A multivariate approach to studying glycosylation gene expression in CHO cells

Faraaz NK Yusufi[1], Niki SC Wong[1], Danny CF Wong[1], Miranda GS Yap[1,2], Dong-Yup Lee[1,2]

[1] Bioprocessing Technology Institute, Biomedical Sciences Institutes, 20 Biopolis Way, #06-01 Centros, Singapore 138668
[2] Department of Chemical & Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117576
Contact: Lee_Dong_Yup@bti.a-star.edu.sg or cheld@nus.edu.sg; Tel: +65-6478-8900

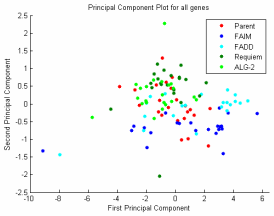
Abstract

Mammalian cell lines such as Chinese Hamster Ovary (CHO) are ideal for producing recombinant proteins that require post-translational modifications such as glycosylation. Since glycosylation has an effect on many biological properties such as folding, stability and efficacy, the quality of secreted proteins is dependent on the consistency of attached glycan structures. Thus far, several cell engineering techniques to improve and optimize glycosylation have been applied to certain gene targets. These targets are selected on the basis of prior biological knowledge of regulatory and metabolic pathways such as known bottlenecks or alternative routes. In addition to prior biological knowledge, multivariate statistics can be applied to cell culture data to discover new gene targets for further manipulation.

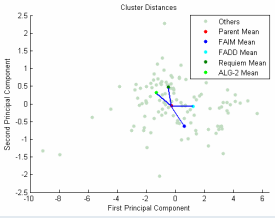
We consider gene expression and glycosylation profiles from five interferon-γ producing CHO cell lines consisting of four apoptotic-resistant gene targeted (GT) lines and one parent CHO line. There are two issues that can be addressed for this data; how to quantifiably differentiate the cell lines from one another, and how to measure the effect of gene expression on glycosylation. Principal component analysis (PCA) is used to identify which genes in the GT cell lines show similar behavior in the parent line and which genes differ significantly. Ridge regression is used to identify genes which have the most impact on sialylation of glycan structures.

PCA Results

PCA

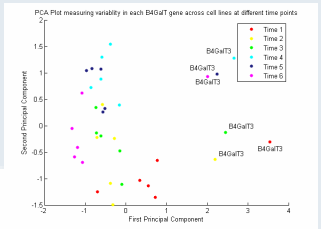


Cluster distances



Cluster distances show how different the glycosylation gene expression profiles are when the CHO GT cells are compared to parental CHO cells. Increasing distance means more dissimilarity.

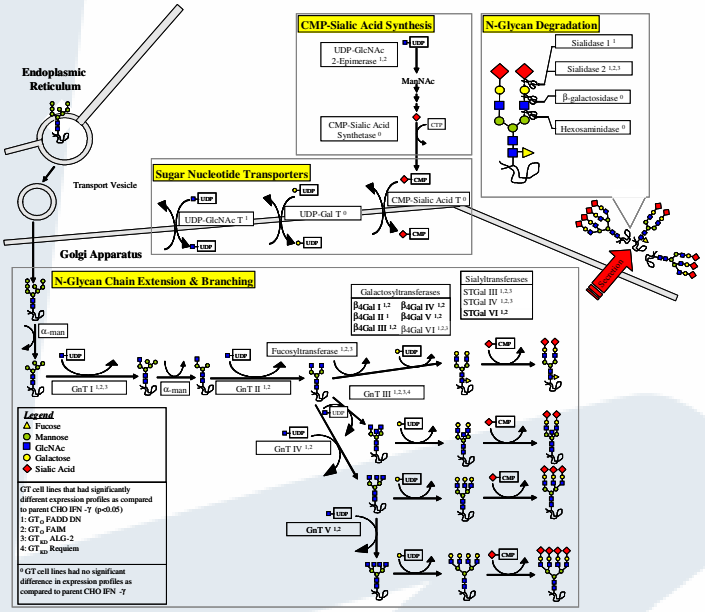
CHO GT_{KD} Req > CHO GT_O FAIM > CHO GT_{KD} ALG-2 > CHO GT_O FADD



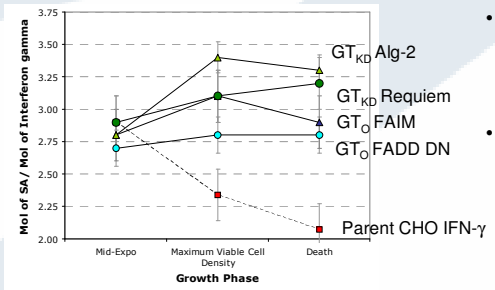
PCA plot of the variance of genes across different growth phases reveals genes which show significant variation from other genes. In this case B4GalTIII shows differential variance throughout the cell culture.

Genes involved in the N-glycosylation pathway

Expression levels of 24 genes involved in N-glycosylation were measured using qRT-PCR



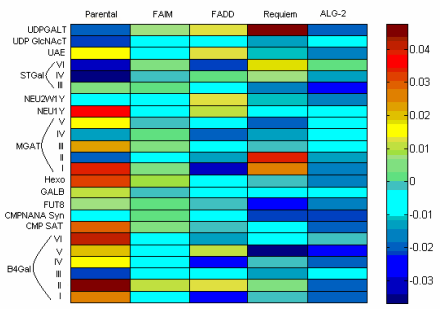
Sialylation profile of IFN-γ during fed-batch cultures



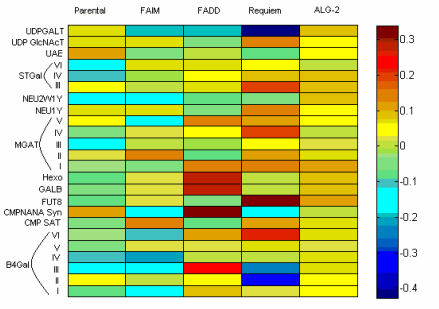
- CHO GT cell lines produced IFN-γ which were more sialylated
- Maintenance of sialic acid content across growth phases

Regression Analysis Results

Ridge regression coefficients for absolute gene expression values



Ridge regression coefficients for gene expression fold-change ratios



Multiple linear regression analysis reveals how much impact each gene has on the outcome (sialylation). Coefficient values obtained can be used to rank genes in order of impact on sialylation.

Since microarray data is highly correlated and the number of genes is often much larger than the number of observations, ordinary regression methods are not sufficient.

Ridge regression has been suggested for use on data sets with high multicollinearity and performs better than PLS and Principal Components Regression (PCR).

Performing regression analysis while using absolute expression values obtained from qRT-PCR show much more similarity between GT cell lines when compared to the parent. This inference is not as clear when fold change ratios are used to perform regression.

Summary

Multivariate statistical methods such as PCA and Ridge Regression were applied to gene expression data from cell engineering experiments. The inferences gained from this analysis, combined with *a priori* biological knowledge, can be used to design further experiments aimed at improving glycosylation in secreted recombinant proteins.

Acknowledgements

The authors would like to thank Dr Lee May May and her team for their excellent technical input and assistance.

Reference

Frank I.E., Friedman J.H. (1993) A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109-135.