RESEARCH ARTICLE

The operating space design of N-linked glycosylation based on PLS inversion

Jingjing Zhang[1], Wei Xi[1], Lei Wang[1], Liming Liu[2, *]

[1]School of Electrical Engineering, Hebei University of Architecture, Zhangjiakou, Hebei, China. [2]Sany Zhangjiakou Wind Power Technology Co., Ltd., Zhangjiakou, Hebei, China

Received: July 14, 2023; **accepted:** August 7, 2023.

With the continuous development of biopharmaceutical market, monoclonal antibody drugs are becoming more and more popular for their specificity, efficiency, targeting, and other advantages. As one of the most important parts in producing monoclonal antibody, the glycan distribution in the N-linked glycosylation process has a very critical influence on the final glycoprotein products. However, N-linked glycosylation is a non-template driven cellular process, which is an extremely complex and nonlinear biochemical reaction system. It's a great challenge to achieve a precise glycan distribution during manufacturing. Therefore, glycosylation control online and glycan distribution prediction have become a hot issue at present in related fields of research. In order to analyze the glycan distributions in quantitative perspective and get the precise glycan prediction values, a novel method of multivariate statistical regression of partial least square (PLS) inversion was developed to predict the desired glycans levels. The operating space design of N-linked glycosylation had been achieved. The PLS model with the data produced by the reaction network was established to achieve the inner relations between the input variable enzymes and the output of glycans. The PLS inversion model was then built to predict the desired 11 kinds of glycans corresponding to the most possible operating range of the input enzymes. The method might provide a foundation for controlling glycosylation on-line and accurate prediction in the future.

Keywords: operating space design; glycosylation; PLS inversion; prediction uncertainty.

[*]**Corresponding author:** Liming Liu, Sany Zhangjiakou Wind Power Technology Co., Ltd., Zhangjiakou 075000, Hebei, China. Email: liulim0216@163.com.

Introduction

With a huge market exceeding $99 billion in 2011 and an expected steady increase in future sales, the increasingly high requirement of biopharmaceutical production and properties of protein are becoming more and more popular [1, 2]. Although many factors may affect the proteins quality and characteristics, one of the most important processes is glycosylation, which is a post-translation modification within the endoplasmic reticulum (ER) of a cell. Glycosylation will construct the structures of proteins and regulate protein functions. Usually, a glycan or carbohydrate chain is added to proteins to rebuild numerous new structures. In reality, the newly formed structure of the glycoproteins may be abundant and uncertain for randomness [3]. From this point, analysis of glycosylation controllability is necessary, which may provide a guiding role to get the desired glycan distribution. Besides, the United States Food and Drug Administration and Pharmaceutical Associations in Europe have made the implementation of quality by design (QbD) paradigm to regulate the

biopharmaceutical production, which proposes biopharmaceutical manufacturers to control glycosylation on-line [4]. However, at present, the on-line glycosylation control, especially the desired glycan states prediction, still has been a challenge work [5, 6]. The related technique skills and the theoretical basis need to be developed to get rid of relying on a large number of experiments for obtaining the desired results. The implementation of QbD strategy requires to establish the inner relations between the quality of the pharmaceutical products and the multiple input variables. And finally, according to the determination of the operating design space, selecting the appropriate manipulated variables is needed to achieve a new desired quality of products.

