Multiblock PLS Analysis of an Industrial Pharmaceutical Process

J.A. Lopes,[1] J.C. Menezes,[1] J.A. Westerhuis,[2] A.K. Smilde[2]

[1]Center for Biological & Chemical Engineering, Technical University of Lisbon, Av. Rovisco Pais, P-1049-001, Lisbon, Portugal; telephone: (+351) 218 417 347; fax: (+351) 218 419 062; joao.lopes@ist.utl.pt
[2]Department of Chemical Engineering, University of Amsterdam, Nieuwe Achtergracht 166, 1018WV Amsterdam, The Netherlands

Received 6 December 2001; accepted 23 April 2002

DOI: 10.1002/bit.10382

Abstract: The performance of an industrial pharmaceutical process (production of an active pharmaceutical ingredient by fermentation, API) was modeled by multiblock partial least squares (MBPLS). The most important process stages are inoculum production and API production fermentation. Thirty batches (runs) were produced according to an experimental planning. Rather than merging all these data into a single block of independent variables (as in ordinary PLS), four data blocks were used separately (manipulated and quality variables for each process stage). With the multiblock approach it was possible to calculate weights and scores for each independent block. It was found that the inoculum quality variables were highly correlated with API production for nominal fermentations. For the nonnominal fermentations, the manipulations of the fermentation stage explained the amount of API obtained (especially the pH and biomass concentration). Based on the above process analysis it was possible to select a smaller set of variables with which a new model was built. The amount of variance predicted of the final API concentration (cross-validation) for this model was 82.4%. The advantage of the multiblock model over the standard PLS model is that the contributions of the two main process stages to the API volumetric productivity were determined. © 2002 Wiley Periodicals, Inc. *Biotechnol Bioeng* **80**: 419–427, 2002.

Keywords: pharmaceutical production; fermentation; PLS; multiblock PLS; multivariate modeling

INTRODUCTION

Most industrial pharmaceutical processes are multistage, batch operated for economic reasons. Many antibiotics are industrially produced by fermentation. Dynamic models for industrial fermentation processes are difficult to identify for a wide variety of reasons: microorganisms complex dynamics, variable and ill-defined raw materials, and dependence on previous process stages (strain selection and preculture production) (Menezes et al., 1994).

Correspondence to: J.A. Lopes
Contract grant sponsor: the Foundation for Science and Technology
Contract grant number: PRAXIS XXI BD/18471/98

The performance of a fermentation is often related to the amount of product obtained at the end of the process (Lopes and Menezes, 1998). Typically, industrial fermentations are sequences of batch inoculum production and fed-batch API fermentation operations. The amount and quality of the final active product ingredient (API) depends on each previous stage, not only on the final fermentation step. Hence, it is essential to consider every process step in order to account for the sources of variability.

Usually, process modeling is based only on fermentation data. Therefore, some hypothetically relevant inoculum properties are not included in the model (Atkinson and Mavituna, 1991). A regression or classification model can be obtained in order to detect inoculum qualities that are most related to high fermentation productivity (Atkinson and Mavituna, 1991; Siimes et al., 1992). When a poor-quality inoculum is detected it should not be used to inoculate the production tank. The following criteria should be respected to obtain a satisfactory inoculum (Stanbury and Whitaker, 1984):

- the inoculum must be in an active and healthy status to minimize the duration of the lag phase in the subsequent fermentation;
- it must be available in sufficiently large volumes to provide an inoculum of optimum size (3–10% of the medium volume);
- the inoculum must be in a suitable morphological form;
- it must be free of contamination; and
- the inoculated biomass must retain its product-forming capabilities.

The major problem in using vegetative inoculum is the difficulty in obtaining a uniform standard inoculum due to the morphological differentiation associated with the growth of this type of microorganism (Stanbury and Whitaker, 1984). Inoculum transfer criteria can also become a critical issue in industrial fermentation processes as scale is increased (Lopes and Menezes, 1998). Inoculation of production tanks is typically based on a schedule age postin-

oculation procedure. Low reproducibility production systems are usually obtained as a result of variable biomass concentration and metabolic activity of inoculum at the instant of inoculation. However, that experimental variability can be reduced and process performance improved by monitoring the activity of the preculture tank, ideally in every fermentation cycle (Buckland, 1984; Neves et al., 2001). In order to improve the performance of a fermentation system, it is necessary to consider variables or states related to the microorganism's metabolic state and not just the environment to which it is exposed in the inoculum or production tanks. In this sense, probably some of the most valuable derived measurements in fermentation technology are provided by off-gas analysis and air-flow measurements (Meyer et al., 1985). These types of measurements are available on-line and are unaffected by solid substrates or other physicochemical properties of the medium. They also provide insight into the metabolic state of the microorganism (Montague, 1997), namely, by giving an indication of shifts from growth to product formation and among different primary substrates. Ignova et al. (1999) investigated the application of Kohonen networks to map the inoculum process according to fermentation productivity. They identified inocula that were more likely to produce high product concentrations. However, little was discovered as to which inoculum properties were more important and were responsible for increased production.

In this article, a regression model based on multiblock partial least squares (MBPLS) was used to obtain the relative contributions of the inoculum production stage and the fermentation (production) stage to API production (fermentation performance). The multiblock approach used considers the inoculum and the fermentation data in the regression model separately (Fig. 1). From the model coefficients for each block, the influence of each stage in the overall production process can be established.

