Biochemical Engineering Journal 103 (2015) 77–84

Contents lists available at ScienceDirect

Biochemical Engineering Journal

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

Regular article

Analysis of time-course gene expression profiles to study regulation of cell growth in fed-batch bioreactors

Yanzhu Lin, Kim Lehmann, Philip Z. Brohawn, Zheng Liu, Nitin Agarwal*

MedImmune LLC, One MedImmune Way, Gaithersburg, MD 20878, USA

ARTICLE INFO

Article history:
Received 14 January 2015
Received in revised form 19 May 2015
Accepted 22 June 2015
Available online 15 July 2015

Keywords:
Bioreactors
Animal cell culture
Fed-batch culture
Amino acids
Gene expression
Sparse partial least squares

ABSTRACT

Compared to other systems biology tools, genomic microarrays represent a mature platform that allows for facile access to the internal biological mechanisms of cell culture. While the large datasets generated by microarrays are a potential goldmine of information, ironically, it is their large size that also makes data-mining a cumbersome task. This can get further complicated by unavoidable batch effects generated when different datasets are combined. Furthermore, gene expression profiles are dependent on combinations of various complex intracellular events and as such identifying the signals primarily related to the phenotype of interest poses a substantial challenge. In this study we addressed these issues by utilizing a workflow that allows adjustment of time-course gene expression datasets for batch effects and incorporates the use of sparse partial least squares analysis to identify specific genes of interest. We were able to identify a set of relevant genes that displayed a strong correlation with cell growth in fed-batch bioreactors under different nutrient compositions. By conducting further biological network analysis, we identified four transcriptional regulators, namely ATP7B, SREBP1, SCAP and INSIG2 that are responsible for regulation of these genes and are likely important drivers for cell growth differences in response to change in nutrient composition.

1. Introduction

Mammalian cell culture remains the major route of production for protein-based therapeutics currently in the market or under development [1–3]. This is mainly due to the superior capacity of mammalian cells to process complex proteins and carry out post-translational modifications that are suitable for use in humans [4]. Since its first implementation, mammalian cell culture has evolved considerably with significant improvements attributable to the advances made in cell line engineering, plasmid design, protein engineering, clone selection and process optimization approaches [5,6]. Most biotech companies choose to operate bioreactors in fed-batch mode due to the simplicity of fed-batch operation as compared to continuous culture [5,7]; this means however that the cell culture microenvironment varies throughout the duration of the culture due to changes in nutrient composition. As a consequence of the constantly changing environment, cell culture dynamics could have a significant impact on the performance of fed-batch culture. While a wealth of empirical knowledge has been generated over time, there still exists a significant gap in the mechanistic understanding of the role that varying nutrient concentrations play in driving the underlying cell biology during cell culture in fed-batch bioreactors [1,8]. In this study, we demonstrate how we can help address this gap by correlating changes in cell biology to changes in observed cell growth in fed-batch bioreactors.

With recent developments in systems biology, new tools are becoming available that can enable rational understanding of how changes in process parameters such as nutrient composition may impact cell culture performance [8]. The mature field of microarray analysis provides one such avenue that can be used to easily generate time-course profiles for mRNA expression in fed-batch bioreactors [8–10]. However, there are some technical hurdles that warrant the use of specialized statistical methods before relevant information can be extracted from microarray data. For instance, the number of genes in the dataset (p) almost always greatly exceeds the number of total biological observations (n), thus leading to the issue of high dimensionality [11]. As a result, methods such as linear regression are unsuitable to analyze such data [12]. Also, gene expression profiles are not independent and are usually covariant with expression of other related genes, which leads to the other issue of multicollinearity [13]. One important strategy to address both the high-dimensionality and multicollinearity problems is to use dimension reduction methods, such as principal

* Corresponding author at: Cell Culture and Fermentation Sciences, MedImmune LLC, One MedImmune Way, Gaithersburg, MD 20878, USA.
E-mail address: AgarwalN@MedImmune.com (N. Agarwal).

http://dx.doi.org/10.1016/j.bej.2015.06.014
1369-703X/

component analysis (PCA) and partial least squares (PLS) [9,14]. These approaches reduce dimensionality by replacing the large number of predictors (genes) with just a few significant principal components that capture the major trends in the data. Furthermore, the principal components are able to group variables that are highly correlated, thus addressing the multicollinearity issue [15]. However, these approaches do not lead to a reduction in the number of predictors (genes) in the dataset. Since microarrays yield information on several thousand expressed genes in the cells, it is especially difficult to identify genes of interest in context of the phenotype under consideration [9,15]. In order to address this, methods that allow for selection of relevant predictors (genes) within the dataset have been proposed. For example, the selection of relevant predictors which is also referred to as implementation of 'sparsity', can be imposed on the dataset by using the L1 penalty, thus allowing for selection of relevant variables [16]. Both the L1 and L2 penalties have been incorporated into the framework of traditional dimension reduction methods to create sparse principal component analysis (SPCA) [17] and sparse partial least squares (SPLS) [13]. Both of these approaches are able to address the high-dimensionality and multicollinearity issues in gene expression datasets while simultaneously selecting for just the relevant genes. Of the two, SPLS is a supervised algorithm, and hence additionally capable of selecting for those predictors (genes) that are correlated to the response (phenotype) of interest while discarding the non-relevant predictors. It is also computationally inexpensive to execute and in simulation studies was shown to outperform other comparable statistical methods such as elastic net [13,15]. In addition to selection of an appropriate algorithm, there exist other issues related to microarrays such as batch-to-batch variability [18–20]. In this study, we discuss the use of a workflow that corrects for batch-to-batch variability in microarray data and incorporates SPLS for the analysis of time-course gene expression data generated from fed-batch bioreactors. Using this workflow we were able to identify specific biological pathways that had a strong correlation with cell growth across several nutrient compositions. This information can be extremely useful to understand the role that individual nutrients play to influence the underlying cell biology.