The definition of the operating design space is "the multidimensional combination and interaction of input variables (e.g., material attributes) of the process parameters that have been demonstrated to provide assurance of quality" [7]. Generally, the operating design space is hard to determine using a first-principal model, most of which rely on heavy experiments [8]. For glycosylation process, to achieve the operating design space corresponding to the desired glycan states may consume enormous resources. Model predictive control is a closed-loop optimization control strategy based on model. The core of its algorithm is the dynamic model that can predict the future, the control action that can be repeatedly optimized and implemented online, and the feedback correction of model errors. Model predictive control has the advantages of good control effect and strong robustness, which can effectively overcome the uncertainty, nonlinearity, and parallelism of the process, and can conveniently deal with various constraints in the controlled and manipulated variables of the process. Starting from the basic principle of model predictive control, there are three common predictive control algorithms including (1) model predictive control based on non-parametric models; (2) predictive control algorithms based on input and output parameterized models such as ARMA or CARIMA; and (3) rolling time domain control developed from linear–quadratic (LQ) and linear quadratic Gaussian (LQG) algorithms. Comparing to the above methods, partial least square (PLS) has fewer sample requirements, does not need to analyze data to conform to normal distribution, can handle complex structural models with multiple facets, can handle both reflection and formation indicators, and is especially suitable for prediction. In addition, PLS is also an effective tool to examine whether causality is significant. PLS can simultaneously achieve regression modeling (multiple linear regression analysis), data structure simplification (principal component analysis), correlation analysis between two groups of variables (canonical correlation analysis), and other three functions. It is the most used method to estimate the parameters of structural equation models with hidden variables. PLS can better solve many problems that cannot be solved by traditional multiple regression method, which can be said to be a leap in multivariate statistical analysis. By comparing with the experimental data, the PLS results show that the method can effectively find the appropriate control variables and feasible operation space. In order to achieve the historical data set, a precise glycosylation process model is critical. So far, there is no report of an accurate mathematical model of N-linked glycosylation in the cell culture environment. Several typical models have been developed in the past decades. A mathematical model was put forward in 1997, which constructed a complicated glycosylation reaction network based on a series of biochemical reactions rules [9]. Through computer simulation, 33 kinds of reactions were included and dozens of N-linked glycan were obtained. In 2005, another superior model of N-linked glycosylation was developed, which expanded the reaction network and added many more rules. As a result, exceeding 2,000 kinds of reactions were taken into account. At the same time, more than 7,000 kinds of N-linked glycan were produced [10]. Then, some other glycosylation process models were subsequently to be presented such as a dynamic mathematical

model for monoclonal antibody N-linked glycosylation in 2011 and so on. At present, most researchers focus on the extracellular environment that affects the glycosylation process such as the composition of the cell culture media. Usually, the method in macro aspect cannot understand the internal reactions deeply. The reactions within the cells may provide the most intuitive relations of the glycan quality, which could provide a guide to achieve online control.

In this study, a methodology was proposed to find the appropriate input variables and their ranges to produce the 11 kinds of glycan states. The method mainly relied on the historical experimental data to develop the new quality products, which was based on the multivariate statistical regression of partial least square (PLS). The N-linked glycosylation reaction network model was applied to generate necessary historical data, which were separated into training data set and the test data set to build the PLS inversion model to predict the proper manipulated variable domain corresponding to the desired glycans.

Materials and Methods

N-linked glycosylation and the data set selection

In order to achieve the historical data set, the model of glycosylation based on reactions network which was put forward in 2005 was employed. The inputs were the intra-Golgi concentrations of 11 glycosylases, while the outputs were the 11 kinds of glycan classes. All of the variables were listed in Table 1. For each input glycosylation enzyme, there was a concentration range marked by a low value and a high value. The outputs *S0* - *S4*, *G0* - *G3*, and *F0* - *F1* represented the number of sialic acid, galactose, and fucose molecules presented in the glycoform, respectively. For example, *S2* indicated that two more sialic acid molecules were attached to the glycoform.

Table 1. The input and output variables.

