0273-2289/02/102–103/0049/$13.50

Fermentation Diagnosis by Multivariate Statistical Analysis

SILVIO BICCIATO,*,[1] ANDREA BAGNO,[1] MARCO SOLDÀ,[1]
RICCARDO MANFREDINI,[2] AND CARLO DI BELLO[1]

*[1]University of Padova, via Marzolo, 9, 35131 Padova, Italy,
E-mail: silvio.bicciato@unipd.it; and [2]Biofin Laboratories Srl,
via F. Petrarca, 16, 46047 Porto Mantovano, Italy*

Abstract

During the course of fermentation, online measuring procedures able to estimate the performance of the current operation are highly desired. Unfortunately, the poor mechanistic understanding of most biologic systems hampers attempts at direct online evaluation of the bioprocess, which is further complicated by the lack of appropriate online sensors and the long lag time associated with offline assays. Quite often available data lack sufficient detail to be directly used, and after a cursory evaluation are stored away. However, these historic databases of process measurements may still retain some useful information. A multivariate statistical procedure has been applied for analyzing the measurement profiles acquired during the monitoring of several fed-batch fermentations for the production of erythromycin. Multivariate principal component analysis has been used to extract information from the multivariate historic database by projecting the process variables onto a low-dimensional space defined by the principal components. Thus, each fermentation is identified by a temporal profile in the principal component plane. The projections represent monitoring charts, consistent with the concept of statistical process control, which are useful for tracking the progress of each fermentation batch and identifying anomalous behaviors (process diagnosis and fault detection).

Index Entries: Fermentation processes; process identification; process diagnosis; multiway principal component analysis; statistical process control; database mining.

Introduction

A major objective in bioprocess development is to create an optimal strategy for maximum productivity and production yield of the desired

*Author to whom all correspondence and reprint requests should be addressed.

product. From a manufacturing standpoint, this is often achieved via control strategies, reactor design, media formulation, and genetic manipulation among others. However, in analyzing biochemical and biotechnological systems, detailed mechanistic data are often limited. It is well known that if measurements of in vivo metabolite concentrations and enzyme activities are readily available, accurate models of system behavior can be developed. To this end, advances in analytical technologies such as microarrays, two-dimensional (2D) electrophoretic gels, and mass spectrometry appear to provide the kind of detailed information that biotechnologists seek. Unfortunately, it may take some time before such devices become mainstream, especially in the industrial sector, and this implies that researchers must make use of the measurements that are available.

In this work the focus is placed on defining monitoring and control strategies based on process measurements. Historically, optimal fermentation control schemes have yielded mixed results owing to the dynamic nature and heterogeneity of bioprocesses. Variability in performances and suboptimal production objectives in monitored and controlled processes may be caused by factors such as changes in the physiologic state of the cells, variations in the initial medium components among the lots, and differences in the inoculum activity of the cells. In the general case, the fault does not lie in the control strategies themselves but has to be ascribed to the difficulty in assessing changes in the state of the system. To minimize such faults, it is fundamental to identify, online and as early as possible, changes in the performance or in the overall state of the fermentation so that a new strategy can be implemented. Unfortunately, previous attempts at direct online evaluation of bioprocess behavior have been limited. Indeed, poor mechanistic understanding of biochemical conversions has hindered the classic approach of creating mathematical models of fermentation processes. This has been further complicated by the lack of appropriate online measurements reflecting the physiologic state of the culture and by the long lag time associated with offline assays. Since the issue of process performance identification cannot be completely addressed mechanistically at the present time, the proposed approach is entirely data driven and requires no *a priori* hypothesis. The basic principle is that the underlying mechanisms and the process dynamics are embedded in the temporal measurement profiles that define characteristic "fingerprints" of different types of process behavior. Processes that have similar outcomes should have a similar fingerprint. The key idea behind this approach is to effectively use the data that are routinely collected during the course of a fermentation to address issues of control, product quality, yield, and productivity *(1,2)*. Thus, the goal is to extract knowledge in the form of patterns from historic data records *(3)*. It is these patterns that will be used to derive data-driven models for the online identification of process state and performance *(4,5)*.

In general, knowledge extraction from historic records requires different data analysis tasks, and a data-mining approach typically employs a wide variety of techniques from fields such as statistics, artificial intelligence, machine learning, and pattern recognition. With the emphasis on knowledge discovery, data mining is concerned with creating empiric models that relate the correlational structure of the data to various physiologic states and process behaviors. One method that can be exploited in pattern extraction is artificial intelligence in the form of neural networks. However, neural networks are normally considered black boxes, and this aspect limits their applications to all those problems in which model interpretation is as important as quality of prediction. These methodologies are extremely powerful and effective at modeling behavior, but therein lies their greatest weakness. Indeed, they are capable of developing input/ output relationships whether they are real or not, and little hypothesis can be formulated on how to reverse their knowledge-encoding process. Since the performance of fermentation runs is the result of complex interactions between the different process-controlling variables, the application of analytical tools that detect the multivariate nature of complex relationships seems a more suitable solution. Using multivariate principal component analysis (MPCA) *(6–8)*, we describe herein how various physiologic and process state changes can be identified early in the process so that the appropriate control paradigm can be activated. Principal component analysis (PCA) is a standard and straightforward data-mining procedure that allows the representation of highly complex sets of variable profiles by a small number of characteristic modes that capture the patterns of process changes *(9)*. This type of analysis yields a dimensional reduction of the data space while preserving information on the variables' interactions. Thus, the identification of the process behavior can be conducted on a lower dimensional space such as a 2D plane where each fermentation can be identified by a single profile.

