Modelling and parameter estimation of gene expression and cell growth in batch cultures [1]

R. Cubarsi
Dept. Matemàtica Aplicada i Telemàtica
Universitat Politècnica de Catalunya, Barcelona, Spain
J. L. Corchero, P. Vila and A. Villaverde
Institut de Biologia Fonamental
Universitat Autònoma de Barcelona, Bellaterra, Spain

[1] This work was presented at the *3ecm Third European Congress of Mathematics*, held 10-14 July, 2000, Barcelona, Spain, and it has been supported by CICYT of Spain under grant No. BIO95-0801, and partially by Generalitat de Catalunya under Grant 1996XT-00030, and CUR (CIRIT) under grant 1995SGHR 00376.

Modelling and parameter estimation of gene expression and cell growth in batch cultures

R. Cubarsi[1], J. L. Corchero[2], P. Vila[2] and A. Villaverde[2]
[1] Dept. Matemàtica Aplicada i Telemàtica, Universitat Politècn ica de Catalunya, Barcelona, Spain; [2] Institut de Biologia Fonamental and Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, Bellaterra, Spain

Abstract

Experimental procedure of CI857ts-controlled recombinant gene expression in bacterial batch cultures is mathematically modelled, and the corresponding minimum variance parameters are estimated from specific statistical or numerical methods, basically by using a global and recursive weighted least squares procedure under some constraints induced by the model. Moreover the numerical techniques proposed in this work act by accumulation of data coming from several runs of the experiment, so that more accuracy is obtained in the parameter estimation. In particular, for the production process, an extra-model parameter depending on an indicator vector is introduced for each run of the experiment in order to globalize the data. The analysis of obtained data leads to an integrated model for both cell growth and gene expression, which describes an asymmetric dynamics between culture growth and protein yield, and can serve to predict the maximal value of accumulated protein and the time required for it to be achieved at any stage of the preinducing cell growth.

CORRESPONDING AUTHOR:
Rafael Cubarsi
Dept. Matemàtica Aplicada i Telemàtica
Universitat Politècnica de Catalunya, Campus Nord
Jordi Girona, 1-3; E08034-Barcelona; Spain
Phone: 34-3-401-5995, Fax: 34-3-401-5981
E-mail: rcubarsi@mat.upc.es

1991 MATHEMATICS SUBJECT CLASSIFICATION: 62, 65, 92.
KEY WORDS: mathematical modelling, parameter estimation, constrained least squares

1 Introduction

One of the mechanisms commonly used to estimulate the expression of the recombinant genes, and consequently, the production of the encoded proteins, is a rapid increase of temperature at which the cells are cultured. At the permissive temperature, 28°C, there is no recombinant gene expression, but when the temperature is shifted to 42°C, cells start the synthesis of the recombinant product while they are also growing in the culture. This temperature-mediated induction of gene expression, which is very convenient for industrial purposes, is achieved by the use of two kind of controllers of the gene expression, which are also introduced into the recombinant cells. They are a positive regulator, the lambda p_{L} and/or p_{R} promoters, and a negative regulator, the repressor CI857, which is active below 32°C but it becomes efficiently inactivated at 42°C (Villaverde et al., 1993). We have developed mathematical procedures to analyze the performance of protein production in cultures of recombinant *E. coli* submitted to heat induction (Cubarsi et al. 1998). For this analysis, two types of mathematical procedures are required. The first type is composed of statistical and numerical methods for parameter and error estimation, fitting curves, etc. But when a function is approximated from a set of data, the problem of what kind of functions must be used always arises. In our case the function must be interpreted from a biological viewpoint, and it must also describe some biological properties of the experimental system. Thus, above techniques can be correctly used only if the biological system has been modeled, and the system properties to be quantified have been focussed. This is the other mathematical aspect of the work, in fact to be done previously to the first one.
Under the assumption that cell growth is not significantly altered by the presence in the cell of the recombinant protein β-galactosidase, the dynamics of the gene expression is modeled in two steps, so that the culture growing model is combined with the gene expression model, a first order differential equation that describes the protein production in terms of cell growth, in order to explain the time evolution of recombinant protein yield along the induction phase. Then, the set of parameters for both models is estimated by using specific least squares techniques subject to constraints from the models, with statistical evaluation of error propagation.
In order to minimize the errors, the numerical algorithms for parameter estimation take advantage of working with a batch culture procedure with multiple induction sequences of the culture, where data from different induction sequences of the same non-induced culture are processed all together, as a single experiment. However, data from several runs of the same experimental process can be pooled in a global data set only under some specific requirements. Thus, for the culture growing process this can be done if the time interval between two consecutive culture samples remains constant along all the process. For the gene expression process, some parameters are constant for all the sequences, namely the model parameters, and others are sequence dependent. In this case, for each run of the experiment, an extra-model parameter depending on an indicator vector is introduced, such as an initial condition for the production process. Hence the total number of estimation parameters is increased by the number of runs of the experiment.
The resulting model for cell growth and synthesis of recombinant proteins reveals an asymmetric distribution of the biosynthetic potential of the cell, which is manifested by a preferential synthesis of recombinant proteins in aged, slowly growing cultures. In other words, both growing and production capabilities of culture cells are not equidistributed, since when the culture growth velocity decreases, the protein production velocity still increases up to its maximum value. Moreover, the proposed model also allows a prediction of the optimal optical density of a batch culture to be temperature-induced in order to get a predetermined amount of recombinant protein with the minimum induction time.