Input (Enzymes)	Low (μM)	High (μM)	Output (Glycan classes)
ManI	0.89	2.67	S0
ManII	0.66	1.98	S1
Fuct	1.25	3.75	S2
GnTI	1.52	4.57	S3
GnTII	0.64	1.93	S4
GnTIII	0.55	1.65	G0
GnTIV	1.18	5.43	G1
GnTV	0.20	0.60	G2
GnTE	1.735	5.20	G3
Galt	0.33	0.99	F0
SiaT	0.50	1.50	F1

The N-linked glycosylation network model was described by a series of sequential enzymatic reactions and kinetic equations. All the enzymatic reactions followed the appropriate reaction rules, which suggested how these glycosylation enzymes acted upon each glycoform. Usually, the transfer process of glycoprotein in the Golgi was modeled as four compartments with each treated like a well-mixed reactor. In this study, the final steady state was only considered. The glycans satisfied the material balance equation:

$$P_{ij} = P_{ij-1} + \tau_j r_{ij} \quad (1)$$

where P_{ij} was concentration of different glycan i in the compartment j of Golgi. τ_j was residence time. r_{ij} was the production rate of glycans. r_{ij} satisfied the following equation:

$$r_{ij} = \frac{k_f[E_t][UDP-S][p_{ij}]}{K_m(K_{md} + [UDP-S](1+\sum \frac{P_{ij}}{K_m})} \quad (2)$$

where k_f, K_m, and K_{md} were the enzyme kinetic constants. UDP-S was the concentration of nucleotide sugar donors. E_t was the enzyme concentration. Because only the steady state was considered, we simply defined the rate $r_i = 0$. In this case, the equation was transferred to the problem that solved the concentrations of the glycans P_{ij}.

Because all the related enzymatic reactions are regulated by a set of reaction rules, meanwhile, the whole reaction network will become larger and more complex. Actually, there would be huge changes in the concentrations of the glycans with many more glycosylation enzymes being added continuously. Thus, in the steady state, the production of glycans might be random, which was hard to identify which kind of glycosylation enzyme would act upon on them. It was even difficult to find out clearly relations between the glycoform and the input variables of glycosylation enzymes. In order to achieve the desired glycan values, the manipulated variable of input and the output glycans model was built in PLS first. Then, the latent variables model was inversed to identify the operating space of the input variables corresponding to the desired glycans. Enough original data was needed to build the multivariable regression model. Based on the reaction network model, we randomly generated 52 sets of the 11 input variables data, which were all limited within the ranges marked by the low and high values shown in Table 1. In this case, all the input variables were ensured in the original experiment ranges. According to the complex reaction rules, 52 sets of outputs of 11 glycans were obtained. A multivariate statistical regression method of PLS was used to establish the mapping relation between the input and output data. Then, the PLS inversion model was used to predict the appropriate operating space of the desired glycans.

PLS modeling and inversion

PLS is a multivariable regression method which is used to build a model of correlating the input data and the output data. For a given input matrix *X (N×M)* and a response matrix *Y (N×k)*, PLS model structure could be shown as follows:

$$X = \sum_{a=1}^{A} t_a p_a^T + \sum_{a=A+1}^{R} t_a p_a^T = TP^T + E \quad (3)$$

$$Y = \sum_{a=1}^{A} t_a q_a^T + \sum_{a=A+1}^{R} t_a q_a^T = TQ^T + F \quad (4)$$

where *T* was the *(N×A)* matrix of score, *P (M×A)* and *Q (k×A)* were loading matrices. Parameter *A* was the dimension of the latent variable space determined by cross validation method. *E* and *F* were residual matrices [11]. In order to deduce the inversion of the PLS, the estimation model of PLS could be simplified as:

$$\hat{X} = TP^T \quad (5)$$

$$\hat{Y} = TQ^T \quad (6)$$

Considering the score *T* was not a square matrix, it was decomposed into two sub- matrices *U* and *S* in Equation (7).

$$T = US \quad (7)$$

where *U* was *(N×A)*, an orthonormal matrix. *S* was an *(A×A)*, diagonal matrix, which could be calculated as:

$$S = (T^T T)^{1/2} \quad (8)$$

According to this conversion, the PLS simplified model of Equation (5) was transferred as:

$$\hat{X} = USP^T \quad (9)$$

$$\hat{Y} = USQ^T \quad (10)$$

where the sub-matrices *S*, *P*, and *Q* contained the features of the correlation structure and the variance presented in the original data *X* and *Y* matrices.