MPCA has been applied to perform a postanalysis on a fermentation run database to identify similarities among eight different batches from an industrial production of erythromycin. Such an analysis has been used to characterize possible sources of batch-to-batch variation and to define optimal operating procedures. The multivariate procedure has also been used to monitor the time progression of the different batches in the reduced plane where statistical control limits and parameters helped identifying anomalous process behaviors and their possible causes.

Materials and Methods

Strain and Growth Media

The culture used throughout this study is a mutant strain of *Streptomyces erythraeus* (SR32/M) supplied by Biofin, Porto Mantovano, Mantova, Italy. The strain was preserved as freeze-dried cultures in a refrigerator at 4°C. The strain was cultivated on shake-flasks containing the following nutrients dissolved in distilled water: sucrose, corn steep liquor (CSL), ammonium sulfate, calcium carbonate, and soybean oil. Each flask, with 400 mL of sterile medium, was inoculated with 2 mL of spore suspension in physiologic solution at 4°C. The inoculated flasks were incubated on alternative shakers at 120 rpm and 34°C for 44–48 h. At the end of the incubation process, the inoculum had a pH of 6.1 and a packed mycelium volume of 5 to 6%.

Batch Fermentations

A series of five 20-L PPS15 fermentors (BioIndustrie Mantova, Gazoldo degli Ippoliti, Mantova, Italy) with a working volume of 10 L were used for the prevegetative, vegetative, and fermentative stages. The fermentors were fitted with 1 Lightning and 1 Rushton turbines. The prevegetative medium contained sucrose, CSL, calcium carbonate, ammonium sulfate, and soybean oil in tap water. The pH of the medium was adjusted to 5.5 with 20% NaOH solution, and after sterilization at 120°C for 60 min, the fermentor was inoculated with 80 mL of grown seed culture. In the prevegetative stage, the temperature was controlled at 33 to 34°C, the pH was at 5.5, the airflow rate was set at 1 vvm with an agitation speed of 400 rpm, the back pressure was at 0.5 bar, and the oxygen partial pressure was >30%. The prevegetative phase was completed when the packed mycelium volume reached a value close to 12%. The vegetative stage was performed in parallel in two fermentors.

The medium of the vegetative fermentation was composed of CSL, calcium carbonate, ammonium sulfate, sodium chloride, dextrin, soybean meal grits, soybean oil, and tap water. The fermentor was inoculated with 1 L of broth coming from the prevegetative phase and during the entire stage the pH was kept at 6.3, with all the other conditions remaining as in the prevegetative stage. The vegetative fermentation normally completes in 20–24 h, reaching a packed mycelium volume close to 20%.

The fermentative stage was performed in parallel in two fermentors. The medium was composed of glucose, soybean meal grits, potato flour, calcium carbonate, CSL, ammonium sulfate, sodium chloride, soybean oil, antifoam, and tap water. After sterilization at 120°C for 60 min, the medium was inoculated with 1 L of broth coming from the vegetative fermentor that reached the higher value of packed mycelium volume. From log 10 on, propionic acid and soybean oil were added continuously according to in-house feeding schemes. The chemical-physical parameters were kept to dynamic set points by the automatic control system of the fermentors.

Analytical Determination of Fermentation Parameters

The PPS15 fermentors are fully equipped for online monitoring and control of the major process variables, such as temperature (measured using a stainless steel temperature probe with two resistances Pt-100), dissolved oxygen (DO) tension (Ingold *in situ* sterilizable polarographic electrode), pH (Ingold 405-DPAS-SC-K8S/120 *in situ* probe), pressure, agitation speed, and flow rate. All the process operations are handled by a personal computer through the bioreactor control system coded in-house.

The packed mycelium volume was determined by a manual reading after centrifugation of 10 mL of fermentation broth at 2300*g* for 10 min. Propionic acid concentration is determined by gas chromatography using an HRGC 53000-Mega Acid 25-m column, operating with helium as carrier fluid at 30 kPa with a flow rate of 4.5 mL/min. Erythromycin concentration was determined by high-performance liquid chromatography using a C_8 column with KH_2PO_4 as the mobile phase at a flow rate of 1 mL/min. Sugar concentrations were determined by absorbance measurements at 420 nm with a Perkin-Elmer spectrophotometer. Ammonia concentration readings were obtained using a Metrohm 692 pH-meter.