2 Basic notation

The mathematical notation used in the work is now introduced by describing the experimental procedure. An initial amount of culture y_0, measured from its optical density at the wave length of 550 nm (OD), is growing at 28°C (initial stage). At this stage the culture growth can be described in terms of a time parameter t by a function $y(t)$, which is the solution of an autonomous first order differential equation generated by a phase velocity field v_y that, as we shall see in the following section, will depend on two parameters A_0 and B_0:

$$\frac{dy(t)}{dt} = v_y(y(t), A_0, B_0) \tag{1}$$

Hence the solution of this equation may be explicitly expressed depending on the parameters, and the initial value $y_0 = y(0)$, as

$$y = y(t, y_0, A_0, B_0) \tag{2}$$

At a time t from the beginning of the experiment, a sample of culture with OD $y(t)$ is transferred to a prewarmed bath at 42°C (induction stage). Then the growth rate of the culture changes and the production of β-galactosidase protein begins. The production is measured in enzymatic units per ml, referred as β-gal in this work.

The growth process under the induction conditions has a similar behaviour as in the initial stage, but with other model parameter values, namely A_1 and B_1. After a time x in the induction stage, the OD of culture $y(t)$, that had been induced at a time t from the beginning of the experiment, varies along what we shall call the t-induction sequence according to a function $y_t(x)$, which is the solution of a differential equation, similar to Eq. 1, such as

$$\frac{dy_t(x)}{dx} = v_y(y_t(x), A_1, B_1) \tag{3}$$

Hence, the solution can be written explicitly as a function of the model parameters, and the initial value $y(t) = y_t(0)$, in the form

$$y_t = y_t(x, y(t), A_1, B_1) \tag{4}$$

On the other hand, for the gene expression process along the induction stage, that is the recombinant protein production, we assume that the protein has not any toxic effect neither for itself, nor for the culture, and depends only on the OD of the growing culture. Details of experimental procedure are given by Corchero et al. (1994), where the functional dependence of β-gal production in terms of the OD of the induced culture has been proved.

Thus, along the t-induction sequence, if $y_t(x)$ is the OD of an induced culture, for a given induction time x, we can evaluate the production process by means of a function $\beta_t(x) = \beta(y_t(x))$, depending on whether the time evolution or the culture OD dependency of the product is emphasized. Thus $\beta_t(x)$ represents the yield of β-gal for the induction time x in the same induction sequence. The functional dependence $\beta(y_t)$ is studied in the following sections from an approximation model given by a first order differential equation depending also on two parameters c_1 and c_2:

$$\frac{d\beta}{dy_t} = \phi(y_t, c_1, c_2) \tag{5}$$

where ϕ is an arbitrary function of the specified arguments, whose solution can be written for each t-induction sequence from an initial condition c_0^t, so that $c_0^t = \beta(0)$, in the form