EXPERIMENTAL

Approximately 1.5 mL of a spore suspension (equivalent to a concentration of about 10^7 cells/mL) was added to 400 mL of inoculum medium in 1,500 mL high-lap baffled flasks. These were incubated at 25°C, 220 rpm, for 48 h. The vegetative flask culture was used to inoculate the vegetative medium of the inoculum tank. A volumetric inoculation rate of 0.4% (v/v) was used for an operating volume of 100 dm^3 (inoculum tank operating volume). Inoculum production is a batch process.

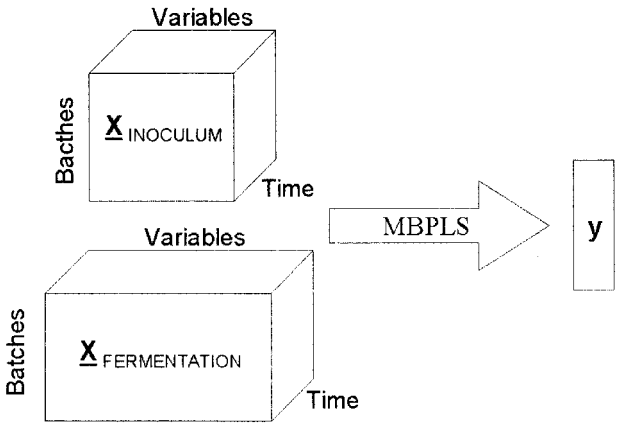

Figure 1. Multiblock PLS scheme used to model the final fermentation API concentration (vector **y**).

Fed-batch cultivation of a *Streptomycetes* strain was carried out using a nondefined (complex) medium containing soy meal and a carbon source. Soy meal is a typical protein-based solid substrate commonly used in industrial fermentation media which contains a nonnegligible amount of insoluble materials. The conditions used were typical of those employed routinely in industry for aerobic microbial growth.

A fully instrumented bioreactor with an operating volume of 200 dm^3 was employed throughout this study as the production tank. This fermenter was fitted with two Rushton turbines of six vertical blades.

Depending on the experiment, 24- and 48-h-old inocula were used to inoculate the tank. When a 24-h-old inoculum was used the inoculation rate is 14% (v/v) and 7% (v/v) for a 48-h-old inoculum. The culture was supplied with a carbon, inorganic nitrogen, and phosphate source. Nitrogen was supplied to the culture as ammonia in the API production phase (after about 24 h) to control and maintain the pH around 6.8. The addition stops before the end of the fermentation (because of downstream operation requirements). Phosphate is used to stimulate bacterial growth and as a precursor to API synthesis. The production fermentation process lasts ~140 h. After this the cultivation media is transferred for downstream processing (Neves et al., 2001).

Inoculum and Fermentation Monitoring

Control modules IL-430-instruments laboratory Systems Spa (UK) were used to measure the on-line temperature, pH, dissolved oxygen, pressure, and air flow (inoculum and production tanks). A gas analyzer connected to the outlet gas stream was used to monitor the on-line oxygen (paramagnetic sensor model PMA-25/M&C) and carbon dioxide (infrared sensor model SIFOR 200/Maihak) concentrations in the exhausted gases (inoculum and production tanks). The remaining monitored variables were only available for the production tanks. Biomass growth was quantified by the standard centrifugation method that gives the packed mycelial volume (PMV)—sediment made up of microorganism cells—or by monitoring daily the cultivation media viscosity (model LVT; Brookfield Engineering, Stoughton, MA, USA). The carbon, nitrogen, and phosphate sources were obtained with a frequency of about 4 h by flow injection analysis (FIA). The final API concentration was measured by HPLC (WellChrom K-2500 spectrophotometer).

Data

Table I summarizes monitored variables both on the inoculum and fermentation tanks. The variables were classified in

Table I. List of process variables

	Inoculum		Fermentation	
	Manipulated I(M)	Quality I(Q)	Manipulated F(M)	Quality F(Q)
1	pH	xO_2	pH	PMV
2	DO	cXO_2	DO	Viscosity
3	Temperature	OUR	Temperature	Carbon source conc.
4	Inlet gas flow	CER	Carbon source feed	Phosphate conc.
5	—	—	Nitrogen feed	Starch conc.
6	—	—	Phosphate feed	Nitrogen conc.
7	—	—	Inlet gas flow	xO_2
8	—	—	—	xCO_2
9	—	—	—	OUR
10	—	—	—	CER

two groups: manipulated and quality variables. Manipulated variables are used to control the process (inputs to the process). Quality variables are descriptive of the process state (process outputs). Quality variables can be viewed as a function of the manipulated variables. The xO_2, xCO_2, temperature, pH, dissolved oxygen (DO), nutrient feeds, oxygen uptake rate (OUR), and carbon dioxide evolution rate (CER) were obtained on-line. The other variables were measured off-line with a frequency of about 4 h except for viscosity, where samples were taken daily.

A total of 16 batches using 24-h-old inocula and 14 batches using 48-h-old inocula were designed with experimental planning in order to improve process modeling. From these, 17 batches were operated under nominal conditions (nine with 24-h-old inocula and eight with 48-h-old inocula). All experiments are production batches. Disturbances in several variables were imposed during the fermentation stage in the remaining 13 batches: feeds, aeration, temperature, pH, and initial amount of soybean flour (the most important nitrogen supply to the culture).