Generally, the PLS method is used to build the inner relation between the input data and the output data, and then, according to a new set of input data, to predict the associate response variables. Usually, in the process industry, the input data represents the operating conditions, while response variables suggest the quality features of the products [12]. So, in the inversion

form of PLS model, we could obtain the new combination of the input operation condition for a new desired quality. In this study, the new desired glycan level was given to find out the appropriate range of the input glycosylation enzymes, which realized the operating space design of N-linked glycosylation. According to the model equation (10), for one-dimension desired quality characteristics, y_{des}^T could be shown as:

$$y_{des}^T = u_{new}^T SQ^T \quad (11)$$

In order to obtain the new operating conditions, x_{new}^T and u_{new}^T should be solved in Equation (11), while u_{new}^T was related to the relative dimension of the product quality space *Y* and the original operating space *X*. Generally, in most common instances, the dimension of the quality space *Y* is less than or equal to the dimension of the original operating variables *X*, which means $k \le A$. So, in this situation, the u_{new}^T could be solved according to the pseudo inversion as:

$$u_{new}^T = y_{des}^T (QS^2Q^T)^{-1}QS \quad (12)$$

The new operating conditions x_{new}^T corresponding to the desired quality y_{des}^T could be achieved by the value of the u_{new}^T . The Equation (9) could be shown as:

$$x_{new}^T = u_{new}^T SP^T \quad (13)$$

The Equation (13) suggested that the new operating conditions x_{new}^T contained the covariance structure and correlation features of the original input data *X*, which were determined by the *S* and P^T [13]. Finally, a comprehensive formula was obtained by combining the Equations (5) and (12) as:

$$x_{new}^T = y_{des}^T (QS^2Q^T)^{-1}QS^2P^T \quad (14)$$

The Equation (14) clearly showed the relations between the operating conditions and the desired quality characters.

(1) The analysis of null space x_{null}^T

When $k < A$, there would be null space x_{null}^T fall on the remaining *(A-k)* dimensional space which should have no influence on the desired product quality y_{des}^T [14]. The u_{null}^T must satisfied the Equation (15).

$$u_{null}^T SQ^T = 0 \quad (15)$$

The singular value definition could be shown as follows. If the matrix was set as $A \in C_r^{m\times n}(r>0)$, the eigenvalues of the matrix A^HA were:

$$\lambda_1 \ge \lambda_2 \ge \mathrm{L}\ \ \lambda_r > \lambda_{r+1} = \mathrm{L}\ = \lambda_n = 0 \quad (16)$$

The $\delta_i = \sqrt{\lambda_i}(i = 1,2,3,...,n)$ was called eigenvalue. It was easy to prove rank (A^HA) = rank (*A*). At the same time, the nonzero eigenvalue of the matrix (A^HA) was the same as the AA^H. Recorded the matrix $P=SQ^T$, thus the eigenvalues of P^HP were $\lambda_1 \ge \lambda_2 \ge ...\lambda_r > \lambda_{r+1} = ... = \lambda_n = 0$. There would be existed unitary matrix *G* satisfied the Equation (17)

$$G^H P^H PG = diag(\lambda_1 \ge \lambda_1 \ge ...\lambda_r) = \begin{bmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{bmatrix} \quad (17)$$

Thus, the matrix *G* could be decomposed as:

$$G = (G_1, G_2),\ (G_1 \in C^{A\times k}, G_2 \in C^{A\times(A-k)}) \quad (18)$$

Therefore,

$$G_1^T P^T PG_1 = \Sigma^2, G_2^T P^T PG_2 = 0 \quad (19)$$

That could be proved $G_2SQ^T = 0$, which was equal to $u_{null}^T SQ^T = 0$. In the function $u_{null}^T = \lambda^T G_2^T$, *λ* was an arbitrary constants *(A-k)* vector and G_2 was an *(A-(A-k))* matrix. Which columns were the left singular vectors of SQ^T associated with the *(A-k)* zero singular values. x_{null}^T determined the others possible operating conditions for a same desired quality. So, in this situation, all the predicted new operating conditions could be shown as:

$$x_{pred}^T = x_{new}^T + x_{null}^T \quad (20)$$

The Equation (20) indicated that there was more than one operating condition to obtain for one desired quality.