$$\beta = \beta(y_t, c_0^t, c_1, c_2) \tag{6}$$

Notice that c_0^t is a function of $y(t)$, that can be implicitly given by $\beta(y(t), c_0^t, c_1, c_2) = 0$, since at the begining of the induction sequence there is no protein yield.
The sub-index referred to the t-induction sequence will be omitted when the context provides sufficient information.
Finally, the production kinetics, the time evolution of the protein production, can be studied by composition of the differential processes expressed in Eq. 3 and Eq. 5, so that the corresponding generating field is

$$\frac{d\beta_t}{dx} = \frac{d\beta}{dy_t}\,\frac{dy_t}{dx} = \phi(y_t(x), c_1, c_2)\; v_y(y_t(x), A_1, B_1) \tag{7}$$

Then, the fuction that describes the protein production in terms of the induction time x can be expressed by recursive substitution of Eq. 2 and Eq. 4 in Eq. 6.

3 Mathematical model

In the working conditions, and for both experiments ((a) **E42**, Table 4, with induction temperature of culture at 42°C, and (b) **E40**, Table 5, with induction temperature at 40°C) the growth rate of culture can be satisfactory described, before and during the induction phase, by using the equation of limited growth of population models (see e.g. Hirsch & Smale, 1974):

$$\frac{dy(x)}{dx} = y(x)(A + By(x)) \tag{8}$$

with A and B arbitrary constants ($A > 0$ and $B < 0$).
At low OD's, the growing rate is nearly constant but, at the same time as the biomass is increasing, the exponential growth stops and the OD of the culture tends to the asymptotic value

$$l = -\frac{A}{B} \tag{9}$$

that depends on the growth conditions.
On the other hand, according to Corchero et al. (1994), a growing culture which has been induced over a time x produces an amount of β-gal $\beta(x)$, that depends nearly in a quadratic way on the biomass $y(x)$ of that culture. Thus, non constant rate of production, with reference to the culture growth, may be reflected along the induction stage, even though toxic effects of the recombinant protein are excluded from this work. Then the relationship between the amount of recombinant protein β-gal and the OD of culture can be written as follows,

$$\beta(y(x)) = c_0 + c_1 y(x) + c_2 y(x)^2 \tag{10}$$

The corresponding differential behaviour, according to Eq. 5, will then have the form

$$\frac{d\beta}{dy} = c_1 + 2c_2 y \tag{11}$$

The meaning of this relationship, from a biological viewpoint, is now investigated by assuming that cell division is not influenced by β-galactosidase protein, and that c_1 and c_2 are parameters of the model.
Along the induction stage the culture is growing according to Eq. 8, with a growth velocity v_y given by

$$v_y(y) = y\;(A + By) \tag{12}$$

Thus, the function $v_y(y)$ is a parabola with vertex at $y = -\frac{A}{2B}$, corresponding to an OD the half of the limit value given by Eq. 9, and also corresponding to the maximum growth velocity. Similarly,

for the production phase, a first approach could assume the same behaviour for the proteins as for the culture, that is, during the induction stage the increase of β-gal is proportional to the increase of biomass, if non-negative. Hence the production velocity of Eq. 7, namely v_β, could be expressed in this simple model as

$$v_\beta(y) = k\, v_y(y) \tag{13}$$

with k a positive constant. In fact, data from Tables 4 and 5 suggest that the increasing of β-gal is always associated with the increasing of biomass (Flickinger & Rouse, 1993). Furthermore, notice that the condition of increasing biomass leads to a working interval $0 \leq y \leq l$ for the culture.
Nevertheless a more complex behaviour, consistent with Eq. 10, must be adopted, since the production protein rate with respect to the culture growth is not constant. More specifically, the OD of culture corresponding to the maximum production velocity, namely m, could be different from the OD of culture for the maximum growth velocity, $y = l/2$.
Therefore, the general case of Eq. 7 must be considered, according to

$$v_\beta(y) = \frac{d\beta}{dy}\, v_y(y); \;\; 0 < y < l \tag{14}$$

taking into account that, if toxicity phenomena are not present in the induction stage the following inequality must be satisfied, in the working interval:

$$\frac{d\beta}{dy} \geq 0 \tag{15}$$

This situation is studied in the first order approximation given by Eq. 11, which enable us to explain in a simple way the asymmetry that the production velocity curve may have with respect to the culture growth velocity curve.

