https://doi.org/10.17979/spudc.9788497498081.0278

Modeling and Identification of ABE fermentation processes

Dominik Hose
Universität Stuttgart (Germany), dominik.hose@outlook.de

Cesar de Prada
Dpto. de Ing. de Sistemas, Univ. de Valladolid
prada@autom.uva.es

Gerardo Gonzalez
Dpto. de Ing. Química, Univ. de Valladolid
gerardo@iq.uva.es

Resumen

Se investigan el modelado y la identificación de procesos de fermentación Acetona-Butanol-Etanol (ABE). El enfoque tradicional intenta ajustar los parametros del modelo mediante una optimización para que las simulacines coincidan con los datos experimentales. Estas optimizaciones a menudo tardan mucho y no siempre consiguen un buen ajuste por la necesidad de imponer una estructura de modelo fijo y por la no-convexidad de la función de coste resultante. Se presenta un enfoque que divide el problema en unos sub-problemas que se pueden resolver de forma más efectiva y que elige el modelo del crecimiento de celulas libremente usando ALAMO (Automatic Learning of Algebraic MOdels). Finalmente, se implementa el algoritmo y realiza una identificación con datos reales y se comprueban los resultados mediante una validación.

The modeling and parameter identification of an Acetone-Butanol-Ethanol (ABE) fermentation process is investigated. The traditional approach tries to adjust the model parameters by means of an optimization such that the simulations fit the experimental data. These optimizations often take a lot of time and do not achieve good fits due to the necessity of imposing a fixed model structure and due to the non-convexity of the resulting cost function. A different approach, which divides the big problem into smaller and more efficiently solvable subproblems, is presented. Its advantage is that it determines a model of the cellular growth term on the go using ALAMO (Automatic Learning of Algebraic MOdels) and thus offers more degrees of freedom. Lastly, the algorithm is implemented, tested and validated with real data.

Palabras clave: ABE Fermentation, Modeling, Parameter Identification, Cellular Growth, Tikhonov Regularization, ALAMO

1. Introduction

Recently, the investigation of the Acetone-Butanol-Ethanol (ABE) fermentation process has experienced a rise in popularity due to its possibilities in the production of bio-butanol which is being used in a lot of products such as bio fuels [9].

Until now a large variety of models of this process has been proposed (e.g. [8], [6], [10], [3]). Even though they all share a certain structure, none of these has been agreed on by the scientific community as the de facto standard model which hints at the difficulties that the modeling of biochemical processes brings along. Experiments can often lead to results that contradict existing models, especially concerning the cell growth.

However, other approaches that try to omit these problems exist and have successfully been used. E.g. Bastin et al. propose an alternative adaptive identification technique in [2] which is able to tackle this problem better.

In this paper a new approach will be presented. It allows to quickly determine a mathematical model which will match the data obtained in experiments as well as an easy model building in a divide and conquer approach by dividing the whole identification process into small and efficiently solvable subproblems. The model for the cellular growth term will be determined on the go.

The results will most likely not be applicable to the general modeling case just like most models of this process. Instead they should rather be used to work within a specific experimental setup to determine a model that describes it reasonably well.

2. Modeling

The ABE fermentation process which is being modeled is a batch process, i.e. it is carried out in a closed fermentor without any input or output flow. Only the temperature and the pH are controlled to maintain a certain value. However, other forms such as the continuous fermentation are possible.

Initially, the substrate, in this case glucose, is given into the reactor along with the so-called inoculum, the microbiological starting culture, in this case *Clostridium acetobutylicum*. After a short lag phase in which the bacteria adjust them-

selves to the new environment, the acid production phase or acidogenesis starts in which mainly acetate, butyrate and lactate are produced. This phase is typically indicated by a rapid growth of cells until the pH has dropped from about 7 to 4.5. In a second step, the solventogenesis, the cell number stalls or even decreases and the production of the solvents starts [9]. The temperature is usually maintained constant at its optimum which lies between 30 and 40°C throughout the whole process. For further insight into the biochemistry of ABE fermentation refer to [1].

The important features to be modeled have been identified as the cell growth and death dynamics as well as substrate utilization for the acid and solvent production and cell maintenance and inhibition mechanisms due to an excess of both substrate and solvents in the broth.

A macroscopic model will be employed trying to capture the quantitative dynamics of the process with the purpose of determining later on the best operating conditions. However, it is possible to use microscopic models based on metabolic pathways as proposed in [10]. For an extensive review of models refer to [9].

The nomenclature of the concentrations (in g/L) is given by X: Biomass (*C. acetobutylicum*), S: Substrate (Glucose), P_a: Solvent (Acetone), P_b: Solvent (Butanol) and P_e: Solvent (Ethanol).