Multivariate Principal Component Analysis

PCA is a statistical data analysis technique that allows reduction of the dimensionality of the system while preserving information on the variable interactions *(9–11)*. PCA transforms the original variables into a set of linear combinations, the principal components (PCs), with special properties in terms of variances. Specifically, it determines an optimal linear transformation $\boldsymbol{y} = \boldsymbol{Wx}$ of an *n-dimensional* data vector $\boldsymbol{x}$ into another *m-dimensional* ($m \leq n$) transformed vector $\boldsymbol{y}$. The $m \times n$ fixed linear transformation matrix $\boldsymbol{W}$ is designed for exploring statistical correlations among variables of the original data matrix and finding reduced compact data representations that retain maximum nonredundant and uncorrelated intrinsic information of the original data. Exploration of the original data set is based on computing and analyzing the data covariance matrix, its eigenvalues, and corresponding eigenvectors organized in descending order. Each element of the *m-dimensional* transformed feature vector $\boldsymbol{y}$ will be statistically independent and in decreasing order according to decreasing information content. This allows a straightforward reduction of the dimensionality by discarding the feature elements with lower information content. Thus, all original *n-dimensional* data patterns can be optimally transformed to data patterns in a feature space with lower dimensionality. In particular, a process measurement matrix can be projected into a space where the principal components are the orthogonal axes that account for the greatest variability displayed by the data *(9)*.

Since several texts cover the calculation of the PCs in details *(9,12–14)*, only theoretical aspects are briefly summarized. The algorithm chosen for this work is based on the singular value decomposition (SVD). Given a data matrix, $\boldsymbol{D}$, that has dimensions $r \times c$, corresponding to r observations of c variables, SVD produces the following decomposition:

$$\boldsymbol{D} \cong \boldsymbol{U} \times \boldsymbol{S} \times \boldsymbol{V}^T \tag{1}$$

in which $\boldsymbol{U}$ and $\boldsymbol{V}$ are $r \times d$ and $c \times d$ matrices, respectively, each of which has orthonormal columns so that $\boldsymbol{U} \times \boldsymbol{U}^T = \boldsymbol{I}_d$, $\boldsymbol{V}^T \times \boldsymbol{V} = \boldsymbol{I}_d$; $\boldsymbol{S}$ is a $d \times d$ diagonal matrix whose nonnegative elements, called singular values, are ordered in decreasing order: $S_{i+1} < S_i$; and d is the true rank of $\boldsymbol{D}$.

If there is no noise in the system and redundancy exists, Eq. 1 changes to equality. Eq. 2 gives the relationship between singular values and PCs, designated as Z_is:

$$\boldsymbol{Z} = \boldsymbol{U} \times \boldsymbol{S} \tag{2}$$

in which $\boldsymbol{Z}$ is an $r \times d$ matrix whose columns correspond to the PCs; that is, the ith column of $\boldsymbol{Z}$ is the ith PC and will be designated $\boldsymbol{Z}_i$.

From Eqs. 1 and 2, each element of $\boldsymbol{D}$ can now be written as follows:

$$d_{ij} = \sum_{n=0}^{d} \boldsymbol{Z}_{in} \times \boldsymbol{V}^T_{nj} + e_{ij} \cong \sum_{n=0}^{d} \boldsymbol{Z}_{in} \times \boldsymbol{V}^T_{nj} \tag{3}$$

or, in matrix form, as

$$\boldsymbol{D} \cong \boldsymbol{Z} \times \boldsymbol{V}^T \tag{4}$$

The approximation sign shown in the second half of Eq. 3 and in Eq. 4 is owing to the fact that not all the PCs are being used. The error term e_{ij} is assumed to be zero for Eq. 4. As mentioned earlier, the first few $\boldsymbol{Z}_i$s capture much of the variance in the data. Should all the generated singular values (which are also equivalent to the square root of the eigenvalues) be used, the resulting PCs would regenerate the data perfectly, including the noise. Since the higher $\boldsymbol{Z}_i$s capture the least amount of variance, they are assumed to reflect system noise; hence, their elimination reduces the noise level and achieves data compaction. The final result is a small set of transformed variables that captures most, if not all, the variability in the data.

In the case of batch-monitoring problems, the experimental data take the form of three-way arrays in which $i = 1, 2, \ldots, I$ batches are described by $j = 1, 2, \ldots, J$ variables measured at $k = 1, 2, \ldots, K$ time intervals throughout the process. In the three-dimensional $\boldsymbol{D}(I \times J \times K)$ array different batch runs are organized along the vertical side, the measured variables are along the horizontal side, while their time profiles occupy the third dimension *(6)*. Thus, the three-way array must be unfolded into a 2D matrix before applying PCA to batch-monitoring problems. Two of the three possible ways of unfolding the $\boldsymbol{D}$ result are particularly valuable to extract useful information from the measurement data matrix. The first unfolding method consists of creating a matrix whose columns generate from each vertical slide $(I \times J)$ of $\boldsymbol{D}$, starting with the one corresponding to the first time interval. The resulting 2D matrix has dimensions $I \times JK$. This unfolding allows identification of possible *batch-to-batch variability*, exploiting the information content of the data with respect to both variables and temporal variation.

