Journal of Biotechnology 179 (2014) 17–25

Contents lists available at ScienceDirect

Journal of Biotechnology

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

Data mining for rapid prediction of facility fit and debottlenecking of biomanufacturing facilities

Yang Yang[a], Suzanne S. Farid[b,*], Nina F. Thornhill[a,**]

[a] Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
[b] The Advanced Centre for Biochemical Engineering, Department of Biochemical Engineering, University College London, Torrington Place, London WC1E 7JE, UK

ARTICLE INFO

Article history:
Received 20 September 2013
Received in revised form 27 February 2014
Accepted 4 March 2014
Available online 15 March 2014

Keywords:
Biopharmaceutical manufacture
Stochastic discrete-event simulation
Decision tree classification
Multivariate data analysis
Data mining

ABSTRACT

Higher titre processes can pose facility fit challenges in legacy biopharmaceutical purification suites with capacities originally matched to lower titre processes. Bottlenecks caused by mismatches in equipment sizes, combined with process fluctuations upon scale-up, can result in discarding expensive product. This paper describes a data mining decisional tool for rapid prediction of facility fit issues and debottlenecking of biomanufacturing facilities exposed to batch-to-batch variability and higher titres. The predictive tool comprised advanced multivariate analysis techniques to interrogate Monte Carlo stochastic simulation datasets that mimicked batch fluctuations in cell culture titres, step yields and chromatography eluate volumes. A decision tree classification method, CART (classification and regression tree) was introduced to explore the impact of these process fluctuations on product mass loss and reveal the root causes of bottlenecks. The resulting pictorial decision tree determined a series of if-then rules for the critical combinations of factors that lead to different mass loss levels. Three different debottlenecking strategies were investigated involving changes to equipment sizes, using higher capacity chromatography resins and elution buffer optimisation. The analysis compared the impact of each strategy on mass output, direct cost of goods per gram and processing time, as well as consideration of extra capital investment and space requirements.

1. Introduction

In recent years, cell culture titres of monoclonal antibodies (mAbs) have increased dramatically as a result of improvements to cell lines, media composition, and feeding strategies (Birch and Racher, 2006; Li et al., 2010). Furthermore, it is common for titres to increase by 50% or more as a product progresses from early to late process development (Kelley et al., 2009). Higher titre processes can pose facility fit challenges for downstream processing (DSP), particularly during tech transfer to legacy biopharmaceutical manufacturing facilities that were constructed with multiple large-volume bioreactors (>10,000 L) and DSP capacities matched to lower titre processes. Legacy facilities can struggle to cope with the resulting higher protein loads onto DSP due to bottlenecks reached in DSP unit operations (e.g. chromatography columns) or tank storage capacities (Aldington and Bonnerjea, 2007; Chang, 2011; Farid, 2008; Kamarck, 2006; Kelley, 2009; Stonier et al., 2012). Thus systematic and rigorous tools for facility fit analysis and debottlenecking are critical to gaining greater understanding of the root causes of suboptimal facility fit and identifying the most promising debottlenecking strategies.

Facility fit analysis and DSP debottlenecking efforts are complicated by the inherent batch-to-batch variability present in bioprocess unit operations (Farid, 2008; Stonier et al., 2013). Facility fit assessments that are based on single point expected values for key process parameters, and hence do not account for process fluctuations, may not identify the correct bottleneck. Certain combinations of worst case values can lead to volumes that exceed equipment capacities and result in having to discard expensive product. The likelihood and consequences of such scenarios would not be captured by facility fit assessments based solely on expected values. Furthermore, large scale facilities often have fixed stainless steel equipment and piping networks. This makes it harder to adopt debottlenecking strategies involving equipment changes in response to fit issues arising from process variability and higher titres.

* Corresponding author at: University College London, Biochemical Engineering, Torrington Place, London WC1E 7JE, UK. Tel.: +44 20 7679 4415.
** Corresponding author at: Imperial College London, Department of Chemical Engineering, London SW7 2AZ, UK Tel.: +44 (0)20 7594 6622.
E-mail addresses: s.farid@ucl.ac.uk (S.S. Farid), n.thornhill@imperial.ac.uk (N.F. Thornhill).

http://dx.doi.org/10.1016/j.jbiotec.2014.03.004

Effective facility fit assessments can benefit from advanced data mining of datasets generated from bioprocess models that can capture the dynamics of bioprocesses as well as the impact of resource constraints and process variability. Commercial bioprocess modelling packages (e.g. Superpro Designer (Intelligen, Inc., Scotch Plains, NJ) tend to be useful for equipment sizing and costing but are not typically designed to capture the consequences of resource delays (e.g. due to buffer storage tank availability) or uncertainties (e.g. titre). In this work, a discrete-event data-driven simulation platform developed by the Advanced Centre for Biochemical Engineering at UCL (Stonier et al., 2012, 2013) was used to model the performance of bioprocesses exposed to uncertainties and facility constraints. The model captures the mass balances, equipment sizing, dynamic resource allocation and process economics of purification sequences. Monte Carlo simulation methods have been used in this work to mimic a batch history record by accounting for key process fluctuations and generating the possible outcomes and their likelihood. These simulations enable predictions of the impact of process fluctuations on the possibility of product loss. Monte Carlo simulation has been used increasingly in various bioprocessing examples to capture the impact of technical, clinical or commercial uncertainties on unit operation models (Sin et al., 2009), whole bioprocess costs (Farid et al., 2005; Pollock et al., 2013) and on portfolio management and capacity planning decisions (George and Farid, 2008).

