference set was divided, based on glycosylation microheterogeneity classification, and plots were generated of input occurrences at 20 residues on either side of the site of glycosylation for complex-type and high mannose glycosylation microheterogeneity. These figures are displayed for the primary sequence in Figure 5.1 and Figure 5.2 for complex-type and high mannose microheterogeneity, respectively. Along the ordinate axis, the site of glycosylation is represented at 0. Positive values represent residues in the direction of the *C*-terminus, and negative values represent residues in the *N*-terminal direction. An observation was made at the site (*n*+2) of the glycosylation sequence. By definition, this site of the N-*X*-S/T glycosylation sequence is always occupied by a serine or threonine residue. Analysis of the reference set yielded a ratio of serine to threonine at the (*n*+2) position of 0.77 for sequences of complex-type glycosylation microheterogeneity classification. Furthermore, a serine to threonine residue ratio of 0.39 was calculated for sequences of high mannose classification. This suggests that threonine in the (*n*+2) position may favor complex-type

glycosylation. In addition, it was found previously in Chapter 4 that threonine in this position favored robust site-occupancy glycosylation. Other observations of primary sequence differences revealed relatively high asparagine occurrences at the (n+6) position of high mannose glycosylated sequences. In addition, a large occurrence of unsubstituted, hydrophobic residues were observed on the immediate *N*-terminal side of the glycosylation site for sequences of complex-type glycosylation. Between positions (n+8) and (n+14) a higher occurrence of hydrophilic, including charged, residues were observed for sequences with high mannose glycosylation. In this region, a majority of residues of complex-type glycosylated sequences were observed to be hydrophobic and unsubstituted.

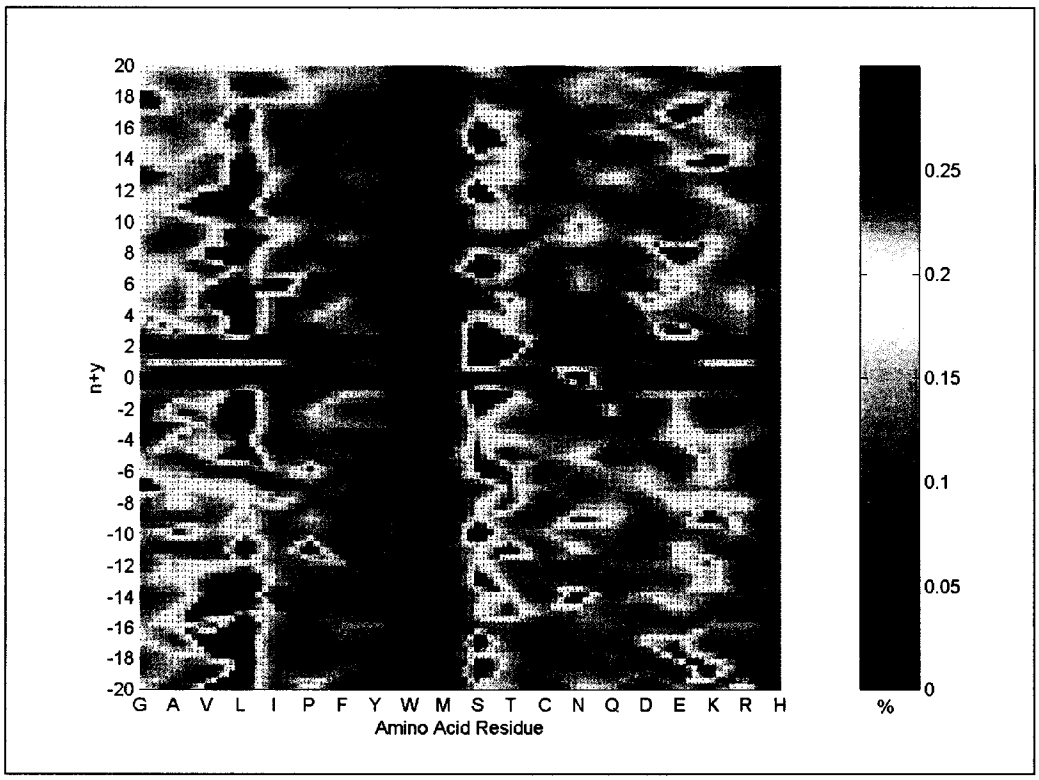

Figure 5.1. Statistical analysis of the primary sequence of polypeptide sequences of the reference set resulting in predominantly complex-type microheterogeneity classification.

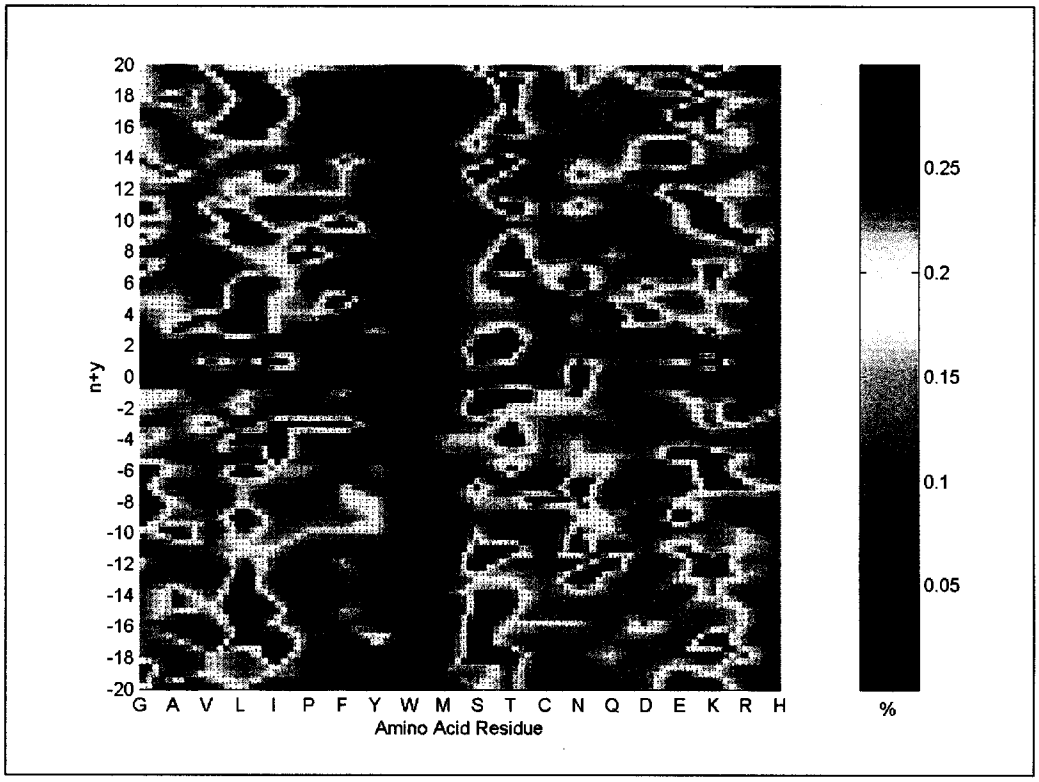

Figure 5.2. Statistical analysis of the primary sequence of polypeptide sequences of the reference set resulting in predominantly high-mannose microheterogeneity classification.

5.3.2 Predicted secondary structure and predicted solvent accessibility

The relative occurrence of predicted secondary structure elements are shown for sequences of complex-type glycosylation microheterogeneity and high mannose classification in Figure 5.3 and Figure 5.4, respectively. The most significant, and obvious, observation is the much larger relative occurrence of predicted helical structures surrounding the site of glycosylation for sequences of high mannose classification. Both classifications displayed a relatively low occurrence of extended, or β-sheet, structures. In addition, in both cases, the site of glycosylation was observed in regions void of defined predicted secondary structure elements. Predicted residue solvent accessibility,

using a 25% threshold, is shown for complex-type and high mannose glycosylation microheterogeneity in Figure 5.5 and Figure 5.6, respectively. While several notable differences are apparent in the two figures, much more order was observed for sequences of high mannose glycosylation. In addition, high occurrences of buried regions were observed on either side of the site of glycosylation. These observations suggest that a high incidence of buried regions may orient the site of glycosylation in a manner that is inhibitory to interaction with glycosyltransferase enzymes in the Golgi apparatus.

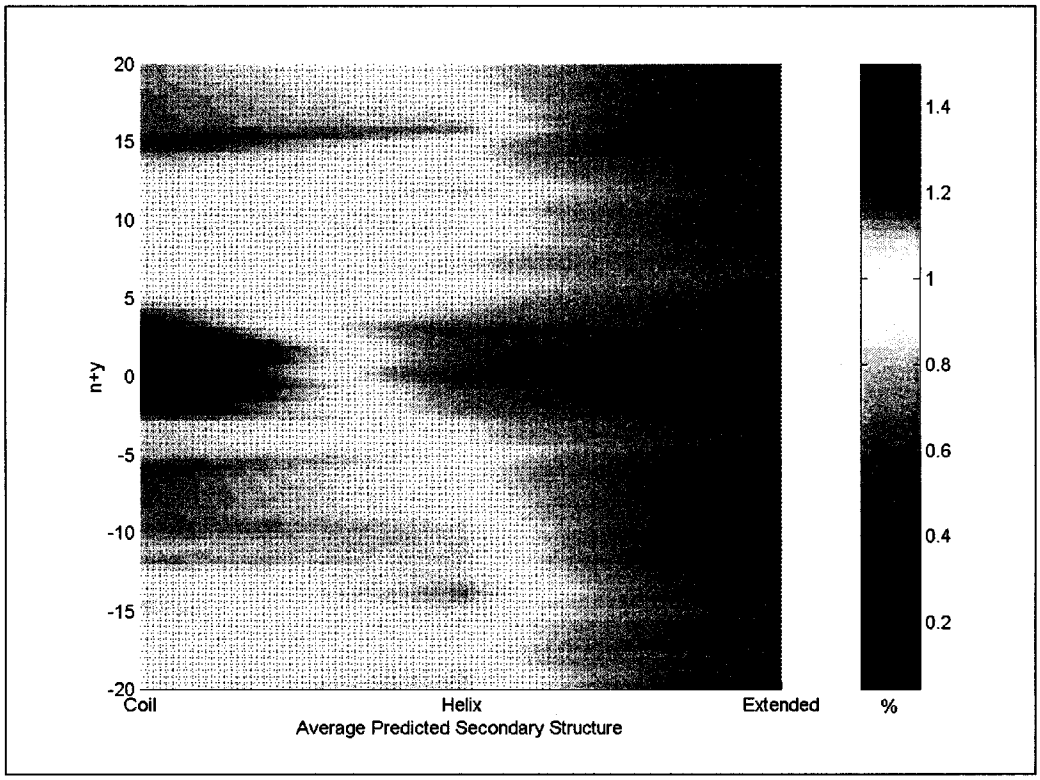

Figure 5.3. Statistical analysis of the average secondary structure prediction of polypeptide sequences of the reference set resulting in predominantly complex-type microheterogeneity classification.

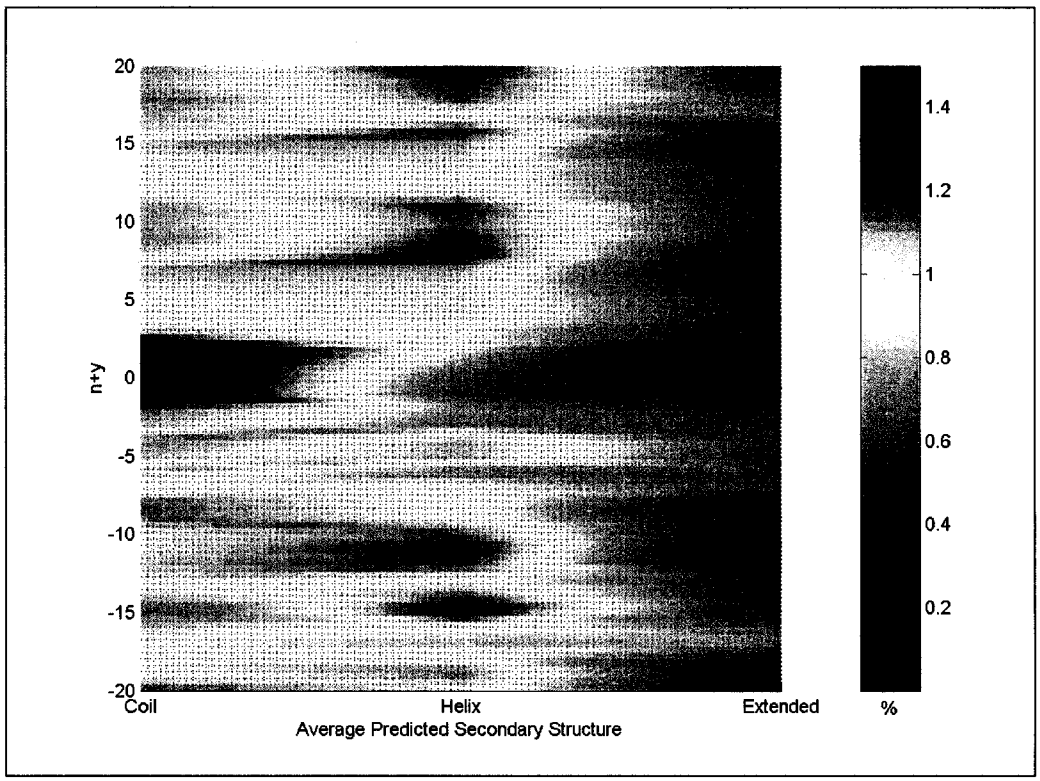

Figure 5.4. Statistical analysis of the average secondary structure prediction of polypeptide sequences of the reference set resulting in predominantly high mannose microheterogeneity classification.

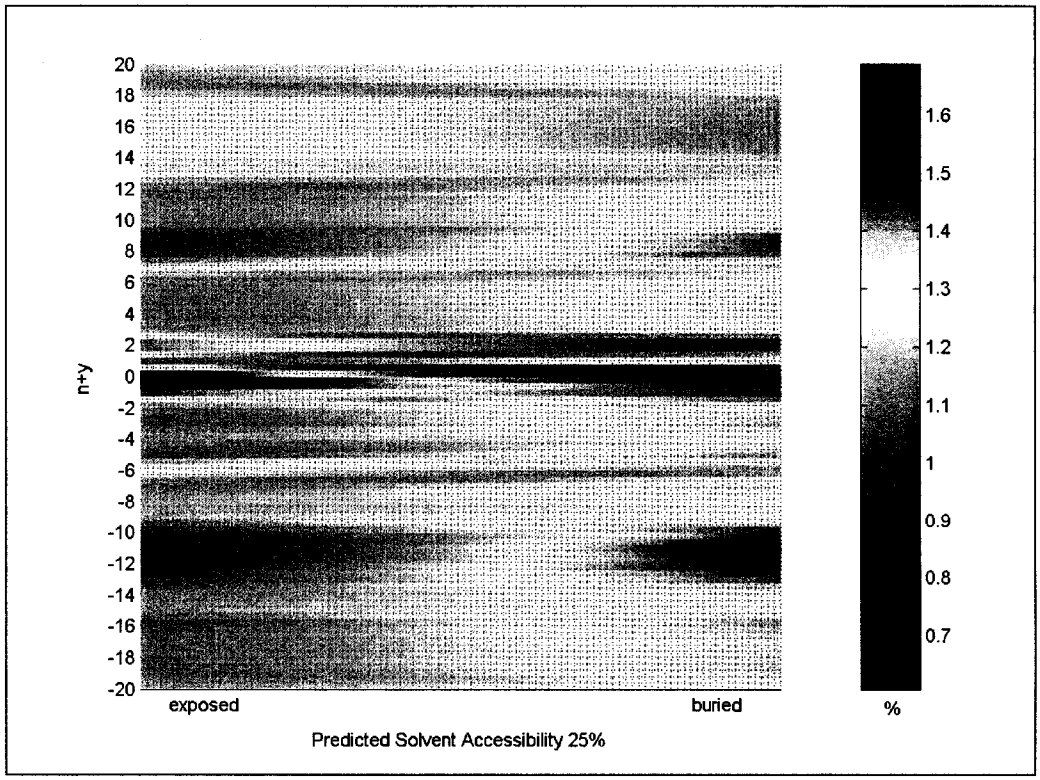

Figure 5.5. Statistical analysis of the predicted solvent accessibility (25% threshold) of polypeptide sequences of the reference set resulting in predominantly complex-type microheterogeneity classification. The site of glycosylation occurs at zero. Positive values along the ordinate axis correspond to the *C*-terminal direction along the polypeptide chain.

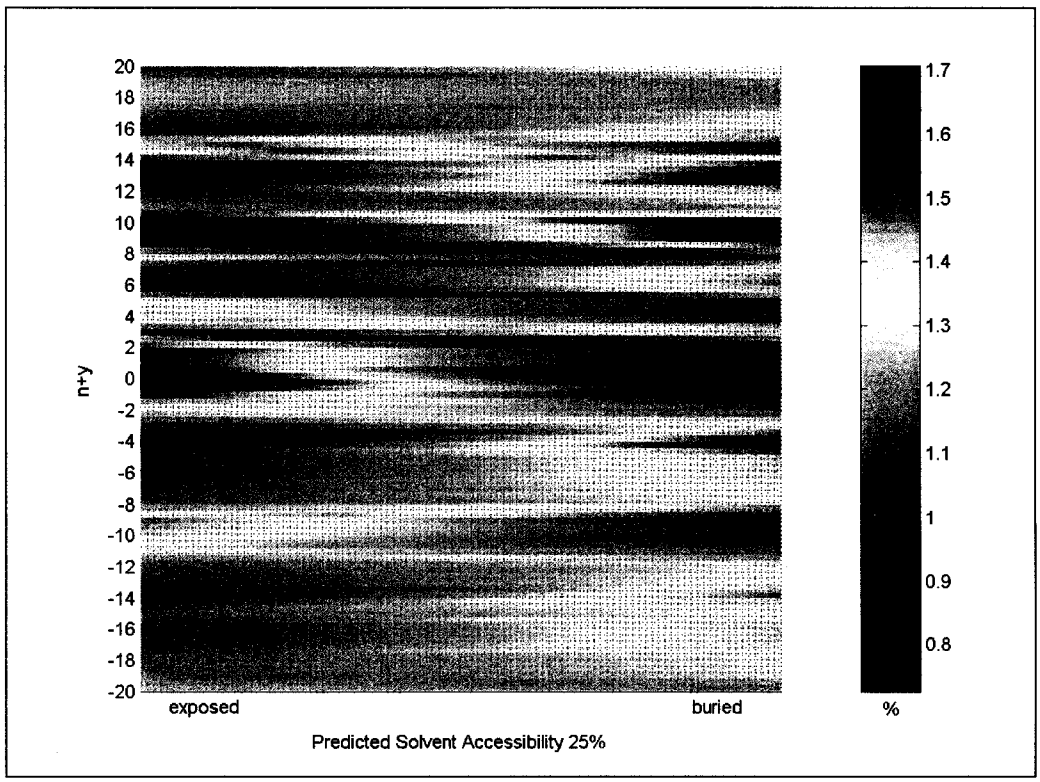

Figure 5.6. Statistical analysis of the predicted solvent accessibility (25% threshold) of polypeptide sequences of the reference set resulting in predominantly complex-type microheterogeneity classification. The site of glycosylation occurs at zero. Positive values along the ordinate axis correspond to the *C*-terminal direction along the polypeptide chain.