With the second unfolding method, the 2D matrix is constructed by adding one after the other of the I blocks $(J \times K)$ that describe each single batch. The resulting matrix has dimensions $IK \times J$, and this unfolding is used to monitor the time progression of the different batches in a reduced PCA space where statistical control limits and parameters help identify anomalous process behaviors and their possible causes.

If the analysis is limited to the projections on a PC plane (defined by the first two PCs), normal or anomalous process behaviors can be identified using control limits defined as 90 and 95% confidence regions based on reference distribution of batch PCs. A batch is defined as anomalous when it scores a value of $Z_{r,k}$ (namely, the rth PC at time interval k) that moves outside the range of variation defined by the control region. Two different statistical parameters are used to quantify the behavior of a batch in the reduced plane defined by the first two PCs: Q and T^2. The sum of the square residual Q is given by Eq. 5:

$$Q_i = \sum_{i=3}^{d} Z_i^2 \tag{5}$$

It represents the square distance of each batch i perpendicular to the space defined by the first two PCs. Similarly, T^2 is defined as

$$T_i^2 = \boldsymbol{Z}_i^T \boldsymbol{S}_{\text{scores}}^{-1} \boldsymbol{Z}_i \tag{6}$$

and gives a measure of the distance in the reduced plane identified by $\boldsymbol{Z}$, the vector of the first two PCs, between the position of batch i and the origin that designates the point with the minimum variation in the batch process behavior *(6)*. $\boldsymbol{S}_{\text{scores}}$ is the variance-covariance matrix of the scores.

Thus, in accordance with Nomikos and MacGregor's work *(6,7)*, the procedure for monitoring the time progress of a multivariate process using PCA charts starts by computing the PCs for each observation vector $\boldsymbol{i}$ (batch $\boldsymbol{i}$). If the batch scores Q and T^2 fall inside the control limits, then the process can be defined as normal. Otherwise, the examination of each variable contribution to the value of Q and T^2 provides some insights into the nature and causes of the anomalous condition.

Results and Discussion

The database analyzed herein was composed of eight production runs from an industrial fed-batch fermentation process courtesy of Biofin. The eight fermentations corresponded to four different experiments (Eri22, Eri23, Eri24, Eri25) carried out in parallel on two identical fermentors (BIM PPS15), named F2 and F5, respectively. Each batch had a duration of 134 h and was described by six online process measurements (temperature, pH, DO, agitation speed, airflow rate, and pressure), by six offline variables (concentrations of reducing sugars, total sugars, ammonia, propionic acid, erythromycin, and packed mycelium volume), and by four calculated parameters (glucose feed rate, propionic acid feed rate, total glucose fed, and total propionic acid fed). Since the analytical and calculated measurements had been available once a day, the complete database results in a three-way array $\boldsymbol{D}$ with dimensions 8 × 16 × 7 (Table 1).

Table 1
Complete Fermentation Database (three-way array ***D***)

I		***J***		***K***	
Batch no.	Batch ID	Variable no.	Variable ID	Time point no.	Time (h)
1	Eri22F2	1	Temperature (°C)	1	0
2	Eri22F5	2	Pressure (bar)	2	14
3	Eri23F2	3	pH	3	38
4	Eri23F5	4	DO (%)	4	64
5	Eri24F2	5	Agitation speed (rpm)	5	86
6	Eri24F5	6	Airflow rate (nL/min)	6	110
7	Eri25F2	7	Packed mycelium volume (%)	7	134
8	Eri25F5	8	Total sugars (g/L)		
		9	Reducing sugars (g/L)		
		10	Ammonia nitrogen (mg/L)		
		11	Propionic acid (mg/L)		
		12	Erythromycin (mg/L)		
		13	Glucose feed rate (g/h)		
		14	Propionic acid feed rate (mg/h)		
		15	Total glucose fed (g)		
		16	Total propionic acid fed (mg)		

Analysis of Batch-to-Batch Variability

The application of the first unfolding method resulted in a 2D matrix with dimensions 8 × (16 × 7). PCA was applied to identify possible batch-to-batch variability. Each batch was projected on the space defined by the first two PCs, accounting for 38 and 16% of the total variance, respectively. Figure 1 shows the projection of the eight fermentations on the PC_1–PC_2 plane. From the plot of Fig. 1, it can be seen how the eight batches tend to group in pairs. This pairing is most probably determined by the common raw materials, sterilization, and prevegetative and vegetative phases that characterized the four parallel fermentations.