One possible model that captures the features explained above and should therefore be able to reproduce the general trajectories of experimental data is the following

$$\dot{X} = \mu X - \lambda X \tag{1}$$
$$\dot{S} = -Y_{XS}\mu X - m_X X \tag{2}$$
$$\dot{P}_a = Y_{XP_a}\mu X \tag{3}$$
$$\dot{P}_b = Y_{XP_b}\mu X \tag{4}$$
$$\dot{P}_e = Y_{XP_e}\mu X. \tag{5}$$

Thus it employs a growth term $\mu = \mu(X, S, P_a, P_b, P_e)$, death and maintenance coefficients λ and m_X as well as the rates Y_{XS}, Y_{XP_a}, Y_{XP_b} and Y_{XP_e} indicating how substrate is converted into cells and product is generated from that.

Without a doubt, the choice of the cellular growth term μ is the most important and at the same time difficult part of the modeling process. Several models exist in literature and each has its justification. Some possible choices are shown in Table 1. Whereas all these models contain substrate inhibition only the models by Hinshelwood and Yang also contains product inhibition which has been shown to play an important role [3]. The model by Yang et al. furthermore considers the pH to include the high sensibility of the cells to the acidity of the broth.

Table 1: Different models for cellular growth, taken from [8], [7], [13] and [12]. In the model by Yang, P_{aa} denotes acetate, P_{ba} denotes butyrate and P_l denotes lactate.

Model	μ
Monod	$\bar{\mu}\frac{S}{S+K}$
Teissier	$\bar{\mu}(1-\exp(-\frac{S}{K}))$
Haldane	$\bar{\mu}\frac{S}{K_1S^2+S+K}$
Hinshelwood	$\bar{\mu}\frac{S}{S+K}\prod_{i\in\{a,b,e,aa,ba\}}(1-K_pP_i)$
Yang	$\bar{\mu}\frac{S}{S+K}(1-(\frac{P_{aa}}{C_{maa}})^{maa}-(\frac{P_{ba}}{C_{mba}})^{mba}-(\frac{P_b}{C_{mb}})^{mb}-m_1(\frac{P_{aa}}{C_{maa}})^{maa}(\frac{C_b}{C_{mb}})^{mb}-m_2(\frac{P_{ba}}{C_{mba}})^{mba}(\frac{P_b}{C_{mba}})^{mb}-m_3(\frac{5.6-pH}{1.6}))$

Note, that the original model by Monod is by far the most popular choice due to its simplicity and the fact that it often suffices to model the general behavior of cell growth according to the Michaelis-Menten kinetics. The other models are often extensions of the one by Monod.

3. Identification approaches

For identification and validation purposes two data sets were provided. The corresponding experiments were carried out in a batch fermentor with Glucose as the substrate and the bacteria *C. acetobutylicum* as biomass. Unfortunately, no information about the temperature or pH during the experiments were provided making it impossible to include these variables in the models.

3.1. Traditional approach

The traditional approach to the parameter identification of such processes considers a given model, for instance model (1)–(5) with states $\mathbf{x}$, parameters $\mathbf{p}$ and outputs $\mathbf{y}$

$$\begin{aligned}\dot{\mathbf{x}} &= f(\mathbf{p}, \mathbf{x}),\\ \mathbf{y} &= g(\mathbf{x})\end{aligned} \tag{6}$$

and a set of experimental output data $(t_k, \tilde{\mathbf{y}}_k)$ for $k = 0 \dots N_k$ to which the outputs of the model

are supposed to be fitted by finding optimal values for $\mathbf{p}$.

Due to the choice of this macroscopic model all the states are outputs $g(\mathbf{x}) = \mathbf{x}$. Furthermore, because they are concentrations, it would make little sense permitting them to be negative, thus $\mathbf{x}(t) \geq 0$, and their initial values are taken from the experimental data $\mathbf{y}(t_0) = \tilde{\mathbf{y}}_0$. The parameters were also chosen in advance such that their sign would be positive, thus $\mathbf{p} \geq 0$. Of course, it is possible to add additional restrictions, e.g. limits on the parameters or states. A weighting matrix W can for instance be used for normalization purposes. The optimization problem is then written as

$$\begin{aligned} \min_{x} \quad & J(\mathbf{p}) = \sum_{k=1}^{N_k} \|W(\tilde{\mathbf{y}}_k - \mathbf{x}(t_k))\|_2^2 \\ \text{s. t.} \quad & \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) = f(\mathbf{p}, \mathbf{x}(t)), \\ & \mathbf{x}(t_0) = \tilde{\mathbf{y}}_0, \\ & \mathbf{x}(t) \geq 0, \\ & \mathbf{p} \geq 0. \end{aligned} \tag{7}$$

To be able to solve this problem only the structure of μ remains to be chosen. The model by Hinshelwood including an inhibitory effect by butanol

$$\mu(X, S, P_a, P_b, P_e) = \bar{\mu}\frac{S}{S + K_\mu}(1 - K_{P_b}P_b) \tag{8}$$

is a reasonable choice.