Respiratory Parameters

CER measures CO_2 production (Eq. [1]), which is produced essentially by cell growth, maintenance, and API production. It is a very important variable because it gives an indication about shifts in the microorganism metabolism during the process (Atkinson and Mavituna, 1991):

$$CER = \frac{F_{air}P_F}{V_F R_G T_F}\left(\frac{0.79 \cdot xCO_2}{1 - xO_2 - xCO_2}\right) \quad (1)$$

where F_{air} is the inlet gas flow rate, P_F the relative pressure, R_G the perfect gas constant, V_F the operating cultivation media volume, T_F the temperature, and xO_2 and xCO_2 are the O_2 and CO_2 molar fractions in the off-gas stream (it is assumed that the inlet air contains 21% O_2, 79% inerts (N_2), and 0% CO_2). The end of the maximum CER plateau indicates the transition from the initial biomass production phase to the API production phase (see units in Nomenclature).

The average CER time profiles for production fermentations using 24- and 48-h-old inocula are depicted in Figure 2 (note that only nominal fermentations were used). The shaded areas are confidence limits for the mean values. Clearly, the microorganism is in a more active state when 24-h-old inocula are used. The exponential growth phase starts earlier (in about 10–20 h) if 24-h-old inocula are used. This means that the production phase is started earlier, which results in increased API production during the first half of the fermentation. The maximum API concentration is reached about 24–48 h earlier with young inocula. In nominal conditions, batches using 24-h-old inocula produced more API than batches using 48-h-old inocula after 140 h. Production tank reproducibility observed for 24-h-old inoculum appears to be substantially better (the standard deviation for nominal batches using 24-h-old inocula is approximately half the SD obtained for nominal batches using 48-h-old inocula).

It was estimated that the use of younger and more active inocula (in terms of cell multiplication) is a more efficient strategy for maximizing productivity. An increase of about

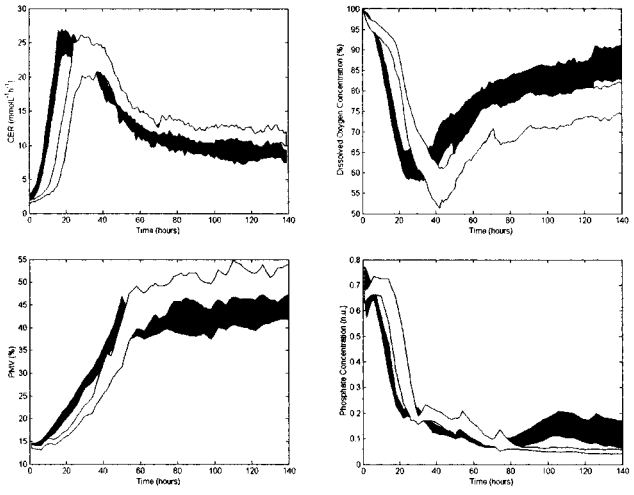

Figure 2. Mean time profiles of carbon exhaustion rate, dissolved oxygen concentration, packed mycelial volume, and phosphate concentration for nominal fermentations using 24- (■) and 48-h-old (□) inocula. The areas are the 95% confidence limits for the means ($\pm t_{1-\alpha/2,v-1}s/n^{0.5}$).

22% on the global rentability of the process was observed at the pilot-plant scale. This was associated with a new operating scheduling scheme based on 96-h (4-day) fed-batch cycles, using late exponential growth phase inocula (Neves et al., 2001).

Data Preprocessing

Data preprocessing included outliers detection and noise reduction. To synchronize laboratory (off-line) and on-line measurements, the time series of all variables were linearly interpolated. This was found appropriate because of the processes slow dynamics. Sampling for the inoculum data was 40 min and fermentation data 60 min. The 30 experiments were separated in four three-way blocks with dimensions batches × variables × time. Two blocks containing inoculum data (manipulated (30 × 4 × 36) and quality (30 × 4 × 30) variables) and two blocks containing fermentation data (manipulated (30 × 7 × 140) and quality (30 × 10 × 140) variables). The first 16 batches in the fermentation blocks were made with 24-h-old inocula and the remaining with 48-h-old inocula. Each data block was mean-centered (removing the variables mean profile over the batches). Iterative double slab scaling across the variables and time modes was necessary to remove magnitude differences (Harshman and Lundy, 1984). The quality block contains the final API concentration of each batch (30 × 1). This vector was mean-centered and divided by variance.

THEORY

Partial least-squares (PLS) is a class of regression models based on the calculation of latent variables or factors (Denham, 1995; Geladi and Kowalsky, 1986; Martens and Naes, 1989; Wold, 1966). In PLS these variables are calculated in order to maximize the covariance between the scores of an independent block (**X**) and the scores of a dependent block (**Y**).

An extension of the PLS algorithm was proposed by Wangen and Kowalsky (1989) to account for the existence of multiple independent or dependent data blocks. The multiblock PLS finds application in processes where data is collected on different process stages or in different process locations (MacGregor et al., 1994; Westerhuis and Coenegracht, 1997). Even if the multiblock PLS algorithm is not superior in terms of pure prediction, it can produce more interpretable results. The scores and weights for each block can provide valuable information about the process (e.g., for fault detection and process diagnosis).