5.3.3 Glycosylation window optimization and reference set cross-validation

Since significant differences between complex-type and high mannose classified input sequences were observed for primary structure, predicted secondary structure and predicted solvent accessibility, all three cases were evaluated in neural network training. Thus, separate input vectors of primary sequence, predicted secondary structure and predicted solvent accessibility data were independently mapped to corresponding glycosylation characteristics using recurrent neural networks. In all three cases, the

objective function was monitored as the number of residues around the glycosylation site (n) was varied. The specified residues leading to the best overall prediction of the testing data sets in cross validation experiments were defined as the optimized glycosylation window for all three cases. Results of glycosylation microheterogeneity classification prediction using primary sequence data are shown in Figure 5.7. The residue corresponding to the start of the glycosylation window is displayed as (n-x), where n corresponds to the site of glycosylation. Thus, along the abscissa axis, a positive value corresponds to the number of residues in the *N*-terminal direction relative to the glycosylation site. On the other hand, the ordinate axis (n+y) defines the ending residue of the glycosylation window. Positive values in this case correspond to the number of residues, relative to the site of glycosylation, in the *C*-terminal direction. For example, the data point (n-20) (n+20) defines a glycosylation window of 41 residues, including the glycosylation site. Thus, it is noted that a portion of the figures of this type have been omitted, as the image forms a mirror image. Two 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. One such region was defined by glycosylation windows starting between 9 and 16 residues to the *N*-terminal side of the glycosylation site and extending to between 16 and 20 residues past the glycosylation site, on the *C*-terminal side. Another significant glycosylation window was observed as starting 4 residues away from the glycosylation site, on the *C*-terminal side, and extending another two residues toward the *C*-terminus. This optimum solution was believed to be an artifact of the data set due to the relatively small size and position of the resulting glycosylation window. Results corresponding to predicted secondary structure inputs are

shown in Figure 5.8. It is noted that many optimum results were observed for glycosylation windows encompassing a large number of residues on the *C*-terminal side of the glycosylation site. Glycosylation windows not containing a significant number of residues on the *C*-terminal side of the glycosylation site resulted in poor prediction of the testing data sets. Optimum results were observed for glycosylation windows ending between 16 and 20 residues to the *C*-terminal side of the glycosylation site. However, the starting residue of the glycosylation window was found to vary between 10 residues on the *N*-terminal side to 10 residues on the *C*-terminal side of the glycosylation site to enable optimum predictions of the testing data sets in cross validation experiments. Finally, results with predicted solvent accessibility inputs are shown in Figure 5.9. Several regions of glycosylation windows appeared optimized for the prediction of the testing set data. A large optimized region was observed between starting residues of 10 and 20 residues on the *C*-terminal side of the glycosylation site and extending to ending sites between 6 and 16 residues on the *C*-terminal side of the glycosylation site. An optimum site was also observed exclusively on the *C*-terminal side of the glycosylation site as starting 7 residues beyond the site and extending between 3 and 9 residues. However, this region was also suspected as an artifact of the data sets. Ideally, use of a glycosylation window starting on the *N*-terminal side of the glycosylation site and extending to residues on the *C*-terminal side was hypothesized as a more ideal glycosylation window. In all cases, error was calculated as approximately ±5%.

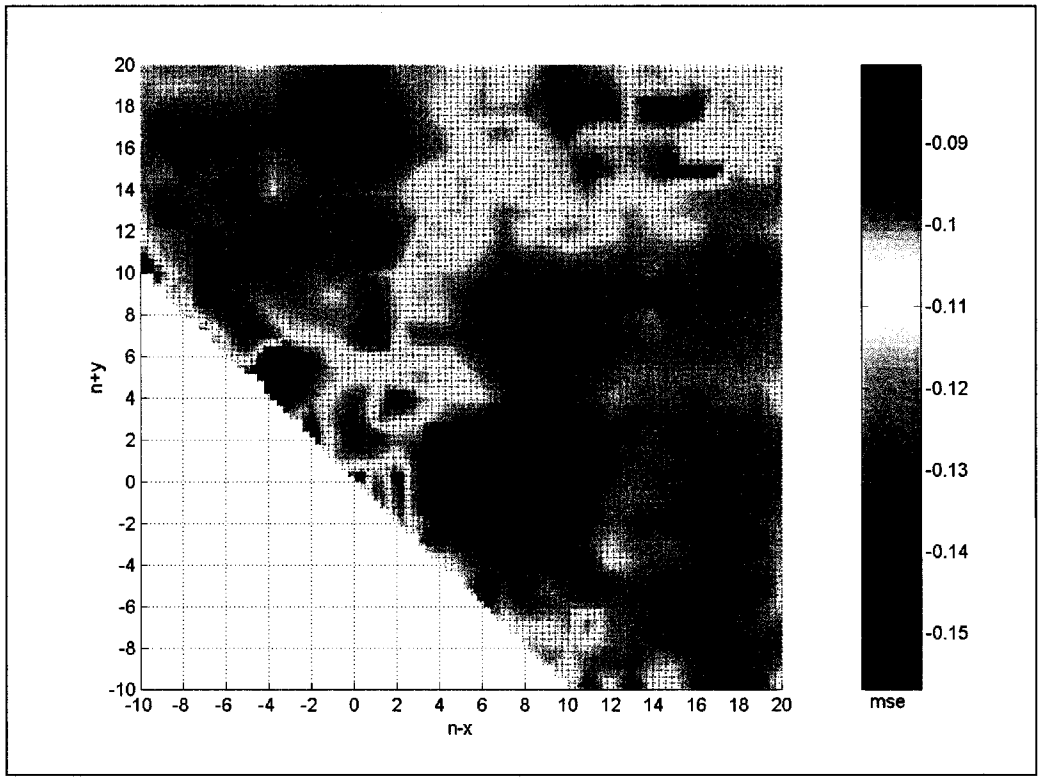

Figure 5.7. Cross-validated neural network predictions of the testing data set for various input window lengths of primary sequence data.

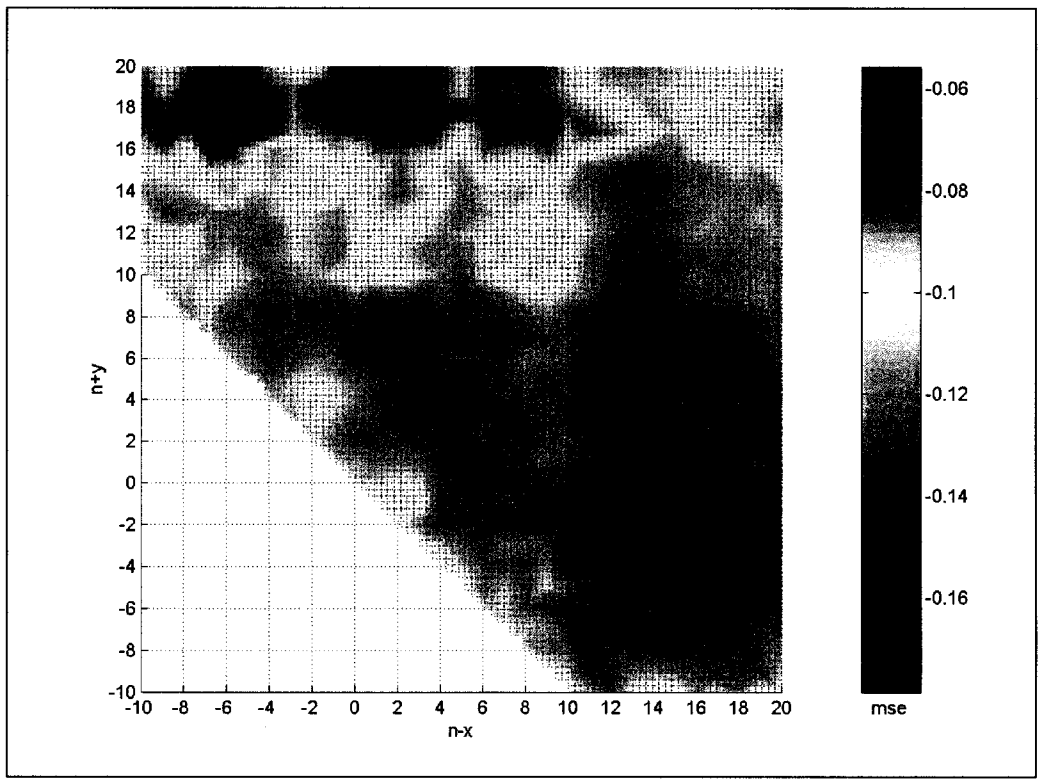

Figure 5.8. Cross-validated neural network predictions of the testing data set for various input window lengths of predicted secondary structure data.

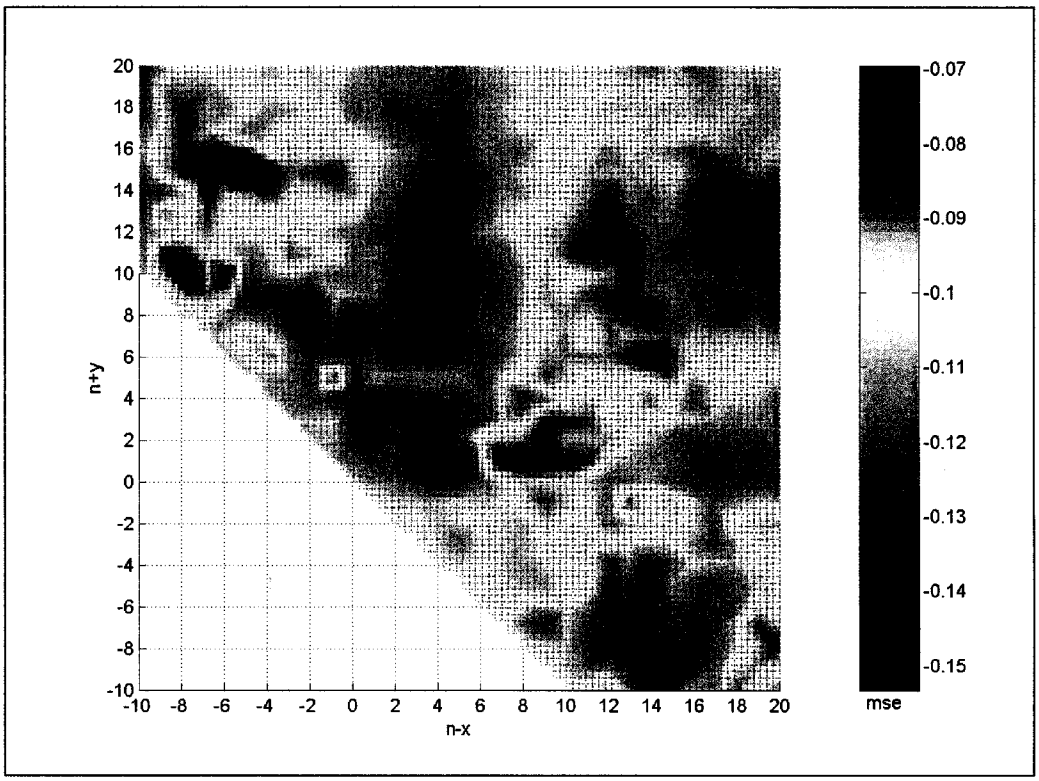

Figure 5.9. Cross-validated neural network predictions of the testing data set for various input window lengths of predicted solvent exposure (25% threshold) data.

5.3.4 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. Two optimized glycosylation windows were chosen for each input vector. One input vector was chosen that spanned the *N*-terminal and *C*-terminal sides of the glycosylation site, and one input vector was chosen from optimized windows solely on the *C*-terminal 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 predictions for each case are shown in Table 5.1. Input vector combinations resulted in decreased prediction accuracy in all cases except for one. This case used input vectors spanning residues of the *N*-terminal side and *C*-terminal side of the glycosylation site with predicted secondary structure and predicted solvent accessibility data. Consequently, this case also dramatically improved prediction accuracy. Thus, the optimized input vector for prediction of glycosylation microheterogeneity classification contained predicted secondary structure data with a glycosylation window starting 9 residues from the glycosylation site on the *N*-terminal side and predicted solvent accessibility data with a glycosylation window starting 16 residues away from the glycosylation site, on the *N*-terminal side. The predicted secondary structure glycosylation window ended 17 residues from the glycosylation site on the *C*-terminal side, and the predicted solvent exposure glycosylation window ended 13 residues from the glycosylation site on the *C*-terminal side.

Primary Sequence Starting Residue (*n*-x)	Primary Sequence Terminating Residue (*n*+y)	Predicted Secondary Structure Starting Residue (*n*-x)	Predicted Secondary Structure Terminating Residue (*n*+y)	Predicted Solvent Exposure Starting Residue (*n*-x)	Predicted Solvent Exposure Terminating Residue (*n*+y)	Resulting average MSE of Testing Data Sets
-4	6	-6	17	-5	15	0.0845
-4	6	-6	17	None	None	0.0860
-4	6	None	None	-5	15	0.1246
None	None	-6	17	-5	15	0.0850
10	18	9	17	16	13	0.0955
10	18	9	17	None	None	0.0800
10	18	None	None	16	13	0.0845
None	None	9	17	16	13	0.0465

Table 5.1. Results of combining primary sequence, predicted secondary structure and predicted solvent exposure (25% threshold) for glycosylation microheterogeneity classification. The glycosylation window for each input is given. A negative value for the starting residue corresponded to a glycosylation window starting on the *C*-terminal side of the glycosylation site. All terminating residues of the glycosylation window resided on the *C*-terminal side of the glycosylation site.

5.3.5 The microheterogeneity classification prediction model

A total of 30 independent recurrent neural networks were combined to make-up the overall predictive model. The network components of the model were selected from all cross-validation experiments. Networks selected for the final predictive model were found to predict the specific testing data set to a minimum MSE value. 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%. 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 classified based on a simple perceptron

classification scheme. In general, average recurrent neural network results less than 0.5 were classified as 0.25, which corresponds to high mannose glycosylation. On the other hand, values greater than, or equal to 0.5, were classified as 0.75, which corresponded to complex-type glycosylation. Consequently, the use of multiple neural networks enabled the calculation of the confidence level of the prediction. The confidence level was defined as the fraction of individual neural networks returning the overall classified result. Thus, the confidence interval maintained a range between 0.5 and 1. A confidence level value close to 0.5 represents a split two-dimensional classification decision by the predictive model. The following protein sequences with specified glycosylation sites, in Table 5.2, were classified by an abbreviated version of the predictive model. Neural networks in which a particular component had been used in training were removed from the predictive model for the following predictions. In all cases, the specific glycosylation site was correctly classified with a maximum confidence level value of 1.

Input Sequence	Average Recurrent Network Result	Overall Model Classification	Confidence Level	Published Classification	Reference
Tissue plasminogen activator (184)	0.8845	0.75	1	0.75	Spellman *et al.*, 1989
Tissue plasminogen activator (448)	0.9502	0.75	1	0.75	Spellman *et al.*, 1989
Thrombopoietin (197)	1.0000	0.75	1	0.75	Hoffman *et al.*, 1996; Inoue *et al.*, 1999
DNase I (40)	0.0279	0.25	1	0.25	Cacia *et al.*, 1998
β-hexosaminidase B (327)	0.4362	0.25	1	0.25	Schuette *et al.*, 2001
β-hexosaminidase B (190)	0.0107	0.25	1	0.25	Schuette *et al.*, 2001

Table 5.2. Examples of recurrent neural network output values and model classification. Examples are given for proteins not appearing in the neural network training data set. Input values of predicted secondary structure and predicted solvent accessibility were used with optimized glycosylation windows. Classification of *0.25* corresponds to high mannose and *0.75* corresponds to complex-type glycosylation microheterogeneity classification.

5.3.6 Further model simulations

Due to the high success rate of model predictions of testing data set components, the predictive model was extended to a set of mutation studies performed by Wilhelm *et al.* (1990) with the tissue-type plasminogen activator (tPA) protein. Results are displayed in Table 5.3. In particular, glycosylation microheterogeneity at N117 was studied in this research with respect to specified domain addition and sequence deletion mutations. In general, the wild-type tPA protein contains a large fraction of high mannose glycan structures at N117 when produced by CHO cell cultures (Spellman *et al.*, 1989). In

addition to evaluation of the wild-type protein, glycosylation microheterogeneity classification was evaluated at N117 for the following deletion mutants by Wilhelm *et al.* (1990): 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 glycosylation microheterogeneity at N117. Glycosylation of tPA at N117 occurs in the kringle I domain of the protein. The kringle I domain is preceded, in the *N*-terminal direction, by a growth factor region and a finger domain (Grossbard, 1987; Bennett *et al.*, 1991). Four hybrid structures of tPA were prepared in the research by Wilhelm *et al.* (1990) 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 wild-type 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 of wild-type tPA. Glycosylation in the extra kringle II domain was blocked by a point mutation in this case. Hybrids A, B and D were found to have complex-type microheterogeneity by Wilhelm *et al.* (1990) through experimental methods, and hybrid C was found to conserve high mannose glycosylation microheterogeneity. 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*-

terminal direction and the *C*-terminal direction in the case of hybrid C. Thus, primary sequence data of the mutant sequences was conserved in all cases in the glycosylation window. However, this was found to not be the case for predicted secondary structure and predicted solvent exposure data as predicted values of the glycosylation window were found affected by mutations far up- and down-stream of the glycosylation site. All hybrid sequences and the Δ44-48 deletion mutant were classified correctly by the predictive model with a relatively high confidence level in most cases. The deletion mutants Δ2-89 and Δ55-62 were classified incorrectly by the model. However, a somewhat low confidence level 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. This case study further 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. Should it be necessary to experimentally determine these quantities, the usefulness of this model is nullified since these properties are much more difficult to determine by experimental methods than glycosylation microheterogeneity classification.

Input Sequence	Average Recurrent Network Result	Overall Model Classification	Confidence Level	Published Classification	In Agreement?
Tissue plasminogen activator (N117)	0.3245	0.25	0.87	0.25	Yes
tPA (N117) Δ2-89	0.4123	0.25	0.63	0.75	No
tPA (N117) Δ44-48	0.6175	0.75	0.80	0.75	Yes
tPA (N117) Δ55-62	0.3855	0.25	0.63	0.75	No
tPA (N117) Hybrid A	0.7915	0.75	0.93	0.75	Yes
tPA (N117) Hybrid B	0.7915	0.75	0.93	0.75	Yes
tPA (N117) Hybrid C	0.4080	0.25	0.63	0.25	Yes
tPA (N117) Hybrid D	0.6175	0.75	0.80	0.75	Yes

Table 5.3. Model predictions of glycosylation microheterogeneity classification for tissue plasminogen activator (tPA) deletion and insertion mutations by Wilhelm *et al.* (1990). Classification of *0.25* corresponds to high mannose and *0.75* corresponds to complex-type microheterogeneity.

5.3.7 Model limitations

As demonstrated in the previous example, model limitations exist for the prediction of glycosylation microheterogeneity classification, but these limitations may be revealed in certain cases by examination of the confidence level. From a theoretical standpoint, other limitations are to be 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. In addition, computational predictions currently cannot account for a situation in which a glycosylation site is shielded by a neighboring domain. Until these effects can be effectively modeled, these limitations will exist within the glycosylation microheterogeneity prediction model, and special care should be taken in these situations.

5.4 Conclusions

N-linked glycosylation microheterogeneity of a protein from CHO culture expression 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 of CHO-derived proteins with *N*-linked glycosylation. Multiple predictions of secondary structure elements and residue solvent accessibility were found to best predict glycosylation microheterogeneity 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. Prediction accuracy was found to be better than 90% based on neural network testing data set prediction. 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 mutant case study. The effective elimination of the primary sequence data from the model input space further reinforces the notion that glycosylation microheterogeneity 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 protein itself. The developed model has immediate implications in pharmaceutical research, as it is CHO cell-based. The intended implications of this model are for glycosylation prediction of theoretical mutations of native protein sequences. Theoretical predictions are commonly used to refine experimental design in combinatorial-type research. In addition, this model has usefulness in predicting glycosylation microheterogeneity of

newly discovered proteins or of those produced in titers insufficient for experimental determination of glycosylation characteristics. To our knowledge, this is the first computational model for the prediction of glycosylation microheterogeneity characteristics that is based on a set expression system and allows for the evaluation of theoretical mutant polypeptide sequences.