(2) Prediction uncertainty

Considering the PLS model is based on data driven method, there must be uncertainty questions. The uncertainty is mainly caused by the original data set, the parameters, the predictions, and other factors. In this study, only the prediction uncertainty was considered. For an observed glycan quality y_{obs}^T , there would be a predicted output $\hat{y}_{obs}$. For each operating condition $\hat{x}_{obs}$, the bias by the mean relative error (MRE) was evaluated [15, 16]. Assuming the bias followed a t-statistic, the 95% confidence interval, *CI*, was calculated as:

$$CI = \hat{x}_{obs} \pm t_{\delta/2,N-d} s \quad (21)$$

where *N* was the number of the samples. *d* was the number of the degrees of the freedom in the model. Often, *d* was equal to the number of latent variables *A*. δ was the significance for the confidence interval, here, δ =0.05. The standard deviation *s* could be shown as:

$$s = SE\sqrt{1 + h_{obs} + 1/N} \quad (22)$$

where *SE* was the standard error of samples. h_{obs} was the leverage of the samples. Both could be calculated by Equations (23) and (24), respectively.

$$s = SE\sqrt{1 + h_{obs} + 1/N} \quad (23)$$

$$h_{obs} = \frac{t_{obs}\Lambda^{-1}t_{obs}^T}{N-1} \quad (24)$$

For the process of the complex N-linked glycosylation, the prediction uncertainty needed to be considered. As for the desired glycan quality, it's better to find out the operation range of the input data rather than an operating point. The operating space design of N-linked glycosylation, *CI* might have better practical guidance on the production.

Results and discussion

In the case of ensuring the quality of sample data and the balanced distribution, the size of sample data determines the precision of training results. The larger the sample data, the higher the accuracy. Since the sample size directly affects the computing time of the computer, we do not need too much sample data when the accuracy meets the requirements. Otherwise, we have to wait for a long training time. As described in the above sections, in order to predict the desired glycan states, 52 sets of samples were generated according to the N-linked glycosylation network. For enough iterations, the accuracy of the training results tended to be consistent, and the method only affected the convergence speed (operation time) of the calculation, which had no direct relationship with the sample size. So, among them, 50 sets of the samples were used to be the training set to build the PLS and inversion model. The left 2 sets were considered as the new desired glycans to test the inversion model. Because the network model was constructed by series of enzymatic reactions, 11 kinds of glycosylation enzymes were selected as

the possible input operating conditions including *X=[ManI ManII FucT GnTI GnTII GnTIII GnTIV GnTV GnTE GalT SiaT]*, while the output were the 11 kinds of main glycans including *Y=[S0 S1 S2 S3 S4 G0 G1 G2 G3 F0 F1].* Each kind of glycan was tested to obtain the desired value through the PLS inversion model. Typically, the desired glycan quality, *S1* and *G1* were taken as examples to make a detailed analysis. The possible space of operating conditions was found out to ensure the product quality characters.

A PLS model was built by using the generated data sets, *X (50×11)* and *Y (50×11)*. The number of latent variables (LVs) used in the model was determined by the cross-validation method. Generally, the main principle to calculate the LVs number is to extract the maximum variance and the related information from the original data set. The more information captured from the data set, the more accurate prediction will be. In this work, the number of LVs was 2, i.e. *A=2* LVs, which explained the 92.16% of the variance of *Y* and 23.27% of the variance of *X*. For one dimension of the desired glycan quality, it should be noted that *k*<*A*, so the x_{null}^{T} would exist. Then, according to the inversion model of Equation (12), the desired glycan state could be obtained.