Notation

Vectors and arrays are represented as bold characters. A three-way array is represented as an underlined uppercase character (e.g., $\underline{\mathbf{X}}$), a two-way array is represented with an uppercase character (e.g., **X**), and a vector is represented by a lowercase letter (e.g., **x**). The column i of an array **X** is represented as $\mathbf{x}_i$. $\mathbf{X}^t$ denotes the transpose of **X**. **X** is used for independent variables and **Y** for dependent variables. In this article, the prediction of a univariate vector (**y**) from blocks of three-way arrays ($\underline{\mathbf{X}}^{(b)}$) will be considered. The vector **y** (I × 1) and the three-way arrays $\underline{\mathbf{X}}^{(b)}$ (I × J_b × K_b) share the batch dimension I.

Unfolded PLS

The standard PLS algorithm can only be applied to two-way data. For this reason, the original three-way data must be reshaped into a two-way array. This operation is called unfolding or matricization ($\underline{\mathbf{X}}(I \times J \times K) \rightarrow \mathbf{X}(I \times JK)$) (Kiers, 2000). If more than one data block exists, each unfolded block information is paired sidewise. Thus, the unfolded PLS model is a simple PLS model with the difference that the data in each block is unfolded and paired horizontally. In practice the time mode and variables mode become mixed (columns), while the observations (rows) are preserved. Alternatively, multilinear PLS could be used to derive the model preserving the trilinear data structure of each block (Bro, 1996). However, it cannot be used for multiblock analysis. Multiway covariates regression can also be used to preserve the trilinear structure (Smilde and Kiers, 1999).

Multiblock PLS (MBPLS)

In a multiblock model, different scores for each data block are calculated. These scores are called block-scores (**s**). With the block-scores it is possible to compute global scores, which are called super-scores (**t**). There are essentially two ways to obtain the multiblock PLS model: the first is to use the block-scores to deflate **X** and **y** (Wangen and Kowalski, 1989). This ensures that the block-scores are orthogonal. The second is to use the super-scores (Westerhuis and Coenegracht, 1997). It can be proven that the former leads to inferior predictive capacity. If the super-scores are used for the deflation of the **X** and **y** blocks, then the predictive capacity is equal to an unfolded PLS model (the same regression vector is obtained). This is the algorithm described and used throughout this work. More recently, Westerhuis and Smilde (2001) proposed that only the **y** block deflation should be used, because when both **X** and **y** blocks are deflated the block-scores became difficult to interpret. The reason for this is that when deflating each $\mathbf{X}^{(b)}$ block, information from the remaining blocks is also used.

Consider each unfolded block b ($\mathbf{X}^{(b)}$) from a total of B blocks. Consider also $\mathbf{X} = [\mathbf{X}^{(1)}|\mathbf{X}^{(2)}|\ldots|\mathbf{X}^{(B)}]$ and **y** for calibration. Make $\mathbf{E}_0 = \mathbf{X}$ and $\mathbf{f}_0 = \mathbf{y}$.

for r = to R

$\mathbf{v}_r = \mathbf{E}_{r-1}{}^t\mathbf{f}_{r-1}$ ($\mathbf{v}_r$ is useful but not needed by the algorithm)

for b = 1 to B

$$\mathbf{u}_r^{(b)} = \frac{(\mathbf{E}_{r-1}^{(b)})^t \mathbf{f}_{r-1}}{\|(\mathbf{E}_{(b)}^{r-1})^t \mathbf{f}_{f-1}\|}$$

$$\mathbf{s}_r^{(b)} = \mathbf{E}_{r-1}^{(b)} \mathbf{u}_r^{(b)}$$

end

$$\mathbf{S}_r = [\mathbf{s}_r^{(1)}|\ldots|\mathbf{s}_r^{(B)}]$$

$$\mathbf{w}_r = \frac{\mathbf{S}_r^t \mathbf{f}_{r-1}}{\|\mathbf{S}_r^t \mathbf{f}_{r-1}\|}$$

$$\mathbf{t}_r = \mathbf{S}_r \mathbf{w}_r \quad (2)$$

$$\mathbf{p}_r = (\mathbf{t}_r^t \mathbf{t}_r)^{-1} \mathbf{E}_{r-1}^t \mathbf{t}_r$$

$$q_r = (\mathbf{t}_r^t \mathbf{t}_r)^{-1} \mathbf{f}_{r-1}^t \mathbf{t}_r$$

$$\mathbf{E}_r = \mathbf{E}_{r-1} - \mathbf{t}_r \mathbf{p}_r^1$$

$$\mathbf{f}_r = \mathbf{f}_{r-1} - \mathbf{t}_r q_r$$

get each block $\mathbf{E}_r^{(b)}$ from $\mathbf{E}_r$

end

Store the model coefficients for predictions; $\mathbf{U}^{(b)} = [\mathbf{u}_1^{(1)}|\ldots|\mathbf{u}_R^{(b)}]$, $\mathbf{P} = [\mathbf{p}_1|\ldots|\mathbf{p}_R]$, $\mathbf{q} = [q_1|\ldots|q_R]$ and $\mathbf{W} = [\mathbf{w}_1|\ldots|\mathbf{w}_R]$. Predictions can be obtained with algorithm 3. Consider $\mathbf{Z} = [\mathbf{Z}^{(1)}|\mathbf{Z}^{(2)}|\ldots|\mathbf{Z}^{(B)}]$ an array with new data scaled as $\mathbf{X}$ and $\hat{\mathbf{y}} = 0$. Make $\mathbf{E}_0 = \mathbf{Z}$.