4 Asymmetry between production and growth

In order to compare both velocity curves of Eq. 14, the function in the right hand side member of Eq. 11 will be denoted, according to Eq. 5, as

$$\phi(y) = c_1 + 2c_2 y \tag{16}$$

Then Eq. 14 becomes

$$v_\beta(y) = \phi(y)\, v_y(y) \tag{17}$$

The condition expressed by Eq. 15 implies $\phi(y) \geq 0$ in the working interval, and then it is easy to deduce that: (a) In any case c_1 must be positive. (b) If $c_2 = 0$ both velocities are proportional and they have a common maximum at $m = l/2$. (c) If $c_2 > 0$ the maximum production velocity is reached after the maximum growth velocity of the culture, and $m > l/2$. (d) If $c_2 < 0$, since $c_1 \geq -2c_2 y$ is fulfilled in the interval $0 \leq y \leq l$, and the maximum value of the right hand side member is held at $y = l$, then the following inequality must be satisfied:

$$c_1 \geq -2c_2 l \tag{18}$$

Thus the maximum production velocity is reached before the maximum growth velocity of the culture, and $m < l/2$ is also held.
The three functions involved in Eq. 17 are non-negative for values of the OD within the working interval, and in its bounds $v_\beta(0)$ and $v_\beta(l)$ are null. Furthermore, taking into account the polynomial form of $v_\beta(y)$, it is easy to see that there is a single maximum on this interval. Thus, by assuming

Exp.	$l/2$	ϵ
E42	1.38 ± 0.05	0
E40	1.71 ± 0.03	0.36 ± 0.03

Table 1: Parameters describing the asymmetry between β-gal production and culture growth from Eq. 20.

$c_2 \neq 0$, the relative position ϵ of the abscissa m referred to the value $l/2$, corresponding to the maximum growth velocity of culture $v_y(y)$, is introduced

$$m = l/2 + \epsilon \tag{19}$$

Then the abcissa of the maximum can be written, in terms of an auxiliary parameter $\alpha = l + \frac{c_1}{c_2}$, as follows

$$\epsilon = sign(c_2)\,\frac{1}{|\alpha| + \sqrt{|\alpha|^2 + 3l^2}}\,\frac{l^2}{2} \tag{20}$$

Above expression is also useful in order to see how far can the maximum moves around the central value $y = l/2$, being consistent with the condition of Eq. 15. Notice that $|\epsilon|$ is a decreasing monotonic function of $|\alpha|$, and from Eq. 18 it is easy to see that $|\alpha| \geq l$. Therefore, from Eq. 20, by substitution of this minimum value of $|\alpha|$, we get the admissible range $|\epsilon| \leq \frac{l}{6}$. Also, taking into account Eq. 19, we can conclude that our model enable us to explain a maximum production velocitiy in the following range of values

$$\frac{1}{3}\,l \leq m \leq \frac{2}{3}\,l \tag{21}$$

Notice that if $\epsilon > 0$ the age for significant production is delayed towards high values of OD, while for low OD's the production would be insignificant. If $\epsilon < 0$ the behaviour is in the opposite way.

5 Production kinetics

In this section, for a given t-induction sequence, we study the time evolution of the product content $\beta_t(x) = \beta(y_t(x))$ that is present in the culture at an age x of the induction stage. Remember that the culture with OD $y_t(x)$ has been induced at a time t from the beginning of the experiment. Following the proposed approximation, according to Eq. 8 and Eq. 11, we can write Eq. 7 as follows

$$\frac{d\beta_t}{dx} = (c_1 + 2c_2 y_t(x))(A + B y_t(x)) y_t(x) \tag{22}$$

where $c_1 > 0$, $A > 0$ and $B < 0$. Then, when the induction time $x \to \infty$, the OD of culture tends to the limit l given by Eq. 9. Hence the function $\beta_t(x)$ tends to the asymptotic value

$$\lim_{x \to \infty} \beta_t(x) = \beta(l) \tag{23}$$

That is, from a sufficient large interval of time, the product concentration becomes nearly stationary. Moreover, since the factor $c_1 + 2c_2 y_t$ of Eq. 22 is non-negative in the working interval $0 \leq y \leq l$, this asymptotic value is the maximum yield that can be reached.
Thus the function $\beta_t(x)$ does not have any maximum before reaching their asymptotic value, and, for any induction sequence, the kinetic of the product has a monotonic increasing curve along all the induction process.
In order to obtain the function $\beta_t(x)$ we must take into account how the culture is growing before and during the induction stage, since the function $y_t(x)$ depends also on the OD of culture just at the beginning of the t-induction sequence, $y(t) = y_t(0)$. Thus we write, according to the solution of

