Journal of Biotechnology 151 (2011) 159–165

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Predicting cell-specific productivity from CHO gene expression

Colin Clarke[a,*,1], Padraig Doolan[a,1], Niall Barron[a], Paula Meleady[a], Finbarr O'Sullivan[a], Patrick Gammell[b], Mark Melville[c], Mark Leonard[c], Martin Clynes[a]

[a] National Institute for Cellular Biotechnology, Dublin City University, Glasnevin, Dublin 9, Ireland
[b] Bio-Manufacturing Sciences Group, Pfizer Inc., Grange Castle International Business Park, Clondalkin, Dublin 22, Ireland
[c] Bioprocess R&D, Pfizer Inc., Andover, MA 01810, USA

ARTICLE INFO

Article history:
Received 28 July 2010
Received in revised form 26 October 2010
Accepted 20 November 2010
Available online 27 November 2010

Keywords:
Chinese hamster ovary
Productivity
Microarray
Partial least squares
Cross model validation
Variable selection

ABSTRACT

Improving the rate of recombinant protein production in Chinese hamster ovary (CHO) cells is an important consideration in controlling the cost of biopharmaceuticals. We present the first predictive model of productivity in CHO bioprocess culture based on gene expression profiles. The dataset used to construct the model consisted of transcriptomic data from 70 stationary phase, temperature-shifted CHO production cell line samples, for which the cell-specific productivity had been determined. These samples were utilised to investigate gene expression over a range of high to low monoclonal antibody and fc-fusion-producing CHO cell lines. We utilised a supervised regression algorithm, partial least squares (PLS) incorporating jackknife gene selection, to produce a model of cell-specific productivity (Qp) capable of predicting Qp to within 4.44 pg/cell/day root mean squared error in cross model validation ($RMSE^{CMV}$). The final model, consisting of 287 genes, was capable of accurately predicting Qp in a further panel of 10 additional samples which were incorporated as an independent validation. Several of the genes constituting the model are linked with biological processes relevant to protein metabolism.

1. Introduction

Cell and process engineering approaches to improve productivity in bioreactors have largely focussed on reactor design and culture strategies such as clonal selection, stability, medium formulation, culture temperature and cell engineering for controlled proliferation and increased resistance to apoptosis (Altamirano et al., 2000; Butler, 2005; Prentice et al., 2007; Wurm, 2004). Using this approach, key cell line characteristics, including cell growth rate, achievable cell densities and correct product processing are identified only following a lengthy labour-intensive screening process. To complement these strategies, previous attempts have been made to modify or improve the performance of these lines in the bioreactor using cellular engineering strategies (reviewed in Mohan et al., 2008). However, these studies have demonstrated only incremental improvements in productivity and the cellular processes underpinning Qp remains poorly understood in Chinese hamster ovary (CHO) and other bioprocess-relevant cell lines.

The development of expression profiling methodologies such as microarrays and proteomics offer the prospect of examining the molecular phenotypes underlying productivity in CHO and their application in bioprocess research has already been extensively reviewed (Griffin et al., 2007). Previous microarray expression profiling studies focussing on productivity in CHO (Doolan et al., 2008; Schaub et al., 2010; Trummer et al., 2008; Kantardjieff et al., 2010; Yee et al., 2007) and in the commercially used mouse myeloma NS0 cell line (Charaniya et al., 2009; Khoo et al., 2007; Seth et al., 2007) have identified several crucial pathways and processes. These microarray-based productivity studies have also been complemented by proteomics studies in CHO (Carlage et al., 2009; Meleady et al., 2008; Nissom et al., 2006) and NS0 (Seth et al., 2007; Smales et al., 2004; Alete et al., 2005; Dinnis et al., 2006).

To date, profiling studies in CHO have been characterised by relatively small numbers of samples (typically < 20) compared in a case/control format. Interesting genes and protein candidates are generally prioritised via the traditional paradigm of differential expression (i.e. fold change). A significant drawback of this approach includes the selection of an appropriate threshold (considering the inherent noisy nature of microarrays) resulting in too few or too many genes identified and providing inconsistent comparison with studies on similar biological systems. This limitation is further compounded by the observation that changes in productivity levels are usually accompanied by only modest changes in gene expression levels (Smales et al., 2004; Yee et al., 2009). Larger sample numbers in combination with more sophisticated algorithms can therefore make a significant contribution to identifying the molecular mechanisms underpinning productivity in CHO.

* Corresponding author. Tel.: +353 1 7005700/5692; fax: +353 1 7005484.
E-mail address: colin.clarke@dcu.ie (C. Clarke).
[1] Both authors contributed equally to this publication.

0168-1656/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbiotec.2010.11.016

Multivariate statistics and machine learning algorithms for classification and regression allow relationships between genes to be considered and have previously been advocated over univariate gene selection methods (Boulesteix and Strimmer, 2007). Partial least squares (PLS) is a statistical modelling technique closely related to principal component analysis (PCA) and is used to construct predictive models for complex multidimensional datasets. PLS components, known as latent variables (LVs), are derived from linear combinations of the original variables to maximise the covariance between a matrix of independent variables (e.g. gene expression) and dependent variable(s) (e.g. productivity). By retaining only those LVs containing the majority of information on the relationship between predictor and response variables (thus removing a substantial amount of noise and measurement error) a model can then be formed between these LVs and cell-specific productivity. Detailed treatments of the PLS algorithm have been previously described (Martens and Naes, 1989).

