Hybrid Modeling Framework for Process Analytical Technology: Application to Bordetella Pertussis Cultures

M. von Stosch
LEPAE, Departamento de Engenharia Quimica, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

R. Oliveria
REQUIMTE, Departamento de Quimica, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

J. Peres and S. Feyo de Azevedo
LEPAE, Departamento de Engenharia Quimica, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

DOI 10.1002/btpr.706
Published online October 31, 2011 in Wiley Online Library (wileyonlinelibrary.com).

In the process analytical technology (PAT) initiative, the application of sensors technology and modeling methods is promoted. The emphasis is on Quality by Design, online monitoring, and closed-loop control with the general aim of building in product quality into manufacturing operations. As a result, online high-throughput process analyzers find increasing application and therewith high amounts of highly correlated data become available online. In this study, an hybrid chemometric/mathematical modeling method is adopted for data analysis, which is shown to be advantageous over the commonly used chemometric techniques in PAT applications. This methodology was applied to the analysis of process data of Bordetella pertussis cultivations, namely online data of near-infrared, (NIR), pH, temperature and dissolved oxygen, and off-line data of biomass, glutamate, and lactate concentrations. The hybrid model structure consisted of macroscopic material balance equations in which the specific reactions rates are modeled by nonlinear partial least square (PLS). This methodology revealed a significant higher statistical confidence in comparison to PLSs, translated in a reduction of mean squared prediction errors (e.g., individual root mean squared prediction errors calibration/validation obtained through the hybrid model for the concentrations of lactate: 0.8699/0.7190 mmol/L; glutamate: 0.6057/0.2917 mmol/L; and biomass: 0.0520/0.283 OD; and obtained through the PLS model for the concentrations of lactate: 1.3549/1.0087 mmol/L; glutamate: 0.7628/0.3504 mmol/L; and biomass: 0.0949/0.0412 OD). Moreover, the analysis of loadings and scores in the hybrid approach revealed that process features can, as for PLS, be extracted by the hybrid method. © 2011 American Institute of Chemical Engineers Biotechnol. Prog., 28: 284–291, 2012

Keywords: hybrid modeling, process analytical technology (PAT), near infrared, partial least square, Bordetella pertussis

Introduction

The process analytical technology (PAT) initiative, published as a nonbinding guidance for industry by the US Food and Drug Administration in 2004,1 was recognized worldwide because it offers the opportunity to cut down product trial time and thus costs. In PAT guidelines, the use of system-integrated approaches, throughout the different stages of product trial (from the development till the manufacture), is encouraged.

The integration of different levels and sources of information requires a framework in which the integrated objects can be adequately linked to establish the desired synergy.

The idea is quality by design that not only starts at the design stage of the manufacturing process but also addresses the need for improved online monitoring and control methods to maintain high product quality during manufacturing operations.2 On the level of process development, the intertwining of process analyzers and adequate data evaluation tools is encouraged in PAT1 to determine the process state at-time and to ultimately manipulate it. Many times, the direct identification of the state is hindered by the fact that either the process key-variables are not at-time measurable or (as undesirable from the process engineering perspective)3 these measurements are invasive or destructive.

At-time knowledge about the key-variables can in principle be derived from noninvasive and nondestructive measurements of other quantities.4 Devices fulfilling these
requirements and that are able to provide information about the physiological state of cells are for instance capacitance probes, infrared spectroscopy, fluorescence spectroscopy and so forth.\(^3\)\(^5\)

Although it is difficult to calibrate the measured physiochemical properties to a meaningful process quantity, the huge amount of generated data poses an additional challenge. Solely for one spectroscopic device, the dimensions might easily reach a number that is unfeasible to be analyzed without the support of very efficient mathematic tools. This is one of the reasons why multivariate data analysis (MVDA) tools are frequently applied in the process analysis and why they are expected to potentially play a central role in PAT.\(^2\)\(^3\)

Multivariate regression, (nonlinear) partial least square (NPLS), evolving factor analysis, support vector machines, or principal component analysis (PCA) are probably the ones that are most commonly applied and most successful.\(^6\)\(^7\) These methods are data driven, and in most cases, they are applied on their own disregarding other valuable process knowledge. As recently highlighted in a review article by Glassey et al.\(^7\) the use of hybrid modeling tools that combine MVDA into a common (hybrid) modeling framework still presents a major challenge to the integration of different layers of information about cells and macroscopic processes.