Eq. 8, the relationship describing the biomass evolution of a culture that has been induced over a time x,

$$y_t(x) = \frac{A_1 y(t) e^{A_1 x}}{A_1 + (1 - e^{A_1 x}) B_1 y(t)} \tag{24}$$

The sub-index 1 is used to distinguish the induction stage, and the value $y(t)$ represents the OD of the culture at the beginning of the induction sequence, according to

$$y(t) = \frac{A_0 y_0 e^{A_0 t}}{A_0 + (1 - e^{A_0 t}) B_0 y_0} \tag{25}$$

In the latter equation y_0 is the initial amount of culture at the beginning of the experiment, and the sub-index 0 is used to distinguish the growing stage before the induction.
Finally, the production of β-gal in terms of the induction time x is obtained from Eq. 10, also combined with Eq. 24 and Eq. 25, by assuming that in the beginning of the t-induction sequence there is not any significative amount of product,

$$\beta_t(x) = c_1(y_t(x) - y(t)) + c_2(y_t(x)^2 - y(t)^2) \tag{26}$$

Sometimes the foregoing equation will be used in order to describe the time evolution of β-gal, and sometimes the production in terms of OD of the induced culture. Some consequences and features of these equations will be pointed out in the last section.

6 Culture growth parameters

The algorithm to estimate the growth parameters A and B of Eq. 8 for the induction stage, as well as in the initial phase, is based on the time equidistribution of the OD samples, as it is shown in Table 4 and Table 5 for both experiments.
By inverting that expression, a linear dependence between the inverse of the OD of two consecutive culture samples with an arbitrary time separation $\Delta x = x - x_0$ is obtained,

$$\frac{1}{y(x)} = e^{-A(x-x_0)} \frac{1}{y(x_0)} - \frac{B}{A}(1 - e^{-A(x-x_0)}) \tag{27}$$

Hence, by defining

$$\begin{array}{l} a = e^{-A\Delta x} \\ b = -\frac{B}{A}(1 - e^{-A\Delta x}) \\ z_k = \frac{1}{y_k} \end{array} \tag{28}$$

and maintaining the time interval Δx constant for any couple of consecutive samples, Eq. 27 can be written in a simple and recursive way as follows:

$$(\xi_k, \eta_k) = (z_{k-1}, z_k),\ \eta_k = a\xi_k + b;\ \ k = 1, \ldots, n-1 \tag{29}$$

Thus, if all the points (ξ_k, η_k) are graphically represented, a straight line is obtained. This equation will be used in order to compute the auxiliary parameters a and b for the culture growth by means of a linear least squares approximation, and by taking into account the covariance matrix of errors. The procedure can be briefly described as follows. The OD measurements y_k are obtained with independent errors Δy_k, with zero means and common variance σ_{OD}^2. The errors of z_k are evaluated according to the linear approximation from Eq. 28, $\Delta z_k \simeq -\Delta y_k / y_k^2$, so that the accuracy of z_k is given by the variance

$$V(\Delta z_k) = z_k^4\ \sigma_{OD}^2 \tag{30}$$

experiment	**E42**				**E40**			
stage	*initial*		*induction*		*initial*		*induction*	
a	0.373	$\pm$ 0.036	0.189	$\pm$ 0.046	0.725	$\pm$ 0.024	0.525	$\pm$ 0.011
b	0.279	$\pm$ 0.029	0.295	$\pm$ 0.024	0.089	$\pm$ 0.046	0.139	$\pm$ 0.005
$cov(a,b)$	-0.0009		-0.0011		-0.0011		-0.00005	
A	0.987	$\pm$ 0.095	1.665	$\pm$ 0.244	0.643	$\pm$ 0.066	1.290	$\pm$ 0.042
B	-0.438	$\pm$ 0.062	-0.605	$\pm$ 0.103	-0.208	$\pm$ 0.1114	-0.377	$\pm$ 0.017
$cov(A,B)$	-0.0057		-0.0249		-0.0071		-0.0007	
s_{OD}	0.09		0.25		0.03		0.13	

Table 2: Culture growth parameters, estimated errors and covariances of the estimates, from Eq. 38 and Eq. 8, with weighted RMS error s_{OD} for culture OD from Eq. 42. The parameters a and b are related to the straight lines of Figures 1 and 2.