Chapter 6

CONCLUDING REMARKS AND RECOMMENDATIONS

6.1 Summary of contributions and relative significance

The focus of this research remained in the development of glycosylation-based mathematical models. However, these types of models were explored for various applications. For instance, not only are glycosylation-dependent models useful for describing rates of recombinant glycoform production and total product activity, but these models were shown vital for describing the inactivation of a heterogeneously glycosylated recombinant product. Identification of glycosylation-dependent and free glucose-dependent inactivation mechanisms of the recombinant tissue-type plasminogen activator (r-tPA) was a novel development of this research. A fed batch optimization application was performed in this research based on an objective function defined by these glycosylation-based models. Of course, the objective function may be re-defined in a number of applications to suit the needs of a particular situation. For instance, minimization of a particular glycoform concentration during recombinant product production may show significant benefit in specified cases as certain glycosylation patterns of particular proteins *in vivo* have been linked to disease. In this case, fed batch feeding profiles resulting in an up-regulation of inactivation, coupled with down-regulated production of undesired recombinant product glycoforms, could prove beneficial. All of these types of applications are possible using the glycosylation-

dependent kinetic model methodology developed in this research. In addition, data-based glycosylation-prediction models were developed using artificial neural-networks in this research as well. As the problem of recombinant product glycoform heterogeneity may be addressed by fed batch optimization, it may also be addressed through site-directed mutations that alter glycosylation characteristics. In this case, the use of predictive models aids in experimental design in mutant protein production and evaluation. In particular, intelligent design, or model-based design of experiments greatly reduces the number of mutants that must be produced, screened and analyzed in a laboratory. This type of experimental design aids in minimizing laboratory costs as well as the amount of time taken to find a suitable recombinant mutant product. Models were created in this research to predict glycosylation characteristics of site-occupancy and microheterogeneity classification.

6.2 Identification of inactivation mechanisms and incorporation into product models

Inactivation mechanisms were identified and effectively modeled for the two major glycoforms of the r-tPA protein that differ in site-occupancy at N184. In particular, a glycosylation-dependent inactivation mechanism was identified through protein-protein interactions. This mechanism was postulated to be due to either autolysis of the serine protease or inactivating protein aggregation. This inactivation mechanism was modeled as a second-order reaction. From a quantitative standpoint, the calculated second-order rate constant for the partially-glycosylated Type II r-tPA protein glycoform was over 190% greater than that for the fully-glycosylated Type I r-tPA glycoform. This suggested an additional role of the glycan at N184 of the Type I r-tPA glycoform in

contribution to the overall stability of the active form of the r-tPA enzyme. In addition, a glucose-dependent inactivation mechanism was identified for the r-tPA protein and was found to be independent of the presence of glycosylation at N184 of r-tPA. Further computational analysis identified four specific lysine residues in the protease domain as possible reaction sites of the inactivating glycation reaction. Based on these findings, separate inactivation models were written for the Type I and Type II r-tPA glycoforms, each containing two modes of inactivation. These models effectively modeled the inactivation of Type I and Type II r-tPA glycoforms when separately incubated with various amounts of free glucose.

6.3 Optimization of a fed batch process using glycosylation-dependent product models with inactivation mechanisms

Rate constants for the inactivation mechanism of Type I and Type II r-tPA glycoforms were obtained for incubation in changing CD-CHO cell culture media compositions. An overall process model was constructed using yield coefficients and model parameters obtained from batch CHO cell cultivations with r-tPA production. Inactivation mechanisms were added to the Type I and Type II r-tPA glycoform product models, and a CHO cell intrinsic death rate model was derived from simple enzyme kinetics correlations. Optimization was performed with respect to maximizing an objective function defined as total r-tPA activity. The design vector consisted of glucose, glutamine and asparagine mass feed flow rates. A design vector was also investigated for metabolite concentration set points in the presence of control algorithms and variable feed flow rates. For fixed feed flow rates, the initiation of feed was investigated for two

cases: as the first derivative of the objective function approached zero and as the second derivative of the objective function approached zero. Dynamic programming methods, using fourth-order Runge-Kutta solutions of the process model, were used to solve the optimization problem. Using metabolite control and variable feed flow rates, a maximum of 5.64 $x10^6$ [IU] was obtained by simulations for the maximum total r-tPA activity when glucose and amino acids set points of 1.51 g L^{-1} and 1.18 g L^{-1} were used, respectively. Using fixed feed flow rates, and a first derivative initiation marker, a maximum that approached 6.0 $x10^6$ [IU] was obtained for mass feed flow rate ratios of glucose to amino acids 3.15. Only slightly better values were obtained for flow initiation as the second derivative of the objective function approached zero. In all, optimized fed batch simulations increased the productivity over batch production by approximately 75%.

6.4 Variable site-occupancy predictions

Data for the training of neural network-based models was obtained from published literature. Glycosylation sites were classified as *variable* and *robust* for neural network training and predictions. In general, variable site-occupancy was defined for glycosylation sites exhibiting both the presence and absence of a glycan moiety during production by CHO cell cultures. This type of glycosylation characteristic commonly results in heterogeneously glycosylated proteins in CHO cultivations. The N184 glycosylation site of r-tPA is an example of variable site-occupancy, as Type I and Type II r-tPA glycoforms were produced during cultivation. In addition, robust glycosylation was defined as either the homogeneous presence or absence of a glycan moiety at a particular glycosylation site during CHO cell production. Neural networks were used to

map the primary sequence of specified residues, surrounding the glycosylation site, to the two-dimensional glycosylation classification. Using a cross-validation scheme, a glycosylation window of five residues on the *N*-terminal side of the glycosylation site to four residues on the *C*-terminal side resulted in the best prediction of designated neural network testing data sets. In further model testing, experimentally validated glycosylation characteristics from site-directed mutations of the rabies virus glycoprotein (Kasturi *et al.*, 1997; Mellquist *et al.*, 1998) were correctly predicted in 95% of the testing cases. Further simulations explored the effect of charged residues on these glycosylation characteristics. Simulations suggested less robust glycosylation in the presence of positively charged residues around the glycosylation site. These simulations also effectively demonstrated the influence of other residues throughout the entire glycosylation window.

6.5 Microheterogeneity classification predictions

A data-based model was constructed, consisting of 30 independent recurrent neural networks, for the prediction of the dominant fraction of glycosylation microheterogeneity classification of a heterogeneously produced recombinant glycoprotein. The glycosylation microheterogeneity classifications predicted were *complex-type* and *high mannose*. Although hybrid structures are recognized as a common type of microheterogeneity classification, none of the 120 glycosylation sites composing the reference set displayed hybrid glycan structures as the dominant glycoform of a heterogeneous mixture when produced by CHO cell culture. Prediction accuracy for the microheterogeneity classification model was found to be better than 90%

when based on predictions of neural network testing data sets during a cross-validation analysis. In addition, inputs of the primary sequence, predicted secondary structure and predicted solvent accessibility were used to optimize the glycosylation window in each case. Combinations of input vectors, with optimized glycosylation windows, were investigated. Based on these results, the primary sequence was effectively eliminated as an input vector component from the predictive model. Thus, as site-occupancy was found to be directly related to the polypeptide sequence, microheterogeneity classification was found dependent upon secondary and tertiary properties of the polypeptide sequence. These results are consistent with previous findings as glycan attachment occurs co-translationally in the endoplasmic reticulum (ER) and most enzymatic reactions corresponding to microheterogeneity determination occur in the Golgi apparatus on a partially or fully-folded polypeptide structure (Kornfeld and Kornfeld, 1985; Roth, 1987; Parekh, 1994). Thus, the optimized prediction model used input components of predicted secondary structure, with a glycosylation window of (n-9) to (n+17), and predicted solvent accessibility, with a glycosylation window of (n-16) to (n+13). Further model testing was performed for the microheterogeneity classification model in prediction of glycosylation characteristics at N117 of r-tPA mutants created and analyzed by Wilhelm *et al.*, (1990). The developed model was capable of predicting glycosylation microheterogeneity changes at N117 mutants for cases of mutations performed over 60 residues away from the site of glycosylation. In the few cases of incorrect model prediction, these cases were easily identified by a low confidence level returned by the model.

6.6 Suggestions for future research

Notable suggestions for future work are in the areas of r-tPA glycoform inactivation by glycation and in models for glycosylation characteristics predictions. First, glycation inactivation of tPA needs to be further explored *in vivo*. With adult-onset diabetes becoming an epidemic world-wide, the full effect of this condition on cardiovascular health needs to be fully realized and understood. Inactivation of plasminogen activator proteins by elevated blood glucose levels may be part of the missing link between hyperglycemia and impaired fibrinolysis. In addition, further work is suggested with the r-tPA protein in fully distinguishing the site of inactivating glycation. Should this site be located, site-directed mutations may be employed to eliminate this inactivation mechanism for the recombinant product. The details of the glucose-independent inactivation mechanism should also be a topic of further research. In addition, this future research should also focus on identifying those residues, hypothesized to reside in the kringle 2 domain near N184, responsible for inactivating r-tPA autolysis or aggregation. Following identification of these residues, site-directed mutagenesis may also be used to minimize this mechanism of inactivation, while conserving r-tPA activity. Site-directed mutation has proven useful in many cases for eliminating product aggregation and dramatically extending the half-life of the active enzyme conformation.

With regards to the glycosylation microheterogeneity classification model, possible further research includes the development of a second-generation model to not only predict microheterogeneity classification, but also predict the complete glycoform

composition of complex-type, high mannose and hybrid fractions of the resulting glycoform mixture during CHO cell production. In addition, other post-translational modifications related to glycosylation may be possibly predicted by this methodology. For example, the degree of sialylation of complex-type glycan structures may be a predictable quantity as well.

Appendix A

NOTATION

The following notation defines variables and units contained in Chapter 1. Note that free glucose concentration in Chapter 1 was commonly referred to with units of [g L^{-1}]. A molecular weight value of 180.16 g $mole^{-1}$ was used for conversion. Similarly, molecular weight values of 73,000 and 70,000 were used for the Type I and Type II r-tPA glycoforms respectively.

A	Active glycoform concentration of r-tPA [moles L^{-1}]
G	Free glucose concentration [moles L^{-1}]
B	Schiff base intermediate r-tPA glycoform concentration [moles L^{-1}]
C	Inactive ketoamine form of r-tPA glycoform concentration [moles L^{-1}]
I	Inactive r-tPA glycoform from protein-protein interaction [moles L^{-1}]
z	Stoichiometric coefficient [dimensionless]
t	Time [hours]
k_0	Active r-tPA glycoform inactivation rate constant [$hours^{-1}$]
k_1	Overall r-tPA glycoform glycation rate constant [L^z $moles^{-z}$ $hours^{-1}$]
k_2	r-tPA glycoform Schiff base formation rate constant [L^z $moles^{-z}$ $hours^{-1}$]
k_{-2}	r-tPA glycoform Schiff base degradation rate constant [$hours^{-1}$]

k_3	r-tPA glycoform ketoamine formation rate constant [hours^{-1}]

Table A.1. Definition of notation and units used in Chapter 2.

The following notation defines variables and units used in Chapter 3. It is noted that the r-tPA inactivation mechanisms of Chapter 3 are identical to those defined in Chapter 2. However, all notation was defined independently of those variables used in Chapter 2 because the solvents of the two studies differed, resulting in different rate constant values.

X_d	Intrinsic dead cell density [x10^5 cells mL^{-1}]
$X_{d,app}$	Apparent dead cell density [x10^5 cells mL^{-1}]
X_l	Intrinsic lysed cell density [x10^5 cells mL^{-1}]
X_t	Intrinsic total cell density [x10^5 cells mL^{-1}]
$X_{t,app}$	Apparent total cell density [x10^5 cells mL^{-1}]
$X_{v,app}$	Apparent viable cell density [x10^5 cells mL^{-1}]
glc	Glucose concentration of culture supernatant [g L^{-1}]
Gln	L-glutamine concentration of culture supernatant [g L^{-1}]
Asn	L-asparagine concentration of culture supernatant [g L^{-1}]
AA	Sum of L-glutamine and L-asparagine concentrations of supernatant [g L^{-1}]
S	Sum of glucose, L-glutamine and L-asparagine concentrations of supernatant [g L^{-1}]
lac	Lactate concentration of culture supernatant [g L^{-1}]

Am	Total ammonia (protonated and unprotonated) concentration in culture supernatant [g L^{-1}]
$TypeI_{total}$	Total Type I r-tPA concentration of culture supernatant [μg mL^{-1}]
$TypeII_{total}$	Total Type II r-tPA concentration of culture supernatant [μg mL^{-1}]
$TypeI_{active}$	Active Type I r-tPA concentration of culture supernatant [μg mL^{-1}]
$TypeII_{active}$	Active Type II r-tPA concentration of culture supernatant [μg mL^{-1}]
F_{Glc}	Glucose feed volumetric flow rate [L h^{-1}]
F_{AA}	Amino acids (L-glutamine and L-asparagine) feed volumetric flow rate [L h^{-1}]
F_T	Sum of glucose and amino acids volumetric flow rates [L h^{-1}]
S_{Glc}	Glucose concentration of glucose feed reservoir [g L^{-1}]
S_{Gln}	L-glutamine concentration of amino acids feed reservoir [g L^{-1}]
S_{Asn}	L-asparagine concentration of amino acids feed reservoir [g L^{-1}]
$M_{Glucose}$	Glucose feed mass flow rate [g h^{-1}]
$M_{Amino\ Acids}$	Amino acids (L-glutamine and L-asparagine) feed mass flow rate [g h^{-1}]
t	Time [hours]
V	Reactor working volume [L]
μ_{int}	Intrinsic growth rate [$hours^{-1}$]
μ_{max}	Maximum growth rate Monod parameter [$hours^{-1}$]
K_{glc}	Monod constant for glucose consumption [g L^{-1}]
$K_{d,glc}$	Monod constant for glucose inhibition [g L^{-1}]
K_{AA}	Monod constant for amino acids (L-glutamine and L-asparagine) consumption [g L^{-1}]

K_{lac}	Monod constant for lactate inhibition [g L^{-1}]
K_{Am}	Monod constant for total ammonia inhibition [g L^{-1}]
k_d	Intrinsic cell death rate [$hours^{-1}$]
$k_{d,max}$	Maximum cell death rate model parameter [$hours^{-1}$]
K_I	Cell death model constant for inhibitor interaction [g L^{-1}]
K_S	Cell death model constant for substrate interaction [g L^{-1}]
k_l	Cell lysis rate constant [$hours^{-1}$]
$Y_{Xv,app/glc}$	Viable cell yield coefficient for glucose [x10^8 cells g^{-1}]
$Y_{Xv,app/Gln}$	Viable cell yield coefficient for L-glutamine [x10^8 cells g^{-1}]
$Y_{Xv,app/Asn}$	Viable cell yield coefficient for L-asparagine [x10^8 cells g^{-1}]
$Y_{Xv,app/lac}$	Viable cell yield coefficient for lactate [x10^8 cells g^{-1}]
$Y_{Xv,app/Am}$	Viable cell yield coefficient for total ammonia [x10^8 cells g^{-1}]
α	Type I r-tPA production rate constant [μg x10^5 $cells^{-1}$]
β	Type II r-tPA production rate constant [μg x10^5 $cells^{-1}$]
k_{deg}	L-glutamine degradation rate constant [$hours^{-1}$]
$k_{I,0}$	Type I r-tPA natural inactivation rate constant [mL μg^{-1} h^{-1}]
$k_{II,0}$	Type II r-tPA natural inactivation rate constant [mL μg^{-1} h^{-1}]
k_{Gly}	r-tPA glycation inactivation rate constant [L g^{-1} h^{-1}]

Table A.2. Definition of notation and units used in Chapter 3.

The following notation was used in Chapter 4 and Chapter 5. An extended explanation of the significant dimensionless variables involved is given.

n	The asparagine residue of the N-*X*-S/T glycosylation sequence. This is also defined as the site of glycosylation.
n-x	The starting residue of the glycosylation window relative to the site of glycosylation (*n*). If (x) is positive, the glycosylation window starts to the *N*-terminus side of the glycosylation site. A negative value of (x) corresponds to a starting residue on the *C*-terminus site of the glycosylation site x-residues from (*n*). For example (*n*+2) corresponds to a glycosylation window starting 2 residues to the *C*-terminus side of (*n*).
n+y	The terminating residue of the glycosylation window relative to the site of glycosylation (*n*). If (y) is positive, the glycosylation window terminates y-residues from (*n*) on the *C*-terminus side of the glycosylation site.

Table A.3. Further explanation of the glycosylation window defined and used in Chapter 4 and Chapter 5.

Appendix B

FED BATCH SIMULATION EQUATIONS

The following kinetic equations were used in fed batch simulations in Chapter 3. It is noted that feed flow values were zero during batch operation modes. All notation and units are summarized in Appendix A.

$$\frac{dX_t}{dt} = \mu_{int} X_{v,app} - X_t\left(\frac{F_T}{V}\right) \tag{B.1}$$

$$\frac{dX_{v,app}}{dt} = (\mu_{int} - k_d)X_{v,app} - X_{v,app}\left(\frac{F_T}{V}\right) \tag{B.2}$$

$$\frac{dX_d}{dt} = k_d X_{v,app} - X_d\left(\frac{F_T}{V}\right) \tag{B.3}$$

$$\frac{dX_l}{dt} = k_l X_{t,app} - X_l\left(\frac{F_T}{V}\right) \tag{B.4}$$

$$\frac{d(glc)}{dt} = \left(\frac{F_{glc}}{V}\right)S_{glc} - (glc)\left(\frac{F_T}{V}\right) - \frac{1}{Y_{X_{v,app}/glc}}\mu_{int} X_{v,app} \tag{B.5}$$

$$\frac{d(Gln)}{dt} = \left(\frac{F_{AA}}{V}\right)S_{Gln} - (Gln)\left(\frac{F_T}{V}\right) - \frac{1}{Y_{X_{v,app}/Gln}}\mu_{int} X_{v,app} - k_{deg}(Gln) \tag{B.6}$$

$$\frac{d(Asn)}{dt} = \left(\frac{F_{AA}}{V}\right)S_{Asn} - (Asn)\left(\frac{F_T}{V}\right) - \frac{1}{Y_{X_{v,app}/Asn}}\mu_{int} X_{v,app} \tag{B.7}$$

$$\frac{d(lac)}{dt} = \frac{1}{Y_{X_{v,app}/lac}}\mu_{int} X_{v,app} - (lac)\left(\frac{F_T}{V}\right) \tag{B.8}$$

$$\frac{d(Am)}{dt} = \frac{1}{Y_{X_{v,app}/Am}} \mu_{int} X_{v,app} + k_{deg}(Gln) - (Am)\left(\frac{F_T}{V}\right) \quad \text{(B.9)}$$

$$\frac{d(TypeI_{total})}{dt} = \alpha\mu_{int} X_{v,app} - (TypeI_{total})\left(\frac{F_T}{V}\right) \quad \text{(B.10)}$$

$$\frac{d(TypeII_{total})}{dt} = \beta\mu_{int} X_{v,app} - (TypeII_{total})\left(\frac{F_T}{V}\right) \quad \text{(B.11)}$$

$$\frac{d(TypeI_{active})}{dt} = \alpha\mu_{int} X_{v,app} - k_{I,0}(TypeI_{active})^2 - k_{Gly}(TypeI_{active})(glc) - (TypeI_{active})\left(\frac{F_T}{V}\right)$$

(B.12)

$$\frac{d(TypeII_{active})}{dt} = \beta\mu_{int} X_{v,app} - k_{II,0}(TypeII_{active})^2 - k_{Gly}(TypeII_{active})(glc) - (TypeII_{active})\left(\frac{F_T}{V}\right)$$

(B.13)

$$\mu_{int} = \frac{\mu_{max}(glc)(AA)}{(K_{glc} + glc)(K_{AA} + AA)\left(\frac{glc}{K_{d,glc}} + 1\right)\left(\frac{lac}{K_{lac}} + 1\right)\left(\frac{Am}{K_{Am}} + 1\right)} \quad \text{(3.10)}$$

$$k_d = \frac{k_{d,max}(I)}{K_I + I + \frac{K_I}{K_S} S} \quad \text{(3.9)}$$

$$F_T = F_{glc} + F_{AA} \quad \text{(B.14)}$$

$$AA = Gln + Asn \quad \text{(B.15)}$$

$$S = glc + AA \quad \text{(B.16)}$$

$$I = lac + Am \quad \text{(B.17)}$$

Appendix C

FED BATCH SIMULATIONS

The following fed batch simulations are summarized in Tables 3.2 through 3.4 in Chapter 3. All simulation results of monitored culture states are grouped together according to fixed feed flow rates or metabolite set points as well as the initiation of fed batch feeding. All simulations were performed for a duration of 500 hours. The following cell density culture states were simulated and are reported with units of [$x10^5$ cells mL^{-1}]: intrinsic total cell density, X_t, apparent viable cell density, $X_{v,app}$, intrinsic dead cell density, X_d, and the lysed cell density, X_l. Glucose, L-glutamine and L-asparagine were the metabolite values simulated and have units of [g L^{-1}]. The byproducts of lactate and total ammonia (both protonated and unprotonated) were simulated and have units of [g L^{-1}]. The total Type I and total Type II r-tPA glycoform concentrations, as well as the active Type I and Type II r-tPA glycoforms, are reported wit units of [μg mL^{-1}]. Finally, reactor volume is reported with units of [L], and the total r-tPA activity is reported with international activity units [IU]. It is noted that the maximum r-tPA total activity determined the reported harvest period for the fed batch simulation.

