Fingerprint Detection and Process Prediction by Multivariate Analysis of Fed-Batch Monoclonal Antibody Cell Culture Data

Michael Sokolov, Miroslav Soos, Benjamin Neunstoecklin, Massimo Morbidelli, and Alessandro Butté
Dept. of Chemistry and Applied Biosciences, ETH Zurich, Institute of Chemical and Bioengineering, Zurich, Switzerland

Riccardo Leardi
Dept. of Pharmacy, University of Genova, Genova, Italy

Thomas Solacroup, Matthieu Stettler, and Hervé Broly
Biotech Process Sciences, Merck Serono S.A., Corsier-sur-Vevey, Switzerland

DOI 10.1002/btpr.2174
Published online October 7, 2015 in Wiley Online Library (wileyonlinelibrary.com)

This work presents a sequential data analysis path, which was successfully applied to identify important patterns (fingerprints) in mammalian cell culture process data regarding process variables, time evolution and process response. The data set incorporates 116 fed-batch cultivation experiments for the production of a Fc-Fusion protein. Having precharacterized the evolutions of the investigated variables and manipulated parameters with univariate analysis, principal component analysis (PCA) and partial least squares regression (PLSR) are used for further investigation. The first major objective is to capture and understand the interaction structure and dynamic behavior of the process variables and the titer (process response) using different models. The second major objective is to evaluate those models regarding their capability to characterize and predict the titer production. Moreover, the effects of data unfolding, imputation of missing data, phase separation, and variable transformation on the performance of the models are evaluated. © 2015 American Institute of Chemical Engineers *Biotechnol. Prog.*, 31:1633–1644, 2015

Keywords: multivariate data analysis, principal component analysis, partial least squares regression, cell culture process, quality by design

Introduction

Cell culture process development for therapeutic proteins

More than two hundred biopharmaceutical products have been approved for clinical use in the European Union and the USA with sales in 2013 over US$140 bln and with a steadily increasing fraction of monoclonal antibodies (mAbs) of 27% until 2014.[1] The clinical and commercial success of those biopharmaceuticals encouraged the large-scale production up to volumes of 20,000 L.[2] To ensure optimal production conditions and robust processes, as well as high product quality and safety, process development must be carried out using multiple stages. This usually starts with the selections of a cell line with a suitable expression system, of a cell medium and of a reactor system. Hereby, optimal ranges for the process variables have to be determined and ensured by characteristic process control methods. Also adequate mixing, mass transfer of the vital gases O_2 and CO_2 as well as nutrient supply have to be ensured and necessary additives such as shear protective agents or anti-foam have to be considered. Finally, an optimal product purification strategy has to be established in the process downstream. Such a process development taking into account all described factors is performed at different scales and may take many months to years.

Multivariate data analysis (MVDA) in biopharmaceutical productions

Despite the stringent control of several process variables, considerable variation in the process outcome is often observed.[3] Generalized deterministic models for dynamic process prediction based on the impact of the process variables on the product quantity and quality attributes do not exist. Moreover, the so-called univariate experimental analysis, comparing a variety of individual process variables, is not able to capture all the important interactions and relationships in the complex and dynamic cell culture system.[4] Therefore, for process development, multivariate data-driven knowledge discovery approaches are needed and they are applied at different stages of the process development to reveal important information on the process.

The frame for the implementation of such approaches and tools for process development has been set by the Pharmaceutical Quality for the 21st century initiative and, successively, by the Quality-by-Design (QbD) initiative,[5] both introduced by the US Food and Drug Administration (FDA) in pharmaceutical industry to improve product consistency

Additional Supporting Information may be found in the online version of this article.

Correspondence concerning this article should be addressed to A. Butté at alessandro.butte@chem.ethz.ch and safety. The key concept of these initiatives is that quality must be built in the product: on the one hand, by an improved knowledge of both the product and the development and manufacturing processes associated to it; on the other hand, by an improved knowledge of the risk associated to manufacturing, together with the tools and strategies to mitigate it. The key tools to achieve such goal are described in the Process Analytical Technology (PAT) initiative. According to PAT, multivariate data acquisition and analysis should be implemented to achieve an advanced knowledge of process and to gain an effective control on risk in manufacturing. Both these two aspects will be the major focus of this work.

Process classification, characterization, and prediction in the space of the process variables by means of multivariate data analysis (MVDA) will be discussed, in particular. MVDA is the simultaneous analysis of the input variables, thus enabling to cope with the complexity of the entire process. Within the last 5 years three review articles[6–8] have covered the use of Principal Component Analysis (PCA)[9] and Partial Least Squares Regression (PLSR)[10] for the analysis of biotechnological processes, including investigations such as peptone screening,[11] batch fault detection and diagnostics,[12] cell culture media screening,[13,14] product titer prediction,[15] scale-down model qualification[4] and evaluation of the lactate consumption as a process indicator.[16] Especially for mammalian cell culture process analysis, other MVDA methods have been applied very rarely so far. For instance, Le et al.[16] showed that in their analysis the results obtained from PLSR and support vector regression (SVR) were comparable, Charaniya et al.[3,17] presented several alternative approaches to access the dynamic process characteristics in the large variable space and Lu et al.[18] as well as Schmidberger et al.[19] extended the process prediction to the product quality attributes utilizing SVR, amongst others.

In this work, we present a sequence of increasingly complex data analysis methods combining univariate and multivariate analysis, while stepwise increasing the process understanding as well as the prediction accuracy. Such techniques are applied to a process qualification data set consisting of more than 100 batches.[20,21] To our best knowledge, such a detailed and versatile analysis has never been applied to cell culture process data on such an extensive set of data using the techniques mentioned above. It will be discussed how this statistical engineering approach allows to access various (dynamic) upstream process characteristics and patterns (fingerprints), both in terms of variable importance, under different perspectives, and in terms of variable interactions and trajectories in the time domain. Different techniques to improve process prediction will be highlighted, together with the impact that enhanced predictability could have on control and reduction of sampling efforts.

Materials and Methods

Batch data set and experimental plan

The analyzed fed-batch data set comprises two published data sets by Rouiller et al.[18] and Neunstoecklin et al.[21] The cell culture process was performed with CHO cells in animal-derived component-free medium in 3.5 L (maximal working volume) glass stirred tank reactors. The process was usually run for 10 days, while few experiments investigated the advantages of harvesting after 12 days. The standard culture conditions for the entire data set embody a set point of dissolved oxygen (DO) at 50% of air saturation, a controlled drift of an initial $\mathrm{pH} = 7.2$ to a standard set point of $\mathrm{pH} = 7.0$, a cell seeding density ($X_{V,0}$) of 1.4×10^6 cells mL^{-1} and a mean hydrodynamic stress value of 2.2 Pa. At the start of the fermentation, the pH was controlled by CO_2 and then by lactic acid and sodium hydroxide. The DO was controlled by a constant air flow. In most cases, glucose was fed daily from day 3. The process evolution was measured in daily samples, whereby not all the variables were quantified at each measurement step. The utilized analytical devices are: Cell density and viability (Vi CELL, Beckman Coulter, USA), pH (Seven Multi, Mettler-Toledo, Germany), pO_2 and pCO_2 (ABL5 blood gas analyzer, Radiometer, Switzerland), metabolites (Nova CRT, Nova Biomedical, USA), turbidity (Turb 550, WTW, Germany) and osmolality (Micro Osmometer, Advanced Instruments, USA), product titer (standard HPLC analysis using Protein A).

In the following, those experimental data will be referred to as batches in accordance with the statistical process analysis concepts applied.[22,23] The first set (published in parts by Rouiller et al.[20]) consists of 104 batches and represents a rigorous approach to analyze the effect of several process parameters, the peptones in the medium as well as the amplification procedure. This set can be classified in three different groups. The first group includes 33 (13 unique, with up to three replicates per experiment) experiments at standard conditions with slight process modifications, which were carried out in intervals throughout the process development. The second group consists of 48 experiments. Of these, 41 designed experiments (23 unique, with up to one replicate) investigated the effects of DO in the range of 10–90% air saturation, the $X_{V,0}$ in the range of 1.0×10^6 cells mL^{-1} to 1.8×10^6 cells mL^{-1} and pH in the range of 6.7–7.2. The remaining seven (four unique, with up to one replicate) experiments were carried out to analyze the effect of the partial pressure of CO_2 (abbreviated as pCO_2) being controlled to a high target level. The third group involves 29 (21 unique, with up to one replicate) experiments aimed at testing process robustness. In these experiments, special aspects such as new lots used for the media or the ratio of wheat and soy grain extract in the peptone media were tested.