for r = 1 to R

b = 1 to B

$$\mathbf{S}_r^{(b)} = \mathbf{E}_{r-1}^{(b)} \mathbf{u}_r^{(B)}$$

end

$$\mathbf{S}_r = [\mathbf{s}_r^{(1)}|\ldots|\mathbf{s}_r^{(B)}] \quad (3)$$

$$\mathbf{t}_t = \mathbf{S}_r \mathbf{w}_r$$

$$\mathbf{E}_r = \mathbf{E}_{r-1} - \mathbf{t}_r \mathbf{p}_r^t$$

$$\hat{\mathbf{y}} = \hat{\mathbf{y}} + \mathbf{t}_r q_r$$

get each block $\mathbf{E}_r^{(b)}$ from $\mathbf{E}_r$

end

Another way is to directly compute a regression vector for the MBPLS model (it is equivalent to the unfolded PLS regression vector) using $\mathbf{V} = [\mathbf{v}_1|\ldots|\mathbf{v}_R]$. The regression vector is given by $\mathbf{b}_{PLS} = \mathbf{V}(\mathbf{P}^t\mathbf{V})^{-1}\mathbf{q}^t$ (Denham, 1995). A more interesting expression is given by Eq. 4, where only the weights and covariances are used (Helland, 1988):

$$\mathbf{b}_{PLS} = \mathbf{V}(\mathbf{V}^t\mathbf{S}_{XX}\mathbf{V})^{-1}\mathbf{V}^1\mathbf{S}_{XY} \quad (4)$$

Covariances in Eq. 4 are given by $\mathbf{S}_{xx} = \mathbf{X}^t\mathbf{X}$ and $\mathbf{S}_{xy} = \mathbf{X}^t\mathbf{y}$. Predictions are obtained merging the data blocks into a single unfolded array ($\mathbf{Z}$), applying the calibration scaling factors to $\mathbf{Z}$ and finally the regression vector:

$$\hat{\mathbf{y}} = \mathbf{Z}\mathbf{b}_{PLS} \quad (5)$$

Equation 6 was used throughout this work to estimate the fitting between original and model predicted final API concentration (note that the vector $\mathbf{Y}$ in the equation is considered to be mean-centered):

$$Q_Y^2 = 1 - \frac{\text{trace}((\mathbf{Y} - \hat{\mathbf{Y}})^t(\mathbf{Y} - \hat{\mathbf{Y}})}{\text{trace}(\mathbf{Y}^t\mathbf{Y})} \quad (6)$$

Model Calibration

In order to build a PLS model with the available batches there are several issues which need to be considered: the variables to include in the model and the number of components. Variable selection will not be considered here because the importance of each block variable has to be determined. The appropriate number of components can be obtained with cross-validation (see MBPLS Model, below).

RESULTS AND DISCUSSION

The differences between the use of 24- or 48-h-old inocula are visible in several fermentation quality variables (similar shape but a visible time shift). Figure 2 shows the average plots of the carbon evolution rate, dissolved oxygen concentration, packed mycelial volume, and phosphate concentration for 24- and 48-h-old inoculum fermentations. The bands correspond to the 95% confidence limits for the estimated averages values. The difference observed in the beginning of the production fermentation (the first 25–30 fermentation hours) is clearly superior. This means that the differences observed between the use of 24- or 48-h-old inocula happens right at the beginning of the fermentation. Hence, important differences are present in the 30 batches, especially due to the inoculum used.

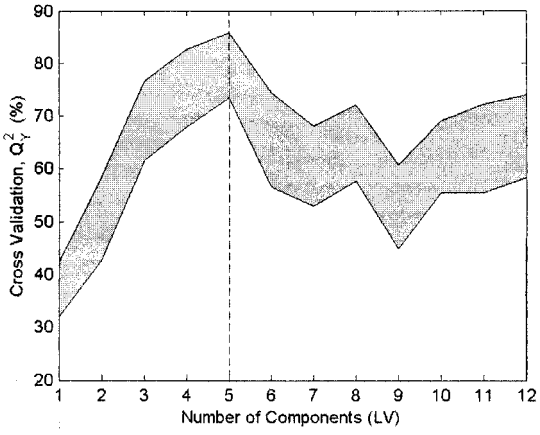

Figure 3. "Leave-one-block-out" cross-validation to determine the best number of components for the process PLS model. Shaded areas are 95% confidence limits calculated by bootstrapping.

Table II. Variance captured for five-component MBPLS model for the API final concentration (%)

LV	I(M)	I(Q)	F(M)	F(Q)	Total **X**	**y**
1	4.1	5.3	11.2	17.3	13.5	78.4
2	9.6	11.7	23.2	28.2	24.1	80.2
3	18.5	25.0	25.0	43.8	37.9	91.5
4	19.4	26.9	26.9	49.5	43.2	95.7
5	23.8	33.9	34.0	54.2	47.9	98.1

MBPLS Model

The MBPLS model was built using four blocks of input variables and a vector of outputs (API final concentration). To investigate the best number of components to use in the MBPLS model, a cross-validation strategy based on a "leave-one-out" procedure was proposed. To increase the robustness of the process, 500 models were calibrated for each number of components using a resampling strategy (bootstrapping). For each model the value of Q_Y^2 was determined. Figure 3 shows the results in terms of the observed average values (note the confidence limits for the means). The five-component model appears to be the best.