The projection on the space defined by the first three PCs (accounting for 38, 16, and 15% of the total variance, respectively) reveals a finer structure of the data clustering (Fig. 2), expanding the knowledge derived from the projection on the PC_1–PC_2 plane. The two Eri25 runs clearly compose a tight group. The fact that these two samples cluster separately seems to indicate that prefermentation conditions were different for these last two runs, since no standard operating procedure was varied. Moreover, these two runs were characterized by unusually high ammonia and glucose consumption, and this event can be interpreted as an indication of anomalous inoculum preparation. Finally, fermentation Eri22F5 constitutes a group by its own, distinct from its parent fermentation Eri22F2. The lower right position of this sample accounts for the fact that this batch experienced a fatal bacterial contamination.

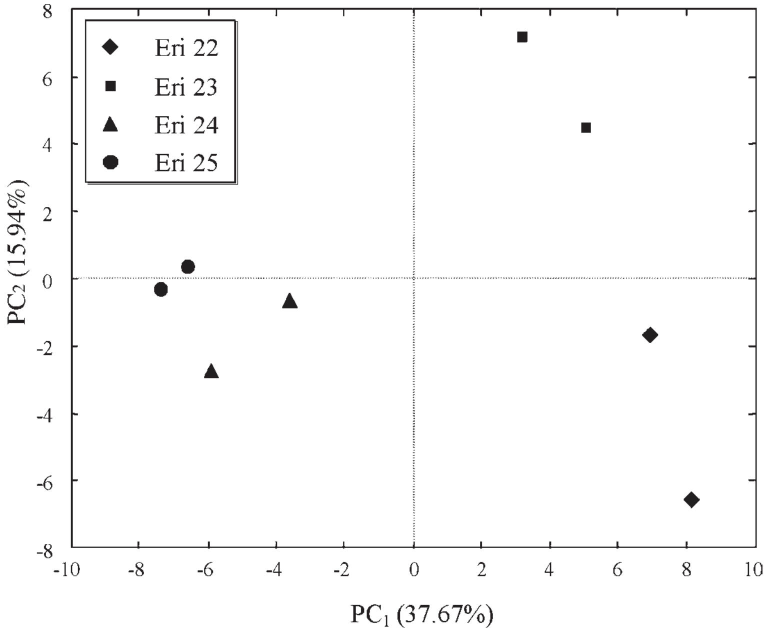

Fig. 1. Projection of eight fermentation runs in PC_1–PC_2 plane.

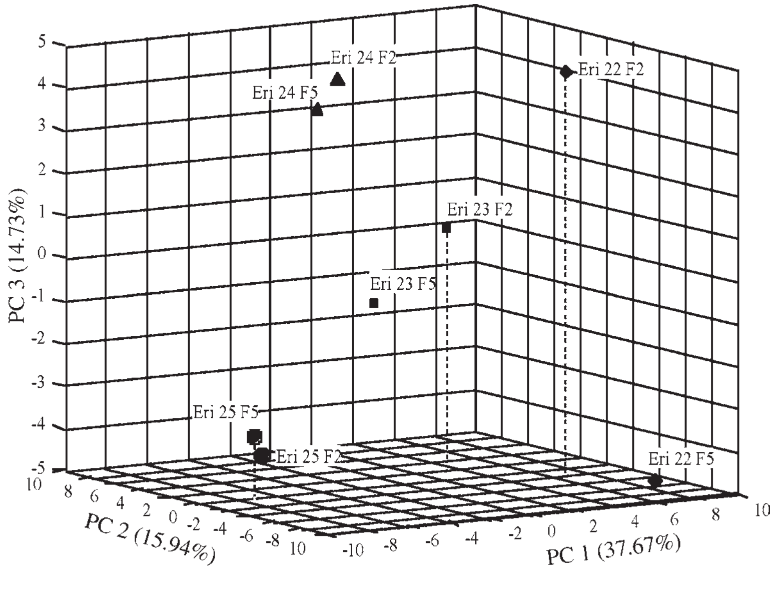

Fig. 2. Projection of eight fermentation runs in PC_1–PC_2–PC_3 plane.