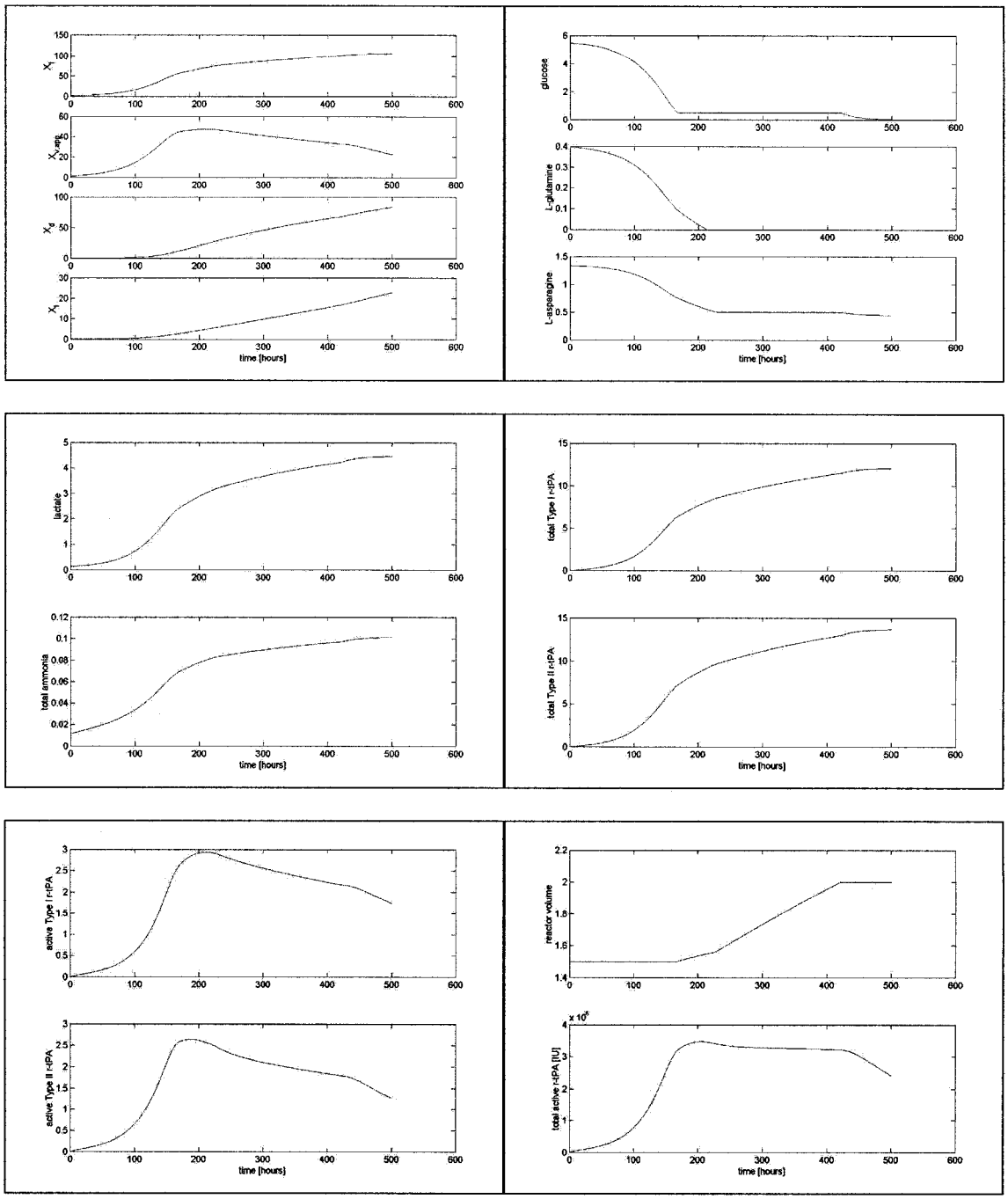

Figure C.1. Simulations of variable feed flow rates. The glucose set point was 0.50 g L^{-1}, and the amino acids set point was 0.50 g L^{-1}. Results are summarized in Table 3.2.

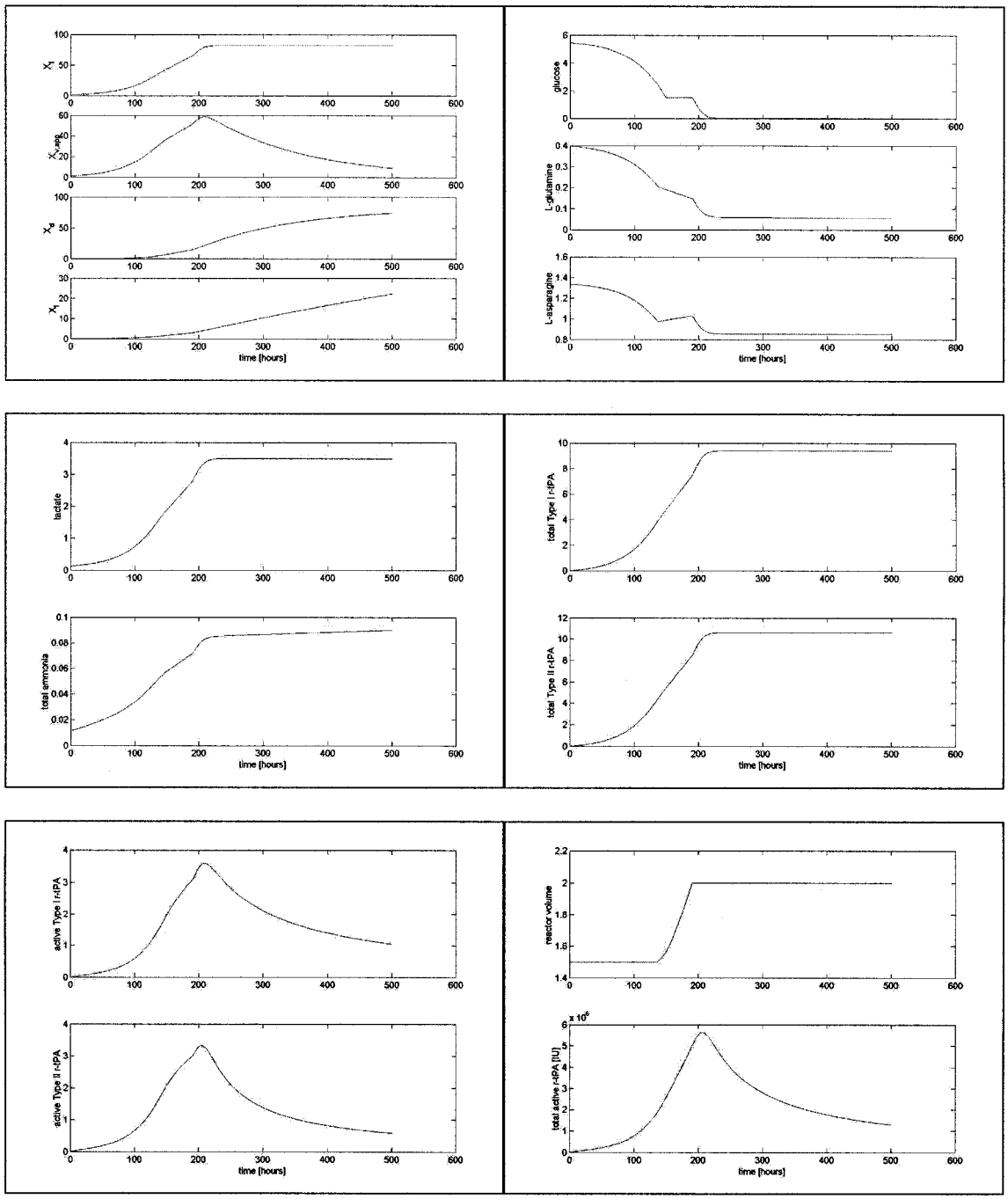

Figure C.2. Simulations of variable feed flow rates. The glucose set point was 1.51 g L^{-1}, and the amino acids set point was 1.18 g L^{-1}. Results are summarized in Table 3.2.

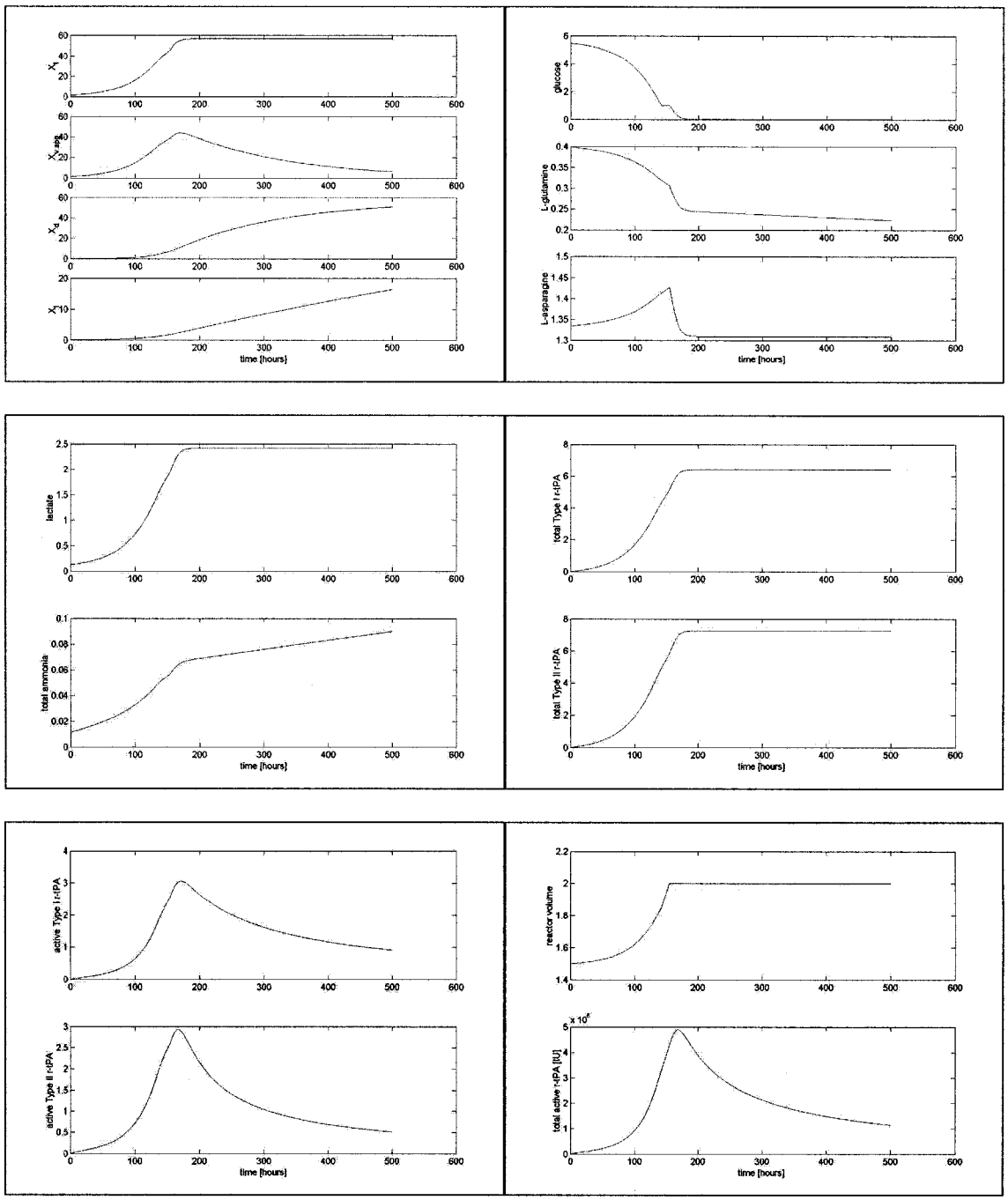

Figure C.3. Simulations of variable feed flow rates. The glucose set point was 1.00 g L^{-1}, and the amino acids set point was 2.70 g L^{-1}. Results are summarized in Table 3.2.

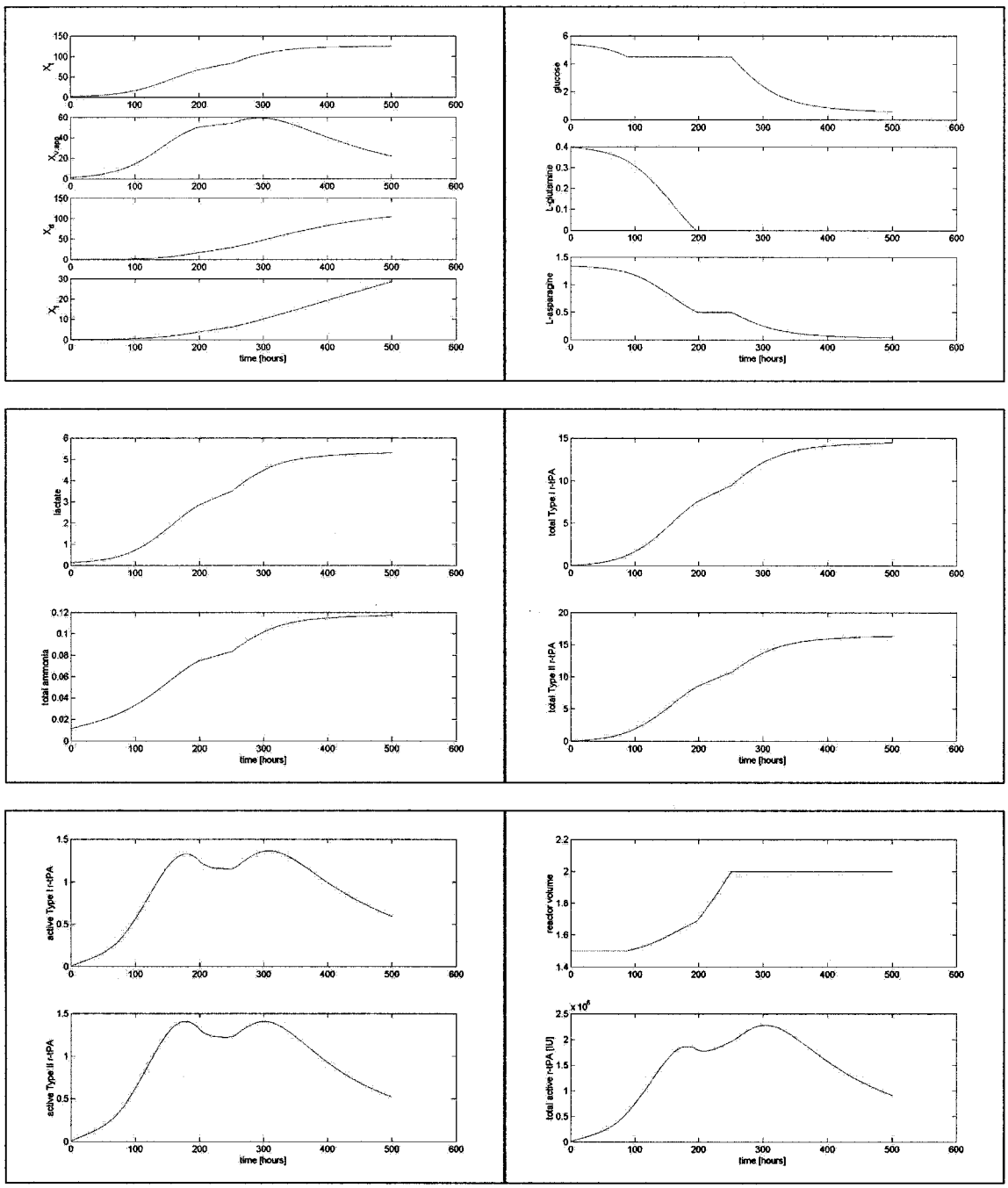

Figure C.4. Simulations of variable feed flow rates. The glucose set point was 4.50 g L^{-1}, and the amino acids set point was 0.50 g L^{-1}. Results are summarized in Table 3.2.

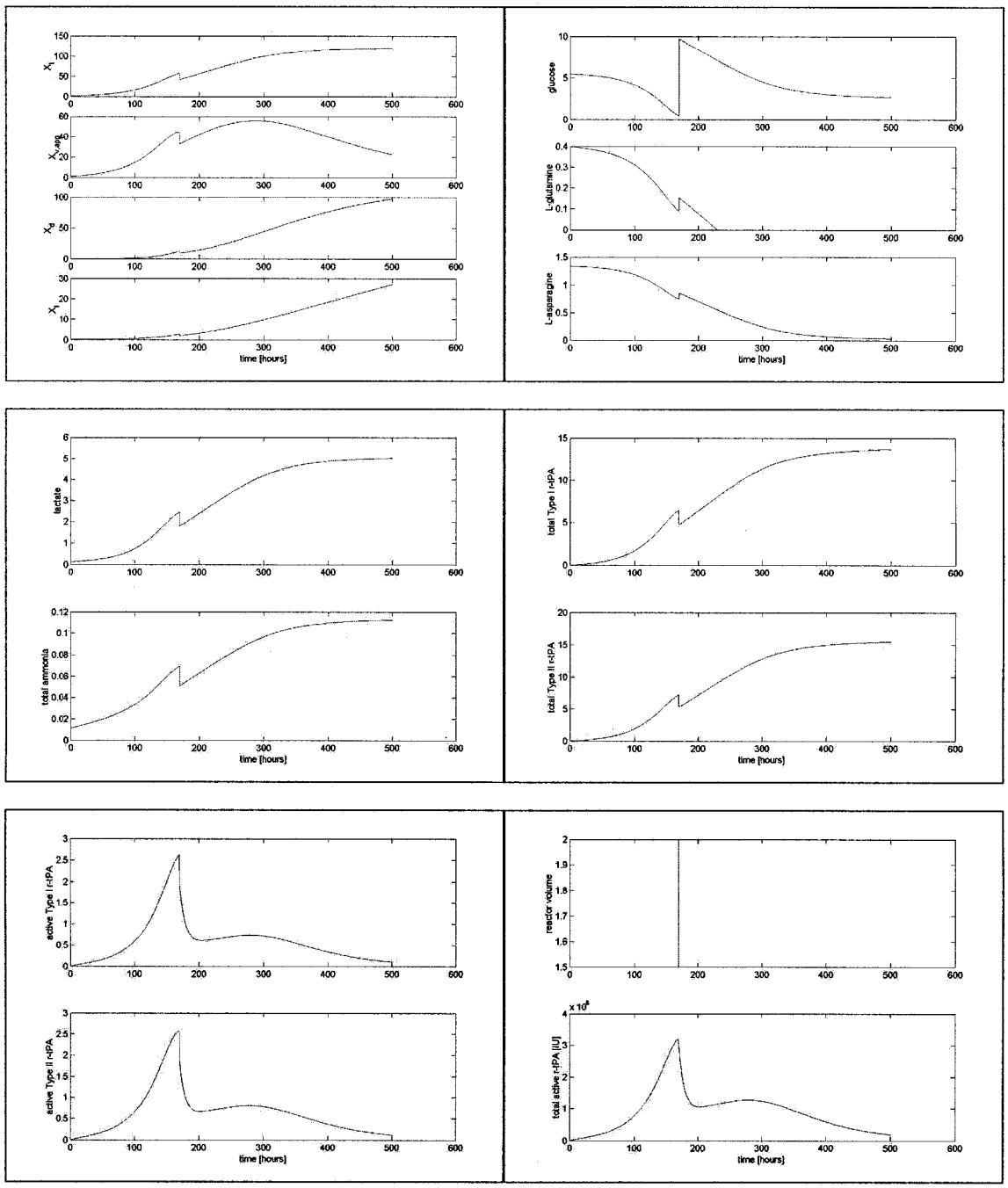

Figure C.5. Simulations of fixed feed flow rates initiated at 170 h. The glucose mass flow rate was 120.0 g h^{-1}, and the amino acids mass flow rate was 5.0 g h^{-1}. Results are summarized in Table 3.3.