Hybrid modeling that can link different types of process knowledge presents a suitable alternative to pure MVDA.\(^8\)\(^–\)\(^14\) The linking of process information helps to understand the interplay between certain key quantities, and it enhances the reliability of the process predictions. It is such an integrated systems framework where the “process understanding” and the “principles and tools,” both defined in the PAT initiative,\(^4\) are brought together to manage the complexity while every time drawing a more complete process picture.

In principle, either parallel or serial hybrid topologies can be adopted. The latter is particularly suitable for complex systems where some internal mechanisms are poorly known, but for which large data sets are available without direct physical interpretation.\(^15\) As artificial neural networks (ANNs) that are traditionally applied in serial hybrid structures\(^8\)\(^–\)\(^14\) are unsuitable for knowledge extraction from large/highly correlated data, an alternative approach is applied,\(^15\) namely a NPLS model.

In this study, the application of such a hybrid methodology is reported for the monitoring of target metabolite concentrations in a *Bordetella pertussis* cultivation, from online available near infrared (NIR), pH, temperature, and dissolved oxygen (DO) measurements. This monitoring system provides critical online process knowledge that can be used for closed-loop control to maintain process quality or maximize its quantity. For comparison, the hybrid methodology is benchmarked against the standard chemometric tool, a static PLS method.

**Materials and Methods**

**Process and data**

*B. pertussis* is cultivated for the production of a vaccine against whooping cough. Different cultivation strategies are reported in which all seek to identify the optimal cultivation conditions ensuring vaccine quality and quantity.\(^16\)\(^–\)\(^18\) The key to ensure quality and quantity is the real-time control of biomass concentration and specific growth.\(^4\)\(^,\)\(^19\) For the at-time identification of the biomass concentration and the specific growth during the process, Soons et al.\(^4\) compared a methodology using a DO sensor to an approach that is based on in situ NIR spectroscopy. The conclusion anent the NIR-based model was however rather disillusioning, in the sense that the DO sensor-based methodology in the situation of fixed path length and limited number of batches is to prefer.\(^4\)

The number of samples along with robustness is a major concern for model calibration from NIR data.\(^6\)\(^,\)\(^20\)\(^–\)\(^22\) Process conditions and the component under study should be varied in such a way that a robust calibration model can be developed from the response in the spectra. Although this is mostly a task for the experimental design before the experiment, it will be shown in the following that it is feasible to obtain enhanced prediction quality from the same data when incorporating additional information using hybrid modeling methods.

The experimental data of *B. pertussis* which find application in this study are the one reported by Soons et al.\(^4\) The process was run in batch mode with the two main carbon sources for cell growth being glutamate and lactate. Variations to the process conditions were made as reported in Soons et al.\(^4\) namely pH temperature, and DO varied considerably from 6.9 to 7.25 log(H\(^+\)), 33.8 to 34.1°C, and 0 to 100%, respectively.

The model input data are, as usually, auto-scaled, that is the inputs are shifted to zero mean and are scaled by the standard deviation. Fluctuations of the wavelength intensities, that is noise, is one of the reported problems of NIR data. These fluctuations are especially problematic when for modeling purposes the event number of the spectral data is reduced to the time dimensions of the counterparts, which are usually infrequent off-line measured concentration data. Thus, the NIR data were pretreated as in Soons et al.\(^4\) by the application of a Savitsky–Golay smoothing with a 45-point window and a second-order polynomial along the dimension of time.

To account for the natural deviations experienced during production runs in the calibration data, two sets of data (A and B) were designed. Each of these sets comprises five batches for the calibration and two batches for validation, wherein one of the batches, the one which exhibits a limitation in DO and a lower pH, was assigned in both sets for validation. A remaining batch, for which no substrate measurements were available, was applied in both sets for testing. Note that the training data span the space of process operating conditions in which the model will reliably work while the validation data are a measure of the performance of the spanned space, that is natural deviations of the process should be reflected in both.