Data mining has been used in the biotech sector to identify trends in large datasets from historical batch records, often applied to fermentation data (Charaniya et al., 2010; Mercier et al., 2013; Rommel and Schuppert, 2004). Principal component analysis (PCA) is a common multivariate analysis method that uses an orthogonal transformation to convert a set of variables into a set of linearly uncorrelated variables. It has been applied in manufacturing process analysis to reveal the internal structure and pattern of historical data (Edwards-Parton et al., 2008; Pate et al., 1999; Thornhill et al., 2006) but cannot generate the potential rules hidden behind the data. Decision tree classification is an effective data mining method that has been applied in fermentation parameter identification (Buck et al., 2002; Ma et al., 2004) and fermentation process optimization (Coleman et al., 2003; Lam and Malik, 2001). In this work, the classification and regression tree (CART) was introduced to analyse the large complex downstream manufacturing bioprocess datasets generated by Monte Carlo simulations and to find the hidden root causes of bottlenecks in existing facilities exposed to batch-to-batch variability and higher titres. The data mining outputs can be used to support better process understanding through rigorous root cause analysis and continuous risk management and hence contribute to effective implementation of quality by design (QbD) principles throughout the lifecycle of a product.

This paper is organized as follows. First, downstream bioprocess facilities used in the case study are described. Second, the methods applied in the case study including stochastic discrete-event simulation, correlation coefficients analysis and CART decision trees are briefly introduced. In Section 4, the Monte Carlo simulation datasets are analysed to identify mismatches in pool volumes resulting in product losses. The key process fluctuations leading to mass loss and threshold values for those process fluctuations are derived using CART decision trees. This work demonstrated that the decision tree classification method can be applied to explore not only the impact of process fluctuations on product mass loss but also the critical combinations of parameter values that lead to mass loss. Furthermore, the pictorial CART tree result with its series of if-then rules of the critical combinations of factors that lead to different mass loss levels can be used to identify debottlenecking solutions worth pursuing. Finally, three different debottlenecking solutions are compared in relation to their impact on three key metrics: mass output, direct cost of goods per gram (COG/g) and processing time. The solutions explored spanned changes to equipment sizing, using more efficient purification resins and reducing the eluate volume fluctuations expected through buffer optimisation.

Table 1
Facility specification for the chromatography and filtration downstream processing steps.

Parameter	Step			
Chromatography	Protein A	AEX	CEX	
Column diameter (m)	1	1	1	
Bed height (m)	0.20	0.25	0.15	
Bed volume (L)	157	196	118	
Load capacity (g/L)	25	50	15	
Linear velocity (cm/h)	450	450	140	
Expected number of cycles	9	3	15	
Expected pool volume (*CV*/cycle)	2	3	2.5	
Pool tank volume (L)	5000	5000	5000	
Expected step yield (%)	88	88	88	
Filtration	Post Protein A UF/DF	Post AEX UF	Final UF/DF	VRF
Retentate tank volume (L)	5000	5000	5000	5000
Expected average flux rate (L/m^2 h)	100/55	110/60	140/80	N/A
Target concentration (g/L)	25	25	38	
Diafiltration volumes (CV)	3	0	10	0
Expected step yield (%)	95	95	95	99

Note: Pool volume refers to the volume of the product stream. In Protein A and CEX steps operated in bind-and-elute mode this refers to the eluate volume collected. In AEX operated in flow-through mode this refers to the load and post wash volumes collected.

2. Problem domain

An existing standard monoclonal antibody (mAb) manufacturing process was considered in this work, as shown in Fig. 1. The volume of bioreactor broth generated during each batch was 10,000 L. Biomass and other debris were removed using centrifugation and depth filtration with step recovery yields of 95%. The mAb downstream processing sequence was defined as: Protein A affinity chromatography capture step, low pH virus inactivation, ultrafiltration/diafiltration (UF/DF), anion exchange chromatography (AEX), ultrafiltration (UF), cation exchange chromatography (CEX), virus reduction filtration (VRF) and a final UF/DF.

The downstream process was originally built to handle titres up to 2 g/L but it now needed to cope with average titres of 4 g/L and hence a harvest kg/batch value of 40 kg rather than 20 kg. The impact of the higher titre feed on the number of cycles required for each DSP step and hence the expected pool volumes from each step were calculated and used to allocate larger product collection tanks where appropriate. The specification of the downstream process equipment sizes and process parameters (e.g. resin binding capacities) is presented in Table 1. This facility configuration, modified to cope with 4 g/L titre feeds, was identified as the base case facility. The aim was to investigate the impact of batch-to-batch variability on its performance and predict facility fit issues.

Facility fit assessments are carried out often with information from a limited number of batches at scale, particularly for new processes or new drug candidates. In the absence of a significant number of batch history records such assessments are typically based on expected or worst case values which do not capture the full range of possible outcomes or their likelihood of occurrence. Hence in this paper stochastic simulation datasets were generated as a mimic of batch record data and then analysed using data mining techniques. The simulation datasets capture typical batch-to-batch variability expected at large scale which can be useful to companies