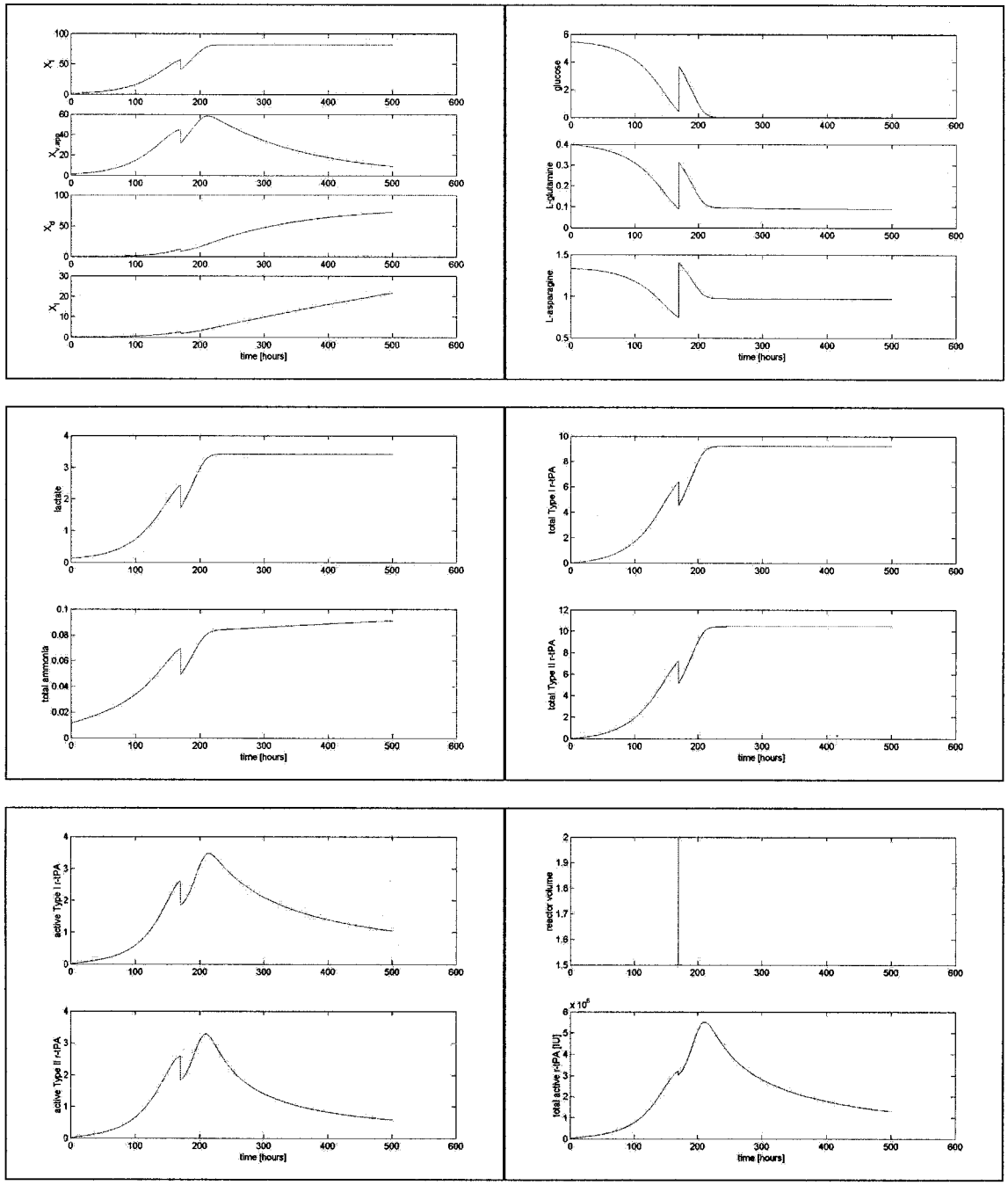

Figure C.6. Simulations of fixed feed flow rates initiated at 170 h. The glucose mass flow rate was 120.0 g h^{-1}, and the amino acids mass flow rate was 40.0 g h^{-1}. Results are summarized in Table 3.3.

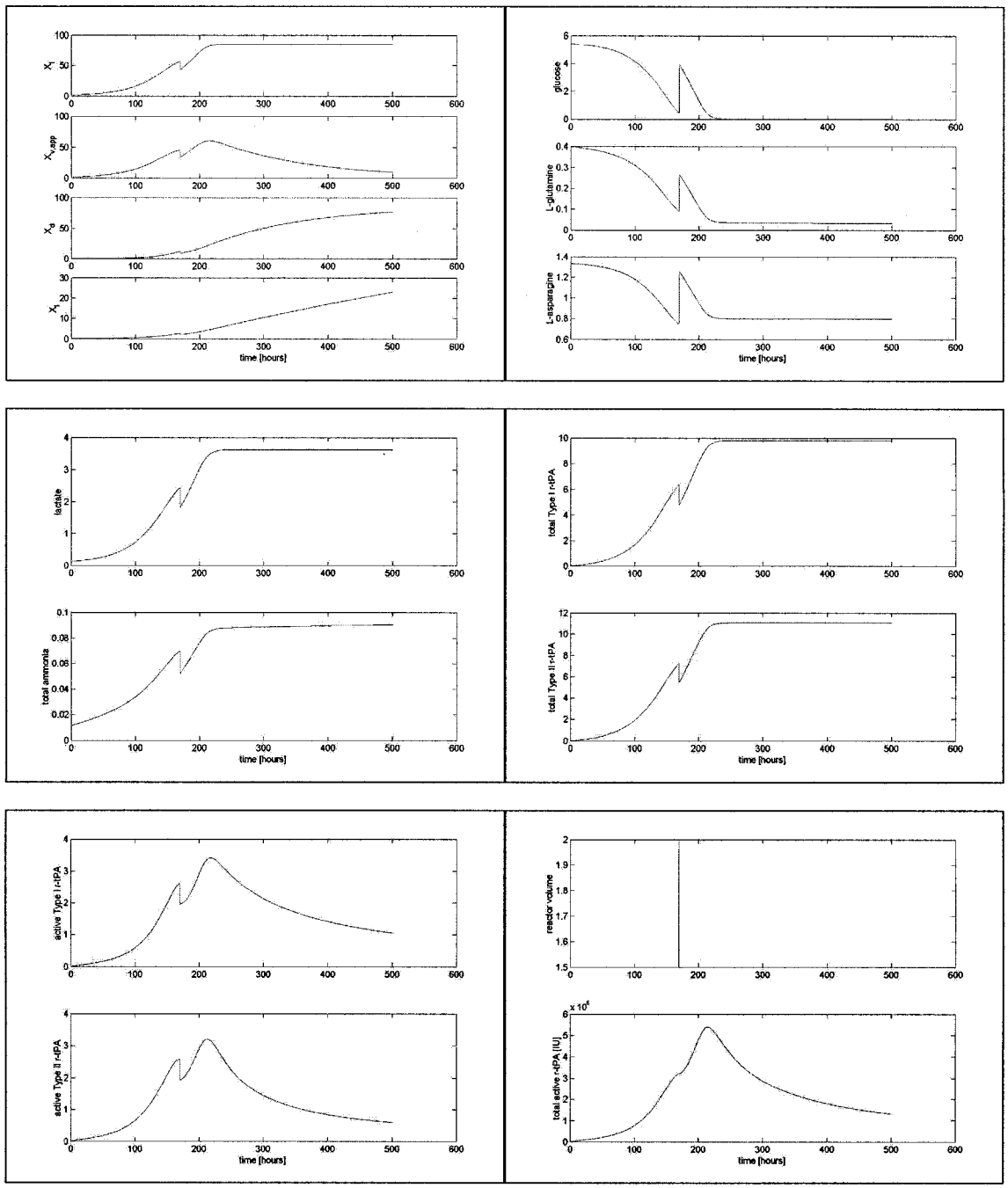

Figure C.7. Simulations of fixed feed flow rates initiated at 170 h. The glucose mass flow rate was 20.0 g h^{-1}, and the amino acids mass flow rate was 5.0 g h^{-1}. Results are summarized in Table 3.3.

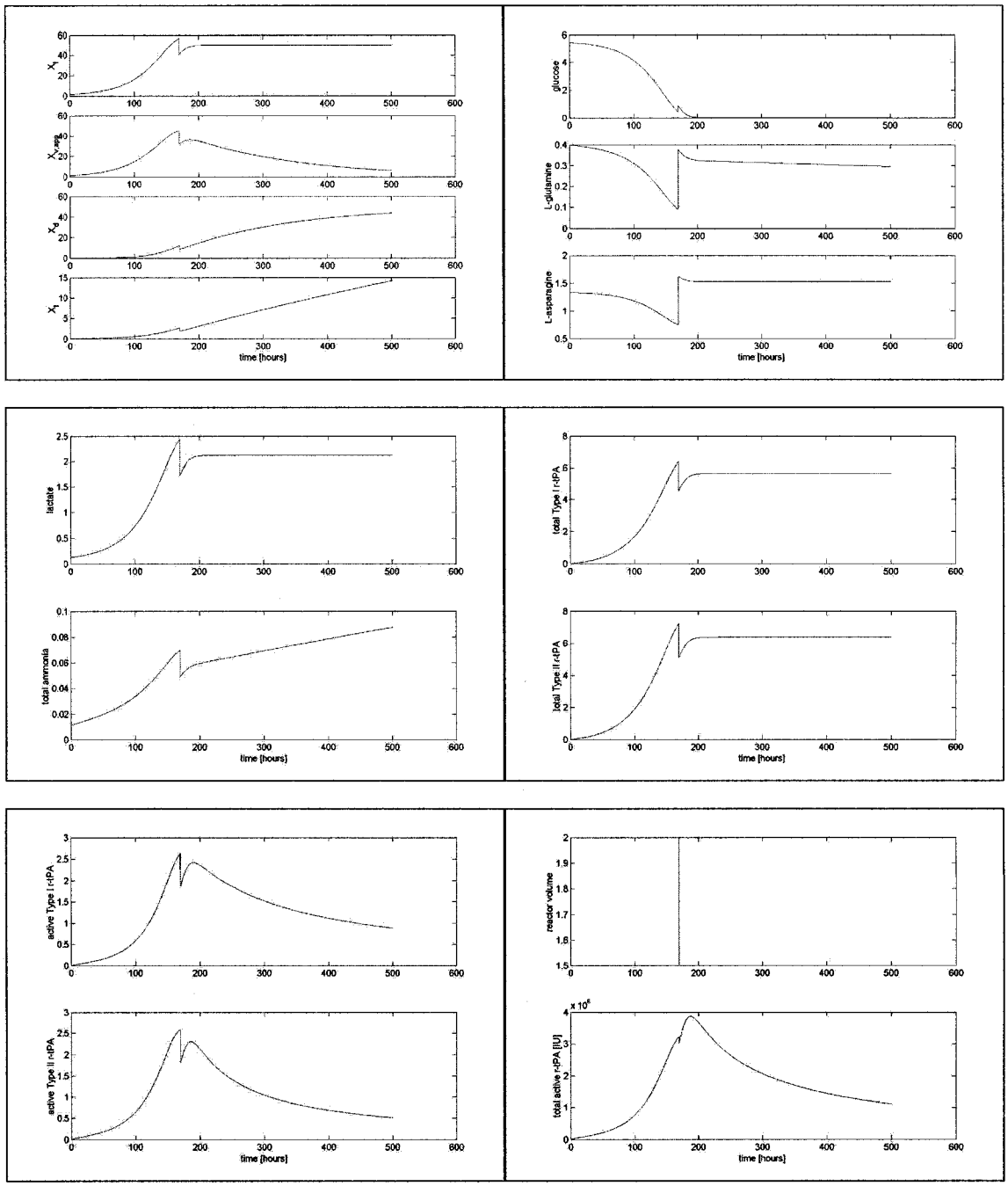

Figure C.8. Simulations of fixed feed flow rates initiated at 170 h. The glucose mass flow rate was 20.0 g h^{-1}, and the amino acids mass flow rate was 40.0 g h^{-1}. Results are summarized in Table 3.3.

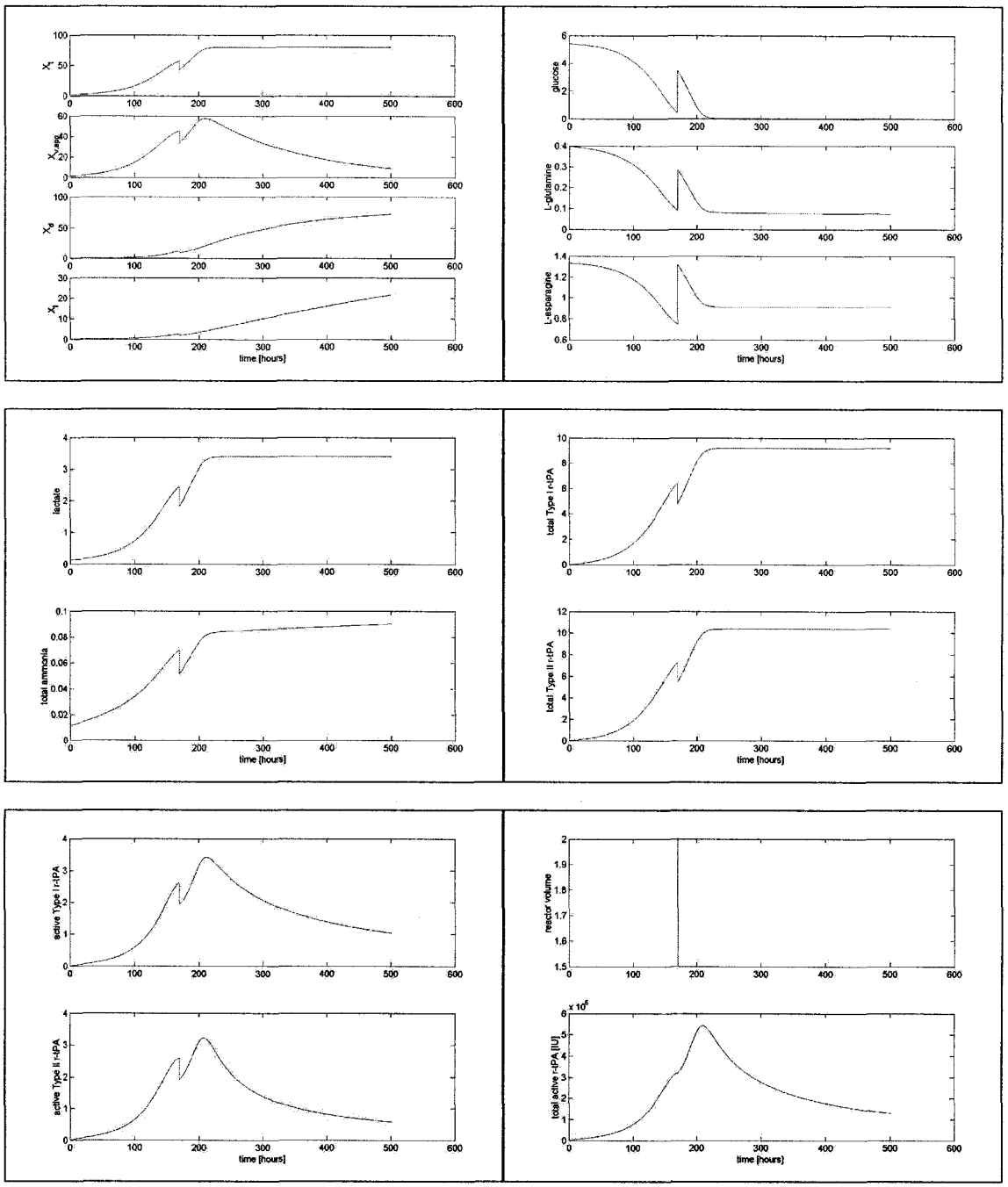

Figure C.9. Simulations of fixed feed flow rates initiated at 170 h. The glucose mass flow rate was 80.0 g h^{-1}, and the amino acids mass flow rate was 25.0 g h^{-1}. Results are summarized in Table 3.3.

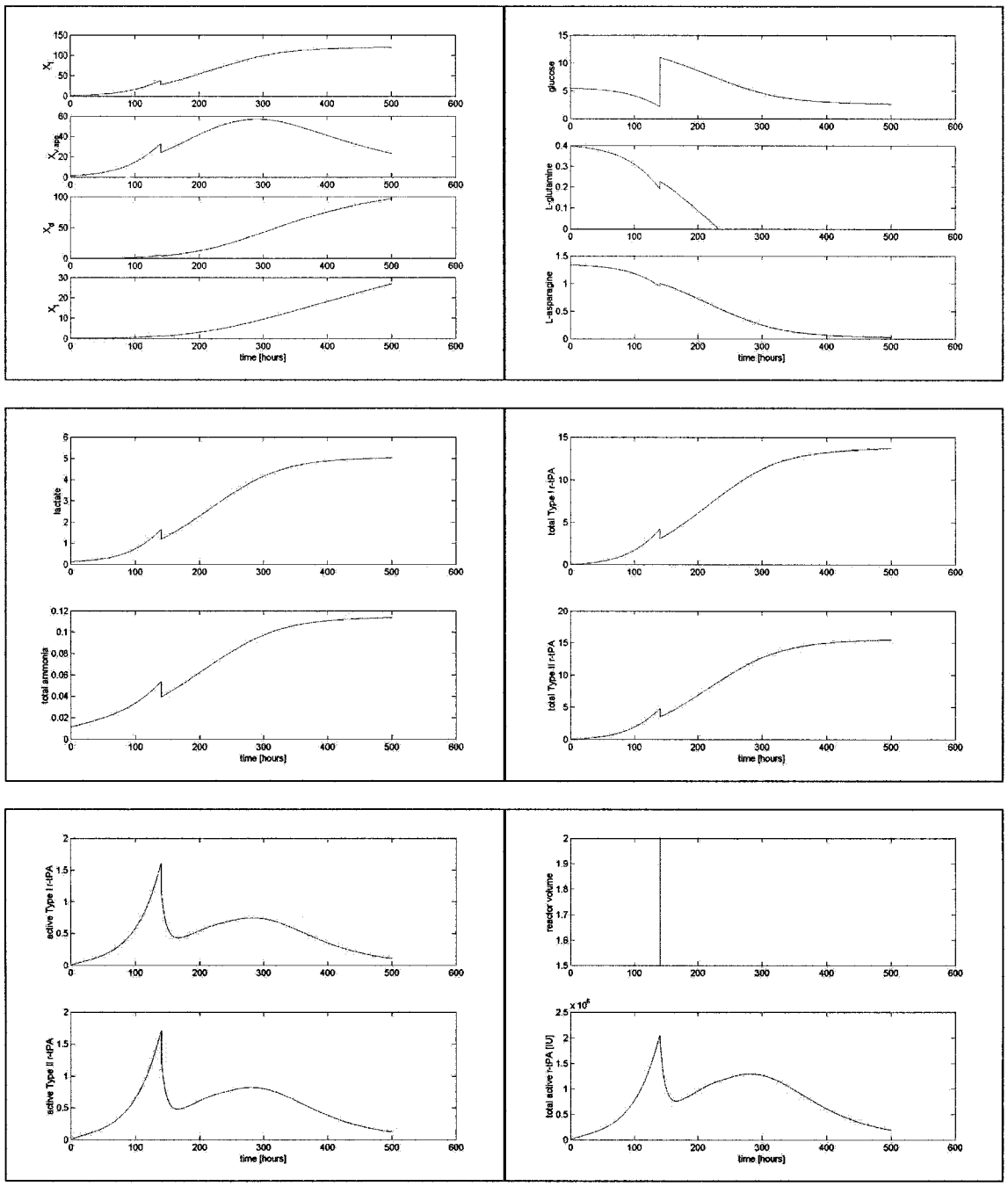

Figure C.10. Simulations of fixed feed flow rates initiated at 140 h. The glucose mass flow rate was 120.0 g h^{-1}, and the amino acids mass flow rate was 5.0 g h^{-1}. Results are summarized in Table 3.4.