The second data set produced by Neunstoecklin et al.[21] contains 12 batch runs and targets the analysis of the maximal operating range for hydrodynamic stress. Those 12 experiments involve four (three unique, maximally one replicate) experiments at standard conditions testing different shear induction mechanisms and eight (six unique, maximally one replicate) experiments investigating the effect of an increased shear stress in the range of 15–105 Pa. This value represents the maximal value of stress in the system as computed by Computational Fluid Dynamics calculations for stress induced by the velocity gradient and turbulent fluctuations.[21]

Multivariate analysis software and methodology overview

The statistical toolbox in MATLAB (MATLAB R2013a, The MathWorks) was used for the analysis. PCA was performed using *princomp* and PLSR with *plsregress* as well as the *libPLS package* provided in the Matlab Central environment offering a more versatile and flexible analysis with PLSR.

The major preparation steps are problem definition, data preparation, definition of the input or independent variables X and output or dependent variables Y, data unfolding of the three-way structure, data recovery to handle missing data and data pretreatment to ensure a comprehensible variable comparison. Those are followed by univariate analyses of the time evolutions of the variables and of the impact of the manipulated variables on the process outcome. Then, PCA is applied for precharacterization of the X variables using a grouping constraint to simplify the interpretation. Finally, PLSR models are built to analyze the variable importance and predict the time varying process response.

Description of analyzed variables and multivariate techniques

The set of X comprises 12 measured variables in total, which are different regarding both their process classification and their dynamic behavior. Four variables, namely pH, DO, pCO_2, and stress can be considered as characteristic environmental conditions of the process. pH, DO, and stress, are controlled (pH starts at an initial value of about 7.2 and converges toward its set point over time), while pCO_2 is controlled only in a few experiments. Viable cell density (X_V) and cell viability (Via) are characteristics of the cells in the reactor. Note that X_V is also a manipulated input variable, since its initial condition ($X_{V,0}$) is varied in the experimental set. The following five variables, namely the concentrations of glucose (GLC), lactate (LAC), ammonium (NH4), glutamine (GLN), and glutamate (GLU), give an insight into the cell metabolism, while the osmolality (OSM) describes the amount of osmotically active molecules in the medium. The concentrations of glutamine and glutamate refer to the ones in the aqueous solution. Apart from glucose, which is fed from day 3, the other six variables as well as the viability were not experimentally controlled, neither at the beginning of the process (initial conditions), nor throughout the cell culture process. The variable osmolality comprised 24% of randomly distributed missing values, while all the other X variables featured <4% of missing values. As most of the product attributes were measured only towards the end of the process, it was not possible to utilize those to model the time evolution of the process response, which is one objective of this work. Therefore, the only applicable response variable Y is the titer, i.e. the cumulated amount of product in solution. It contained 56% of missing values, as it was usually measured on even days.

The two most common ways to unfold a three-way data structure spanned by the dimensions of the variables, the batches and the time, were used.[22,23] Variable- or observation-wise unfolding results in a long data structure where the columns represent the variables and all the batches are distinguished with respect to the time points along the rows (i.e., variables of one batch at two time points being represented by two rows). Batch-wise unfolding results in a wide data structure where the rows represent the batches and the variables are distinguished with respect to the time along the columns (i.e., variables of one batch at two time points being represented by two columns).

Three strategies to deal with missing values were considered in this work: (i) Eliminating all rows which comprise missing X (and Y) values for PCA (and PLSR); (ii) Interpolating the missing values separately for all the X and Y evolutions; and (iii) Imputing the missing data in the entire X space using the iterative algorithm by Walczak et al.,[24] and imputing the entire missing Y (titer) data by regressing Y with a logistic function (i.e., also extrapolation). This is possible due to the characteristic profile of titer (shown in Figure 1B). The reason for performing the imputations separately in the X and Y space is to eliminate any information exchange in order to ensure unbiased prediction.

The data was auto-scaled prior to the multivariate analysis, i.e. subtracting the mean of each column and dividing each column by its standard deviation.

The model performance is evaluated using k-fold cross-validation, which is one of the standard tools in PLSR analysis.[10] In our case, the data set is first divided into $k = 5$ groups and five separate PLSR models are built based on a data set with one of the different groups being deleted. Then, the Y values in the left-out groups are predicted and the corresponding root mean squared error in cross-validation (RMSECV) is calculated from the summed squared differences of the actual and predicted Y values. The (relative) variance explained in cross-validation (Q^2), i.e. the ratio of the explained variance to the total variance, is used as a further characteristic of single models. In general, those measures enable to quantify the prediction capability of a model with the aim of reducing overfitting.

The interrelationships between the X and Y variables are widely characterized by the computation of β-coefficients as well as Variable Importance in the Projection (VIP)[13,25–27] and W^*-loadings.[4,10,27] β-coefficients are the regression coefficients of the auto-scaled X variables corresponding to the Y variable. The VIP values of the X variables represent their relative importance to explaining both the X and Y space in the latent variable (LV) model. Compared to the other two measures those are always positive and the value of one reflects the average impact of an X variable. The W^*-loadings quantify the respective weight of the X variables on each LV. If for instance, the first LV accounts for a large portion of the explained variance, the X variables featuring an absolutely high loading on this LV, are likely to be influential.

Results and Discussion

Univariate preanalysis

Figure 1 shows the evolution of the product titer and of three selected X variables, namely X_V, the glucose concentration and pH, as functions of time for all the investigated 116 batches. The time evolutions of the remaining X variables are shown in the Supporting Information in Figure S1 (apart of DO and stress, for which the constant controlled value was used throughout the entire culture). To simplify the interpretation of the results, batches were *a priori* classified into two groups: 91 high titer batches producing more than half of the maximal harvest titer, and 25 low titer batches producing less. This classification is evident in Figure 1B, where the distinction between high (black curves) and low titer batches (red curves with dots) is imposed. Note that the high titer batches show in most cases a steeper curvature in the titer evolution, whereas the low titer batches tend to flatten out earlier. The evolution of the cell viability, X_V, in Figure 1A shows that good performing batches exhibit a similar characteristic pattern, with a maximal viable cell density usually around day 6, indicating the transition from cell growth to cell death. As expected low titer batches also feature a lower level of viable cells. Figure 1C shows the time evolution of glucose concentration. This is a direct