Analysis of Time Profiles for Process Diagnosis and Fault Detection

With the second unfolding method, it is possible to monitor the time progression of the different batches in a reduced PCA plane where statistical control limits and parameters help identify anomalous process behav-

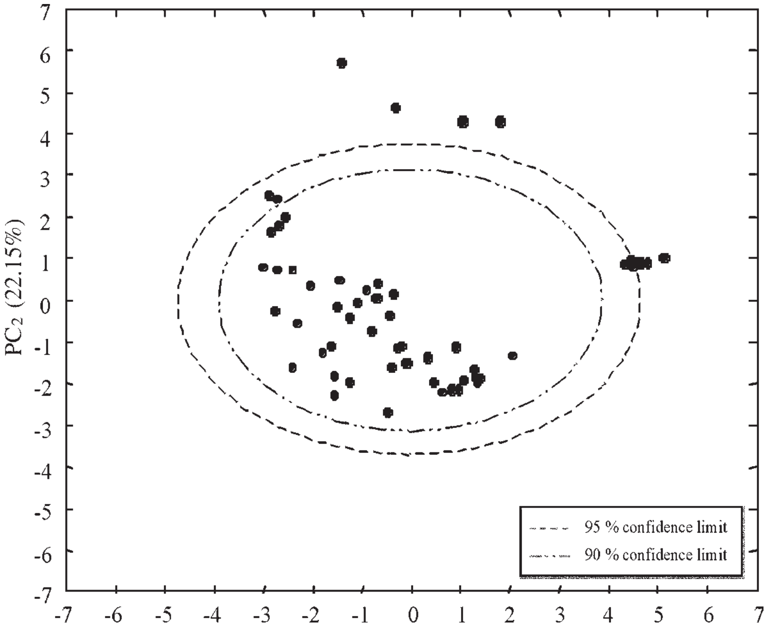

Fig. 3. Time course projection of eight fermentation runs in PC_1–PC_2 plane.

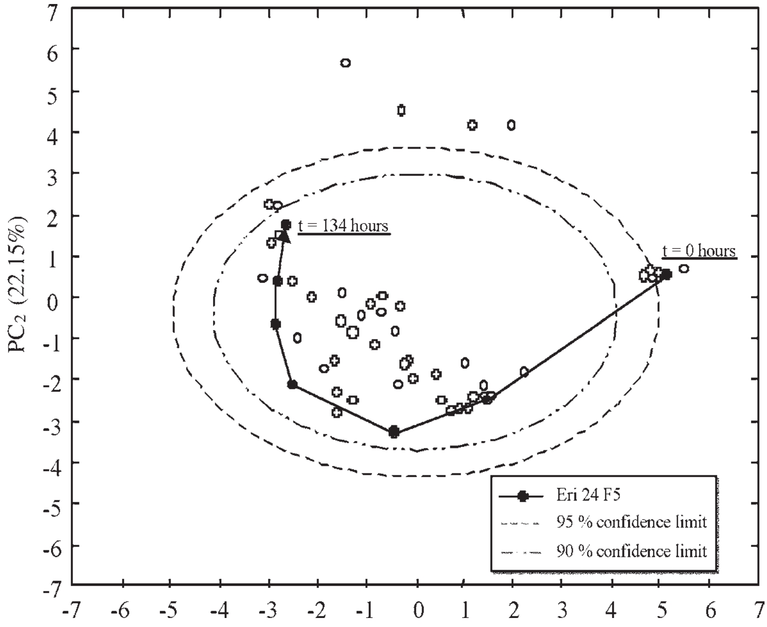

Fig. 4. Time course projection of Eri24F5 run in PC_1–PC_2 plane.

iors and their time intervals and possible causes. From the original database array, a 2D matrix was constructed by adding one after the other of the eight blocks (16 × 7) that describe each single batch. The resulting matrix had dimensions (8 × 7) × 16.