Resolving this optimization problem theoretically allows us to find the parameters that best fit the experimental data. However, these problems are far from convex and it is often hard to find good initial values or even know the range they should be in. Therefore, global optimization methods have to be considered. Unfortunately, those take a lot of time due to the fact that in each step of the dynamic optimization a simulation has to be executed.

The fit which this technique provides using the genetic algorithm from MATLAB is shown in Figure 1. The cell dynamics are approximated reasonably well and although the substrate utilization yields some error this could be considered an adequate fit.

Unfortunately, the validation results shown in Figure 2 comparing other experimental data with different initial values to their simulation using the same parameters demonstrate how difficult it is to find universal models and parameters that apply in the general case as they do not fit at all.

The underlying problem with this technique is that a model for the cellular growth has to be chosen in advance which may not apply in this special case. However, because all the models provided in Table 1 are heuristic, they cannot generally be considered applicable and have to be chosen carefully.

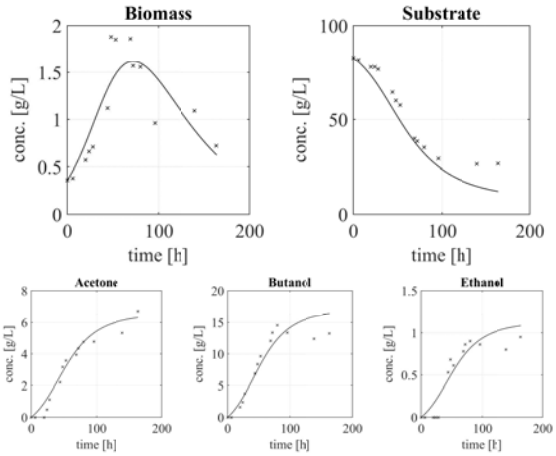

Figure 1: The resulting fit of the global parameter optimization for model (1)–(5).

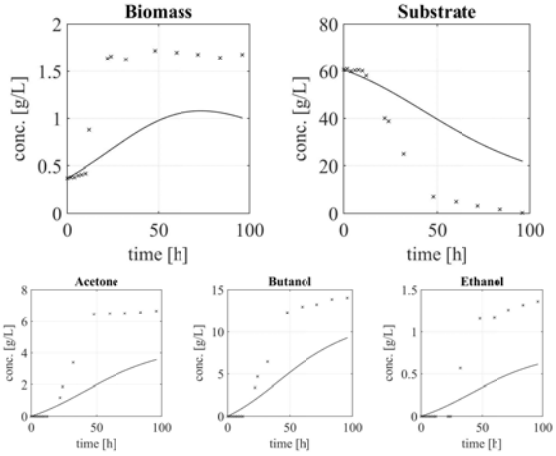

Figure 2: The validation of the global optimization for model (1)–(5).

3.2. Proposed approach

In this section a divide and conquer technique for the model identification will be presented. Instead of trying to tackle the entire problem (7) at once, smaller subproblems will be solved successively. It allows for arriving at a broader range of models as it is not necessary to impose any specific model of the cell growth and the resulting optimization problems are mostly convex thus greatly reducing the computational cost.

Step I Consider Equation (1) for the biomass X. It can be summarized as

$$\dot{X}(t) = g^X(t)X(t), \qquad X(0) = X_0 \tag{9}$$

with an arbitrary time dependent gain $g^X(t)$ and the solution

$$X(t) = \exp\left(\int_0^t g^X(\tau)\mathrm{d}\tau\right) X_0. \tag{10}$$

From the experimental data one would like to extract this gain. Suppose g^X is discretized into an equally distributed piece wise continuous function $\gamma^X = (\gamma_k^X)_{k=1}^{N_k}$ with nodes $(t_k)_{k=1}^{N_k}$. One then gets the approximate piecewise continuous solution of $X(t)$ given by

$$\begin{aligned} X_k = X(t_k) &= \exp(\gamma_k^X \Delta t) X(t_{k-1}) \\ &= X_0 \prod_{j=1}^{k} \exp(\gamma_j^X \Delta t). \end{aligned} \tag{11}$$