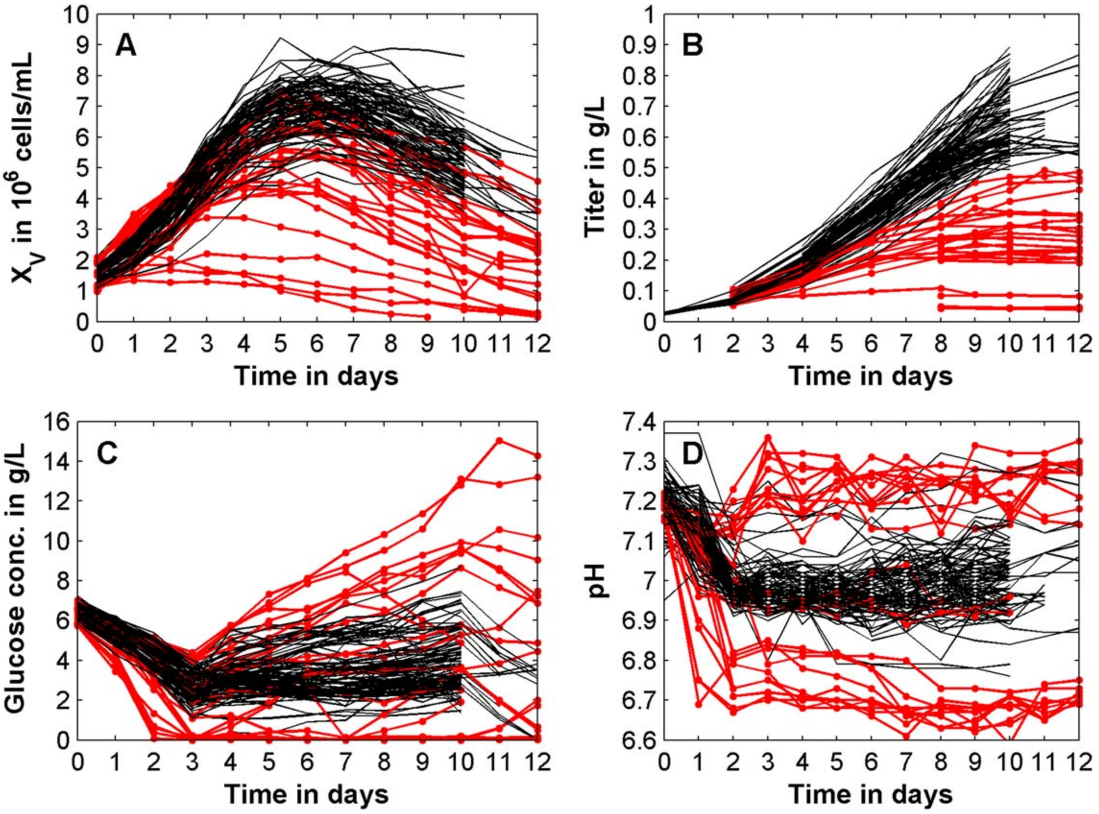

Figure 1. Time evolutions of X_v (A), titer (B), the concentration of glucose (C), and pH (D). The low titer batches are visualized by dot-markers. The remaining evolutions are shown in Figure S1 in the Supporting Information.

consequence of the glucose feeding strategy, according to which no glucose is fed until day 3. Most of the high titer batches perform in a similar way, with a roughly stationary glucose concentration after day 3. In few of them and in several low titer batches, the daily glucose feed (usually constant at 6 g L^{-1}, adjusted in cases when glucose concentration was significantly outside of desired range) caused a steady increase in glucose concentration, i.e. glucose accumulated due to a low consumption caused by early cell death. On the other hand, glucose starvation, where glucose is fully depleted, was observed only in low titer batches. Nonetheless, some low titer batches fall within the characteristic glucose pattern. Finally, the pH evolution shown in Figure 1D reveals that the vast majority of low titer batches are represented by runs with low (below 6.8) or high (above 7.2) pH values.

The classical univariate analysis above has enlightened that the evolution of most of the variables reveals some characteristic patterns. This analysis was simplified by the introduction of an arbitrary quality criterion with respect to the titer value, which allowed distinguishing between high and low performance batches. Although this empirical grouping is helpful to gain some basic understanding of the process, it is not possible to evaluate the overall variable interactions and differences within those two groups.

The observations of the pH profile suggest applying an additional univariate investigation technique. In Figure 2, the scatter plot of the titer at day 10 vs. the four manipulated process variables (pH, DO, stress, and $X_{V,0}$) is shown. To simplify result visualization, the same classification between high titer (empty circles) and low titer (filled circles) is applied. Figures 2A,B show that almost all low titer batches can be characterized by experimental set points for pH and DO at the lowest or highest values in the investigated range. The investigated ranges of the cell seeding density and stress (shown in Figures 2C,D) seem to not cover a large enough range which includes the so-called edge of failure. In fact, no evident separation between high and low titer batches based on those two variables can be observed. Note that due to difficulty in precise control the cell seeding density is rather scattered 1.0 × 10^6 cells mL^{-1} and 2.0 × 10^6 cells mL^{-1}. Therefore, this value should be regarded as a poorly controlled initial process condition.

The analysis above has revealed the existence of an optimum at standard conditions of pH and DO, which, in turn, is suggesting the existence of a nonlinear relation between final product titer and these to variables. This is for instance in agreement with the pH dependency presented by Villadsen et al.[28] In spite of this, this result also highlights the limitation of univariate approaches, since, for example, from a purely visual investigation it is impossible to differentiate among the 91 high titer batches, as these are all following similar trends.

Preregression analysis with PCA

The major objective of this analysis is to visualize the evolutions and the correlations among the 12 input variables by multivariate analysis and to infer characteristic patterns and deviations among the different batches. To analyze the evolutions, the data set was unfolded variable-wise and the rows incorporating missing values were deleted. This resulted in a matrix consisting of 1,020 observations (rows) and 12 variables (columns).

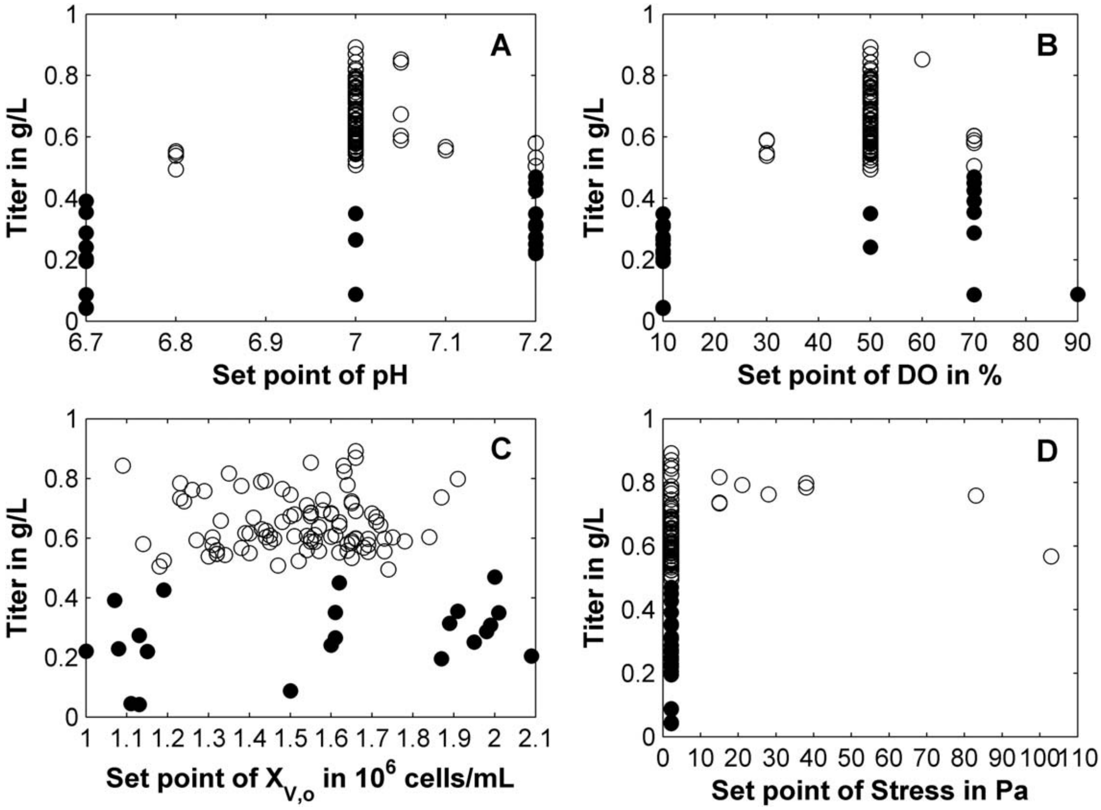

Figure 2. **Impact of set points of pH (A), DO (B), $X_{v,0}$ (C), and stress (D) on product titer at day 10. The filled symbols represent low titer batches and empty symbols represent high titer batches.**

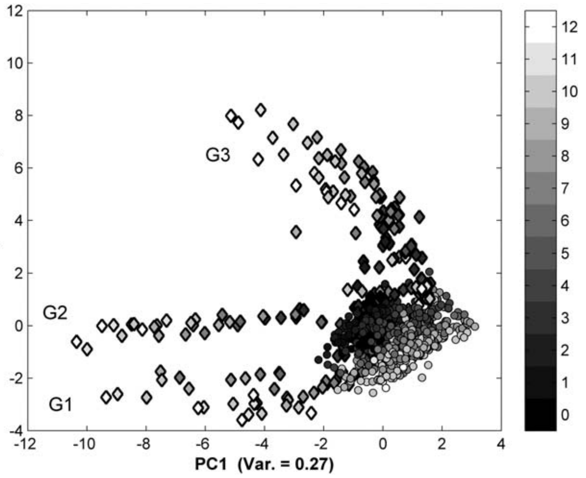

Figure 3. **PCA of the *X* variables. Score plot showing time evolution of batches from day 0 (coded black) to 12 (coded white). The groups G1, G2, and G3 mark the evolutions for low pH and DO, low pH and high DO, high pH and low DO, respectively. The numbers in brackets indicate the *X* variance explained by each PC.**