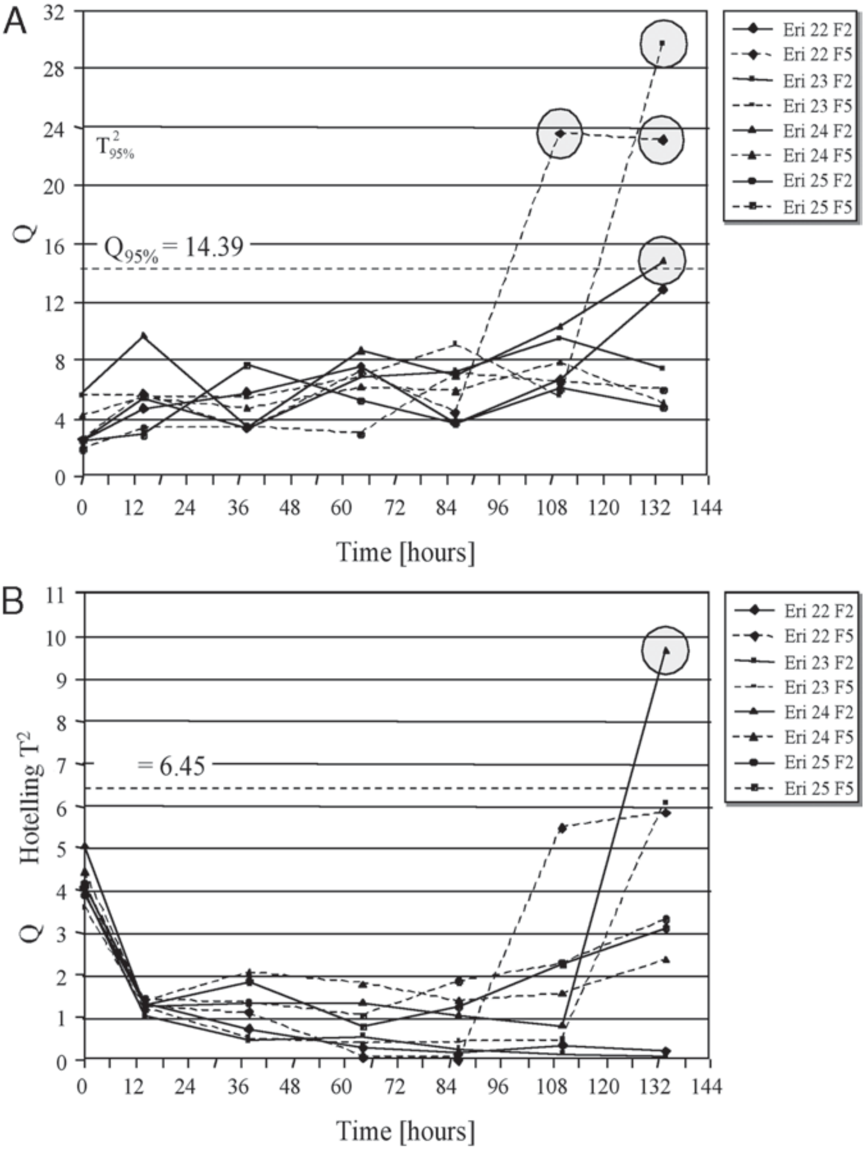

Fig. 5. Time course of *Q* **(A)** and T^2 **(B)** for fermentation runs. The time intervals where the value of these two statistical indexes exceeds the 95% control limit are circled.

Figure 3 represents the time course projections of all the batches in the plane defined by the two first PCs PC_1–PC_2, explaining 34 and 22% of the total variance, respectively. In Fig. 3, three major clusters can be identified. The first one (right-side group) comprises the initial time point (t = 0 h) of the fermentations and confirms the statistically uniform setup of all the batches. The second cluster constitutes the central group inside the confidence limits, while the third is composed by the ending time intervals of the anomalous runs. Figure 4 highlights the time course of a normal fermentation: batch Eri24F5. The time direction is highlighted as a solid line.

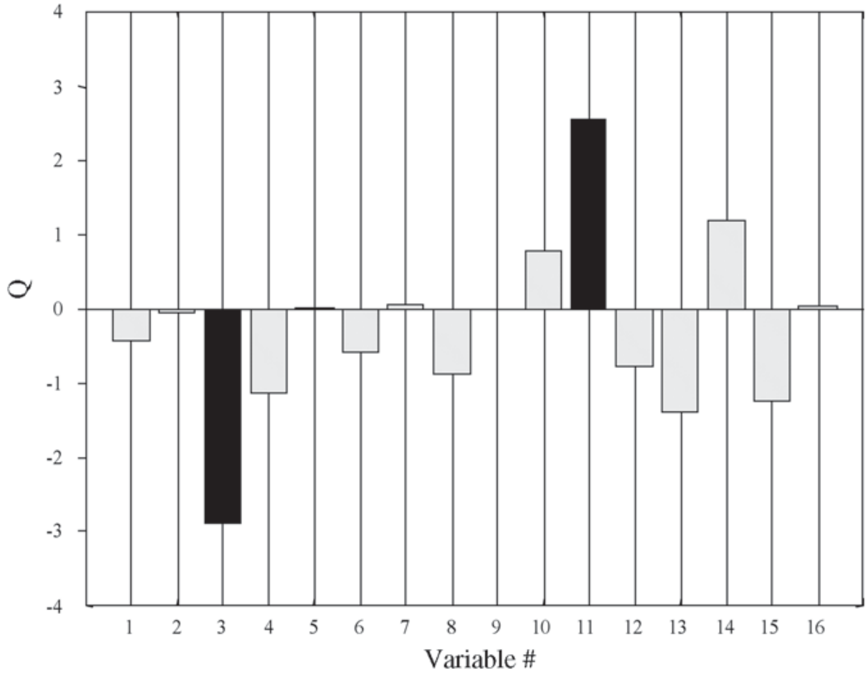

Fig. 6. Variable contributions to *Q* for fermentation batch Eri22F5.