2. Materials and methods

2.1. Growth of cells in bioreactors and sampling

This study utilized a clonal cell line that was generated from NS0 host cells transfected with a heterologous gene to enable secretion of a monoclonal antibody. The clonal cell line was expanded using a proprietary growth medium in baffled shake flasks incubated on an agitated plate in a humidified incubator set at 37 °C with an overlay of 6% carbon dioxide for buffering. Once enough culture was obtained, the cells were used to inoculate bench-top bioreactors that were set at 37 °C. The pH of the bioreactors was controlled between 7.1 and 7.3 by use of sodium carbonate (base) and carbon dioxide (acid) additions. The dissolved oxygen level was maintained at 76 mm Hg throughout the run by supplementing oxygen using a combination of air and oxygen sparge in the bioreactors. The cultures were allowed to run for 14 days with pre-scheduled intermittent addition of nutrient feeds, as well as supplementation of glucose in the cultures as required. Depending on the condition, one of three feed schemes was implemented during these runs. The three feed schemes differed primarily in nutrient composition. A platform in-house proprietary nutrient composition was used for feed scheme 1. Feed scheme 2 utilized a more concentrated version of feed 1. Feed scheme 3 utilized a feed that had similar distribution of most nutrients as in feed 2, with the exception of key amino acids. While the total molar content for amino acids was constant between feed schemes 2 and 3, feed 3 had increased levels of asparagine, aspartic acid, arginine and glutamine and reduced levels of isoleucine, leucine, cysteine and proline when compared to feed scheme 2. Samples were extracted from the bioreactors on a routine basis and analyzed for cell growth using Cedex (Roche Diagnostics Corporation, Indianapolis, IN) cell counters. Cell pellets were also collected in order to generate mRNA for use in microarrays. Approximately 2 ml of these samples from the bioreactors were centrifuged and decanted to remove the supernatant. The cell pellets, which consisted of anywhere from 4 to 100 million cells depending on culture day and condition, were then snap-frozen on dry ice and immediately transferred to −80 °C storage until further use.

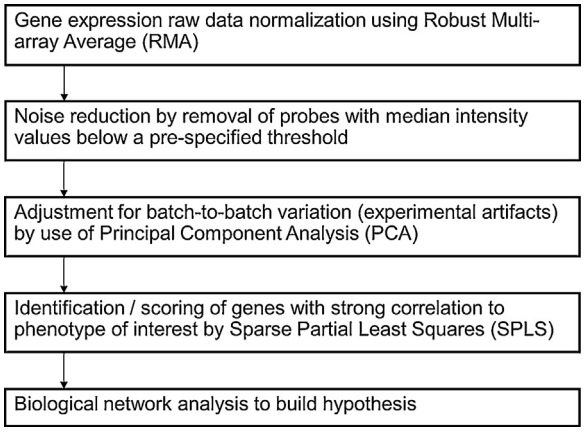

Fig. 1. Gene expression data analysis workflow used in this study.

2.2. Generation of microarray data

Frozen cells were lysed and RNA was extracted using the Qiagen RNeasy Mini kit (Qiagen, Hilden, Germany) following the manufacturers' protocol. RNA concentration was determined spectrophotometrically on a NanoDrop 2000 (Thermo Fisher Scientific, Waltham, MA) and quality of the intact RNA was assessed on a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) using the RNA 6000 Nano LabChip (Agilent Technologies, Santa Clara, CA). Sharp bands were obtained for 28S and 18S RNA at the expected sizes with no evident degradation as indicated by the RNA Integrity (RIN) score [21], which ranged from 8 to 10 for all samples (see Supplementary Fig. 1 for an example of the RNA quality assessment data). No correlation was found between the duration of storage of the cell pellets and the RIN scores, thus confirming that cell pellet handling and storage had minimal impact on RNA quality. Generation of biotin-labeled aRNA amplified from 3 μg of total RNA was accomplished with the Invitrogen Superscript Double-strand cDNA kit (Life Technologies, Carlsbad, CA) and Affymetrix GeneChip IVT labeling kit (Affymetrix, Santa Clara, CA) following the manufacturers' protocols. Fifteen micrograms of each biotin-labeled aRNA was fragmented and hybridized on Affymetrix Mouse Genome 2.0 GeneChip arrays (Affymetrix, Santa Clara, CA). All GeneChip washing, staining, and scanning procedures were performed with Affymetrix standard equipment following the manufacturers' protocols. Data capture and array quality assessments were performed with the GeneChip Operating Software (GCOS) tool (Affymetrix, Santa Clara, CA).

2.3. Gene expression data analysis

The workflow used for analysis of the gene expression data obtained is shown in Fig. 1. The steps involved in the data anal-