Thus Eq. 29, a system of $n-1$ equations may be explicitly written with the corresponding errors as follows

$$z_k = a\, z_{k-1} + b + \delta_k; \quad \delta_k = a\, \Delta z_{k-1} - \Delta z_k; \quad k = 1, \ldots, n-1 \tag{31}$$

Hence the error associated with each equation depends on the unknown parameter a and, taking into account Eq. 30 and Eq. 31, the errors of Eq. 31 system are correlated according to the following covariance matrix of error vector $\vec{\delta}$,

$$E(\vec{\delta}\,\vec{\delta}^{\,t}) = \sigma_{OD}^2 \mathbf{V}_z = \sigma_{OD}^2 \begin{pmatrix} z_1^4 + a^2 z_0^4 & -a z_1^4 & 0 & \ldots & 0 \\ -a z_1^4 & z_2^4 + a^2 z_1^4 & -a z_2^4 & \ldots & \vdots \\ 0 & -a z_2^4 & \ldots & & 0 \\ \vdots & & & \ddots & -a z_{n-2}^4 \\ 0 & \ldots & 0 & -a z_{n-2}^4 & z_{n-1}^4 + a^2 z_{n-2}^4 \end{pmatrix} \tag{32}$$

where E denotes the expectation and the symbol t means transpose.
It is well known that the value σ_{OD}^2 is not necessary in order to estimate the parameters a and b. However, since the matrix $\mathbf{V}_z$ depends on a, the estimation must be done iteratively, by revaluating $\mathbf{V}_z$ in each step, and by assuming the initial covariance matrix as the identity matrix. On the other hand, σ_{OD}^2 must be known in order to evaluate the covariance matrix of estimators $\mathbf{V}_{(a,b)}$, then an unbiased estimator s_{OD}^2 of σ_{OD}^2 (Stuart & Ord, 1991, pp.723 and 737) is given by the following inner product, where $\vec{\eta}$ is the vector of experimental values, and $\vec{\eta}^{\,*}$ is the predicted values vector,

$$s_{OD}^2 = \frac{1}{n-3} (\vec{\eta} - a\vec{\eta}^{\,*})^t\, \mathbf{V}_z^{-1} (\vec{\eta} - a\vec{\eta}^{\,*}) \tag{33}$$

Note that this is equivalent to evaluate a weighted root mean square (RMS) error with respect to the inverted covariance matrix, over the number of degrees of freedom, $n-3$, of the problem. Finally, the covariance matrix of parameters A and B, $\mathbf{V}_{(A,B)}$, is obtained by error propagation approximation from the jacobian matrix $\mathbf{J}$, and the covariance matrix $\mathbf{V}_{(a,b)}$ (see e.g. Barlow, 1989)

$$\mathbf{V}_{(A,B)} = \mathbf{J}\, \mathbf{V}_{(a,b)}\, \mathbf{J}^t; \quad \mathbf{J} = \frac{\partial(A,B)}{\partial(a,b)} \tag{34}$$

The estimates and corresponding errors of a and b, as well as of the parameters A and B , are listed in Table 2. Also the weighted RMS error s_{OD}^2 is given for OD estimations. Note that least squares approximation provides us with unbiased estimators of the parameters and of sampling variances and covariances of the estimators, without assumptions concerning the forms of the error distribution (Stuart & Ord, 1991, p.716). This is only necessary when testing hypotheses about the

parameters are discussed. In the Figure **??** the culture growth before and during the induction stage, from Eq. 29, is represented for both experiments **E40** and **E42**.
On the other hand, when the production kinetics are studied for different induction sequences from Eq. 24, Eq. 25 and Eq. 26, it is worth noticing that the initial value y_0 is a very small quantity with a great uncertainty. Hence this constant can not be fixed from the in situ measurement, but it may be estimated from the curve of the culture growth, Eq. 25, whose parameters A_0 and B_0 have been previously computed. The way to do this is also by another linear least squares approximation from Eq. 25, expressed in the form of Eq. 27, for different OD values in the initial stage, according to,

$$\frac{1}{y(t)} = e^{-A_0 t}\,\frac{1}{y_0} - \frac{B_0}{A_0}(1 - e^{-A_0 t}) \tag{35}$$