Applying a logarithm yields

$$\begin{aligned} \log X(t_k) &= \gamma_k^X \Delta t + \log X(t_{k-1}) \\ &= \log(X_0) + \Delta t \sum_{j=1}^{k} \gamma_j^X \end{aligned} \tag{12}$$

which is linear in the coefficients γ_k^X. Thus, it is possible to formulate the quadratic optimization problem

$$\sum_{k=1}^{N_k} \left\| X_k^{\exp} - X_k(\gamma^X) \right\|_2^2 \to \min. \tag{13}$$

which is equivalent to

$$\sum_{k=1}^{N_k} || \log X_k^{\exp} - \log X_0^{\exp} - \Delta t \sum_{j=1}^{k} \gamma_j^X ||_2^2 \to \min. \tag{14}$$

where $X_k^{\exp}$ are the experimental data at time t_k. Note, that the time discretization has to be chosen such that this is possible. In order to express this in matrix vector form define the data vector

$$\mathbf{y}^{\exp} := (\log X_k^{\exp} - \log X_0^{\exp})_{k=1}^{N_k} \tag{15}$$

and the integration matrix

$$M := \Delta t \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 1 & \dots & 1 & 1 & 1 & 0 \\ 1 & \dots & 1 & 1 & 1 & 1 \end{bmatrix}. \tag{16}$$

If the experimental data at time t_k are not available just leave out the kth row in M and $\mathbf{y}^{\exp}$. Now, write (14) as

$$||\mathbf{y}^{\exp} - M\gamma^X||_2^2 \to \min. \tag{17}$$

One would expect that this minimization problem could be solved by means of the standard regression technique solving the normal equation resulting in

$$\gamma^X = (M^*M)^{-1} M^* \mathbf{y}^{\exp}. \tag{18}$$

However, in most cases this problem is underdetermined and (M^*M) is not regular if N_k is greater than the number of samples. In order to remedy this one can make additional demands on γ^X.

The traditional approach of the Tikhonov or L_2 regularization [11] would be to additionally minimize the L_2 norm of the solution vector

$$||\mathbf{y}^{\exp} - M\gamma^X||_2^2 + \alpha||\gamma^X||_2^2 \to \min. \tag{19}$$

with a regularization parameter α. The solution to the normal equation then takes the form

$$\gamma^X = (M^*M + \alpha \mathbb{1})^{-1} M^* \mathbf{y}^{\exp}. \tag{20}$$

With this modification the problem can be solved. Some exemplary results are shown in Figure 3.

Upon taking a closer look it becomes evident that this regularization is ill-suited because either the resulting gain takes a form that is not typical as it exhibits discontinuities and plateaus (compare e.g. to the gains presented in [12]) or it provides poor consistency with the real experiment depending on α.

Demanding a certain continuity in the gain has proven to be more effective. This can be achieved by minimizing

$$||\mathbf{y}^{\exp} - M\gamma^X||_2^2 + \beta||D\gamma^X||_2^2 \tag{21}$$

where the matrix D is given by

$$D := \frac{1}{\Delta t} \begin{bmatrix} -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -1 & 1 & 0 \\ 0 & \dots & 0 & 0 & -1 & 1 \end{bmatrix} \tag{22}$$

and is an approximation of the first derivative. The corresponding solution is given by

$$\gamma^X = (M^*M + \beta D^*D)^{-1} M^* \mathbf{y}^{\exp}. \tag{23}$$

Figure 4 exhibits that this regularization is more suitable to solve the problem as it produces smoother gains while maintaining a better consistency with the experimental data. A combination is also possible and can be used to combine the desired effects. Other minimizations with respect to other norms such as $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ can successfully be applied, too, without losing the convex properties of the problem. This can achieve several objectives. For instance, a minimization with

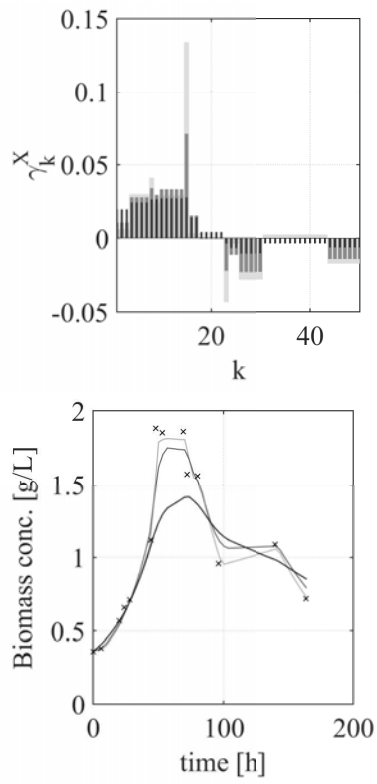

Figure 3: Solution to (19); top: The gains resulting from different regularization parameters (light gray $\alpha = 1$, dark gray $\alpha = 10$, black $\alpha = 100$); bottom: The corresponding simulations using the gains from the top in comparison with the experimental data (x). It is evident that with α one can control if one would like the data to fit the simulation (low values) or would like a more regular gain (high values).