**Partial least square/projection to latent structures (PLS)**

PLS is commonly applied to correlate spectroscopic data to chemical compound concentration data.\(^23\) The correlation is established through maximization of the data covariances. The fundamentals for this maximization are provided by the model structure of PLS. Therein, in the so-called “outer model,” the matrix of input values and the matrix of output values are decomposed into loading matrices, score matrices, and matrices of residual. Through the “inner” model, which is also referred to as “latent structure,” the score matrices are then linked.\(^20\)\(^,\)\(^23\) In the perspective of statistical process monitoring, quality prediction and fault diagnosis, these
latent structures are of special interest because it can reveal important process information.28,24,25

Two settings of PLS model inputs are investigated in this study. Setting (a) comprises the online measured data of pH, temperature, percentage of DO, and a selection of wavelength of the spectra (833–2,400 nm). Setting (b) contains the online measured data of pH, and a selection of wavelength of the spectra (1,111–1,397 and 1,587–1,852 nm). This wavelength selection originates from a priori knowledge about the kinetics and is chosen in this study because the excluded wavelengths correspond to saturated intensities due to water. Both settings, (a) and (b), are augmented by the initial component concentration values of every batch to compare the PLS model to the hybrid model providing the exact same data. The PLS model outputs comprise the concentrations of lactate, glutamate, and biomass. This PLS structure implies that the identification of the correlations between the inputs and the outputs can only be established when measurements of the input as well as the output are available for the same time instant, implying that the high number of sampled input data is significantly reduced namely to the sampling rate of the concentration data.

Hybrid (nonlinear) PLS model

The adopted serial semiparametric hybrid model structure is schematized in Figure 1. This structure consists of two major modules, namely a module assigned to the macroscopic material balances and another assigned to the biological fluxes. The formulation of material balances is straightforward, yet the balances take a central role, as (i) they build the frame of the system and (ii) they link the macroscopic reactor system to the microscopic cell factory. The material balances written for a batch reactor in the state space form are

$$\frac{dc}{dt} = r.$$  

(1)

Therein, $c$ is a vector of concentrations (in the present case comprising the concentrations of lactate, glutamate, and biomass, $c = \{\text{Lac, Glu, } X\}$) and $r$ is the vector of kinetic rate functions. The vector of kinetic rates is the link to the biological system (see Figure 1); and it describes the rate of consumption or production of the particular compound. In the displayed hybrid model, the biological system is mimicked by a semiparametric model proposed by Oliveira.12 For the present case, a set of kinetic constraints are assumed a priori, namely that (i) the substrate uptake is zero if substrate depletes and (ii) the uptake rates are proportional to the concentration of biomass, and thus the semiparametric model reads as

$$r = \begin{bmatrix} \text{Lac} \cdot X & 0 & 0 \\ 0 & \text{Glu} \cdot X & 0 \\ 0 & 0 & X \end{bmatrix} \phi \begin{bmatrix} r_{\text{Lac}} \\ r_{\text{Glu}} \\ \mu \end{bmatrix},$$  

(2)

where $\phi$ comprises the a priori knowledge about the kinetics and $\rho$ is a vector that comprises the unknown kinetic rates, that is the specific reaction rates. These rate functions are of the form $c_i \cdot r_{c_i}$, and some parameters, $w_A$, i.e., $\rho = \rho(L_0, w_A)$. The vector of inputs, $L_0$, may in general comprise (i) the concentrations, $c$, and/or (ii) at-time measurements $X_{\text{experimental, } 1,\ldots,k-m}$ (see Figure 1). Thus, the rates $\rho = [r_{\text{Lac}}, r_{\text{Glu}}, \mu]^T$ might not only depend on the presently modeled concentrations, but also depend (i) on the physiochemical properties, such as pH or temperature and (ii) on the concentrations of metabolites which are not comprised by the model but whose traces are for instance contained in measured spectra. In the present case, only at-time measurements are comprised in the inputs, that is settings (a) and (b) are used as defined above in the section on PLS. Because of the high numbers and the nature of the information comprised by $L_0$, the adoption of ANNs is infeasible, as (i) this would lead into a highly underdetermined system of equations and (ii) these ANNs are unsuitable for knowledge extraction from large/highly correlated data. Instead, a NPLS alike nonparametric model is adopted.26 The nonparametric model, as illustrated in Figure 1, consists of $o$ independent submodels, i.e.,

$$\rho_{1, \ldots, m}(L_k, w_A) = \sum_{i=1}^{0} \rho_{i, 1, \ldots, m}(L_k, 1, \ldots,k, w_A).$$  