The resulting score plot shown in Figure 3 clearly differentiates between the evolutions of low titer (diamond symbols) and high titer batches (circles). The dynamic behavior is visualized based on culture time, from initial conditions (day 0 in black) to final ones (day 12 in white). The initial conditions are all located in a distinct region, close to the plot origin. Then high titer batches are twirling around the origin, while low titer batches diverge from it. This trend is also visualized in Supporting Information Figure S2, where the first two PCs are plotted against time. In other terms, PCA is revealing that in low titer batches all process variables are moving together in a coherent way and follow a trajectory different from the high titer batches. In particular, one can identify, e.g., using the corresponding loadings (shown in Supporting Information Figure S2), three characteristic groups, which were labeled G1, G2, and G3, respectively. The first two groups correspond to experiments at pH = 6.7 as well as high (70 and 90%, G1) or low (10%, G2) conditions for DO, respectively. Group G3 corresponds to the experiments performed at high pH (7.2) and low DO (10%) levels. The five remaining low titer batches could be identified around the borderline of the bulk region with group G3 as fed-batch processes performed at high pH or low DO and standard conditions for the other parameter. This means that using this analysis the strong effect of pH, and in particular the combinatory effect of pH and DO at extreme conditions could be revealed in a single plot. For completeness, the plot of PC3 against PC4 could be considered, which in total explain further 24% of the variance. However, the information provided by those PCs does no longer reflect the different behavior of low and high titer batches, while showing the characteristic evolution of X_V and the effect of stress. The corresponding score and loading plots are provided in Supporting Information Figure S2.

Therefore, the results obtained with PCA go beyond the findings of the univariate analysis and give important information to characterize the *X* evolutions prior to regressing those to the response *Y*. The advantage of such an analysis for data featuring a dynamic behavior is the additional visualization of the characteristic multivariate evolutions, which, for instance, is the basis for batch monitoring.

Batch evolution modeling with PLSR

The major objective of this analysis is to characterize the interrelation of the evolutions of the 12 input variables with the evolution of the output variable, the product titer. Moreover, the effect of missing data imputation strategies and elimination of the low titer batches on the model performance shall be evaluated. The analysis corresponding to the latter goal is presented in the section S1 in the Supporting Information. For the first goal, a PLSR model was built based on a variable-wise unfolded data set, comprising 555 observations, where information was available for the 12 *X* and the *Y* variable, i.e., all incomplete rows were eliminated. The model quality is presented in the first row in Table 1. Considering the RMSECV and the titer range in Figure 1B (up to 0.9 g L^{-1}), the prediction accuracy goes beyond the possibility to distinguish high and low titer batches. Nonetheless, it is clearly above the experimental error, which is about 0.030 g L^{-1}. This value was determined as the average standard deviation of repetition experiments, for the final titer only. The reason for this inaccuracy is the steady state assumption in the PLSR models based on variable-wise unfolded data. Such models presume a time-invariant correlation structure among variables, i.e. the loadings (and regression coefficients) are constant for each variable at every point of time.

Table 1. RMSECV (in g L^{-1}) in Phase-Separated PLSR Observation Evolution Models Distinguished by Glucose Feeding or Beginning of Cell Death (at $X_{v,max}$) Compared to Results for Full Data Set*

		# LVs		
Data Set	*N*	1	2	3
Full	555	0.138	0.097	0.094
Phase 1 (no feeding)	159	0.028	0.021	0.020
Phase 2 (with feeding)	396	0.125	0.091	0.089
Phase 1 ($\leq X_{v,max}$)	252	0.079	0.070	0.069
Phase 2 ($>X_{v,max}$)	303	0.104	0.096	0.091

*The first column represents the number of observations *N* in the model and the three numbered columns are related to the number of latent variables used in the model.

However, the major characteristic of models based on variable-wise unfolded data is their capability to distinguish different observations according to their time location in an evolution. This characteristic has a central role for applications in monitoring of the process evolution. Similarly to Figure 3, Figure 4A discriminates high (circles) and low titer batches (diamonds) along the time course of the process (coded by color). Therefore, Figure 4A along with Figure 3 shall be summarized as the observation evolution fingerprint of the process. Figure S3 in the Supporting Information shows the time evolution of the scores of the first two LVs. The plot corresponding to LV1 clearly identifies the different evolutions of the low and high titer batches. Depending on the experimental accessibility of the response variables and the strategy for process control, those results can be used as a basis to determine corresponding batch control plots. In other words, while the trajectories given by the PCA model in Figure 3 and Supporting Information Figure S2 are singly governed by the interrelationship of the *X* variables, the trajectories in Figure 4A and Supporting Information Figure S3

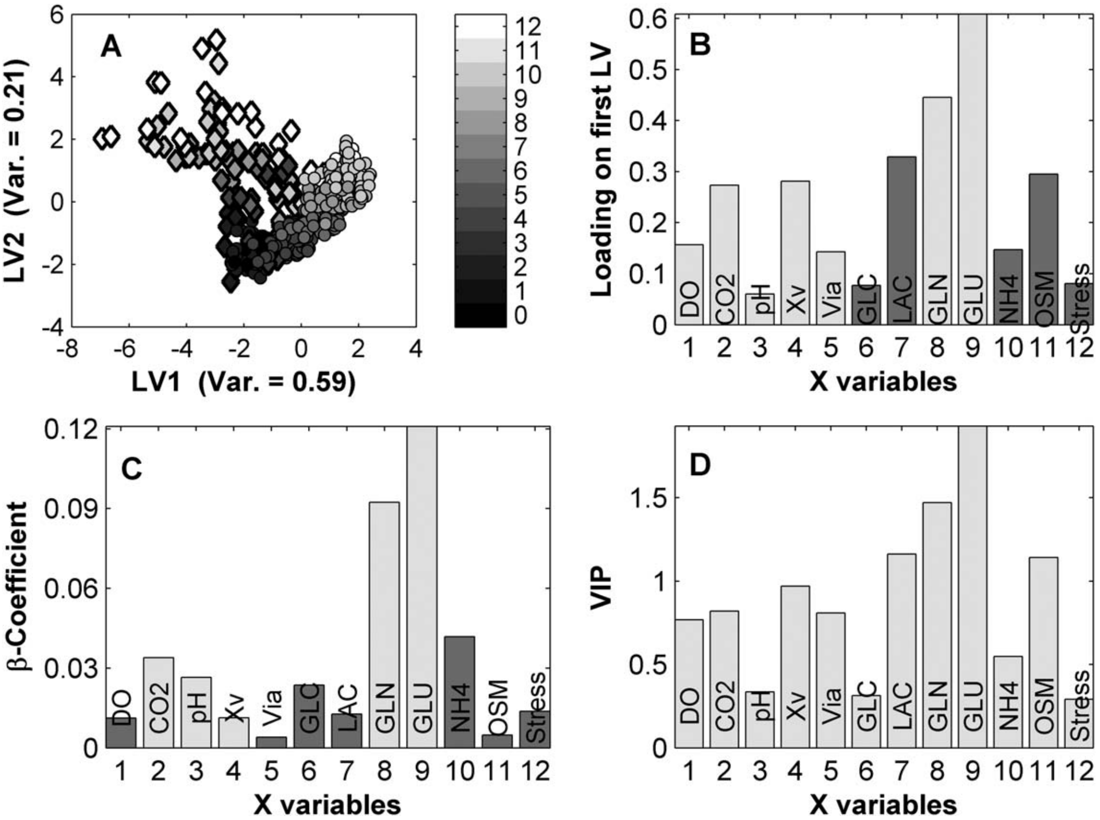

Figure 4. Observation evolution fingerprint based on PLSR model for variable-wise unfolded, nonimputed data set. A: Score plot showing time evolution of batches from day 0 to 12 (diamonds represent low titer batches, circles represent high titer batches, values in brackets represent the *Y* variance explained by each LV), B: *W-loadings on first LV, C: β-coefficients (for auto-scaled data), D: VIP plot. Deep gray bars indicate negative values.**

are based on constantly predicting the *Y* variable with the *X* variables along the time.