For a new desired glycan *S1=144*, i.e. $y1_{des}$*=144*, which was not included in the original history dataset, it was achieved to find out the possible operating conditions. The two latent variables scores of x_{new}^{T} were shown in Figure 1. The dashed ellipse was the 95% confidence limit, The score pointed out of the ellipse (the little blue circles) should be ignored, while the red circles represented the original dataset scores. The green triangle was the predicted score values of the x_{new}^{T} corresponding to the desired glycan *S1*, $y1_{des}$*=144*. For more intuitive comparison, the real point score (the green circle) value was reported. The predicted scores were close to the real scores, i.e. *k*<*A*. Actually, *A*-*k*=1 suggested that there would be one dimension of x_{null}^{T} existed (the blue points). Besides, considering the prediction uncertainty of the model, the possible operating conditions were achieved, which was shown as the magenta area within the ellipse. The magenta area determined by the prediction uncertainty in PLS model was an operating design space, which narrowed the scope of the experiment manipulation of variables to a certain degree. The design of the operating space provided a relatively much smaller feasible region than the original experiment space. Absolutely, it would save more cost and resources to obtain the certain desired glycoprotein product.

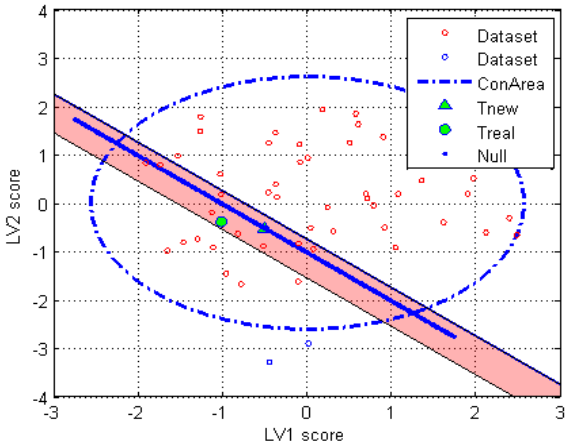

Figure 1. Glycan *S1,* $y1_{des}$*=144.*

For the desired glycan *S1,* $y1_{des}$=144, the 11 kinds of new input combination of the x_{new} were obtained, which were compared with the real values (Table 2). The column of relative errors (*RE*) were the relative errors of each glycosylation enzyme and were calculated as:

$$RE = \frac{|x_{new} - x_{real}|}{x_{new}} \times 100\% \quad (25)$$

The fifth column of *CI* suggested the prediction uncertainty in the inversion of PLS. Most of the input manipulated variables achieved were close to the real values (Table 2). Taking glycosylation

Table 2. Comparison with the real values.

Input	x_{real}	x_{new}	RE	CI
Fuct	1.52	1.86	22.37%	(1.767,1.953)
Galt	1.5	1.3	13.33%	(1.235,1.365)
GnTE	2.4	2.5	4.17%	(2.375,2.625)
GnTI	3.55	3.35	5.63%	(3.1825,3.5175)
GnTII	1.43	1.23	13.99%	(1.1685,1.2915)
GnTIII	0.83	1.11	33.73%	(1.0545,1.1655)
GnTIV	3.37	3.47	2.97%	(3.2965,3.6435)
GnTV	0.43	0.34	20.93%	(0.323,0.357)
ManI	2.41	3.6	49.38%	(3.42,3.78)
ManII	0.82	0.64	21.95%	(0.608,0.672)
SiaT	1.39	1.07	23.02%	(1.0165,1.1235)

Table 3. Different glycans prediction.