(3)

Each submodel $\rho_{i, 1, \ldots, m}(L_k, 1, \ldots,k, w_A)$ can further be decomposed into an outer and an inner model. The outer model reduces the high dimension of the inputs

$$L_k, 1, \ldots,k = W_{x,j} \cdot t_j,$$  

(4)

by the application of input loadings, $W_{x,j}$, to the input latent variable $t_j$ and decompresses the output latent variable $u_t$ through the application of output loadings, $W_{y,i}$, to

$$\rho_{i, 1, \ldots, m} = W_{y,j} \cdot u_i.$$  

(5)

The inner model links (non)linearly the input latent variable $t_j$ with the output latent variable $u_t$, that is in this study, a ANN representation is chosen:

$$u_t = w_{1,i} \cdot g(w_{1,i} \cdot h(t_j) + b_{1,i}) + b_{2,i},$$  

(6)

where $w_{1,i}$ and $w_{2,i}$ are weights, $b_{1,i}$ and $b_{2,i}$ are biases, and $h(.)$ and $g(.)$ are transfer functions, which are in this study,
The sensitivity equations have to be integrated along with the sensitivity equation technique. The sensitivity time and the subsequent application of for example the non-linear iterative partial least squares (NIPALS) algorithm or (ii) by the sensitivity equation technique. The sensitivity equation technique in the context of fluctuating or sparse or noisy concentration data definitely is to prefer and was therefore adopted.

The sensitivity equations have to be integrated along with the reactor material balances, wherefore a time inexpensive Euler integration scheme was applied. It is convenient to fit the time-steps of this scheme to the sampling rate of the online measurements, for example spectral measurements, to circumvent the interpolation between those.

**Model assessment criteria**

Model assessment criteria are required to objectively assess the model performances and to select an appropriate number of latent variables. For the latter, cross-validation is applied, that is (i) in the case of PLS, the number of latent variables is increased till the desired level of sophistication is reached, that is the best number of latent variables, is selected according to the lowest mean square error (MSE) calculated for the validation data and (ii) in the case of the hybrid approach, the number of latent variables needs to be determined a priori, wherefore an heuristic search of numbers of latent variables that produce the best performing hybrid model in terms of the Bayesian information criterion (BIC) value (defined below) obtained for the validation data, is performed.

The MSE is a qualitative measure of the model performance. Its calculation bases on the number of samples and the distance between the prediction and the measured data value:

\[ \text{MSE} = \frac{1}{P \cdot n} \sum_{j=1}^{n} \frac{(c_{j,\text{mes}}(t) - c_j(t, w_A))^2}{\sigma_j^2}, \]

(7)

where \( P \) signifies the number of samples, \( c_{1,\ldots,n,\text{mes}} \) are the \( n \) off-line measured concentration values and \( \sigma_{1,\ldots,n} \) are the standard deviations of the measured concentrations. However, a criterion for model selection should not only based on the quality of the model estimates. It should also account for both the complexity of the structure in the form of the number of parameters and the number of measured events.

Two criteria regarding these requirements find wide application, namely the Akaike information criteria and the BIC. In the context of the addressed processes, the BIC is reported to be more appropriate, and therefore is adopted in this work for model comparison. The BIC is defined as

\[ \text{BIC} = \left( - \frac{n \cdot P}{2} \ln \left( \frac{1}{P} \sum_{j=1}^{P} \left( c_{j,\text{mes}}(t) - c_j(t, w_A) \right)^2 \right) \right) \]

(8)

where the term in the first bracket is the logarithmic maximum-likelihood and \( n_w \) is the total number of parameters/weights. In the sense of the BIC, the model to prefer is the one that exhibits the larger BIC value for the validation set.

**Results and Discussion**

Comparing PLS and hybrid modeling

An overview of the best model performances in terms of MSE and BIC is compiled in Table 1.
The BIC values obtained by the PLS models for the validation and test data are therein found to be disproportionally high in comparison to the ones obtained for the hybrid models. This significant difference is due to the much higher number of parameters in the PLS models, which is indicated by the respective higher number of latent variables. For the calculation of the BIC, a model with a higher number of parameters is penalized, as this indicates a model structure which is more complex and less robust. Especially for control purposes, model robustness is important, as uncertainty and model-plant mismatch compromise the controller performance. Model robustness can be addressed through the statistical confidence of the estimates and thus the BIC is the measure of such. Therefore, it is concluded that all PLS models presented in Table 1 have a lower statistical confidence than the hybrid models presented.