In an MBPLS model, the weights are very useful to interpret block importance. Considering the inoculum quality and manipulated variables (I(M) and I(Q)) and the fermentation manipulated and quality variables (F(M) and F(Q)) a five-component multiblock PLS model was built using the 30 batches (five components yields the optimum model according to the cross-validation strategy adopted). Weights, super-weights, and block-weights were stored. Table II summarizes the percentage captured in each block for each of the five model components. The variance associated with each independent data block (I(M), I(Q), F(M), and F(Q)) and with the dependent data (**y**) is presented. The global variance captured by all independent data blocks is also depicted (Total **X**). Note that even with five components there is 47.9% of the **X** blocks explained (this low amount is common for batch data). The model is able to explain the **y** variable as expected. Production fermentation variables are better explained by the model than inoculum variables. Quality variables are also better explained than manipulated variables. A value of $Q_Y^2 = 80.2\%$ was obtained ("leave-one-out" cross-validation) with five components. The actual vs. predicted API concentration for the model with the selected batches is presented in Figure 4.

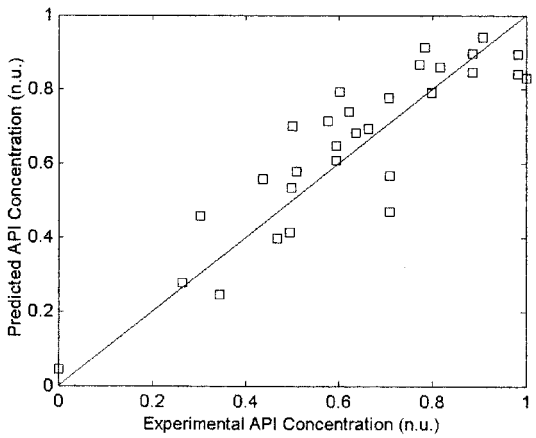

Figure 4. Experimental against predicted final API concentration in normalized units (scaled between 0 and 1). Predictions are obtained by cross-validation ("leave-one-out") with a five-component PLS model ($Q_Y^2 = 80.2\%$).

The super-weights for this model are shown in Table III. Super-weights for the production fermentation variables (F(M) and F(Q)) are in general greater than the super-weights for the inoculum variables (I(M) and I(Q)). The fermentation data is more appropriately modeled by the MBPLS model than the inoculum data, especially in the case of the production fermentation quality variables (F(Q)). This is not surprising, since differences in the inoculum development will be reflected in the quality of the fermentation. The importance of the inoculum, however, is not negligible. The differences observed in the block-scores for the quality variables of the fermentation are clearly visible in Figure 5. The main reason for the observed separation in the experiments is the inoculation time.

There was also some correlation between the manipulated and quality variables block-scores for each process stage. The plot in Figure 6 shows the correlation obtained for the first component for the inoculum data. The correlation coefficient is $Q_Y^2 = 0.73$. The chart shows that manipulation of the inoculum growth process is still nonoptimized, since there is variability among inoculum scores that is also expressed in the quality variables scores. There is mutual correlation in the blocks to describe the final API concentration.

A detailed analysis of this model can be done by analyzing the weights of every variable (vector **v**). Averaging the square of these weights, the relative importance of each individual variable can be inspected for each component. Figure 7 shows that there are some variables that are more

Table III. Super-weights for the five-component MBPLS model (**W**)

	Latent variable (LV)				
Blocks	1	2	3	4	5
I(M)	0.15	0.21	0.22	0.19	0.22
I(Q)	0.13	0.22	0.28	0.13	0.18
F(M)	0.61	0.70	0.52	0.69	0.61
F(Q)	0.77	0.64	0.77	0.61	0.74

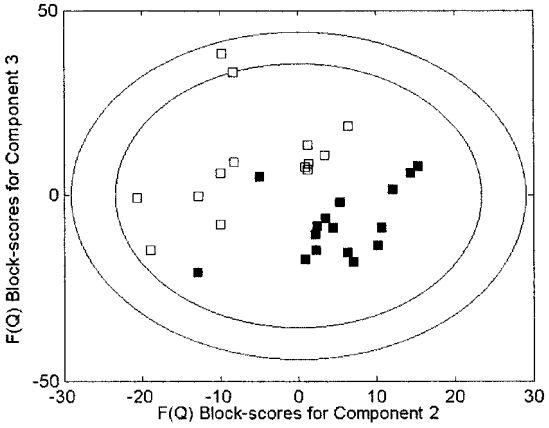

Figure 5. Fermentation quality (F(Q)) block-scores for components 2 and 3 for the MBPLS model obtained with the selected batches. The points correspond fermentations using 24- (■) and 48-h-old (□) inocula.

relevant for the model than others. In the first principal component, pH and inorganic nitrogen concentration in the fermentation are the most important variables. This is consistent with the fact that when the pH is above or below the optimal range at the API production phase (~6.8 ± 0.2), problems in process volumetric productivity in API productivity occur. Components 2 and 3 show that the inoculum pH and respiratory activity have consequences on the microorganism activity. Component 4 reflects the importance of the microorganism on the API concentration. The last component could be explained similarly to components 2 and 3.

The more important variables found for the five components were selected and grouped in two blocks: one block for the inoculum data and one block for the fermentation data (Table IV). A total of 15 variables was selected and a new multiblock model was obtained.