Figures 4B–D show the three utilized measures of *X* variable importance (β-coefficients, VIP and W*-loadings on the first LV) applied to the PLSR model. Chong et al.[29] concluded that the VIP values and the β-coefficients can be used complementarily and that in the case of both being small a variable can be considered as unimportant for the model. In this case, all three measures reveal that the concentrations of glutamate, glutamine, and lactate as well as the viable cell density and the osmolality are important in the overall observation evolution model of the titer. Hereby, the large explanatory power of the concentrations of glutamine and glutamate stands out in the analysis, which will be further discussed in Phase Separation and Time Grouping section.

The analysis so far highlighted the role of DO and pH and the consequences of deviation from the respective standard conditions. However, the results in Figure 4 indicate a minor role of these two variables. One way to understand this result is to split all variables into those which refer to controlled conditions and those which correspond to the time evolution of the process. Depending on the controlled environmental conditions (pH and DO), the cell culture process behaves differently, as indicated by the different process trajectories of the monitored variables (Figures 3 and 4A). In other words, such important monitoring variables are internal process responses of the controlled environmental conditions and tend to better represent the statistical variance in the data caused by the manipulated variables with respect to titer production. Like this, the variables which can be used as process levers can be characterized on the one hand, and the variables providing a significant response over time and, hence, are essential to be measured, on the other hand. The identification and evaluation of those two sets of variables is an important step toward process monitoring and control reducing the subsequent experimental effort in process development as well as the analytical requirements.

Phase separation and Time Grouping

The analysis in Batch Evolution Modeling with PLSR section evaluated the general importance of process variables for monitoring the overall batch evolution under the assumption of a time-invariant interrelationship of the *X* and *Y* variables. In this section the possibility to unveil characteristic properties of specific phases and time points shall be considered. Two phase separations were analyzed: a process related separation into two phases distinguished by the beginning of glucose feeding (around day 3, evident from glucose profile in Figure 1C) and a biological separation into two phases distinguished by the transformation from cell growth to increasing cell death (at maximal cell density of the X_v profiles shown in Figure 1A). Table 1 shows that the precision of the titer prediction can be improved in the first phase, while it remains similar to the overall observation evolution model for the second phase. Therefore, for this process those phase separations can be advantageous for the understanding and monitoring of the first phases. For instance, applying the feeding-based separation, pH (due to the pH-shift) and the glucose concentration demonstrate a higher importance in the PLSR model (not shown) compared to the fingerprint for the overall evolution shown in Figure 3.

Table 2. RMSECV (in g L^{-1}) and Number of Observations *N* in Time-specific PLSR Models at Days 2–10 (using two LVs) Compared to Full Observation Evolution Model

	Full	Day = 2	4	6	8	10
RMSECV	0.097	0.010	0.021	0.039	0.063	0.085
N	555	53	96	64	108	110

To analyze the dynamic change of the correlation structure, a separate PLSR model was developed for every second day (as the titer was usually measured on even days). Table 2 shows that the accuracy of the time-specific models worsens in time but is always better compared to the overall observation evolution model from Batch Evolution Modeling with PLSR section. This is also reflected in Figure S4 in the Supporting Information comparing the observed and predicted titer values for every time-specific model. The increase of the RMSECV in the time-specific models can be attributed to the growing system variance in time, which is evident from the increasingly broader evolution profiles of the variables shown in Figure 1 as well as in Supporting Information Figure S1. Relating the RMSECV to the respective titer range at that point of time, one can conclude that the relative prediction accuracy is quite comparable in the time-grouped models. Those models can be considered as local instantaneous models predicting the titer using a characteristic pattern at each point of time.

The corresponding variable importance can be evaluated for each of the time-specific models revealing a dynamic variable importance pattern based on instantaneous system characteristics. Figure 5 shows the VIP values for four selected variables in the time-specific models in comparison to the full evolution model (represented by the horizontal line). The remaining VIP plots are visualized in Figure S5 in the Supporting Information. Figures 5A–C show that the importance of the variables OSM, Via, and GLC in explaining the titer variance at each point of time is variant and is larger than the time-averaged importance of these variables in explaining the variance during the entire process. Hence, decoupling the model from the entire evolution is advantageous to access the dynamic variable properties. Figure 5D and Figure S5 in the Supporting Information show that the concentrations of glutamate and glutamine are far more important in the observation evolution model compared to the time-specific ones. This suggests that the latter models are capable of explaining the titer even without these *X* variables. Removing these two variables does not considerably change the performance of the time-grouped models, but significantly reduces the prediction accuracy and the variance explained (with two LVs: RMSECV = 0.185 g L^{-1}, $Q^2 = 0.40$) of the observation evolution model. Therefore, those two variables are likely to capture the time trend of the system. This is also reflected in the evolutions of glutamate and glutamine (shown in Supporting Information Figure S1), which very well resemble the one of the titer (shown in Figures 1B). The instantaneous (time-specific) fingerprint presented in Figure 5 and Supporting Information Figure S5 can be used to analyze a detected deviation of a batch evolution at a certain point of time comparing the sample to the respective instantaneous fingerprint of high titer batches or to detect characteristic dynamic transitions of variable interactions such as presented by Xing et al.[30]

Predictive batch modeling

The analysis in Phase Separation and Time Grouping section highlighted the time-variant correlation structure among

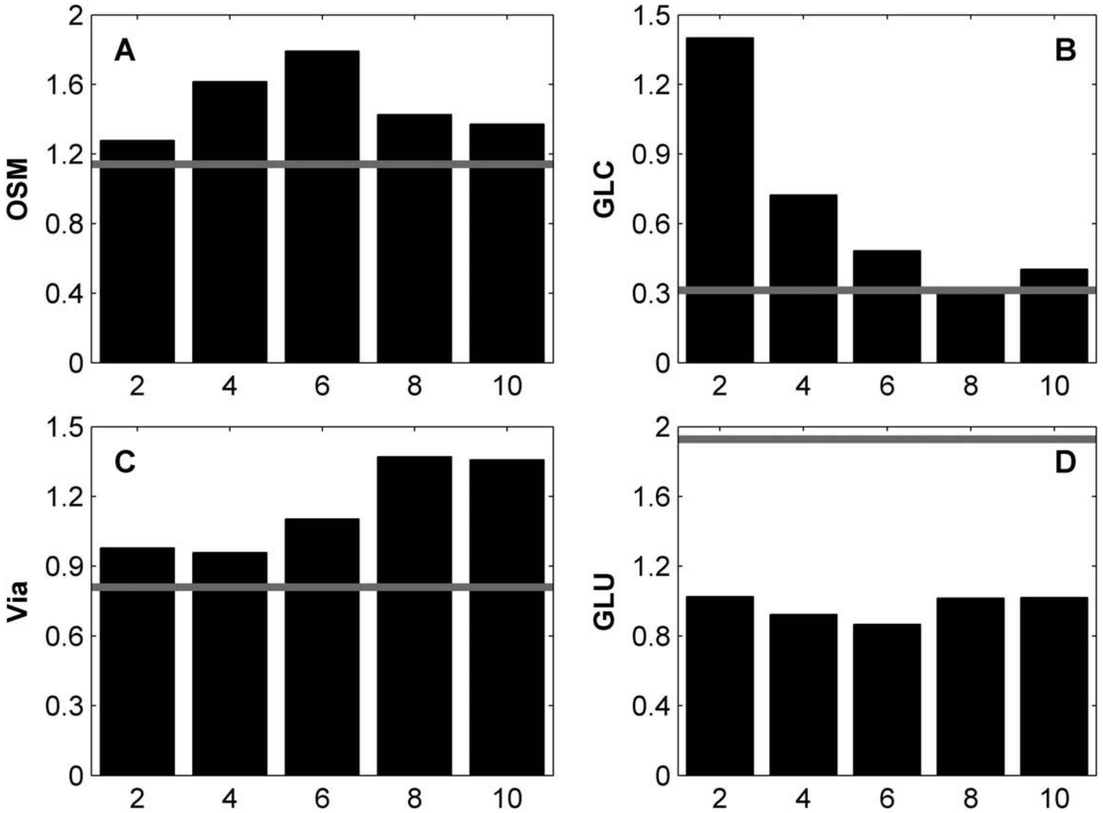

Figure 5. Instantaneous process fingerprint: Comparison of VIP values variable importance in overall (represented by solid line) and time-grouped or instantaneous PLSR models for four selected variables at five points of time (day 2–10): osmolality (A), glucose concentration (B), viability (C), and glutamate concentration (D). The results for the remaining models are presented in Figure S5 in the Supporting Information.