Previous examples of PLS predictive model generation from microarrays include regression (Gidskehaug et al., 2007; Huang et al., 2004; Misra et al., 2007), the development of models for classification (Aaroe et al., 2010; Nguyen and Rocke, 2002a) and proportional hazard models for survival analysis (Nguyen and Rocke, 2002b). Apart from microarrays, the technique is utilised across a variety of fields and has previously been applied to various aspects of bioprocessing including mass spectrometry-based proteomic profiling, process monitoring and process analytical technology (PAT) (Sellick et al., 2010; Stansfield et al., 2007; Thomassen et al., 2010).

In this paper, we construct a regression model using the PLS algorithm to capture the relationship between gene expression and a quantitative phenotypic variable (cell-specific productivity). We aim to produce a model for prediction of Qp from gene expression measurements with a potential application in bioprocess development. The use of a gene selection routine coupled with rigorous statistical validation was incorporated to reduce PLS model complexity and decrease the error rate. The algorithm may also provide a vehicle for the identification of subsets of genes relevant to the biology underlying productivity of recombinant proteins in CHO. This work represents one of the largest studies of CHO transcriptomic datasets published to date.

2. Materials and methods

2.1. Determination of cell-specific productivity

The concentration of recombinant protein product in conditioned media samples (volumetric titre) was determined by Protein-A HPLC. Cell viability was determined using the trypan blue dye-exclusion viability assay and hemocytometer counting (for shake flask samples) or a Cedex Automated Cell Culture Analyzer (Roche Innovatis) (for bioreactor samples). Cell specific productivity was determined as shown below.

$$\mathrm{Qp(pg/cell/day)} = \left[\frac{\mathrm{titre\ \ 2} - \mathrm{titre\ \ 1}}{(\mathrm{density\ \ 2} - \mathrm{density\ \ 1})}\right] \times \mathrm{daily\ \ growth\ \ rate} \quad (1)$$

where

$$\mathrm{daily\ \ growth\ \ rate} = \frac{(\ln(\mathrm{density\ \ 2}) - \ln(\mathrm{density\ \ 1}))/(\mathrm{time\ \ 2} - \mathrm{time\ \ 1})}{24}$$

2.2. Cell line selection and experimental design

A total of 80 fed-batch, temperature-shifted CHO production cell line samples displaying a range of cell-specific productivity values (0.81–50.4 pg protein/cell/day) were selected for transcriptional profiling using a proprietary (Wye2aHamster) CHO-specific affymetrix microarray. All cell line samples were grown in serum-free suspension culture in the temperature-shifted range of 29.5 °C to 31 °C (culture temperature shift time-point varied between 24 and 72 h according to process design) and were collected during the stationary growth phase (5–10 days) at the following time-points: Day 5 (23 samples), Day 7 (42 samples), Day 8 (7 samples) and Day 10 (8 samples). The entire sample set comprised 42 CHO DUX and 38 CHO K1 samples, from 10 production cell lines expressing a variety of monoclonal antibody (60 samples) and fc-fusion protein products (20 samples). 18 of the samples were isolated from a total of 14 shake flasks (11 of which were carefully maintained to a pH setpoint using CO_2 and base addition as required); the remaining samples were isolated from 40 individual bioreactor cultures. The sample set was split into 70 microarrays for PLS model construction and validation (*calibration data*). 10 samples from 5 CHO DUX and 5 CHO K1 cultures producing monoclonal antibody and fc-fusion proteins were held back from model building and gene selection to serve as an independent test set evaluation (*test data*).

2.3. Microarray analysis and data preprocessing

The methods and criteria used for total RNA purification, cRNA sample processing and hybridisation to hamster microarrays have been previously described (Doolan et al., 2008). The study presented here utilises a proprietary WyeHamster2a oligonucleotide microarray, which has been described previously (Doolan et al., 2008), representing an estimated 10–15% of the CHO transcriptome. All microarray data were pre-processed in the statistical software environment R (www.r-project.org) and the *aroma.affymetrix* package using the robust multichip average (RMA) algorithm (Bolstad et al., 2003; Irizarry et al., 2003a,b).

2.4. Partial least squares implementation

PLS model construction and jackknife variable selection was carried out within R using the '*pls*' package (Mevik and Wehrens, 2007). Cross model validation was implemented using a script written in-house (available on request).

2.5. Jackknife gene selection

The elimination of genes which do not contribute significantly to the model should simplify and improve the accuracy of PLS and possibly reveal biologically important genes related to cell-specific productivity. During the construction of the PLS model, each gene is interrogated within the inner loop of model validation (Fig. 1) as to its importance during the model building process. The resampling method known as 'jackknifing' (JK) (Efron and Stein, 1981) was employed to assess the significance of variables and to remove uninformative or "noisy" genes which have no contribution to the final model. The selection of important genes from the analysis is achieved by initially considering the entire complement of the array and constructing a model. Each PLS regression coefficient is perturbed and its approximate "significance" determined using a *t*-test (as the distribution of PLS regression coefficients and the degrees of freedom are unknown, it is recommended to treat the resulting *p*-value as measure of non-significance (Mevik and Wehrens, 2007)). The least "significant" gene (i.e. the gene with the largest *p*-value) in the model is eliminated from the dataset. The backward elimination of genes from the model continues until all remaining