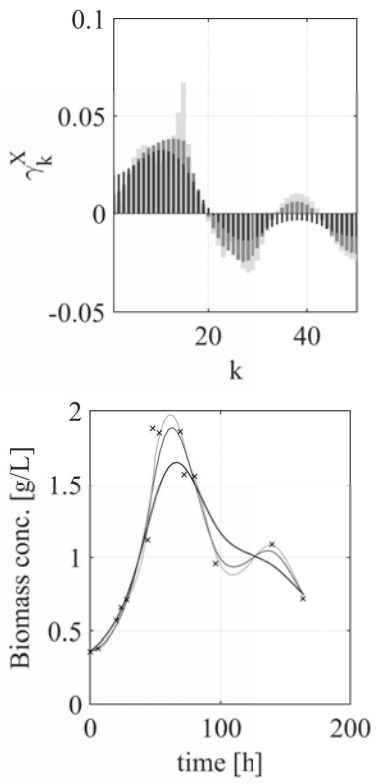

Figure 4: Solution to (21); top: The gains resulting from different regularization parameters (light gray $\beta = 100$, dark gray $\beta = 1000$, black $\beta = 10000$); bottom: The corresponding simulations using the gains from the top in comparison with the experimental data (x). It is evident that with β one can control if one would like the data to fit the simulation (low values) or would like a smoother gain (high values).

respect to the ∞ norm will minimize the peaks of the solution vector γ^X.

From this point on, in the exemplary calculations the regularization parameters $\alpha = 1$ and $\beta = 1000$ will be used. These values are not necessarily as small or as big as they seem because no further scaling has been performed.

This way one can successfully obtain γ^X and also a time evolution of $X^{\text{sim}}(t)$ from the simulations – seen in Figures 3 and 4 – which will be used in the next step to extract the cellular growth μ and the other model parameters.

Step II Again, considering model (1)–(5) one finds that if the time evolution of the cell growth term $\mu(X(t), S(t), P_a(t), P_b(t), P_e(t)) = \mu(t)$ was known beforehand, it would immediately follow from (2) that

$$S(t) = S_0 - Y_{XS} \int_{\tau=0}^{t} \mu(\tau) X(\tau) \mathrm{d}\tau + - m_X \int_{\tau=0}^{t} X(\tau) \mathrm{d}\tau. \tag{24}$$

Discretizing again yields

$$S(t_k) = S_0 - Y_{XS} \Delta t \sum_{j=1}^{k} [\mu_j X(t_j)] + - m_X \Delta t \sum_{j=1}^{k} X(t_j). \tag{25}$$

If (25) is divided by Y_{XS} and written in matrix vector form defining $\mu \in \mathbb{R}^{N_k}$ and

$$\mathbf{S}^{\text{exp}} := (S_k^{\text{exp}} - S_0)_k, \tag{26}$$

$$\mathbf{X}^{\text{sim}} := (X_k^{\text{sim}})_{k=1}^{N_k} \tag{27}$$

one gets

$$\frac{1}{Y_{XS}} \mathbf{S}^{\text{exp}} + Q\mu + \frac{m_X}{Y_{XS}} M \mathbf{X}^{\text{sim}} = 0. \tag{28}$$

The convolution matrix Q is defined by (32) in Equation set 1. Again, if S_k^{exp} is not available, leave out the k-th row of Q and $\mathbf{S}^{\text{exp}}$.

Similarly, for the products $P_i\,(i \in [a,b,e])$ one arrives at

$$\frac{1}{Y_{XP_i}}\mathbf{P_i} - Q\mu = 0 \qquad i \in [a,b,e]. \tag{29}$$

Additionally, it can be supposed that the biomass gain is a composition of cellular growth and death $\gamma^X(t) = \mu(t) - \lambda$, written in vector notation

$$\gamma^X - \mu + \lambda\mathbb{1} = 0. \tag{30}$$

Equations (28), (29) and (30) are altogether linear in

$$\mu, \quad \lambda, \quad \tfrac{1}{Y_{XS}}, \quad \tfrac{m_X}{Y_{XS}}, \quad \tfrac{1}{Y_{XP_a}}, \quad \tfrac{1}{Y_{XP_b}} \quad \text{and} \quad \tfrac{1}{Y_{XP_e}}$$