the variables as well as the corresponding instantaneous fingerprints for specific time points of the process. However, for prediction purposes, it is highly important, to include the process history in order to improve the prediction accuracy. Therefore, the data set is unfolded batch-wise resulting in a data matrix made by 116 rows (batches) and 112 columns (2 controlled variables: Stress and DO, 10 variables at days 0–10). To incorporate the entire historic information, the third imputation strategy using the iterative algorithm in the *X* space and the logistic function fit in the *Y* space was applied. This complete data set was used to predict the titer at a given point of time using the *X* history until that point of time. Specifically, for each day, different PLSR models were built based on the historic information starting from initial conditions and sequentially adding further historic data for each variable until the given day. These models were tested upon their ability of predicting the titer from the selected day to the end (day 10). Table 3 shows the quality of the PLSR models represented by the RMSECV (A) and the variance explained in cross-validation (B). A maximum of 10 LVs was used as a constraint in the model development.

The comparison of the results in Table 3 to the ones obtained with the PLSR model for the corresponding variable-wise unfolded data presented in Supporting Information Table S1 (model A3) demonstrates that this dynamic model, where correlations among variables change in time, has a significantly higher accuracy for modeling the final titer (RMSECV 0.051 compared to 0.104 g L^{-1}) and describes almost the entire variance of the titer (92% compared to 65%). It has to be noticed that the RMSECV is close to the process error ($\pm$0.030 g L^{-1}). This means that the prediction accuracy for this data set cannot be significantly improved. It is simple to calculate that the relative error of the prediction, i.e. the ratio between the RMSECV and the titer, is continuously decreasing as a function of time, in contrast to the absolute error reported in Table 3. This clarifies the fact that the variance explained in Table 3 is also improving and eventually stagnating with increasing time. When focusing on the prediction of the titer at a given day, the prediction is improving for increasing process time (moving down a column). However, it is interesting to notice that, after a few days, the error in prediction is already close to the analytical error. In particular, the variables encompassing day 0–3 already predict the titer at day 10 with an acceptable accuracy (RMSECV = 0.084 g L^{-1}). Therefore, the models provide a very early and reliable evaluation of the process outcome.

The results presented along the diagonal in Table 3 deviate from the ones presented in Table 2. On the one hand, the models presented in Table 3 incorporate the entire history and not only the measurements at the considered point of time so that a higher precision can be obtained. On the other hand, data recovery is likely to introduce inconsistencies into the data set, which especially is in the beginning of the process, where the titer was extrapolated in some cases, results in a higher variance compared to the raw data.

Given the unsatisfactory prediction accuracy of the models incorporating little process history, the extension of the variables pH and DO toward a profile with a maximum suggested by the observations in Figure 2 shall be analyzed.

Table 3. Prediction of Process Titer at Different Points of Time (Columns) Using Historic Data for the Process Variables (Rows)*

	Y_0	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
A: RMSECV in g L^{-1}											
X_0 (12)	0.055	0.054	0.057	0.049	0.046	0.050	0.060	0.082	0.106	0.127	0.140
$X_{0\text{-}1}$ (22)		0.049	0.053	0.048	0.044	0.047	0.054	0.074	0.092	0.113	0.125
$X_{0\text{-}2}$ (32)			0.053	0.045	0.041	0.044	0.049	0.070	0.077	0.107	0.117
$X_{0\text{-}3}$ (42)				0.043	0.040	0.038	0.036	0.050	0.055	0.072	0.084
$X_{0\text{-}4}$ (52)					0.039	0.036	0.033	0.042	0.044	0.061	0.074
$X_{0\text{-}5}$ (62)						0.035	0.029	0.037	0.041	0.062	0.068
$X_{0\text{-}6}$ (72)							0.026	0.033	0.035	0.057	0.063
$X_{0\text{-}7}$ (82)								0.033	0.034	0.056	0.060
$X_{0\text{-}8}$ (92)									0.032	0.054	0.053
$X_{0\text{-}9}$ (102)										0.051	0.052
$X_{0\text{-}10}$ (112)											0.051
	Y_0	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
B: Variance explained in cross-validation											
X_0 (12)	0.17	0.18	0.11	0.21	0.24	0.36	0.39	0.36	0.29	0.33	0.37
$X_{0\text{-}1}$ (22)		0.31	0.24	0.26	0.31	0.42	0.51	0.48	0.47	0.47	0.49
$X_{0\text{-}2}$ (32)			0.25	0.35	0.40	0.50	0.60	0.53	0.62	0.53	0.56
$X_{0\text{-}3}$ (42)				0.41	0.44	0.64	0.79	0.76	0.81	0.79	0.78
$X_{0\text{-}4}$ (52)					0.46	0.67	0.82	0.83	0.88	0.85	0.82
$X_{0\text{-}5}$ (62)						0.69	0.86	0.87	0.89	0.84	0.85
$X_{0\text{-}6}$ (72)							0.89	0.90	0.92	0.87	0.87
$X_{0\text{-}7}$ (82)								0.90	0.93	0.87	0.89
$X_{0\text{-}8}$ (92)									0.94	0.88	0.91
$X_{0\text{-}9}$ (102)										0.89	0.91
$X_{0\text{-}10}$ (112)											0.92

*The rows indicate the time history employed for all the *X* variables. The subscript and the number in brackets represent the time point or period and the number of predictors incorporated in the model, respectively.

Table S3 in the Supporting Information shows the obtained results using two additional variables: $pH_{ext}(t) = (pH(t)\text{-}mean(pH(t)))^2$ and $DO_{ext} = (DO\text{-}50)^2$. The model predictability early in the process (up to day 3) increases significantly, to become comparable to that without the addition of the two new variables at longer process times. This demonstrates that in the limit of little process history incorporated such as in black box models, where final conditions are predicted by initial conditions and process settings only, introducing some variable transformation is mandatory to explain the nonlinear process response. When process history is included, other variables can better explain the process variability and the effect of such transformations becomes irrelevant.

The special arrangement of the variables for this analysis enables to evaluate the dynamic variable importance fingerprint, i.e. the importance of all the variables distinguished in time along their history. As an example, the variable importance for the model incorporating the entire history (model for Y_{10} based on $X_{0\text{-}10}$ in Table 3) was analyzed. Figure 6 presents the score plot where low titer batches (smaller than $0.5 \cdot Titer_{max,day10}$), high titer batches (between $0.5 \cdot Titer_{max,day10}$ and $0.75 \cdot Titer_{max,day10}$), and very high titer batches (higher than $0.75 \cdot Titer_{max,day10}$) are distinguished. Regarding the variance explained by each LV (shown in brackets on the axis), the major variability is featured between the low and (very) high titer batches, which can be discriminated along the axis of the first score. Moreover, compared to high titer batches many very high titer batches tend to feature a higher value of the second score besides a slightly higher first score value.