The analysis of performance in terms of MSE point in the direction just stated, that is the results in Table 1 show that the PLS model performance with the best number of latent variables is worse than that observed for hybrid models.

Not only from another point of view but also supporting the statistical confidence results, the PLS models exhibit performance inconsistencies between themselves, in that the best model structure on the basis of validation data is significantly different from the best structure that would be obtained on the basis of the test data (discussed below).

Finally, comparing results with different model input settings, (a) and (b), it is seen that, as expected, excluding the NIR wavelength with saturated intensities due to water, cases (b), lead to an increase in performance of all hybrid models.

Analysis of model structural differences

The observed discrepancy in the MSE performance raises the question about the possible structural reasons for the better hybrid model performance in comparison to the PLS models.

One main structural difference arises from the nature of the models—the input spectral information is linearly correlated to the concentrations in the PLS model, whereas it is correlated to the kinetic rates in the hybrid model. This issue was subject of analysis, where it was concluded that, in the case of PLS, the estimation quality of the kinetic rates is poor for two main reasons: (1) the calculation of the kinetic rates is prone to error and consequently the identification of the correlation becomes more difficult and (2) the correlation between the spectral intensities and the rates is most probably nonlinear.

A further issue analyzed was the effect that the noise in the input NIR intensities had on estimates of both hybrid and PLS model state variables and parameters.

Fluctuations in the hybrid model estimates are less distinct due to two main reasons: (1) in the serial hybrid modeling framework, the estimated kinetic rates are integrated, this leading to a smoothing effect to the noise in the kinetic rate estimates and (2) the application of the sensitivities approach for parameter identification enables the utilization of input data at each integration time step, as such diminishing the impact of punctual fluctuations on the kinetic rate estimations. This is in contrast to the standard PLS, because for PLS, the identification of the correlations can only be established when both input and output data exist for the same instance of time (see The “Partial Least Square/Project to Latent Structures (PLS)” Section).

This huge difference in the number of input data, that are used for the parameter identification, is exemplified in Figure 2. Therein, it can be seen that in the case of PLS, the number of available 89 input samples decreases to 15, namely to the time instances for which both, input and output samples exist. Note that the number of output samples for both models, the hybrid and the PLS, are exactly the same.

Effect of latent variables

The observation, in this study, of a relatively high number of latent variables (mostly seven) in the case of PLS models, is in agreement with Soons et al. and van Sprang et al. An additional PCA on the inputs revealed however that the variance in the inputs can be captured by only three latent variables (~97.9% of the variance explained). Furthermore, it is observed that the input variance captured by PLS with three latent variables was ~97%, whereas the corresponding captured output variance was ~82%. This low number of latent variables was also observed for the PLS model performance with the test data, in terms of MSE (see Table 1).

The ~97% of input variance captured in L1 with three latent variables gives rise to another question, namely what these latent variables are due to and whether the correlations between inputs and outputs are biunique. It can be seen in Figure 2 that the trajectory of the first input latent variable t1 of the PCA, PLS, and the hybrid models is almost identical, and that, further, their shape is similar to the trajectories of lactate, glutamate, and especially biomass, which are shown in Figure 3. The shapes of the trajectory of the second input latent variable, t2, is observed to be influenced by the DO measurements, as (1) the respective input loading value is usually high and (2) the characteristics of the trajectories partially coincide (see Figure 2). The third input latent variable scores, t3 only shown in Figure 2 for PCA and PLS, can however not be directly related to any specific input quantity.

The observation that the shapes of the trajectories of t1 are similar to the ones of the concentrations is very interesting in connection to the observation that the spectral intensities increase toward the end of each batch. This points at a unique correlation of biomass and the spectral intensities. Therefore, the correlations for glutamate and lactate concentrations would represent a stoichiometric relation to the biomass concentration. The observation further scrutinizes the wavelength selections for lactate, glutamate, and biomass made by Ref. 7 and questions the reason or need of any number of latent variables that is greater than one.