Cross-validation determined a four-components model. Table V contains the super-weights to the models up to four components. It can be compared with Table III. Again, the

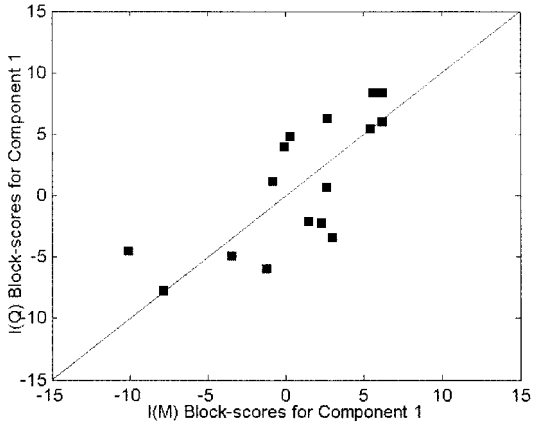

Figure 6. Inoculum manipulated variables (I(M)) block-scores vs. inoculum quality variables (I(Q)) block-scores for component 1 for a five-component MBPLS model ($Q_Y^2 = 0.73$).

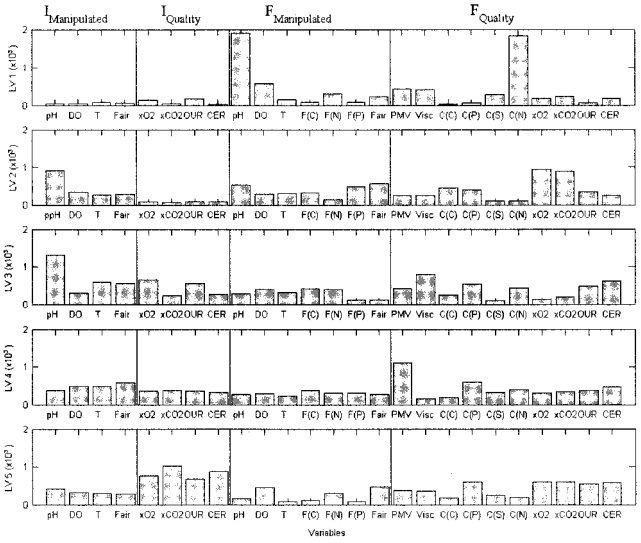

Figure 7. Average squared weights for each block variable and each model component. For this five-component MBPLS model the cross-validation ("leave-one-out") amount of variation predicted (Q_Y^2) is 80.2%.

relative importance of fermentation variables is higher. The "leave-one-out" type amount of variance predicted for this model is 82.4%, which is very similar to the value obtained with all variables and a five-component model.

CONCLUSIONS

A multiblock PLS method was applied to model an industrial pharmaceutical production process. Data from two process stages was available: inoculum growth and production fermentation. From the data it was found that the inoculum growth duration is very important in fermentation productivity. Even if the inoculum growth duration parameter does not seem to affect the amount of API obtained, it is important because it reduces the production fermentation stage variability in a significant way. The inoculum production and the fermentation data were used to model the production fermentation final API concentration. In 30 pilot plant experiments an amount of variance predicted of 80.2% was

Table IV. Selected variables

	Inoculum	Fermentation
1	pH	pH
2	Temperature	DO
3	Inlet gas flow	Nitrogen feed
4	OUR	Inlet gas flow
5	CER	PMV
6	—	Viscosity
7	—	Phosphate conc.
8	—	Nitrogen conc.
9	—	OUR
10	—	CER

Table V. Super-weights for the four-component MBPLS model built with the selected variables (**W**)

	Latent variables (LV)			
Block	1	2	3	4
Inoculum	0.17	0.27	0.45	0.56
Fermentation	0.98	0.96	0.96	0.85

obtained for a five-component model. From the model block-weights and block-scores, it was possible to draw some conclusions about the influence of the variables measured in the two stages on overall API production. Two influences on fermentation quality were found through the analysis of the model block-scores: the first was the influence of the inoculum growth duration in the model block-scores, which was more evident for the nominal fermentations, and the second was changes in operating conditions during fermentation (feeds, temperature, pH control, and inlet gas flow rate control). This was possible because the dataset contained designed experiments. Under nominal conditions the amount of API produced is strongly dependent on the inoculum development duration. From the analysis of the model weights it was possible to select a smaller set of variables that are more important and a new model was obtained with these variables. An amount of variance predicted for the concentration of API at 140 h of 82.4% was obtained for a four-component model with the new set of variables. A model of this type could be used in the early detection of a poor-quality inoculum (or the selection of the best inoculum among several possible candidates), leading to improved fermentation performance and reduced process variability.

The authors thank Companhia Industrial Produtora de Antibióticos SA (CIPAN) in Portugal for providing the data.