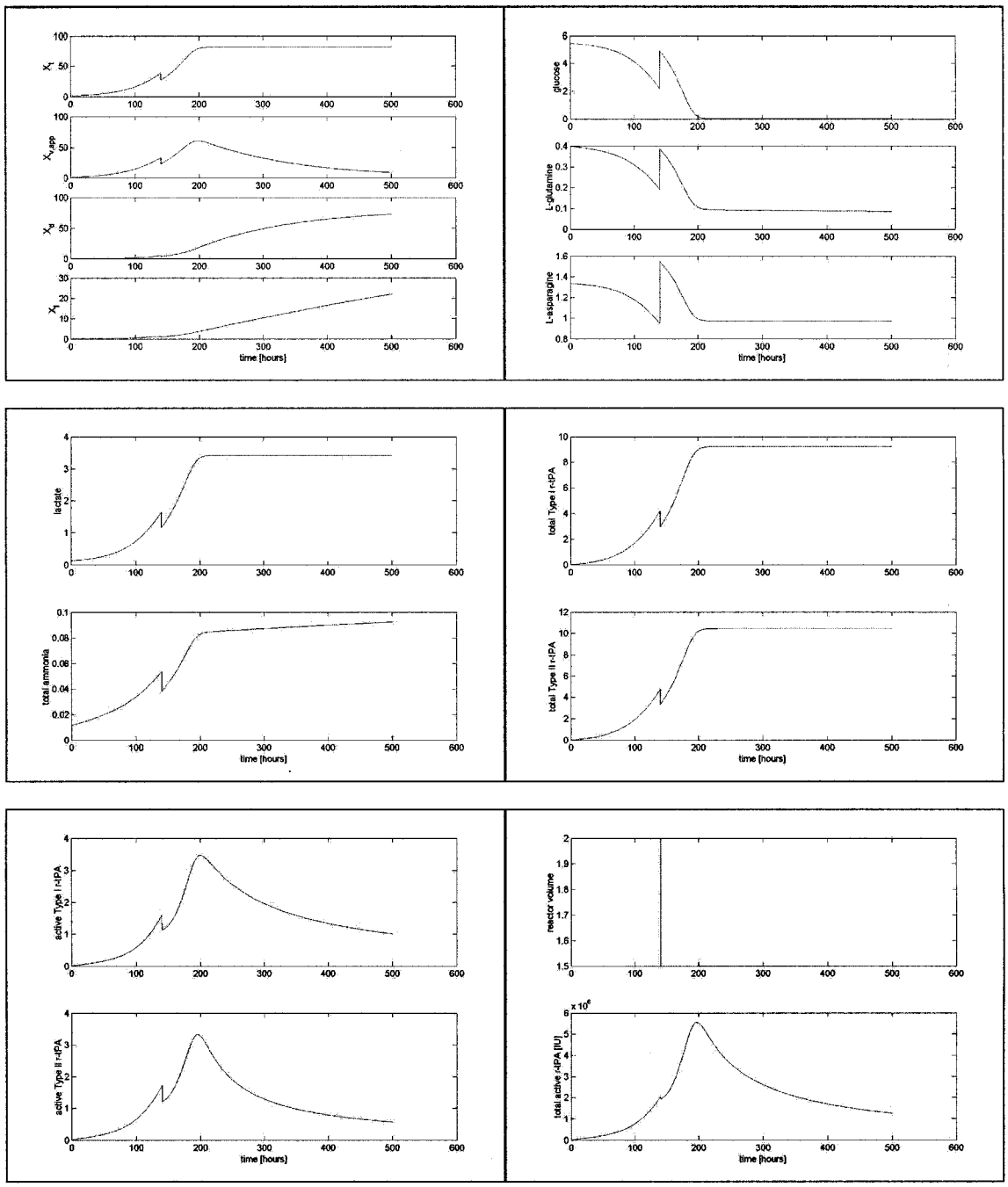

Figure C.11. Simulations of fixed feed flow rates initiated at 140 h. The glucose mass flow rate was 120.0 g h^{-1}, and the amino acids mass flow rate was 40.0 g h^{-1}. Results are summarized in Table 3.4.

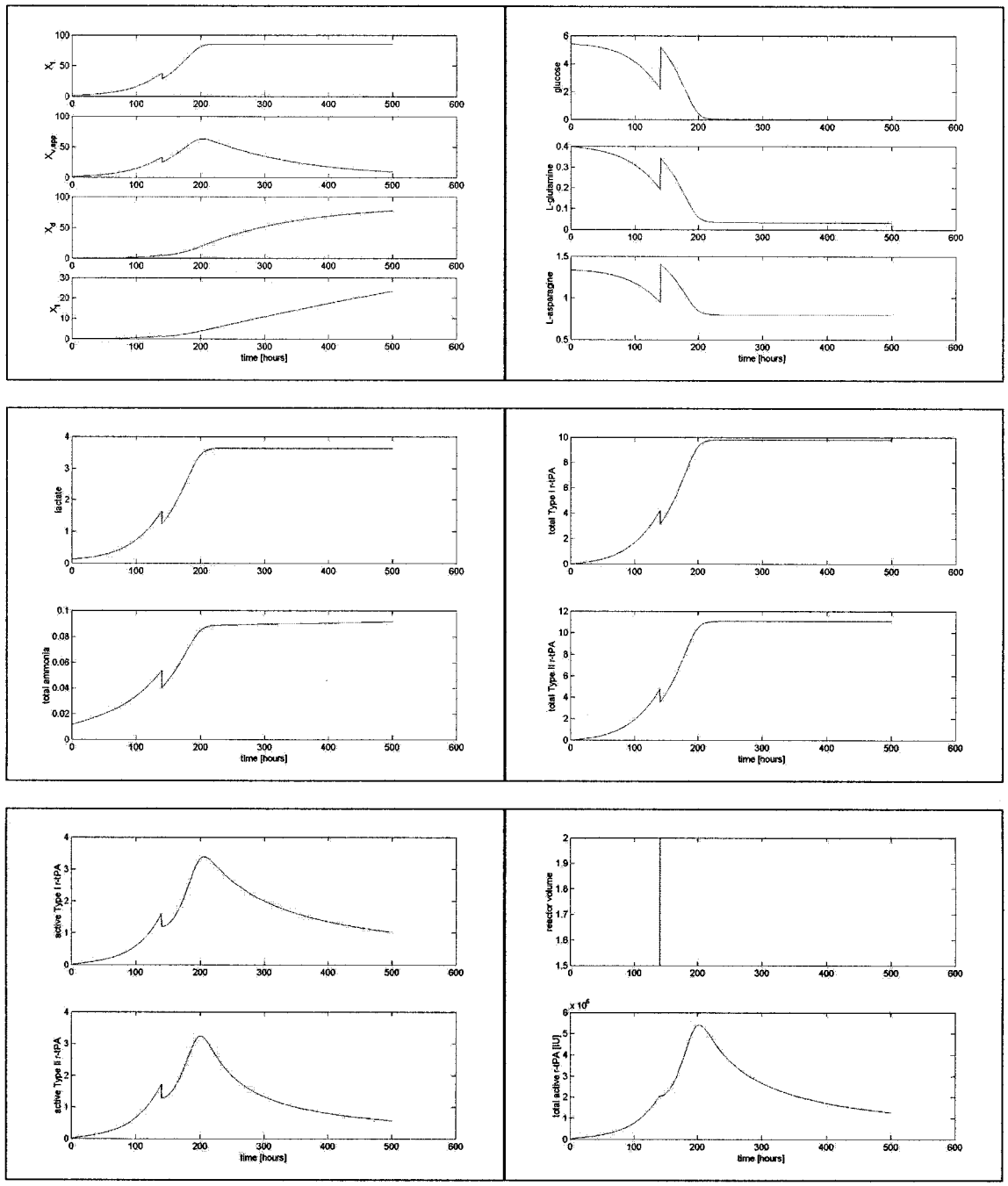

Figure C.12. Simulations of fixed feed flow rates initiated at 140 h. The glucose mass flow rate was 20.0 g h^{-1}, and the amino acids mass flow rate was 5.0 g h^{-1}. Results are summarized in Table 3.4.

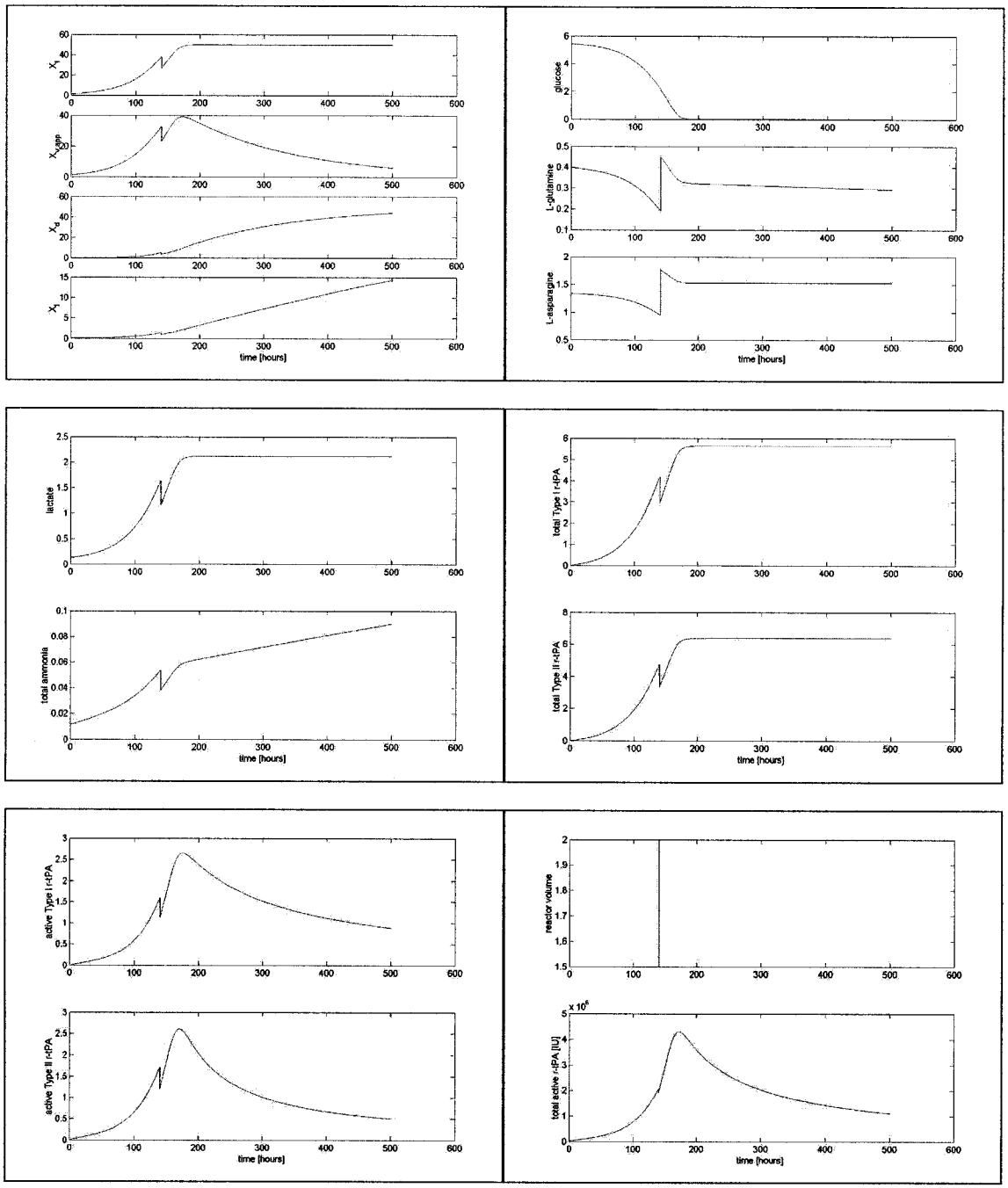

Figure C.13. Simulations of fixed feed flow rates initiated at 140 h. The glucose mass flow rate was 20.0 g h^{-1}, and the amino acids mass flow rate was 40.0 g h^{-1}. Results are summarized in Table 3.4.

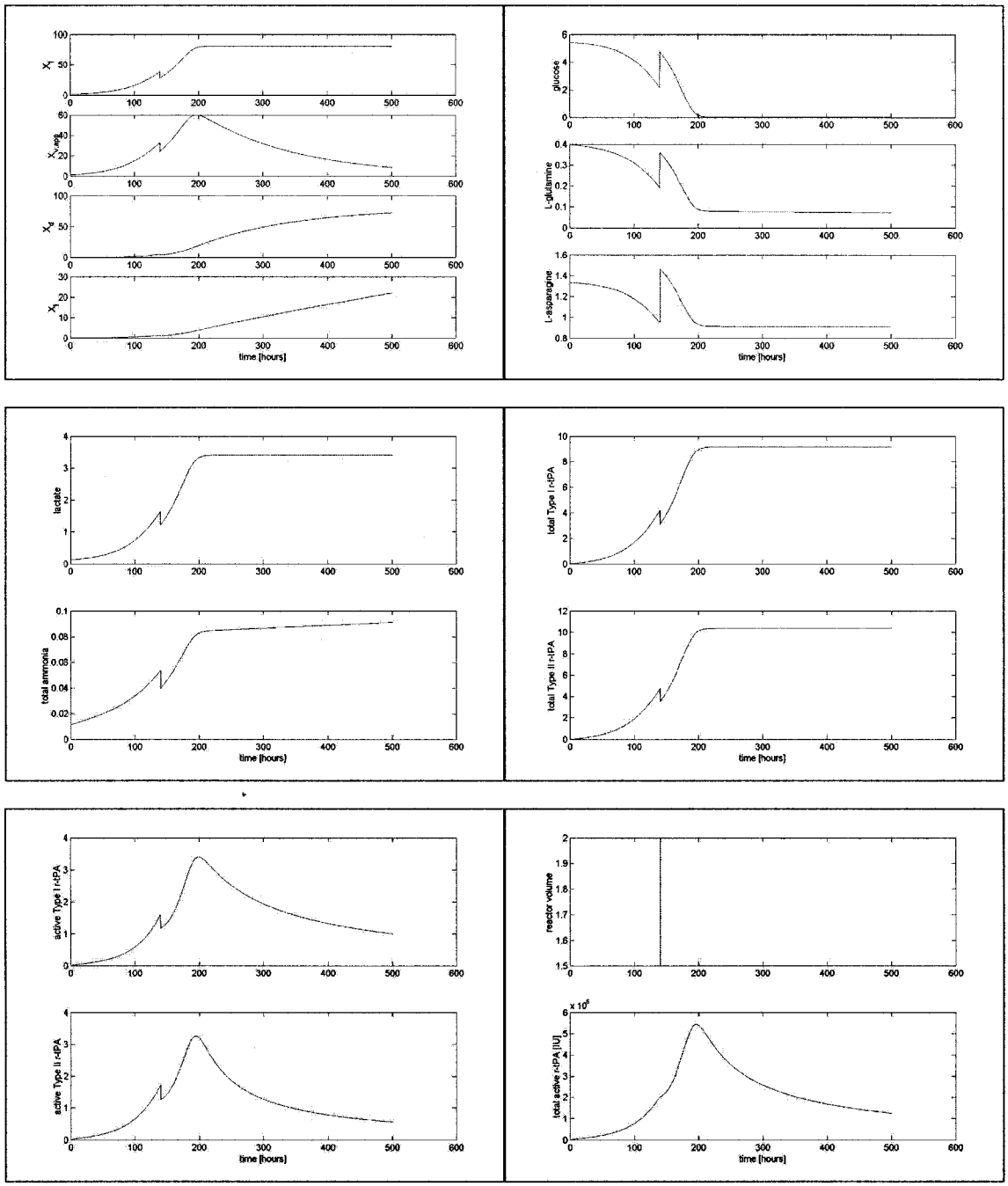

Figure C.14. Simulations of fixed feed flow rates initiated at 140 h. The glucose mass flow rate was 80.0 g h^{-1}, and the amino acids mass flow rate was 25.0 g h^{-1}. Results are summarized in Table 3.4.

Appendix D

SITE-OCCUPANCY REFERENCE SET

The following is the complete data set used to compose neural network training and testing data sets for the neural network models in Chapter 3. The specific protein name is given along with the specific site of glycosylation. A classification of *1* corresponds to variable site-occupancy, and a classification of *0* corresponds to robust glycosylation.

Protein	Site	Classification	Reference
Alpha-galactosidase B	124	0	Ohta *et al.*, 2000
Alpha-galactosidase B	177	0	Ohta *et al.*, 2000
Alpha-galactosidase B	201	0	Ohta *et al.*, 2000
Alpha-galactosidase B	359	0	Ohta *et al.*, 2000
Alpha-galactosidase B	385	0	Ohta *et al.*, 2000
Antithrombin III	96	0	Picard *et al.*, 1995; Fan *et al.*, 1993
Antithrombin III	135	1	Picard *et al.*, 1995; Fan *et al.*, 1993
Antithrombin III	155	0	Picard *et al.*, 1995; Fan *et al.*, 1993

Antithrombin III	192	0	Picard *et al.*, 1995; Fan *et al.*, 1993
Bovine Pancreatic Ribonuclease	34	1	Rudd *et al.*, 1995
Cholesterol ester transfer protein (CETP)	88	0	Stevenson *et al.*, 1993
Cholesterol ester transfer protein (CETP)	240	0	Stevenson *et al.*, 1993
Cholesterol ester transfer protein (CETP)	341	1	Stevenson *et al.*, 1993
Cholesterol ester transfer protein (CETP)	396	0	Stevenson *et al.*, 1993
Coagulation factor VIIa	145	1	Thim *et al.*, 1988
Coagulation factor VIIa	322	0	Thim *et al.*, 1988
GM1 Synthase	149	0	Martina *et al.*, 2000
GM1 Synthase	235	0	Martina *et al.*, 2000
Human interferon gamma	26	1	Nyberg *et al.*, 1999
Human interferon gamma	98	1	Nyberg *et al.*, 1999
Insulin-like growth factor binding protein	116	0	Firth and Baxter, 1999
Insulin-like growth factor binding protein	136	0	Firth and Baxter, 1999
Insulin-like growth factor binding protein	199	1	Firth and Baxter, 1999
Interleukin-1 beta	123	1	Livi *et al.*, 1991
Plasminogen	308	1	Rudd *et al.*, 1995
Procathespin L	204	0	Kane, 1993
Prolactin	59	1	Shelikoff *et al.*, 1994; Shelikoff *et al.*, 1996

Rabies Virus Glycoprotein	37	1	Shakin-Eshleman, 1996; Kasturi *et al.*, 1997; Mellquist *et al.*, 1998
Rabies Virus Glycoprotein	247	0	Shakin-Eshleman, 1996; Kasturi *et al.*, 1997; Mellquist *et al.*, 1998
Rabies Virus Glycoprotein	319	0	Shakin-Eshleman, 1996; Kasturi *et al.*, 1997; Mellquist *et al.*, 1998
Rabies Virus Glycoprotein	465	0	Shakin-Eshleman, 1996; Kasturi *et al.*, 1997; Mellquist *et al.*, 1998
Rat alpha 1,3 frucosyltransferase IV	117	0	Baboval *et al.*, 2000
Rat alpha 1,3 frucosyltransferase IV	218	0	Baboval *et al.*, 2000
Serum Transferrin	432	1	Landberg et al., 1995; Iourin *et al.*, 1996
Serum Transferrin	632	1	Landberg et al., 1995; Iourin *et al.*, 1996
sThy-1	23	1	Devasahayam *et al.*, 1999
sThy-1	60	1	Devasahayam *et al.*, 1999
sThy-1	100	1	Devasahayam *et al.*, 1999
Thrombopoietin	176	1	Hoffman *et al.*, 1996

Thrombopoietin	185	0	Hoffman *et al.*, 1996
Thrombopoietin	213	1	Hoffman *et al.*, 1996
Thrombopoietin	234	0	Hoffman *et al.*, 1996
Tissue Plasminogen Activator	117	0	Grossbard, 1987
Tissue Plasminogen Activator	184	1	Grossbard, 1987
Tissue Plasminogen Activator	448	0	Grossbard, 1987
Transferrin receptor (G724S)	251	0	Williams and Enns, 1993
Transferrin receptor (G724S)	722	1	Williams and Enns, 1993
Transferrin receptor (G724S)	727	0	Williams and Enns, 1993

Table D.1. The entire glycosylation site reference set used for construction of neural network training and testing data sets for glycosylation site-occupancy prediction.