In the case of the hybrid models, a number of two latent variables are justifiable on the basis of the observation that the second latent variable scores can be linked to the DO measurements. In contrast to the PLS models, the hybrid model therefore seems to profit from the known relation between DO and biomass production. Variations in the intensities of the NIR spectra which are due to other properties than concentrations, that is pH, temperature, DO and so on, do not seem to strongly effect the best hybrid model identification.

Qualitative analysis of the performance

A final qualitative analysis is presented based on the comparison of the individual prediction errors for lactate, glutamate, and biomass, presented in Table 2, and the results
presented in Figure 3. Globally, it can be observed that the differences in results between PLS and hybrid models are the most significant for the lactate and glutamate concentrations, as (i) the respective prediction errors (Table 2) of both substrates, lactate and glutamate obtained for the hybrid model cases are in general improved when compared with the ones obtained for PLS and (ii) these improvements are also visually observable (Figure 3). In the case of the biomass concentrations, no such general trend is observable; however, the partially worse performances for biomass in the hybrid model cases (Table 2) might be explained through the compromise between overall performance and individual performance.

The expected increase in performance, when excluding the wavelengths from the inputs whose intensities are saturated, due to water, can be observed for the estimates of lactate, glutamate, and biomass in cases of both hybrid and PLS models (Table 2).
Table 2. Individual Prediction Errors in Form of MSEs (Eq. 7 in which the Standard Deviation Term is Dropped) for Lactate, Glutamate, and Biomass Concentrations Obtained for Training, Validation, and Test Data over Model Types, Data Sets (See the "Process and Data" Section), Model Inputs (See the "Partial Least Square/Projection to Latest Structures (PLS)" Section), and Latent Variables

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Input Setting</th>
<th>Set lv*</th>
<th>MSE Lac† Train (mmol/L)</th>
<th>MSE Lac† Valid (mmol/L)</th>
<th>MSE Lac† Test (mmol/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HYB (a)</td>
<td>A</td>
<td>2</td>
<td>0.7597</td>
<td>0.3044</td>
<td>–</td>
</tr>
<tr>
<td>HYB (b)</td>
<td>A</td>
<td>2</td>
<td>0.5734</td>
<td>0.1304</td>
<td>–</td>
</tr>
<tr>
<td>PLS (a)</td>
<td>A</td>
<td>7</td>
<td>1.1267</td>
<td>1.6728</td>
<td>–</td>
</tr>
<tr>
<td>PLS (b)</td>
<td>A</td>
<td>7</td>
<td>1.5852</td>
<td>4.0445</td>
<td>–</td>
</tr>
<tr>
<td>HYB (a)</td>
<td>B</td>
<td>2</td>
<td>4.8814</td>
<td>0.4231</td>
<td>–</td>
</tr>
<tr>
<td>HYB (b)</td>
<td>B</td>
<td>2</td>
<td>0.7568</td>
<td>0.5169</td>
<td>–</td>
</tr>
<tr>
<td>PLS (a)</td>
<td>B</td>
<td>7</td>
<td>1.0978</td>
<td>1.5008</td>
<td>–</td>
</tr>
<tr>
<td>PLS (b)</td>
<td>B</td>
<td>5</td>
<td>3.2456</td>
<td>3.2692</td>
<td>–</td>
</tr>
<tr>
<td>PLS (b)</td>
<td>B</td>
<td>7</td>
<td>1.8357</td>
<td>1.0174</td>
<td>–</td>
</tr>
</tbody>
</table>

* Number of latent variables. † Lactate. ‡ Glutamate. § Biomass. HYB, Hybrid model.

Figure 4. Input latent variable scores obtained for set A with inputs (b) by the hybrid model that comprises two latent variables.

As can be exemplarily seen in Figure 3, the different hybrid model structures tested in general lead to smoother estimates than those obtained by corresponding PLS models.

The relatively worse quality of the biomass estimates at the beginning of the batches, obtained from PLS models, may be explained with the relatively low NIR spectral intensities due to low biomass concentrations at the beginning of the batch (low signal to noise ratio).