NOMENCLATURE

R	number of components
I	number of experiments (batches)
J	number of variables
K	number of time points
B	number of blocks
LV	latent variables
$\mathbf{X}^{(b)}$	Data for block b (three-way array)
$\overline{\mathbf{X}}^{(b)}$	unfolded data for block b
$\mathbf{X}$	unfolded data for all blocks
$\mathbf{Z}^{(b)}$	unfolded unseen data for block b
$\mathbf{Z}$	unfolded unseen data for all blocks
$\mathbf{E}$	deflated $\mathbf{X}$
$\mathbf{y}$	vector of final API concentrations
$\hat{\mathbf{y}}$	predicted vector of final API concentrations
$\mathbf{f}$	deflated $\mathbf{y}$
$\mathbf{S}_{xx}$	covariance matrix ($\mathbf{X}^t\mathbf{X}$)
$\mathbf{S}_{xy}$	covariance matrix ($\mathbf{X}^t\mathbf{y}$)
D_k	discrimination importance for time point k
$\mathbf{u}_r^{(b)}$	block-weight (block b, component r)
$\mathbf{s}_r^{(b)}$	block-score (block b, component r)
$\mathbf{v}_r$	weight for component r
$\mathbf{S}_r$	array of block-scores (component r)
$\mathbf{W}$	super-weights
$\mathbf{T}$	super-scores
$\mathbf{P}$	independent data loadings
q	dependent data loadings
$\mathbf{V}$	weights
$\mathbf{b}_{PLS}$	PLS regression vector
P_F	fermenter absolute pressure (atm)
V_F	fermenter volume (dm^3)
T_F	fermenter temperature (K)
R_G	perfect gas constant (0.082 $atm.dm^3.mol^{-1}K^{-1}$)
PMV	packed mycelial volume (% v/v)
DO	dissolved oxygen (%)
F(C)	carbon-source solution feed rate ($dm^3.h^{-1}$)
F(N)	ammonia (nitrogen) feed rate ($dm^3.h^{-1}$)
F(P)	phosphate solution feed rate ($dm^3.h^{-1}$)
C(C)	carbon-source residual concentration ($g.dm^{-3}$)
C(N)	nitrogen residual concentration ($g.dm^{-3}$)
C(P)	phosphate residual concentration ($g.dm^{-3}$)
C(S)	starch residual concentration (% w/v)
F_{air}	inlet air flow ($dm^3.h^{-1}$)
xO_2	molar fraction of O_2 on the exhaust gas
xCO_2	molar fraction of CO_2 on the exhaust gas
CER	carbon dioxide evolution rate ($mol.dm^{-3}.h^{-1}$)
OUR	oxygen uptake rate ($mol.dm^{-3}.h^{-1}$)

References

Atkinson B, Mavituna F. 1991. Biochemical engineering and biotechnology handbook. New York: Stockton Press.

Bro R. 1996. Multiway calibration: multilinear PLS. J Chemomet 10: 47–61.

Buckland B. 1984. The translation of scale in fermentation processes: the impact of computer process control. Bio-Technol 2:875–883.

Denham M. 1995. Implementing partial least squares. Stat Comput 5: 191–202.

Harshman R, Lundy M. 1984. Data preprocessing and the extended PARAFAC model. In: Law HG, Snyder CW Jr, Hattie J, McDonald RP, editors. Research methods for multimode data analysis. New York: Praeger. p 216–284.

Helland I. 1988. On the structure of partial least squares regression. Communications in statistics-elements of simulation and computation. 17: 581–607.

Geladi P, Kowalsky B. 1986. Partial least squares regression: a tutorial. Anal Chim Acta 185:1–17.

Ignova M, Montague G, Ward A, Glassey J. 1999. Fermentation seed quality analysis with self-organising neural networks. Biotechnol Bioeng 64:82–91.

Kiers H. 2000. Towards a standardized notation and terminology in multiway analysis. J Chemomet 14:105–122.

Lopes J, Menezes J. 1998. Faster development of fermentation processes. early stages process diagnostics. AIChE Symp Series 94:391–396.

MacGregor J, Jaeckle C, Kiparissides C, Koutoudi M. 1994. Process monitoring and diagnosis by multiblock PLS methods. Proc Syst Eng 40826–838.

Martens H, Naes T. 1989. Multivariate calibration. Chicester, UK: John Wiley & Sons.

Menezes C, Alves S, Lemos J, Azevedo S. 1994. Mathematical modelling of industrial pilot-plant penicillin-G fed-batch fermentations. J Chem Technol Biotechnol 64:123–138.

Meyer H, Kappeli O, Fiechter A. 1985. Growth control in microbial cultures. Annu Rev Microbiol 39:299–319.

Montague G. 1997. Monitoring and control of fermenters. IchemE, UK.

Neves A, Vieira L, Menezes J. 2001. Effects of preculture variability on clavulanic acid fermentation. Biotechnol Bioeng 72:628–633.
Siimes T, Nakajima M, Yada H, Asama H, Nagamune T, Linko P, Isao E. 1992. Knowledge based diagnosis of inoculum properties and sterilization time in lactic acid fermentation, Biotechnol Tech 6:385–390.
Smilde A, Kiers H. 1999. Multiway covariates regression models. J Chemomet 13:31.48.
Stanbury P, Whitaker A. 1984. The development of inocula for industrial fermentations. In: Principles of fermentation technology. Oxford: Pergamon. p 108–119.
Wangen L, Kowalski B. 1989. A multiblock PLS algorithm for investigating complex chemical systems. J Chemomet 3:3–20.
Westerhuis J, Coenegracht P. 1997. Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. J Chemomet 11:379–392.
Westerhuis J, Smilde A. 2001. Deflation in multiblock PLS. J Chemomet 15:485–493.
Wold H. 1966. Nonlinear estimation by iterative least squares procedures. In: David F, editor. Research papers in statistics. New York: John Wiley & Sons. p 411–444.