Figure 5A, B reports the time course of *Q* and T^2 for each fermentation run; the time intervals where the value of these two statistical indexes exceeds the 95% control limit are highlighted. These monitoring charts not only help to identify anomalous process behaviors but also can shed light on their possible causes. By analyzing the contribution of every single variable to the final values of *Q* or T^2 for those time intervals where the parameter exceeds the control limit, it is possible to have some insights into the reasons that determined the faulty condition. For example, Fig. 6 reports the contribution of each single variable to the value of *Q* at the final time interval (134 h) for fermentation batch Eri22F5. Variable no. 3, corresponding to pH, and variable no. 11, propionic acid, give the larger negative and positive contributions accounting for an anomalous accumulation of propionic acid in the broth with a consequent lowering of pH. This situation clearly indicates a critical condition for the biomass, which is unable to metabolize the fed acid. Indeed, the Eri22F5 fermentation batch experienced a fatal bacterial contamination that caused the death of the entire *Streptomyces* population. Similarly, the bar graph of the variable contributions to the value of *Q* at the final time interval (134 h) for fermentation batch Eri24F2 indicates that the glucose feed rate accounts for the larger contribution to *Q* (data not shown). Since the glucose feed rate is in cascade with the pH-measuring probe to adjust the pH value to a constant set point, its contribution seems to indicate an anomaly related to the pH control. In fact, an investigation at the end of the run revealed that the pH probe had a one-unit drift that induced the control system to raise the glucose feed rate in the attempt to lower the faulty higher pH value. Note that the contribution of the pH profile to the value of *Q* is not significant, and this is owing to the fluctuations of the signal around the set point value determined by the faulty reading and the control system action.

Conclusion

An approach for monitoring the progress of batch fermentation processes has been presented. Because of the poor mechanistic understanding of most biologic systems, this procedure utilizes information contained in historic databases of process measurements to extract knowledge in the form of patterns. These patterns are used to derive data-driven models for the online identification of process state and performance. Specifically, a multivariate statistical procedure originally developed by Nomikos and MacGregor *(6,7)* was applied for analyzing the measurement profiles acquired during the monitoring of several fed-batch fermentations for Erythromycin production. MPCA was used to extract information from the multivariate historic database by projecting the process variables onto low-dimensional spaces defined by the PCs. The projections allowed identification of similarities among eight different batches and characterization of possible sources of batch-to-batch variation for the definition of optimal operating procedures. Groups of similar fermentation batches could also be identified by usual cluster analysis algorithms that group feature vectors on the basis of a distance or similarity metric. However, when, as in this case, the sample is described by a matrix of measurements with rows representing the time points and columns the variables, measures of similarity for sample clustering become a more demanding and partially unsolved task *(15)*. Principal component projection thus represents a straightforward tool for identifying fermentation batches with common characteristics.

The multivariate procedure was also used to track the time progression of the different batches in the reduced plane where statistical control limits and parameters helped identifying anomalous process behaviors and their possible causes.

This monitoring framework is currently under integration in the bioreactor control system and software equipping BioIndustrie Mantovane fermentors. The new controlling scheme will initially rely on the model generated using the data collected during the present research work (eight batches). However, the simplicity and computational efficiency of the SVD algorithm allow the model to be automatically adapted every time the historic database expands through the acquisition of new data. The possibility of reformulating the model on a broader database, eventually accounting for changes in the standard operating procedures, will result in a more reliable and robust representation of the system.

References

1. Royce, P. N. (1993), *Crit. Rev. Biotechnol.* **13(2),** 117–149.
2. Lübbert, A. and Simutis, R. (1994), *TIBTECH* **12,** 304–311.
3. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996), *Advances in Knowledge Discovery and Data Mining*, MIT Press, Cambridge, MA.
4. Stephanopoulos, G. N., Locher, G., Duff, M. J., Kamimura, R. T., and Stephanopoulos, G. (1997), *Biotechnol. Bioeng.* **53(5),** 443–452.
5. Kamimura, R. T., Bicciato, S., Shimizu, H., Alford, J., and Stephanopoulos, G. N. (2000), *Metab. Eng.* **2(3),** 218–227.
6. Nomikos, P. and MacGregor, J. F. (1994), *AIChE J.* **40(8),** 1361–1375.
7. Nomikos, P. and MacGregor, J. F. (1995), *Technometrics* **37(1),** 41–59.
8. Glassey, J., Montague, G., and Mohan, P. (2000), *TIBTECH* **18,** 136–141.
9. Jolliffe, I. (1986), *Principal Components Analysis*, Springer-Verlag, New York.
10. Brereton, R. G. (1992), *Multivariate Pattern Recognition in Chemometrics*, illustrated by Case Studies, Elsevier, New York.
11. Massart, D. L., Vandeginste, B. G. M., Deming, S. N., Michotte, Y., and Kaufman, L. (1988), *Chemometrics: a Textbook*, Elsevier, New York.
12. Dillon, W. R. and Goldstein, M. (1984), *Multivariate Analysis: Methods and Applications*, John Wiley & Sons, New York.
13. Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, Academic, San Diego.
14. Cios, K., Pedrycz, W., and Swiniarski, R. (1998), *Data Mining—Methods for Knowledge Discovery*, Kluwer Academic, Boston.
15. Kamimura, R. T., Bicciato, S., Shimizu, H., Alford, J., and Stephanopoulos, G. N. (2000), *Metab. Eng.* **2(3),** 228–238.