and can be written as the minimization problem

$$\left\|\Sigma(\tilde{\mathbf{b}} - \tilde{V}\mathbf{z})\right\|_2^2 = \left\|\underbrace{\mathbf{b}}_{:=\Sigma\tilde{\mathbf{b}}} - \underbrace{V}_{:=\Sigma\tilde{V}}\mathbf{z}\right\|_2^2 \to min. \tag{31}$$

with the vectors $\tilde{\mathbf{b}}$ and $\mathbf{z}$ and the matrix $\tilde{V}$ defined by (33), (34) and (35) in Equation set 1 and some scaling matrix Σ. Since the growth term and the parameters have a physical meaning, it is postulated that these parameters are to be greater than zero

$$\mathbf{z} \geq 0. \tag{36}$$

The minimization is then rewritten as the quadratic cost function

$$\begin{aligned} J(\mathbf{z}) &= \|\mathbf{b} - V\mathbf{z}\|_2^2 \\ &= (\mathbf{b} - V\mathbf{z})^T(\mathbf{b} - V\mathbf{z}) \\ &= \mathbf{z}^T V^T V\mathbf{z} - 2\mathbf{b}^T V\mathbf{z} + \mathbf{b}^T\mathbf{b} \end{aligned} \tag{37}$$

which is supposed be minimized under the restriction (36).

The resulting problem can be solved by quadratic programming. From the solution vector $\mathbf{z}$ it is possible to determine $\mu = (z_k)_{k=1}^{N_k}$, $\lambda = z_{N_{k+1}}$, $Y_{XS} = \frac{1}{z_{N_{k+2}}}$, $m_X = \frac{z_{N_{k+3}}}{z_{N_{k+2}}}$, $Y_{XP_a} = \frac{1}{z_{N_{k+4}}}$, $Y_{XP_b} = \frac{1}{z_{N_{k+5}}}$ and $Y_{XP_e} = \frac{1}{z_{N_{k+6}}}$.

In the exemplary calculations the parameters were determined to be

$$\begin{aligned} \lambda &= 0.008, & Y_{XP_a} &= 2.625, \\ m_X &= 0.106, & Y_{XP_b} &= 6.869, \\ Y_{XS} &= 19.224 \quad \text{and} & Y_{XP_e} &= 0.458. \end{aligned}$$

The vector μ with offset $-\lambda$ can be seen in Figure 5.

Until now all the parameters have successfully been identified by means of a linear regression or quadratic programming. Note also, that with all the information available it is already possible to simulate the whole system and obtain time series data of X, S, P_a, P_b and P_e according to (1)–(5). These time series can be seen in Figure 5 and will be used below.

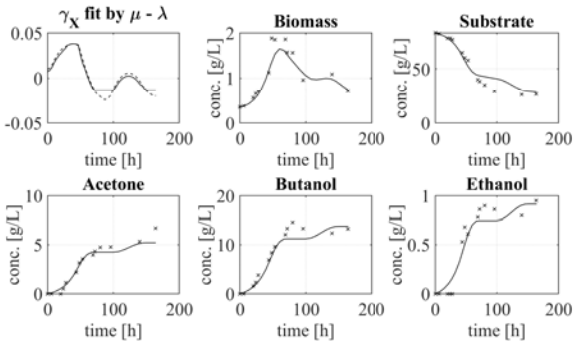

Figure 5: Top left: The gain $\gamma^X(t)$ (dashed) with its approximation $\mu(t) - \lambda$ (continuous). The part where the approximation is a straight line as opposed to the curve of the real gain stems from the restriction (36).; other: Simulation results using the gain approximation in comparison with the experimental data (x). A nice side effect of this technique is that it also provides a filter for the experimental data which can contain a lot of noise. Additionally, restriction (36) ensures that the product only increases and the substrate only decreases.

Step III The remaining step is to find a meaningful expression for μ similar to those from Table 1. Of course, this could also be accomplished by means of a linear regression in order to find the coefficients for a set of given basis functions. However, in order to do so, ALAMO (Automatic Learning of Algebraic MOdels) has proven to be a stronger tool.

ALAMO is an easy to use software that allows to generate algebraic surrogate models of black-box systems [4]. For a set of given in- and output data u and y it can find a relationship $y = f(u)$. For this purpose, it provides some standard basis functions, such as monomials, logarithms and exponentials. Nevertheless, in this context its strengths lie in the possibility of including user-defined basis functions. While a linear regression would assign coefficients to all these basis functions, thus including all of them, ALAMO only picks a few and focuses on simplicity. A detailed explanation of ALAMO can be found in [4] and [5].