Note that in this expression, after the estimation of A_0 and B_0, the only unknown is $\frac{1}{y_0}$.

7 Production parameters

In this section the coefficients of the Eq. 10 will be computed. The parameters c_1 and c_2 depend on the experimental conditions of the induction stage, and they are specific of a given culture, but the parameter c_0 must be interpreted as the initial OD of culture at the beginning of the corresponding induction sequence. For this reason, we had written c_0^t in Eq. 6, by referring this parameter to the t-induction sequence.
The set of pairs (y, β), corresponding to OD of culture and amount of protein respectively, could be fitted for every sequence of sub-culture subjected to induction from different initial conditions in consequent stages of the culture growth. Then, some similar values of c_1 and c_2 would have to be obtained for all the sequences, but different estimations of c_0^t in each induction sequence. However the uncertainty when measuring the OD of culture and the amount of β-gal is quite notorious and, on the other hand, the number of samples in each induction sequence is small due to methodological reasons. Therefore, if the estimation is carried out for each individual sequence, the parameters from each sequence are poorly estimated. In order to avoid this problem we propose a new scheme of data, with a special strategy for the algorithm of computing the least squares solution. The whole set of induction sequences must be joined in a global set of data, and they must be treated like a single experiment, although the parameter c_0^t has different values in each sequence alone. Thus a set of new auxiliary variables, the components of an indicator vector, must be added to the pair (y, β) in order to label them according to their own induction sequence. Then, if the experiment is composed of n induction sequences, with initial induction times t_k $(k = 1 \div n)$, the valuation of the indicator vector components, that will be noted as u_k, is as follows,

$$u_k = \begin{cases} 1 & \text{for the } k\text{-th sequence} \\ 0 & \text{otherwise} \end{cases} \tag{36}$$

Then, for any induction sequence, the expression to be fitted in order to estimate the production parameters becomes

$$\beta = c_2 y^2 + c_1 y + \sum_{k=1}^{n} c_0^k u_k \tag{37}$$

where, to simplify the notation, c_0^k denotes the parameter c_0 corresponding to the sequence induced at a time t_k.
The total number of parameters to be determined is $n + 2$, which are involved in an overdetermined system of equations, according to the Eq. 37 evaluated for each sample of culture in the induction stage. The respective data for each equation of the system are represented in the following scheme,

where the sub-index indicates the induction sequence, and the super-index the sample number in the corresponding sequence:

Seq.	β	$\mathbf{y}$	$\mathbf{u}_1$	$\mathbf{u}_2$	$\cdots\cdots$	$\mathbf{u}_n$
	β_1^1	y_1^1	1	0	$\cdots\cdots$	0
1	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
	$\beta_1^{m_1}$	$\mathrm{y}_1^{m_1}$	1	0	$\cdots\cdots$	0
	β_2^1	y_2^1	0	1	$\cdots\cdots$	0
2	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
	$\beta_2^{m_2}$	$\mathrm{y}_2^{m_2}$	0	1	$\cdots\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
	β_n^1	y_n^1	0	0	$\cdots\cdots$	1
n	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
	$\beta_n^{m_n}$	$\mathrm{y}_n^{m_n}$	0	0	$\cdots\cdots$	1