Figures 7A,B show the *W**-loadings of the 112 predictor variables on the first two LVs, whereby DO and Stress are single variables and the remaining variables are shown as different predictors ordered by increasing time (10 variables times 11 process times). To obtain a high titer, the variables having large positive loadings on the first two latent variables favor high values of the titer and, therefore, have to be maintained at a high level. On the contrary, those with large negative loadings have to be maintained at a low level. Analyzing Figure 7, it is possible to say that the general importance pattern is quite comparable to the time-averaged pattern shown in Figure 4. In both cases, lactate, glutamine, and glutamate are very important to explain process variability, with CO_2, viable cell density and osmolarity in a second line. Depending on the chosen variable importance method, one could argue that this analysis emphasizes more the role of other variables, like ammonia and cell viability. On the other hand, using a time-variant variable correlation structure due to batch-wise unfolding, the concentrations of glutamine and glutamate are decoupled from their

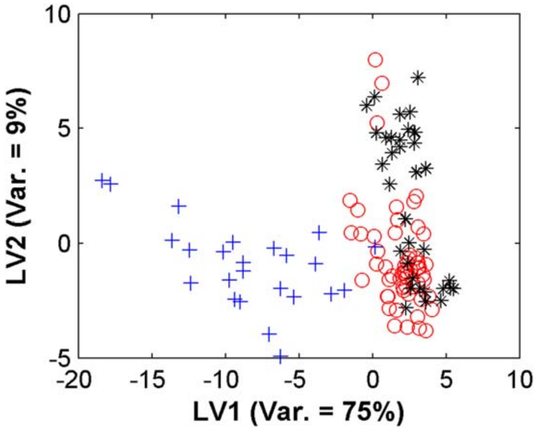

Figure 6. Score plot for the first two LVs of PLSR model for batch-wise unfolded data set incorporating the full *X* history. Low titer batches (producing <50% of maximal harvest titer) are represented by plus symbols, high titer batches (producing between 50 and 75% of maximal harvest titer) by circles and very high titer batches (producing more than 75% of maximal harvest titer) by asterisks. The numbers in brackets indicate the *Y* variance explained by each LV.

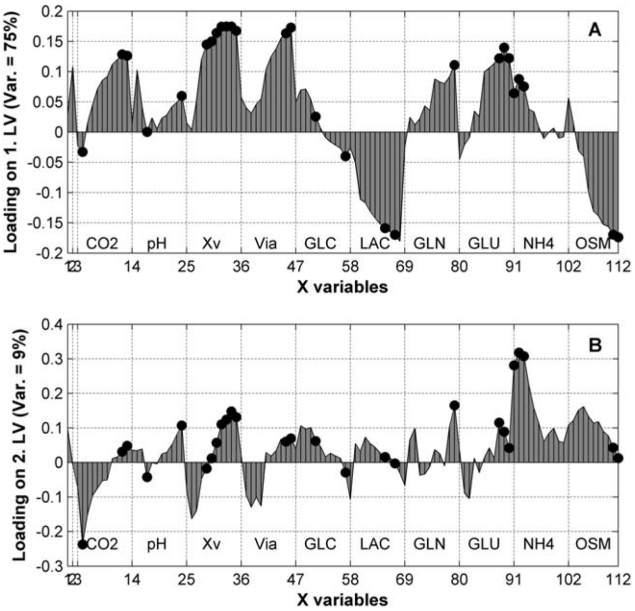

Figure 7. Dynamic variable importance fingerprint: *W-Loadings of 112 *X* variables on first (A) and second (B) LVs in PLSR model for the Titer at day 10 incorporating the entire variable history (ordered by increasing time). The vertical gridlines indicate a new variable at day 0. The first two *X* variables are stress and DO. The filled circles represent the variables selected by the genetic algorithm of Leardi et al.[31]**

strong correlation with the titer evolution and given the information by the other variables provide unique input toward the final days only.

The genetic algorithm elaborated by Leardi et al.[31] was tried out as a method to select the important variables (providing unique information) in the full sample history (112 *X* variables), while maintaining the prediction performance. Although such an analysis without a validation on a separate test set might be biased by overfitting, the results shall be considered as a possibility to gain a further insight into the process properties. The algorithm was run for ten times and those variables selected six times or more were considered as important. Table S2 in the Supporting Information shows the selection pattern and in Figure 7 the selected variables are presented by filled circles in the loading plots. Those plots demonstrate that the variance in the harvest titer is expressed by a specific subset of historic variables along the importance (*W**-loading) profile. For instance, the glucose concentration and pH feature importance at specific process points, the adaptation to the controlled pH set point and the beginning of glucose feeding, respectively. It is clear that once the pH set point is reached this information is not needed any longer. Regarding the loadings on the second LV, the conditions of the concentration of ammonium on the first days and the p_{CO2} at day 1 are likely to be the most important variables. The first effect can be attributed to characteristically high concentrations of ammonium in some of the very high titter batches, which after a further analysis was lined to a special amplification scheme. The latter effect can be explained by a special pH control methodology in some of the very high titter batches. In general, to better access and understand those discrimination criteria further process variables such as the split ratio in the inoculation and the added acid and base quantities or rates for pH control should be used in the *X* set, which however were not always available in the data set.

The variable importance obtained with this analysis is different from the patterns depicted with time grouping since the respective VIP plots shown in Figure 5 reveal the importance considering "snapshots" of the reactor at certain points of time. The variable importance in the whole batch model uses all the historic information measured throughout the process. The two controlled variables, DO and stress, were never selected by the genetic algorithm. Especially, regarding DO, which has shown to be an important variable in the Univatirate Preanalysis and Preregression Analysis with PCA sections, this does not mean that it is not an important variable. The variable selection depicts those variables, which represent the time-variant system most accurately. As most of the time-variant variables adapt to the set points of the manipulated variables, those are also more likely to better represent the process conditions regarding their monitoring and evaluation. However, in the case of a model incorporating only a short sequence of the process history, the time-variant process variables will not bear enough information on the process characteristics so that variable transformation of the manipulated variables can be of great advantage.

Summary and Conclusion

This work presents an application of linear multivariate methods to cell culture process fed-batch data. The three elaborated process fingerprints demonstrate different insights, which can be gained for process understanding and applied at different stages of process development. More specifically, first univariate plotting showed the distinguished effect of the controlled parameters pH and DO. Such an analysis is important to design and refine the subsequent analysis steps. Refinements include the reconsideration of the selection of the variables to be analyzed as well as an appropriate transformation, which is especially meaningful in black box models as well as hybrid approaches. The multivariate observation evolution analysis with PCA and PLSR revealed that the process features totally different trajectories if both, DO and pH, are at an extreme level at the same time. Moreover, the PLSR analysis distinguished the process variables important for process monitoring and process control. Therefore, the corresponding fingerprint provides a deeper process understanding building the basis for applications in process monitoring. While due to characteristically long (variable-wise unfold) data matrix the presence of missing data can be handled by eliminating the corresponding rows, the major limitation of those approaches is the time-invariant interrelationship assumption among the variables. To access time- or phase-specific effects, the instantaneous fingerprint was elaborated decoupling the variables from the process time. This approach reveals a more versatile insight into the dynamic influence of process variables, which can be used a basis for phase definition, fault detection and improvement of the monitoring models. The predictive model accuracy could be significantly improved in the last analysis, which uses the entire process history for PLSR. To incorporate the full process history, due to the broad (batch-wise unfolded) data set an appropriate strategy to deal with missing data has to be established. Moreover, the corresponding dynamic variable importance fingerprint revealed the unique information provided by certain variables along the time and can, hence, be used for simplification of the process analytics. The linear multivariate prediction models clearly improved the understanding of the process showing that the final process titer can be adequately forecasted based on the process data from the first 3–4 days and that the process history is capable of comprising the non-linear process behavior. These findings enable to make early process evaluation building the basis for process control.

For future work, it is interesting to apply such an analysis to cell culture data using further variables, which, for instance, account for the expansion process, the feeding strategy, the cell metabolism, the cell culture medium and the cell line. From the application point of view, monitoring and control studies incorporating online data will be of great interest. In general, it will be advantageous to perform the process analysis according to an experimental design in order to derive the maximal amount of process information using the least experimental effort. The performance of the linear models is satisfactory for this process data set, which already included a long standing development experience, so that not all of the variables were analyzed at the range of failure. Especially in the case of less robust process data sets at an earlier stage of process development, it will be interesting to compare the model performance to more complicated statistical models such as Support Vector and Random Forest Regression. Even more interestingly, will be the extension of those linear models to so called hybrid models combining the flexibility from the statistical models and the engineering knowledge incorporated in deterministic models to build robust analysis and prediction tools. Of course, the broadening of the analysis towards product quality attributes is very important to meet the goals of QbD.