As no further dependency on X is assumed, the cellular growth μ is suspected to only depend on S, P_a, P_b and P_e. These will be the input data and μ the output data. However, it suffices to let it

$$Q := \Delta t \begin{bmatrix} X_0 & 0 & 0 & \dots & 0 \\ X_0 & X_1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ X_0 & \dots & X_{N_k-2} & X_{N_k-1} & 0 \\ X_0 & \dots & X_{N_k-2} & X_{N_k-1} & X_{N_k} \end{bmatrix}, \tag{32}$$

$$\tilde{\mathbf{b}} := \begin{pmatrix} \gamma^X & 0 & 0 & 0 & 0 \end{pmatrix}^T, \tag{33}$$

$$\mathbf{z} := \begin{pmatrix} \mu & \lambda & \frac{1}{Y_{XS}} & \frac{m_X}{Y_{XS}} & \frac{1}{Y_{XP_a}} & \frac{1}{Y_{XP_b}} & \frac{1}{Y_{XP_e}} \end{pmatrix}^T \tag{34}$$

$$\tilde{V} := \begin{bmatrix} \mathbb{1} & -\mathbb{1} & 0 & 0 & 0 & 0 & 0 \\ Q & 0 & \mathbf{S}^{\exp} & M\mathbf{X}^{\text{sim}} & 0 & 0 & 0 \\ -Q & 0 & 0 & 0 & \mathbf{P_a}^{\exp} & 0 & 0 \\ -Q & 0 & 0 & 0 & 0 & \mathbf{P_b}^{\exp} & 0 \\ -Q & 0 & 0 & 0 & 0 & 0 & \mathbf{P_e}^{\exp} \end{bmatrix} \tag{35}$$

Equation set 1: Definitions of vectors $\tilde{\mathbf{b}}$ and $\mathbf{z}$ matrices Q and $\tilde{V}$ from Section 3.2.

depend only on S and one product, e.g. P_b. Since the product gains only differ by their coefficient it is easily deduced that e.g.

$$\begin{aligned} P_b(t) &= \int_{\tau=0}^{t} \underbrace{\dot{P}_b(\tau)}_{=\frac{Y_{XP_b}}{Y_{XP_a}}\dot{P}_a(\tau)} \mathrm{d}\tau + \underbrace{P_b(0)}_{=0} \\ &= \frac{Y_{XP_b}}{Y_{XP_a}} \underbrace{\int_{\tau=0}^{t} \dot{P}_a(\tau)\mathrm{d}\tau}_{P_a(t)} = \frac{Y_{XP_b}}{Y_{XP_a}} P_a(t) \end{aligned} \tag{38}$$

and therefore obvious that their ratios are constant and that the inclusion of all three products in the input data does not provide any more information.

Furthermore some Monod-type basis functions of the form

$$f(Y) = \frac{S}{1 + {}^{Y}/_{K_y}} \qquad Y \in [S, P_b]$$

are provided along with the standard basis functions ALAMO already brings along.

Note, that the results obtained can vary strongly depending on experiment, regularization parameters and basis functions. A thorough validation of the results obtained is therefore highly recommended.

In this example, ALAMO arrives at

$$\begin{aligned} \mu =& 3.7 \frac{S}{300 + S} - 0.46 \cdot 10^3 S P_b + \\ & - 0.95 \cdot 10^{-02} \frac{S}{1 + {}^{P_b}/_{14.041904}}. \end{aligned} \tag{39}$$

Thus, the whole model and its parameters can successfully be identified. The resulting fit is shown in Figure 6.

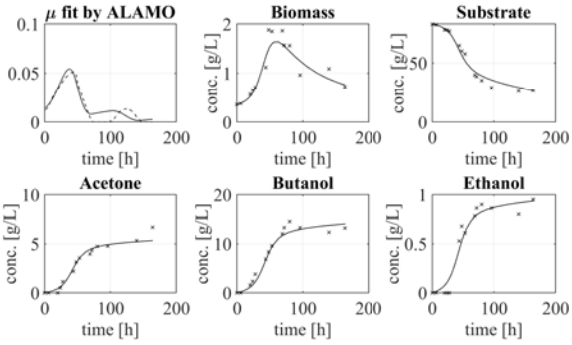

Figure 6: Identification results using the ALAMO model; top left: The cellular growth $\mu(t)$ (dashed) with its approximation by ALAMO (continuous); other: Simulation in comparison with the experimental data (x).

Step IV In a last step the model is again used to predict other experimental data as in the section above. The validation results can be seen in Figure 7. It becomes evident that the accuracy of the simulation is highly dependent on the exactness of the μ model, since the small error in the approximation of the actual gain γ^X by ALAMO still leads to some error.