Appendix E

MICROHETEROGENEITY CLASSIFICATION REFERENCE SET

The following is the complete data set used to compose neural network training and testing data sets for the neural network models in Chapter 4. The specific protein name is given along with the specific site of glycosylation. A classification of *0.25* corresponds to high mannose glycosylation, and a classification of *0.75* corresponds to complex-type glycosylation microheterogeneity classification.

Protein	Site	Classification	Reference
Alpha-1-acid glycoprotein	72	0.75	Fournet *et al.*, 1978; Yoshima *et al.*,1981; Sutton *et al.*, 1994
Alpha-1-antichymotrypsin	33	0.75	Wilson *et al.*, 2002
Alpha-1-antichymotrypsin	271	0.75	Wilson *et al.*, 2002
Alpha(2)-HS-glycoprotein	156	0.75	Wilson *et al.*, 2002
Alpha(2)-HS-glycoprotein	176	0.75	Wilson *et al.*, 2002
Antithrombin III	167	0.75	Mizuochi *et al.*, 1980
Antithrombin III	187	0.75	Mizuochi *et al.*, 1980
Apolipoprotein D	65	0.75	Schindler *et al.*, 1995
Apolipoprotein D	98	0.75	Schindler *et al.*, 1995

Beta-hexosaminidase B	84	0.25	Schuette *et al.*, 2001
Beta-hexosaminidase B	190	0.25	Schuette *et al.*, 2001
Beta-hexosaminidase B	327	0.25	Schuette *et al.*, 2001
Campath-1H	27	0.75	Sheeley *et al.*, 1997; Lifely *et al.*, 1995
Campath-1H	40	0.25	Sheeley *et al.*, 1997; Lifely *et al.*, 1995
CD59	43	0.75	Wheeler *et al.*, 2002
Chronic gonadotropin (beta subunit)	33	0.75	Moriwaki *et al.*, 1997
Cystic fibrosis transmembrane conductance regulator	48	0.75	O'Riordan *et al.*, 2000
Cystic fibrosis transmembrane conductance regulator	396	0.75	O'Riordan *et al.*, 2000
Cystic fibrosis transmembrane conductance regulator	597	0.75	O'Riordan *et al.*, 2000
Cystic fibrosis transmembrane conductance regulator	894	0.75	O'Riordan *et al.*, 2000
Cystic fibrosis transmembrane conductance regulator	1148	0.75	O'Riordan *et al.*, 2000
Cystic fibrosis transmembrane conductance regulator	1229	0.75	O'Riordan *et al.*, 2000
DNAse I	40	0.25	Cacia *et al.*, 1998

DNAse I	128	0.75	Cacia *et al.*, 1998
Epidermal growth factor receptor	128	0.25	Smith *et al.*, 1996
Epidermal growth factor receptor	175	0.75	Smith *et al.*, 1996
Epidermal growth factor receptor	352	0.75	Smith *et al.*, 1996
Epidermal growth factor receptor	413	0.75	Smith *et al.*, 1996
Epidermal growth factor receptor	444	0.25	Smith *et al.*, 1996
Epidermal growth factor receptor	528	0.75	Smith *et al.*, 1996
Epidermal growth factor receptor	568	0.75	Smith *et al.*, 1996
Erythropoietin	65	0.75	Watson *et al.*, 1994; Rush *et al.*, 1993
Erythropoietin	110	0.75	Watson *et al.*, 1994; Rush *et al.*, 1993
Fibroblast growth factor	16	0.75	Asada *et al.*, 1999
Fibroblast growth factor	48	0.75	Asada *et al.*, 1999
Fibroblast growth factor	80	0.75	Asada *et al.*, 1999
Follicle stimulating hormone	14	0.75	Flack *et al.*, 1994; Chin *et al.*, 1996
Follicle stimulating hormone	314	0.75	Flack *et al.*, 1994; Chin *et al.*, 1996
Follicle stimulating hormone	334	0.75	Flack *et al.*, 1994; Chin *et al.*, 1996

Follicle stimulating hormone	368	0.75	Flack *et al.*, 1994; Chin *et al.*, 1996
Follicle stimulating hormone	613	0.75	Flack *et al.*, 1994; Chin *et al.*, 1996
Follicle stimulating hormone	631	0.75	Flack *et al.*, 1994; Chin *et al.*, 1996
Follicle stimulating hormone	707	0.75	Flack *et al.*, 1994; Chin *et al.*, 1996
Haptoglobin	125	0.75	Wilson *et al.*, 2002
Haptoglobin	148	0.25	Wilson *et al.*, 2002
Haptoglobin	152	0.75	Wilson *et al.*, 2002
Haptoglobin	182	0.75	Wilson *et al.*, 2002
Hepatic lipase	42	0.25	Boedeker *et al.*, 1999
Hepatic lipase	78	0.25	Boedeker *et al.*, 1999
Hepatic lipase	362	0.25	Boedeker *et al.*, 1999
Hepatic lipase	397	0.25	Boedeker *et al.*, 1999
HIV SF2 GP120	57	0.75	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	99	0.75	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	124	0.75	Karlsson *et al.*, 1993; Zhu *et al.*, 2000

HIV SF2 GP120	128	0.75	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	160	0.75	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	170	0.25	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	203	0.25	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	214	0.25	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	235	0.25	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	249	0.75	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	262	0.25	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	268	0.25	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	274	0.75	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	304	0.25	Karlsson *et al.*, 1993; Zhu *et al.*, 2000

HIV SF2 GP120	311	0.25	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	328	0.25	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	334	0.25	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	358	0.25	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	364	0.25	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	370	0.25	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	378	0.25	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	428	0.75	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
HIV SF2 GP120	431	0.75	Karlsson *et al.*, 1993; Zhu *et al.*, 2000
IgG2 (heavy chain)	297	0.75	Lund *et al.*, 1993
IgG3 (heavy chain)	291	0.25	Lund *et al.*, 1993
IgG4 (heavy chain)	239	0.75	Lund *et al.*, 1993

Interferron gamma	26	0.75	Hooker *et al.*, 1995; James *et al.*, 1996; Gu *et al.*, 1997
Interleukin 4	53	0.75	Rajan *et al.*, 1995
Interleukin 4	98	0.75	Rajan *et al.*, 1995
Interleukin 4	134	0.75	Rajan *et al.*, 1995
Interleukin 4	176	0.75	Rajan *et al.*, 1995
Lecithin cholesterol acyltransferase	32	0.75	Schindler *et al.*, 1995
Lecithin cholesterol acyltransferase	96	0.75	Schindler *et al.*, 1995
Lecithin cholesterol acyltransferase	396	0.75	Schindler *et al.*, 1995
Merozoite surface protein 1	218	0.25	Yang *et al.*, 1999
Merozoite surface protein 1	258	0.25	Yang *et al.*, 1999
Merozoite surface protein 1	336	0.25	Yang *et al.*, 1999
Merozoite surface protein 1	816	0.25	Yang *et al.*, 1999
Merozoite surface protein 1	1076	0.25	Yang *et al.*, 1999
Merozoite surface protein 1	1167	0.25	Yang *et al.*, 1999
Merozoite surface protein 1	1350	0.25	Yang *et al.*, 1999
Merozoite surface protein 1	1634	0.25	Yang *et al.*, 1999
Merozoite surface protein 1	1083	0.25	Yang *et al.*, 1999
Merozoite surface protein 1	1570	0.25	Yang *et al.*, 1999
Merozoite surface protein 1	1466	0.25	Yang *et al.*, 1999
Pancreatic ribonuclease	34	0.75	Liang *et al*, 1980

Ribonuclease B	34	0.25	Liu *et al.*, 2001
Secreted alkaline phosphatase (SEAP)	271	0.75	Kaufmann *et al.*, 2001
Stanniocalcin	62	0.75	Zhang *et al.*, 1998
Thrombopoietin	197	0.75	Hoffman *et al.*, 1996; Inoue *et al.*, 1999
Thrombopoietin	234	0.75	Hoffman *et al.*, 1996; Inoue *et al.*, 1999
Thrombopoietin	255	0.75	Hoffman *et al.*, 1996; Inoue *et al.*, 1999
Thy-1	79	0.75	Williams *et al.*, 1993
Thy-1	119	0.25	Williams *et al.*, 1993
Thy-1 (mouse)	42	0.25	Williams *et al.*, 1993
Thy-1 (mouse)	94	0.75	Williams *et al.*, 1993
Thy-1 (mouse)	118	0.25	Williams *et al.*, 1993
Thy-1 (rat)	93	0.75	Williams *et al.*, 1993
Thy-1 (rat)	117	0.25	Williams *et al.*, 1993
Tissue inhibitor of metalloproteinases (TIMP)	53	0.75	Sutton *et al.*, 1994
Tissue Plasminogen Activator	117	0.25	Spellman *et al.*, 1989
Tissue Plasminogen Activator	184	0.75	Spellman *et al.*, 1989
Tissue Plasminogen Activator	448	0.75	Spellman *et al.*, 1989
TSL Lecithin	28	0.25	Liu *et al.*, 2001

Tumor necrosis factor receptor	83	0.75	Gawlitzek *et al.*, 2000
Tumor necrosis factor receptor	257	0.75	Gawlitzek *et al.*, 2000
Tumor necrosis factor receptor	278	0.75	Gawlitzek *et al.*, 2000
Tumor necrosis factor receptor	439	0.75	Gawlitzek *et al.*, 2000
Tumor necrosis factor receptor	573	0.75	Gawlitzek *et al.*, 2000

Table E.1. The entire glycosylation site reference set used for construction of neural network training and testing data sets for glycosylation microheterogeneity classification prediction.

Appendix F

COMPUTER PROGRAMS

Examples are given in this section of selected computer programs written and executed using the Matlab version 7.0.1 computer program. The programs include only the non-trivial programs written for applications in Chapter 3, Chapter 4 and Chapter 5. More information about functions not fully explained in this section, can be accessed by using the "help" command in Matlab. In Matlab notation, the (%) symbol denotes comments not executed by the program. Many programs make use of pre-defined raw matrix workspaces. These have been fully defined where appropriate. In addition, in many cases programs make reference to imported data sets. These were all compiled in a spreadsheet application prior to importing.

The create_mesh.m program creates the color three-dimensional contour surfaces observed in Chapter 3, Chapter 4 and Chapter 5. Corresponding data points were imported as a matrix of 3 columns, corresponding to x, y and z data points. The program was written to automatically perform interpolation of given data points.

```
%create_mesh.m
%the following program creates 3-D mesh countour plots observed
%in Chapter 3, 4 and 5
%the program creates a 3-D mesh surface and the plot is rotated
%to obtain the contour plot.

%import x by 3 data as 'clipboarddata'

x=clipboarddata(:,1);
```

```
y=clipboarddata(:,2);
z=clipboarddata(:,3);

%higher number=higher resolution
xres=100;
yres=100;

xmin=min(x);
xmax=max(x);
ymin=min(y);
ymax=max(y);

xv=linspace(xmin,xmax,xres);
yv=linspace(ymin,ymax,yres);

[Xinterp,Yinterp]=meshgrid(xv,yv);

Zinterp=griddata(x,y,z,Xinterp,Yinterp);

hf4=figure;

%can also use 'mesh' here:
surf(Xinterp,Yinterp,Zinterp);
colormap(jet);
colorbar;

%input desired axes labels:
xlabel('fglc');
ylabel('faa');
zlabel('Active tPA [IU]');

%use the property editor to enter the following values:
%Face Properties:
%Color: Mapped CData
%Lighting: Smoother Rounded (Phong)
%Transparency: 1.0 (opaque)
%Mesh Properties:
%Line Style: dotted line
%Line Width: 0.5
%Color: Black
%Mesh Style: Rows and Columns (both)
%Lighting: Smoother Rounded (Phong)
%Transparency: 1.0 (opaque)
%Marker Properties:
%Style: No Marker (none)
```

The following programs were executed in Chapter 3 in simulations of fed batch CHO cell culture in the presence of metabolite control by variable feed flow rates. The same program may be easily manipulated to obtain a program for fed batch simulations with fixed feed flow rates. Commonly, a controller program was written and executed in