The hybrid model does not suffer from this type of effect because more data are incorporated during the parameter identification and therefore fluctuations are damped. This is an important feature, as for instance, for the control of specific biomass growth rate, a reliable estimation of it is required. In case that this estimate would be derived from fluctuating state estimates it is prone to error. In the hybrid model case, the specific growth rate estimate is (i) directly accessible, as a result of the chosen structure and (ii) in comparison to PLS less noisy, wherefore enhanced control performance would be enabled.

The slight overestimation for biomass by the hybrid model with input setting (a) at about 15 h, which remains till the end of the batch, can be explained by error propagation.

**Extracting process knowledge from latent variable scores**

The structure of the NPLS submodel, described, by Eqs. 3–6, is similar to the structure of PLS models. For such PLS models, the analysis of the input scores represents a relevant source of information concerning characteristics and features of the processes. This important PLS feature is present in the applied hybrid NPLS models. Figure 4 shows the scores plot of the final hybrid model. By analyzing the relative position of the scores to each other as a measure of their similarity, one can extract important process information, online. For instance, it is possible to detect outlying data samples, which enables automatic fault detection. In the present case, it was possible to pinpoint in a single plot NIR, temperature, and DO data outliers. The latent variable time trajectories carry information about distinct process phases and also batch-to-batch variability. Certain process regions can be classified, that is a region of inoculation and a region of high biomass and low substrate concentrations. Thus, from the time-course of the latent scores of complex spectral data for a certain batch, conclusions anent its performance can be effectively extracted using the hybrid modeling approach.

**Conclusions**

In what is called the PAT initiative,1 the Food and Drug Administration proposes an integrated system approach. On the process level, the intermeshing use of process analyzers and adequate tools for the incoming data evaluation is recommended to accurately determine the process state at-time and to ultimately manipulate it.

Hybrid modeling provides an integrated system approach whereby different sources of knowledge can be linked. When applied to chemical or biochemical processes, such a framework can be build on material balances wherein the specific reaction rates are modeled through the combination of both fundamentals and models, typically adopted in PAT, such as (N)PLS.

This methodology was applied to process data of a *B. pertussis* cultivation to correlate online NIR, pH, temperature, and DO measurements to off-line biomass, glutamate, and lactate concentration measurements. Thus, during the process, the state identification would be feasible by using only the at-time available measurements. Benchmarking is provided by the classical PLS methodology.

The following was observed and can be stated:

(i). Results revealed that the statistical confidence in terms of the BIC of the hybrid method in comparison to the PLS method improved by several orders of magnitude (from ~(-1,000) to ~(-10)), an evidence that was supported by the analysis of concentration trajectories, as shown in Figure 3.
(ii). The higher statistical confidence traces back not only to a significantly lower number of latent variables (from 7 to 2) but also to enhanced quality of estimates, which is observed in the form of lower overall and individual mean square errors.

(iii). The lower number of latent variables results from the fact that the scheme proposed could incorporate the existing correlations between main state variables (in this case, DO and specific growth), thus lowering dimensionality.

(iv). The improved quality of the state estimates was essentially the result of two factors: (1) the smoothing effect that the integration procedure had on noise contained in the kinetic rates estimates and (2) the incorporation of a wider range of input data, viz. at each integration step, which was feasible only due to the applied parameter identification procedure.

(v). The extraction of valuable process information from the analysis of the latent scores is enabled, equivalently to the case of PLS.

All in all, the better performance of the hybrid model is worth the higher computational load, in comparison to the PLS method. Further, it provides a more consistent interpretation of the process data in terms of fundamental mechanisms, thus enhancing the level of sophistication of knowledge generated. Finally, as a result of applying, this hybrid structure, the trajectories of the estimated fluxes are directly accessible online, which allows for their control.

Acknowledgments

The authors thank Zita I.T.A Soons, Mathieu Streelfland and the Netherlands Vaccine Institute for providing data and the Fundacao para a Ciencia e a Tecnologia (reference scholarship to the Netherlands Vaccine Institute for providing data and the Biotechnol. Prog., directly accessible online, which allows for their control. hybrid structure, the trajectories of the estimated fluxes are knowledge generated. Finally, as a result of applying, this mechanisms, thus enhancing the level of sophistication of knowledge generated. Finally, as a result of applying, this hybrid structure, the trajectories of the estimated fluxes are directly accessible online, which allows for their control.

Literature Cited


Manuscript received Mar. 3, 2011, and revision received Jun. 13, 2011.