If you compare Figure 1 to Figure 6 and Figure 2 to Figure 7, though, the advantages of this method become very clear as both the identification and the validation of the proposed approach yield a much better, although not perfect, fit.

4. Conclusions

In this paper an identification approach specifically tailored to the modeling and identification of an ABE fermentation process has been presented.

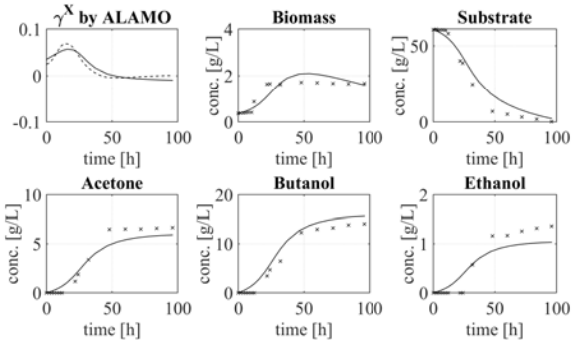

Figure 7: Validation results using the ALAMO model; top left: The gain $\gamma^X(t)$ (dashed) with its approximation by the ALAMO model $\mu(t) - \lambda$ (continuous); other: simulation in comparison with the experimental data (x).

Its strengths compared to the traditional approach are that fewer assumptions about the model have to be made leaving additional degrees of freedom and that the resulting optimization problems are more easily solvable. It has been tested and validated using experimental data. In comparison to the traditional approach the resulting fit was clearly better and the computational cost was significantly lower. In the exemplary calculations it took less than a minute compared to an hour for the global optimization. Further advantages include that neither starting values nor ranges of the parameters have to be considered and the cellular growth term does not need to be modeled.

This technique can certainly be generalized to be applicable to a broader range of models which will likely be presented in a subsequent paper. Especially the identification of models where experimental time series can not be provided for every state would be an interesting subject to investigate.

Agradecimientos

Agradecemos los datos experimentales al departamento de Ingeniería química de la Universidad de Valladolid y al proyecto DPI2015-70975-P (MINECO/FEDER).

References

[1] Hubert Bahl, Wolfram Andersch, and Gerhard Gottschalk. Continuous production of acetone and butanol by clostridium acetobutylicum in a two-stage phosphate limited chemostat. *European journal of applied microbiology and biotechnology*, 15(4):201–205, 1982.

[2] G Bastin and JF Van Impe. Nonlinear and adaptive control in biotechnology: a tutorial. *European Journal of Control*, 1(1):37–53, 1995.

[3] Mathieu Bouville. Fermentation kinetics including product and substrate inhibitions plus biomass death: a mathematical analysis. *Biotechnology letters*, 29(5):737–741, 2007.

[4] Alison Cozad, Nikolaos V Sahinidis, and David C Miller. Learning surrogate models for simulation-based optimization. *AIChE Journal*, 60(6):2211–2227, 2014.

[5] Alison Cozad, Nikolaos V Sahinidis, and David C Miller. A combined first-principles and data-driven approach to model building. *Computers & Chemical Engineering*, 73:116–127, 2015.

[6] JBS Haldane. Enzymes longmans. *Green and Co, UK*, 1930.

[7] JJ Heijnen and B Romein. Derivation of kinetic equations for growth on single substrates based on general properties of a simple metabolic network. *Biotechnology progress*, 11(6):712–716, 1995.

[8] Cyril Norman Hinshelwood. The chemical kinetics of the bacterial cell. Technical report, 1946.

[9] Rahul Mayank, Amrita Ranjan, and Vijayanand S Moholkar. Mathematical models of abe fermentation: review and analysis. *Critical reviews in biotechnology*, 33(4):419–447, 2013.

[10] Hideaki Shinto, Yukihiro Tashiro, Mayu Yamashita, Genta Kobayashi, Tatsuya Sekiguchi, Taizo Hanai, Yuki Kuriya, Masahiro Okamoto, and Kenji Sonomoto. Kinetic modeling and sensitivity analysis of acetone–butanol–ethanol production. *Journal of Biotechnology*, 131(1):45–56, 2007.

[11] AN Tikhonov and V Ya Arsenin. *Methods for solving ill-posed problems*. John Wiley and Sons, Inc, 1977.

[12] Víctor Manuel Trejos, Javier Fontalvo Alzate, and Miguel ángel Gómez Garcia. Descripción matemática y análisis de estabilidad de procesos fermentativos, mathematical description and stability analysis of fermentative processes. *Dyna*, 76(158):111–121, 2009.

[13] Xiaoping Yang and George T Tsao. Mathematical modeling of inhibition kinetics in acetone-butanol fermentation by clostridium acetobutylicum. *Biotechnology progress*, 10(5):532–538, 1994.