addition to the major simulation program. The purpose of implementing the controller program was to enable the evaluation of multiple set points or fixed feed flow rates with minimal user interface, as the controller program executes the simulation program multiple times and saves results with separate file names. As an example of this application, the controller program runprogram1_fedbatch.m and major simulation program RungeKuttaCulture.m are displayed below.

```
%runprogram1_fedbatch.m

clear
load 'experimental data.mat'

%specify the total fed batch values:
%(list all combinations to be evaluated)

%glucose feed rate in litres/hour
glcsetpoint_total=[0.5 1.51 1 4.5];

%AA feed rate in litres/hour
aasetpoint_total=[0.5 1.18 2.7 0.5];

r=size(aasetpoint_total);
r=r(2);

for q=1:r

   aasetpoint=aasetpoint_total(:,q);
   glcsetpoint=glcsetpoint_total(:,q);

   Sglc=50;    %glucose concentration of glucose feed
   SAA=5;      %AA concentration of AA feed

   Sgln=SAA./4.5;  %gln concentration of AA feed
   Sasn=SAA-Sgln;  %asn concentration of AA feed

   RungeKuttaCulture

   s3=['results' int2str(q) 'runprogram1_fedbatch.mat'];
   save(s3, 'X', 't', 'maxtpa', 'glcsetpoint', 'aasetpoint', '-v4');

   q

end
```

```
%RungeKuttaCulture
%Solves coupled ode's for CHO cell batch/fed batch culture for given feed
%flow rates of glucose and amino acids feeds using a 4th order Runge Kutta
%numerical method.
```

```
%This method assumes glc and AA concentrations are controled at a given setpoint

disp('running simulation')

clear t X maxtpa

h=0.05;  %defines time step

t(1)=0; %initiate simulation at 0 hours

%specify the initial fed batch values:

FGLC=0;  %glucose feed rate in litres/hour
FAA=0;  %AA feed rate in litres/hour

F=FGLC+FAA; %total feed flow rate for volume calculation

%constants and model parameters:
%cell growth model:
%uint=(umax.*GLC.*AA)./((GLC+Kglc).*(AA+Kaa).*(GLC./Kdglc+1).*(LAC./Kilac+1).*(AM./Kiam+1
));
%where AA=Gln+Asn
umax=0.23;
Kglc=2.5;
Kdglc=10;
Kaa=1.5;
Kilac=14.5;
Kiam=3.5;

%cell death model:
%kd=(kdmax.*I)./(Kdi+I+Kdi./Kds.*S)
%where I=lac+Am and S=Glc+Gln+Asn
kdmax=0.008;
Kdi=0.625;
Kds=6.5;

%yield coefficients:
Yxglc=11.11;
Yxgln=190;
Yxasn=94.76;
Yxlac=24.08;
Yxam=1300;
alpha=0.115;
beta=0.13;

%cell lysis rate constant:
kl=9.0e-4;

%glutamine degradation rate constant:
kdeg=3e-4;

%tPA glycoform natrual degradation rate constants:
kI0=1.74e5;
kII0=3.48e5;

%tPA glycoform glycation degradation rate constants:
```

```
kIg=2;
kIIg=2;

%molecular weight values:
MWTypeI=73000;
MWTypeII=70000;
MWglc=180.16;
MWgln=146;

%tPA glycoform specific activities:
ActTypeI=363;  %IU/ug
ActTypeII=459;

%initial values:
xt(1)=rawdata(1,6);       %intrinsic total cell density
xv(1)=rawdata(1,2);       %viable cell density
xd(1)=rawdata(1,7);       %intrinsic dead cell density
xl(1)=rawdata(1,5);       %lysed cell density
glc(1)=rawdata(1,8);      %glucose concentration
gln(1)=rawdata(1,9);      %glutamine concentration
asn(1)=rawdata(1,10);     %asparagine concentration
lac(1)=rawdata(1,11);     %lactate concentration
am(1)=rawdata(1,12);      %total (protonated + unprotonated) ammonia concentration
typeIa(1)=rawdata(1,13);  %active type I tPA concentration
typeIIa(1)=rawdata(1,14); %active type II tPA concentration
typeIt(1)=rawdata(1,13);  %total type I tPA concentration
typeIIt(1)=rawdata(1,14); %total type II tPA concentration
v(1)=1.5;                 %reactor volume
AtPA(1)=0;                %total active tPA (IU)

%final time cutoff for simulation:
tfinal2=500;

%volume restriction:
vfinal=2;

%creation of a vector for variable updating:
X=[xt; xv; xd; xl; glc; gln; asn; lac; am; typeIa; typeIIa; typeIt; typeIIt; v; AtPA];

a=1;
while t(a)<=tfinal2

   a=size(t);
   a=a(2);

   %calculate k-coefficients for runge-kutta numerical method

   %k1 coefficients:
   XT=X(1,a);
   XV=X(2,a);
   XD=X(3,a);
   XL=X(4,a);
   GLC=X(5,a);
   GLN=X(6,a);
   ASN=X(7,a);
   LAC=X(8,a);
```

```
AM=X(9,a);
TYPEIA=X(10,a);
TYPEIIA=X(11,a);
TYPEIT=X(12,a);
TYPEIIT=X(13,a);
V=X(14,a);

AA=GLN+ASN;
S=GLC+GLN+ASN;
I=LAC+AM;


uint=(umax.*GLC.*AA)./((GLC+Kglc).*(AA+Kaa).*(GLC./Kdglc+1).*(LAC./Kilac+1).*(AM./Kiam+1));
kd=(kdmax.*I)./(Kdi+I+Kdi./Kds.*S);

%Fed Batch feed flow rates:

if X(5,a)<=glcsetpoint
  FGLC=(FAA.*GLC+V./Yxglc.*uint.*XV)./(Sglc-GLC);
else
  FGLC=0;
end

if (X(6,a)+X(7,a))<=aasetpoint
  FAA=(FGLC.*(AA)+V.*kdeg.*GLN+V.*(1./Yxgln+1./Yxasn).*uint.*XV)./((SAA-AA));
else
  FAA=0;
end

%repeat of the iterative process for flow rate calculation:

if X(5,a)<=glcsetpoint
  FGLC=(FAA.*GLC+V./Yxglc.*uint.*XV)./(Sglc-GLC);
else
  FGLC=0;
end

if (X(6,a)+X(7,a))<=aasetpoint
  FAA=(FGLC.*(AA)+V.*kdeg.*GLN+V.*(1./Yxgln+1./Yxasn).*uint.*XV)./((SAA-AA));
else
  FAA=0;
end

%volume restriction:

if V>=vfinal
  FGLC=0;
  FAA=0;
end

F=FGLC+FAA;

kxt1=uint.*XV-(F./V).*XT;
kxv1=(uint-kd).*XV-(F./V).*XV;
kxd1=kd.*XV-(F./V).*XD;
kxl1=kl.*(XT-XL)-(F./V).*XL;
```

```
   kglc1=(FGLC./V).*Sglc-1./Yxglc.*uint.*XV-(F./V).*GLC;
   kgln1=(FAA./V).*Sgln-1./Yxgln.*uint.*XV-kdeg.*GLN-(F./V).*GLN;
   kasn1=(FAA./V).*Sasn-1./Yxasn.*uint.*XV-(F./V).*ASN;
   klac1=1./Yxlac.*uint.*XV-(F./V).*LAC;
   kam1=1./Yxam.*uint.*XV+kdeg.*GLN-(F./V).*AM;
   ktypeIa1=alpha.*uint.*XV-(MWTypeI.*1e3).*kI0.*(TYPEIA./(MWTypeI.*1e3)).^2-
(MWTypeI.*1e3).*kIg.*TYPEIA.*GLC./(MWglc.*MWTypeI.*1e3)-(F./V).*TYPEIA;
   ktypeIIa1=beta.*uint.*XV-(MWTypeII.*1e3).*kII0.*(TYPEIIA./(MWTypeII.*1e3)).^2-
(MWTypeII.*1e3).*kIIg.*TYPEIIA.*GLC./(MWglc.*MWTypeII.*1e3)-(F./V).*TYPEIIA;
   ktypeIt1=alpha.*uint.*XV-(F./V).*TYPEIT;
   ktypeIIt1=beta.*uint.*XV-(F./V).*TYPEIIT;
   kv1=F;

   %k2 coefficients:
   XT=X(1,a)+0.5.*h.*kxt1;
   XV=X(2,a)+0.5.*h.*kxv1;
   XD=X(3,a)+0.5.*h.*kxd1;
   XL=X(4,a)+0.5.*h.*kxl1;
   GLC=X(5,a)+0.5.*h.*kglc1;
   GLN=X(6,a)+0.5.*h.*kgln1;
   ASN=X(7,a)+0.5.*h.*kasn1;
   LAC=X(8,a)+0.5.*h.*klac1;
   AM=X(9,a)+0.5.*h.*kam1;
   TYPEIA=X(10,a)+0.5.*h.*ktypeIa1;
   TYPEIIA=X(11,a)+0.5.*h.*ktypeIIa1;
   TYPEIT=X(12,a)+0.5.*h.*ktypeIt1;
   TYPEIIT=X(13,a)+0.5.*h.*ktypeIIt1;
   V=X(14,a)+0.5.*h.*kv1;

   AA=GLN+ASN;
   S=GLC+GLN+ASN;
   I=LAC+AM;


uint=(umax.*GLC.*AA)./((GLC+Kglc).*(AA+Kaa).*(GLC./Kdglc+1).*(LAC./Kilac+1).*(AM./Kiam+1));
   kd=(kdmax.*I)./(Kdi+I+Kdi./Kds.*S);

      %Fed Batch feed flow rates:

   if X(5,a)<=glcsetpoint
      FGLC=(FAA.*GLC+V./Yxglc.*uint.*XV)./(Sglc-GLC);
   else
      FGLC=0;
   end

   if (X(6,a)+X(7,a))<=aasetpoint
      FAA=(FGLC.*(AA)+V.*kdeg.*GLN+V.*(1./Yxgln+1./Yxasn).*uint.*XV)./((SAA-AA));
   else
      FAA=0;
   end

   %repeat of the iterative process for flow rate calculation:

   if X(5,a)<=glcsetpoint
      FGLC=(FAA.*GLC+V./Yxglc.*uint.*XV)./(Sglc-GLC);
   else
```

```
    FGLC=0;
  end

  if (X(6,a)+X(7,a))<=aasetpoint
    FAA=(FGLC.*(AA)+V.*kdeg.*GLN+V.*(1./Yxgln+1./Yxasn).*uint.*XV)./((SAA-AA));
  else
    FAA=0;
  end

  %volume restriction:

  if V>=vfinal
    FGLC=0;
    FAA=0;
  end

  kxt2=uint.*XV-(F./V).*XT;
  kxv2=(uint-kd).*XV-(F./V).*XV;
  kxd2=kd.*XV-(F./V).*XD;
  kxl2=kl.*(XT-XL)-(F./V).*XL;
  kglc2=(FGLC./V).*Sglc-1./Yxglc.*uint.*XV-(F./V).*GLC;
  kgln2=(FAA./V).*Sgln-1./Yxgln.*uint.*XV-kdeg.*GLN-(F./V).*GLN;
  kasn2=(FAA./V).*Sasn-1./Yxasn.*uint.*XV-(F./V).*ASN;
  klac2=1./Yxlac.*uint.*XV-(F./V).*LAC;
  kam2=1./Yxam.*uint.*XV+kdeg.*GLN-(F./V).*AM;
  ktypeIa2=alpha.*uint.*XV-(MWTypeI.*1e3).*kI0.*(TYPEIA./(MWTypeI.*1e3)).^2-
(MWTypeI.*1e3).*kIg.*TYPEIA.*GLC./(MWglc.*MWTypeI.*1e3)-(F./V).*TYPEIA;
  ktypeIIa2=beta.*uint.*XV-(MWTypeII.*1e3).*kII0.*(TYPEIIA./(MWTypeII.*1e3)).^2-
(MWTypeII.*1e3).*kIIg.*TYPEIIA.*GLC./(MWglc.*MWTypeII.*1e3)-(F./V).*TYPEIIA;
  ktypeIt2=alpha.*uint.*XV-(F./V).*TYPEIT;
  ktypeIIt2=beta.*uint.*XV-(F./V).*TYPEIIT;
  kv2=F;

  %k3 coefficients:
  XT=X(1,a)+0.5.*h.*kxt2;
  XV=X(2,a)+0.5.*h.*kxv2;
  XD=X(3,a)+0.5.*h.*kxd2;
  XL=X(4,a)+0.5.*h.*kxl2;
  GLC=X(5,a)+0.5.*h.*kglc2;
  GLN=X(6,a)+0.5.*h.*kgln2;
  ASN=X(7,a)+0.5.*h.*kasn2;
  LAC=X(8,a)+0.5.*h.*klac2;
  AM=X(9,a)+0.5.*h.*kam2;
  TYPEIA=X(10,a)+0.5.*h.*ktypeIa2;
  TYPEIIA=X(11,a)+0.5.*h.*ktypeIIa2;
  TYPEIT=X(12,a)+0.5.*h.*ktypeIt2;
  TYPEIIT=X(13,a)+0.5.*h.*ktypeIIt2;
  V=X(14,a)+0.5.*h.*kv2;

  AA=GLN+ASN;
  S=GLC+GLN+ASN;
  I=LAC+AM;


uint=(umax.*GLC.*AA)./((GLC+Kglc).*(AA+Kaa).*(GLC./Kdglc+1).*(LAC./Kilac+1).*(AM./Kiam+1));
  kd=(kdmax.*I)./(Kdi+I+Kdi./Kds.*S);
```

```
    %Fed Batch feed flow rates:

  if X(5,a)<=glcsetpoint
    FGLC=(FAA.*GLC+V./Yxglc.*uint.*XV)./(Sglc-GLC);
  else
    FGLC=0;
  end

  if (X(6,a)+X(7,a))<=aasetpoint
    FAA=(FGLC.*(AA)+V.*kdeg.*GLN+V.*(1./Yxgln+1./Yxasn).*uint.*XV)./((SAA-AA));
  else
    FAA=0;
  end

  %repeat of the iterative process for flow rate calculation:

  if X(5,a)<=glcsetpoint
    FGLC=(FAA.*GLC+V./Yxglc.*uint.*XV)./(Sglc-GLC);
  else
    FGLC=0;
  end

  if (X(6,a)+X(7,a))<=aasetpoint
    FAA=(FGLC.*(AA)+V.*kdeg.*GLN+V.*(1./Yxgln+1./Yxasn).*uint.*XV)./((SAA-AA));
  else
    FAA=0;
  end

  %volume restriction:

  if V>=vfinal
    FGLC=0;
    FAA=0;
  end

  kxt3=uint.*XV-(F./V).*XT;
  kxv3=(uint-kd).*XV-(F./V).*XV;
  kxd3=kd.*XV-(F./V).*XD;
  kxl3=kl.*(XT-XL)-(F./V).*XL;
  kglc3=(FGLC./V).*Sglc-1./Yxglc.*uint.*XV-(F./V).*GLC;
  kgln3=(FAA./V).*Sgln-1./Yxgln.*uint.*XV-kdeg.*GLN-(F./V).*GLN;
  kasn3=(FAA./V).*Sasn-1./Yxasn.*uint.*XV-(F./V).*ASN;
  klac3=1./Yxlac.*uint.*XV-(F./V).*LAC;
  kam3=1./Yxam.*uint.*XV+kdeg.*GLN-(F./V).*AM;
  ktypeIa3=alpha.*uint.*XV-(MWTypeI.*1e3).*kI0.*(TYPEIA./(MWTypeI.*1e3)).^2-
(MWTypeI.*1e3).*kIg.*TYPEIA.*GLC./(MWglc.*MWTypeI.*1e3)-(F./V).*TYPEIA;
  ktypeIIa3=beta.*uint.*XV-(MWTypeII.*1e3).*kII0.*(TYPEIIA./(MWTypeII.*1e3)).^2-
(MWTypeII.*1e3).*kIIg.*TYPEIIA.*GLC./(MWglc.*MWTypeII.*1e3)-(F./V).*TYPEIIA;
  ktypeIt3=alpha.*uint.*XV-(F./V).*TYPEIT;
  ktypeIIt3=beta.*uint.*XV-(F./V).*TYPEIIT;
  kv3=F;

  %k4 coefficients:
  XT=X(1,a)+h.*kxt3;
  XV=X(2,a)+h.*kxv3;
```

```
XD=X(3,a)+h.*kxd3;
XL=X(4,a)+h.*kxl3;
GLC=X(5,a)+h.*kglc3;
GLN=X(6,a)+h.*kgln3;
ASN=X(7,a)+h.*kasn3;
LAC=X(8,a)+h.*klac3;
AM=X(9,a)+h.*kam3;
TYPEIA=X(10,a)+h.*ktypeIa3;
TYPEIIA=X(11,a)+h.*ktypeIIa3;
TYPEIT=X(12,a)+h.*ktypeIt3;
TYPEIIT=X(13,a)+h.*ktypeIIt3;
V=X(14,a)+h.*kv3;

AA=GLN+ASN;
S=GLC+GLN+ASN;
I=LAC+AM;


uint=(umax.*GLC.*AA)./((GLC+Kglc).*(AA+Kaa).*(GLC./Kdglc+1).*(LAC./Kilac+1).*(AM./Kiam+1));
kd=(kdmax.*I)./(Kdi+I+Kdi./Kds.*S);

   %Fed Batch feed flow rates:

if X(5,a)<=glcsetpoint
   FGLC=(FAA.*GLC+V./Yxglc.*uint.*XV)./(Sglc-GLC);
else
   FGLC=0;
end

if (X(6,a)+X(7,a))<=aasetpoint
   FAA=(FGLC.*(AA)+V.*kdeg.*GLN+V.*(1./Yxgln+1./Yxasn).*uint.*XV)./((SAA-AA));
else
   FAA=0;
end

%repeat of the iterative process for flow rate calculation:

if X(5,a)<=glcsetpoint
   FGLC=(FAA.*GLC+V./Yxglc.*uint.*XV)./(Sglc-GLC);
else
   FGLC=0;
end

if (X(6,a)+X(7,a))<=aasetpoint
   FAA=(FGLC.*(AA)+V.*kdeg.*GLN+V.*(1./Yxgln+1./Yxasn).*uint.*XV)./((SAA-AA));
else
   FAA=0;
end

%volume restriction:

if V>=vfinal
   FGLC=0;
   FAA=0;
end
```

```
   kxt4=uint.*XV-(F./V).*XT;
   kxv4=(uint-kd).*XV-(F./V).*XV;
   kxd4=kd.*XV-(F./V).*XD;
   kxl4=kl.*(XT-XL)-(F./V).*XL;
   kglc4=(FGLC./V).*Sglc-1./Yxglc.*uint.*XV-(F./V).*GLC;
   kgln4=(FAA./V).*Sgln-1./Yxgln.*uint.*XV-kdeg.*GLN-(F./V).*GLN;
   kasn4=(FAA./V).*Sasn-1./Yxasn.*uint.*XV-(F./V).*ASN;
   klac4=1./Yxlac.*uint.*XV-(F./V).*LAC;
   kam4=1./Yxam.*uint.*XV+kdeg.*GLN-(F./V).*AM;
   ktypeIa4=alpha.*uint.*XV-(MWTypeI.*1e3).*kI0.*(TYPEIA./(MWTypeI.*1e3)).^2-
(MWTypeI.*1e3).*kIg.*TYPEIA.*GLC./(MWglc.*MWTypeI.*1e3)-(F./V).*TYPEIA;
   ktypeIIa4=beta.*uint.*XV-(MWTypeII.*1e3).*kII0.*(TYPEIIA./(MWTypeII.*1e3)).^2-
(MWTypeII.*1e3).*kIIg.*TYPEIIA.*GLC./(MWglc.*MWTypeII.*1e3)-(F./V).*TYPEIIA;
   ktypeIt4=alpha.*uint.*XV-(F./V).*TYPEIT;
   ktypeIIt4=beta.*uint.*XV-(F./V).*TYPEIIT;
   kv4=F;

   %calculate the new values:

   X(1,(a+1))=X(1,a)+h./6.*(kxt1+2.*kxt2+2.*kxt3+kxt4);
   X(2,(a+1))=X(2,a)+h./6.*(kxv1+2.*kxv2+2.*kxv3+kxv4);
   X(3,(a+1))=X(3,a)+h./6.*(kxd1+2.*kxd2+2.*kxd3+kxd4);
   X(4,(a+1))=X(4,a)+h./6.*(kxl1+2.*kxl2+2.*kxl3+kxl4);
   X(5,(a+1))=X(5,a)+h./6.*(kglc1+2.*kglc2+2.*kglc3+kglc4);
   X(6,(a+1))=X(6,a)+h./6.*(kgln1+2.*kgln2+2.*kgln3+kgln4);
   X(7,(a+1))=X(7,a)+h./6.*(kasn1+2.*kasn2+2.*kasn3+kasn4);
   X(8,(a+1))=X(8,a)+h./6.*(klac1+2.*klac2+2.*klac3+klac4);
   X(9,(a+1))=X(9,a)+h./6.*(kam1+2.*kam2+2.*kam3+kam4);
   X(10,(a+1))=X(10,a)+h./6.*(ktypeIa1+2.*ktypeIa2+2.*ktypeIa3+ktypeIa4);
   X(11,(a+1))=X(11,a)+h./6.*(ktypeIIa1+2.*ktypeIIa2+2.*ktypeIIa3+ktypeIIa4);
   X(12,(a+1))=X(12,a)+h./6.*(ktypeIt1+2.*ktypeIt2+2.*ktypeIt3+ktypeIt4);
   X(13,(a+1))=X(13,a)+h./6.*(ktypeIIt1+2.*ktypeIIt2+2.*ktypeIIt3+ktypeIIt4);
   X(14,(a+1))=X(14,a)+h./6.*(kv1+2.*kv2+2.*kv3+kv4);

   X(15,(a+1))=X(10,(a+1)).*ActTypeI.*X(14,(a+1)).*1000+X(11,(a+1)).*ActTypeII.*X(14,(a+1)).*1000;

   %forcing functions (lower limits):
   for j=1:14
      if X(j,(a+1))<0
         X(j,(a+1))=0;
      end
   end

   %update the time step:

   t(a+1)=t(a)+h;

end

%optimum value:
maxtpa=max(X(15,:));

%make plots:
figure(1)
subplot(4,1,1)
% errorbar(rawdata(:,1),rawdata(:,6),errordata(:,6),'o')
```

```
% hold on
plot(t,X(1,:))
ylabel('X_t')
subplot(4,1,2)
% errorbar(rawdata(:,1),rawdata(:,2),errordata(:,2),'o')
% hold on
plot(t,X(2,:))
ylabel('X_{v,app}')
subplot(4,1,3)
% errorbar(rawdata(:,1),rawdata(:,7),errordata(:,7),'o')
% hold on
plot(t,X(3,:))
ylabel('X_d')
subplot(4,1,4)
% errorbar(rawdata(:,1),rawdata(:,5),errordata(:,5),'o')
% hold on
plot(t,X(4,:))
ylabel('X_l')
xlabel('time [hours]')

figure(2)
subplot(3,1,1)
% errorbar(rawdata(:,1),rawdata(:,8),errordata(:,8),'o')
% hold on
plot(t,X(5,:))
ylabel('glucose')
subplot(3,1,2)
% errorbar(rawdata(:,1),rawdata(:,9),errordata(:,9),'o')
% hold on
plot(t,X(6,:))
ylabel('L-glutamine')
subplot(3,1,3)
% errorbar(rawdata(:,1),rawdata(:,10),errordata(:,10),'o')
% hold on
plot(t,X(7,:))
ylabel('L-asparagine')
xlabel('time [hours]')

figure(3)
subplot(2,1,1)
% errorbar(rawdata(:,1),rawdata(:,11),errordata(:,11),'o')
% hold on
plot(t,X(8,:))
ylabel('lactate')
subplot(2,1,2)
% errorbar(rawdata(:,1),rawdata(:,12),errordata(:,12),'o')
% hold on
plot(t,X(9,:))
ylabel('total ammonia')
xlabel('time [hours]')

figure(4)
subplot(2,1,1)
% errorbar(rawdata(:,1),rawdata(:,13),errordata(:,13),'o')
% hold on
plot(t,X(12,:))
```

```
ylabel('total Type I r-tPA')
subplot(2,1,2)
% errorbar(rawdata(:,1),rawdata(:,14),errordata(:,14),'o')
% hold on
plot(t,X(13,:))
ylabel('total Type II r-tPA')
xlabel('time [hours]')

figure(5)
subplot(2,1,1)
plot(t,X(10,:))
ylabel('active Type I r-tPA')
subplot(2,1,2)
plot(t,X(11,:))
ylabel('active Type II r-tPA')
xlabel('time [hours]')

figure(6)
subplot(2,1,1)
plot(t,X(14,:))
ylabel('reactor volume')
subplot(2,1,2)
plot(t,X(15,:))
ylabel('total active r-tPA [IU]')
xlabel('time [hours]')
```

Similarly, this set-up of an execution program governed by a controller program was used with neural networks, particularly when evaluating different lengths of the glycosylation window. The following programs runprogram1_nn.m and rnetprogram.m were used in evaluation of the glycosylation window for the primary sequence data inputs. A saved workspace was loaded for each case of the cross-validation procedure, and information was contained for the entire (n-20) to (n+20) glycosylation window. First, the glycosylation window was modified and the correct number of hidden layer neurons were specified. This number was calculated by correlations performed outside of the program itself. The input workspace consisted of training and testing input data contained as separate variables. In addition, the target vectors for input and testing data sets were also specified as separate variables. These variables are specified in rnetprogram.m program code.

```
%runprogram1_nn.m

%input the corresponding data set:
%(different data sets were defined and saved accoring
%to the cross-validation study)

load start13.m;
dataset=['start13'];

%input the starting residues in stotal:
stotal=[-10      -10      -10      -10      -10      -10];

%input the ending residues in etotal:
etotal=[20       18       16       14       12       10];

%input the number of hidden layer neurons in n1total:
n1total=[24      23       21       18       14       9];

sizetotal=size(stotal);
sizetotal=sizetotal(2);

for beta=1:sizetotal

   s=stotal(beta);
   e=etotal(beta);
   n1=n1total(beta);

   rnetprogram

   %simulating a simple classification to classify
   %recurrent netowrk output values:
   Y4=Y3;
   for i=1:50
   for t=1:10
      if Y3(i,t)<0.5
         Y4(i,t)=0.25;
      end
      if Y3(i,t)>=0.5
         Y4(i,t)=0.75;
      end
   end
   end


   for i=1:50
      e4(i,:)=Y4(i,:)-TestT1;
   end

   for i=1:50
      mse4(i,:)=mse(e4(i,:));
   end

   s3=['n' int2str(s) 'n' int2str(e) 'x' int2str(n1) '_' description '_01.mat'];
   save(s3, 'mse1', 'mse2', 'mse4', 'e1', 'e2', 'e4', 'Y3', 'Y4', 'w1', '-v4');

   disp('workspace saved')
```

```
  disp(s3)

end
```

```
%rnetprogram
%for recurrent neural networks of primary data

clear i Y1 Y2 Y3 Y4

%primary1all refers to the training data set
primary1=primary1all((20-s+1):(41-(20-e)),:);

%Testprimary1 refers to the testing data set
Testprimary1=Testprimary1all((20-s+1):(41-(20-e)),:);

primary=primary1;
Testprimary=Testprimary1;

%Tset1 refers to the training target set
Tset=Tset1;

%TestT1 refers to the testing target set
TestT=TestT1;

size1=size(primary1);
size1=size1(1);
clear a

%sets the limits of the input vectors
for i=1:size1
  a(i,:)=[0 11];
end
clear size1

%performs 100 independent iterations of network initation/training
j=100;

for i=1:j
  %recurrent network
  rnet1p=newelm([a],[n1 1],{'tansig','logsig'});
  rnet1p.trainParam.epochs=2000;
  rnet1p.trainParam.goal=1e-6;
  rnet1p=train(rnet1p,primary,Tset);

  Y1=sim(rnet1p,primary);
  e1(i,:)=Tset-Y1;

  mse1(i,:)=mse(e1(i,:));

  Y2=sim(rnet1p,Testprimary);
  e2(i,:)=TestT-Y2;

  mse2(i,:)=mse(e2(i,:));

  Y3(i,:)=Y2;
```

```
    X1=getx(rnet1p);
    w1(:,i)=X1;

    J1=1:i;
    [J1' mse1 mse2]
    disp('training primary network')
end
```