Output	E_PLS	$y1_{des}$	EE_inv	$y2_{des}$	EE_inv
S0	2.21%	388.03	6.67%	370.00	12.12%
S1	5.27%	144.00	1.77%	203.47	7.61%
S2	1.59%	728.16	1.84%	831.12	5.70%
S3	6.03%	208.60	5.68%	205.92	5.62%
S4	39.80%	33.47	4.77%	24.59	9.52%
G0	15.12%	7.26	4.53%	4.75	5.80%
G1	7.90%	175.90	3.70%	185.59	3.63%
G2	4.31%	528.53	6.39%	730.03	2.35%
G3	1.80%	504.69	16.70%	424.09	6.56%
F0	7.93%	195.93	3.68%	185.48	8.28%
F1	0.80%	1,359.00	8.27%	1,348.00	3.68%

transfer enzymes *GnTE* and *GnTIV* as examples, the values of the *RE* were 4.17% and 2.97% respectively. However, for some of the enzymes, the values of the *RE* were too large to explain the accuracy of the prediction. The main reasons might be the extreme correlations within the 11 kinds of input variables. But overall, the magnitude of the predicted values was as the same as the real values, besides the scores of the latent variables were projected in the same *CI* areas compared with the real ones, which indicated that the method of this study was effective to predict the desired glycans.

Figure 2 showed another desired glycan quality of *G1, y2des=175.9*. The scores of the new operating conditions were near to the real values, which indicated that the prediction performance was good. In the same way, all the 11 kinds of output glycans were predicted to find out the feasible operating conditions. The errors in the PLS model and in the inversion model were listed in Table 3, respectivily. The *E_PLS* represented the mean relative error *(MRE)* in PLS model, which was calculated by:

$$E\_PLS = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_{obs} - y_{real}}{y_{real}}\right) \quad (26)$$

EE_inv was the mean relative error in the inversion of the PLS, which could be calculated by:

$$EE\_inv = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{new} - x_{real}}{x_{real}}\right) \quad (27)$$

Each of the glyacns were made two set of prediction, $y1_{des}$ and $y2_{des}$. Most of them were raliable. Howerer, some of them might not be accurate and the main reasons might be (1) PLS was a linear modeling technique, the process of N-linked glycosylation was a nonliear model; (2) there were uncertannty of the original data sets generated by the glycosylation reaction network model; and (3) the accumulated errors in the PLS model and in the reverse transfer process.