The results in this work reveal how important process understanding can be obtained with various generally applicable univariate and multivariate process analysis methods. Those shall show that especially in cell culture process development, which is characterized by time demanding experiments, a large number of influential and interacting factors as well as high-cost raw materials and process analytics, multivariate data analysis offers attractive and versatile tools for efficient process development. We believe that the implementation of a sequential data analysis procedure providing an extensive quantitative feedback to the investigator will clearly simplify the subsequent decision making process such as the design of the next experiments. Hence, the experimental effort for the process development team can be remarkably reduced, which in turn yields lower costs and faster time to market.

Acknowledgment

The authors are thankful to Dr. Gianmarco Polotti who supported the project with fruitful discussions.

Literature Cited

1. Walsh G. Biopharmaceutical benchmarks 2014. *Nat Biotechnol.* 2014;32:992–1000. doi:10.1038/nbt.3040.
2. Birch JR, Racher AJ. Antibody production. *Adv Drug Deliv Rev.* 2006;58:671–685. doi:10.1016/j.addr.2005.12.006.
3. Charaniya S, Hu W-S, Karypis G. Mining bioprocess data: opportunities and challenges. *Trends Biotechnol.* 2008;26:690–699. doi:10.1016/j.tibtech.2008.09.003.
4. Tsang VL, Wang AX, Yusuf-Makagiansar H, Ryll T. Development of a scale down cell culture model using multivariate analysis as a qualification tool. *Biotechnol Prog.* 2014;30:152–160. doi:10.1002/btpr.1819.
5. US Food and Drug Administration R. *Pharmaceutical CGMPs for the 21st Century—a Risk-Based Approach.*; 2004. Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Pharmaceutical+cGMPs+for+the+21st+Century+-+A+Risk-Based+Approach#0. Accessed September 9, 2014.
6. Rathore AS, Mittal S, Lute S, Brorson K. Chemometrics applications in biotechnology processes: predicting column integrity and impurity clearance during reuse of chromatography resin. *Biotechnol Prog.* 2012;28:1308–1314. doi:10.1002/btpr.1610.
7. Ündey C, Ertunç S, Mistretta T, Looze B. Applied advanced process analytics in biopharmaceutical manufacturing: challenges and prospects in real-time monitoring and control. *J Process Control.* 2010;20:1009–1018. doi:10.1016/j.jprocont.2010.05.008.
8. Mercier SM, Diepenbroek B, Wijffels RH, Streefland M. Multivariate PAT solutions for biopharmaceutical cultivation: current progress and limitations. *Trends Biotechnol.* 2014;32:329–336. doi:10.1016/j.tibtech.2014.03.008.
9. Jolliffe IT. Principal Component Analysis, 2nd ed. Springer Verlag, New York, Inc.; 2002.
10. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst.* 2001;58:109–130. doi:10.1016/S0169-7439(01)00155-1.
11. Luo Y, Chen G. Combined approach of NMR and chemometrics for screening peptones used in the cell culture medium for the production of a recombinant therapeutic protein. *Biotechnol Bioeng.* 2007;97:1654–1659. doi:10.1002/bit.21365.
12. Gunther JC, Conner JS, Seborg DE. Fault detection and diagnosis in an industrial fed-batch cell culture process. *Biotechnol Prog.* 2007;23:851–857. doi:10.1021/bp070063m.
13. Rouiller Y, Périlleux A, Collet N, Jordan M, Stettler M, Broly H. A high-throughput media design approach for high performance mammalian fed-batch cultures. *MAbs.* 2013;5:1-11. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23563583.
14. Ryan PW, Boyan L, Shanahan M, Ryder AG. Prediction of cell culture media performance using fluorescence spectroscopy. *Anal Chem.* 2010;82:1311–1317. Available at: http://aran.library.nuigalway.ie/xmlui/handle/10379/2321. Accessed February 25, 2014.
15. Gunther JC, Conner JS, Seborg DE. Process monitoring and quality variable prediction utilizing PLS in industrial fed-batch cell culture. *J Process Control.* 2009;19:914–921. doi:10.1016/j.jprocont.2008.11.007.
16. Le H, Kabbur S, Pollastrini L, Sun Z, Mills K, Johnson K, Karypis G, Hu W-S. Multivariate analysis of cell culture bioprocess data–lactate consumption as process indicator. *J Biotechnol.* 2012;162:210–223. doi:10.1016/j.jbiotec.2012.08.021.
17. Charaniya S, Le H, Rangwala H, Mills K, Johnson K, Karypis G, Hu W-S. Mining manufacturing data for discovery of high productivity process characteristics. *J Biotechnol.* 2010;147:186–197. doi:10.1016/j.jbiotec.2010.04.005.
18. Le H, Castro-Melchor M, Hakemeyer C, Jung C, Szperalski B, Karypis G, Hu W-S. Discerning key parameters influencing high productivity and quality through recognition of patterns in process data. *BMC Proc.* 2011;5:P91.doi:10.1186/1753-6561-5-S 8-P91.
19. Schmidberger T, Posch C, Sasse A, Gülch C, Huber R. Progress toward forecasting product quality and quantity of mammalian cell culture processes by performance-based modeling. *Biotechnol Prog.* 2015;31:1119–1127.
20. Rouiller Y, Solacroup T, Deparis V, Barbafieri M, Gleixner R, Broly H, Eon-duval A. Application of quality by design to the characterization of the cell culture process of an Fc-fusion protein. *Eur J Pharm Biopharm.* 2012;81:426–437. doi:10.1016/j.ejpb.2012.02.018.
21. Neunstoecklin B, Stettler M, Solacroup T, Broly H, Morbidelli M, Soos M. Determination of the maximum operating range of hydrodynamic stress in mammalian cell culture. *J Biotechnol.* 2015;194:100–109. doi:10.1016/j.jbiotec.2014.12.003.
22. Nomikos P, MacGregor JF. Monitoring batch processes using multiway principal component analysis. *AIChE J.* 1994;40:1361–1375. doi:10.1002/aic.690400809.
23. Kourti T. Abnormal situation detection, three-way data and projection methods: robust data archiving and modeling for

industrial applications. *Annu Rev Control*. 2003;27:131–139. doi:10.1016/j.arcontrol.2003.10.004.

24. Walczak B, Massart DL. Dealing with missing data: Part I. *Chemom Intell Lab Syst*. 2001;58:15–27. doi:10.1016/S0169-7439(01)00131-9.
25. García-Muñoz S, Kourti T, MacGregor JF, Mateos AG, Murphy G. Troubleshooting of an industrial batch process using multivariate methods. *Ind Eng Chem Res*. 2003;42:3592–3601. doi: 10.1021/ie0300023.
26. Kirdar AO, Green KD, Rathore AS. Application of multivariate data analysis for identification and successful resolution of a root cause for a bioprocessing application. *Biotechnol Prog*. 2008;24:720–726. doi:10.1021/bp0704384.
27. Kirdar AO, Conner JS, Baclaski J, Rathore AS. Application of multivariate analysis toward biotech processes: case study of a cell-culture unit operation. *Biotechnol Prog*. 2007;23:61–67. doi:10.1021/bp060377u.
28. Villadsen J, Nielsen J, Lidén G. Bioreaction Engineering Principles. 3rd ed. Springer Verlag, New York, Inc.; 2011. Available at: http://books.google.com/books?hl=en&lr=&id=Sq2VJ3D_QOwC&oi=fnd&pg=PR5&dq=Bioreaction+Enginnering+Principles&ots=w7Pi-rh9CK0&sig=-ugSWLV4tqGofIAyvg7m0kb1a-o. Accessed August 14, 2014.
29. Chong I-G, Jun C-H. Performance of some variable selection methods when multicollinearity is present. *Chemom Intell Lab Syst*. 2005;78:103–112. doi:10.1016/j.chemolab.2004.12.011.
30. Xing Z, Li Z, Chow V, Lee SS. Identifying inhibitory threshold values of repressing metabolites in CHO cell culture using multivariate analysis methods. *Biotechnol Prog*. 2008;24:675–683. doi:10.1021/bp070466m.
31. Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature selection. *J Chemom*. 1992;6:267–281. doi:10.1002/cem.1180060506.

Manuscript received May 5, 2015, and revision received July 24, 2015.