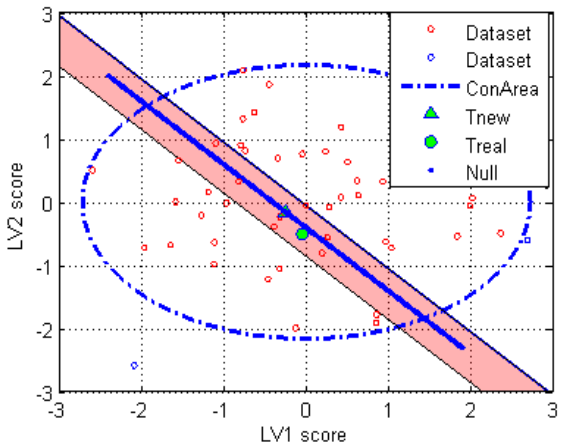

Figure 2. Glycan *G1, $y2_{des}$=175.9.*

Conclusion

The N-linked glycosylation is a complex biochemical process. The glycan level prediction is a critical aspect to guide the glycoprotein production. Currently, most of the researchers are focusing on the extracellular environment from a macro perspective that influences glycoprotein product. Few people consider the intracellular glycosylation reactions, especially the critical glycosylation enzymes that have a direct and crucial influence on the different glycans. In this study, a novel method of multivariate statistical regression of partial least square (PLS) inversion was developed to predict the desired glycan levels corresponding to the manipulated variables and the possible operating range. Realizing the operating space design of N-linked glycosylation could reduce the cost and resources to a certain degree. The results of this study suggested that the method developed in this study demonstrated a good performance on prediction of the glycan qualities, which might provide a guide to the glycosylation control online in the future. The next step will be the implementation of a control strategy to guide this process and improve the product quality. However, there are still problems that need to be solved for more accurate methods or theory technology development to overcome the nonlinear and extremely complex process, such as how to improve the accuracy of multivariate statistical regression of PLS inversion method and how to get the more accurate glycan level operating range and so on. In addition, it is hoped that the algorithm can be improved to accurately operate the space, which involves the collection of sample data. Further, sample training of the algorithm can be conducted to improve the accuracy of the model, and the PLS inversion algorithm can also be combined with other nonlinear system analysis algorithms, such as the variable structure control based on sliding mode or nonlinear system analysis modeling based on neural network and so on. Furthermore, only one kind of glycan quality was considered in this study. Multidimensional quality of product should be expanded. We will refer to the complex biological reaction process to expand the nonlinear degree of the system model and consider a variety of glycan reactions to improve the prediction algorithm in further research.

Acknowledgement

This work was supported by the Science and Technology Project of Hebei Education Department under Grant QN2022108.

References

1. Geyer H, Geyer R. 2006. Strategies for analysis of glycoprotein glycosylation. BBA-Proteins Proteom. 1764:1853–1869.

2. Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, Liu X, *et al*. 2011. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. Nat Biotechnol. 29(8):735–741.
3. Goochee CF, Gramer MJ, Andersen DC, Bahr JB, Rasmussen JR. 1991. The oligosaccharides of glycoproteins - bioprocess factors affecting oligosaccharide structure and their effect on glycoprotein properties. Nat Biotechnol. 9:1347–1355.
4. FDA. 2009. Q8(R1) Pharmaceutical development revision 1. In: CDER C, editor. Rockiville, MD, USA.
5. Butler M. 2006. Optimisation of the cellular metabolism of glycosylation for recombinant proteins produced by mammalian cell systems. Cytotechnology. 50:57–76.
6. St. Amand M, Tran K, Radhakrishnan D, Robinson AS, Ogunnaike BA. 2014. Controllability analysis of protein glycosylation in Cho cells. Plos One. 9(2):e87973.
7. Peterson JJ. 2008. A Bayesian approach to the ICH Q8 definition of design space. J Biopharm Stat. 18:959.
8. Pantelides CC, Shah N, Adjiman CS. 2009. In: Comprehensive quality by design in pharmaceutical development and manufacture. 1st Edition. Edited by Reklaitis GV, Seymour C, Garcia-Munoz S. Wiley-AIChE (Nashville, TN, USA). p417f.
9. Umana P, Bailey JE. 1997. A mathematical model of N-linked glycoform biosynthesis. Biotechnol Bioeng. 55:890–908.
10. Krambeck FJ, Betenbaugh MJ. 2005. A mathematical model of N-linked glycosylation. Biotechnol Bioeng. 92:711–728.
11. Zhong B, Wang J, Zhou JL, Wu HY, Jin QB. 2016. Quality-related statistical process monitoring method based on global and local partial least squares projection. Ind Eng Chem Res. 55(6):1609-1622.
12. Jaeckle CM, MacGregor JF. 2000. Industrial applications of product design through the inversion of latent variable models. Chemom Intell Lab Syst. 50(2):199-210.
13. Pantelides C, Pinto M, Bermingham SK. 2010. Optimization-based design space characterization using first-principles models. In: 2010 AIChE Annual Meeting (Salt Lake City, UT, USA). 2010:358b.
14. MacGregor JF, Bruwer MJ. 2008. A framework for the development of design and control spaces. J Pharm Innov. 3:15-22.
15. Jaeckle CM, Macgregor JF. 1998. Product design through multivariate statistical analysis of process data. AIChE J. 44(5):1105-1118.
16. Jaeckle CM, MacGregor JF. 2000. Product transfer between plants using historical process data. AIChE J. 46(10):1989-1997.