For both experiments, the least squares estimators of above parameters, with the respective errors, are shown in the Table 3. Notice that in the linear model represented by Eq. 37, the experimental values of β are the observations, and their errors can be also assumed with zero means, uncorrelated, and with the same variance σ_β^2 for each independent measurement. Moreover, in this case, due to the central limit theorem, the errors can be assumed normal distributed. Thus it is also possible to give an unbiased estimator of σ_β^2 (Stuart & Ord, 1991, p.715) according to the following expression

$$s_\beta^2 = \frac{1}{N-n-2}\sum_{i=1}^{N}(\beta_i - \beta_i^*)^2 \tag{38}$$

where N represents the total number of samples ($N-n-2$ is the number of degrees of freedom), β_i is the experimental measurement, and β_i^* is the value provided by the model.
The significance of the resulting parameters can be analyzed if Eq. 10 is rewritten as follows,

$$\beta = [c_1 + c_2(y + y_0)](y - y_0) \tag{39}$$

where y_0 is the initial OD of culture for a given induction sequence. Thus, the contribution to the total β value from the terms c_1 and $c_2(y + y_0)$ can be evaluated for all the induction sequences. If this comparison criterion is used, for the experiment **E42** we obtain that the contribution of the c_1 term to β is not less than the 96%. Hence, in this case the value of c_2 is virtually zero. However, for the experiment **E40**, the contribution to β of the term containing c_2 can reach the 58% for the last induction sequences, hence both parameters are totally significant. The graphics corresponding to these regression curves are displayed in the Figure **??** for the respective experiments. Furthermore, the experimental values of the parameters $l/2$, the OD corresponding to the age of maximum growth velocity of culture, and ϵ, the OD shift for the maximum production velocity, are compared in Table 1. Finally the global evolution of recombinant protein production provided by the present model can be described from the production kinetic curves in Figures **??**, and **??**.

	E42		E40	
c_2	47	$\pm$ 333	179	$\pm$ 23
c_1	6194	1132	627	76
c_0^1	-78	810	56	74
c_0^2	-703	870	-36	74
c_0^3	-2720	906	-214	76
c_0^4	-4529	977	-190	77
c_0^5	-7367	1035	-146	79
c_0^6	-10815	1098	-184	82
c_0^7	-12372	1149	-299	84
c_0^8	-13189	1158	-427	85
c_0^9	-	-	-376	91
c_0^{10}	-	-	-661	93
c_0^{11}	-	-	-666	90
c_0^{12}	-	-	-1024	91
s_β	1284		193	

Table 3: Estimated parameters for the production model $\beta(y)$, Eq. 46, and estimation of the standard deviation s_β from the observations, Eq. 47.

8 Discussion

The β-galactosidase production process in a CI857-based recombinant system is modeled by studying the relationship between the velocities of product synthesis and of culture growth. In order to explain this relationship, a first order model is assumed. The model provides us with an explanation of the asymmetry between the evolution of product and the OD of culture during the induction stage. An important consequence is that the synthesis of recombinant β-galactosidase is a quadratic function of the cell density of the culture. This non linear dependency is made evident, in particular, from the experiment **E40** where the highest production velocity is reached at an age near to the stationary phase of the culture, that is, when the biomass is close to its maximum value. Then, the production increases at a faster velocity than biomass. Notice that the production is always associated with the culture growth so that, if the culture induction is made under conditions allowing a further culture growth, then the production kinetics never vanishes. Hence, the protein production in the stationary growing stage is possible only if the asymptotic biomass in the induction stage is greater than the highest cell density allowed by the initial stage. This situation is represented in the graphics of Figure **??**.

Also, the Figure **??** shows that at a fixed age x of the induction stage, the amount of product yield varies depending on the initial OD of culture at the beginning of the induction sequence. Then, on a section of constant x, it is possible to calculate the initial OD that produces the maximal β-gal for that time. This result is displayed in the Figures **??** and **??** for both experiments. We can also affirm, from Figure **??**, that the maximal production is obtained for cultures that have been induced during the phase of exponential growth, and for these induction sequences, the less the initial OD of culture, the greater the maximal protein production. However these high yield levels are reached after a long time of induction. For this reason it is necessary to compare the production of the first sample of induced culture with the curve of maximal productions, Figures **??** and **??**, where it is shown that the absolute maximum of product always corresponds to the initial induction sequence. Nevertheless, an interesting result, from an experimental point of view, could be the following one: if only a production level of 75% referred to the absolute maximum have to be obtained, for example for the experiment **E42**, the culture with an initial OD of 0.4 must be induced only for a period of 2 hours, while the first sequence needs to be induced about 4 hours in order to get the same production level. If the same comparison is done with the experiment **E40**, that production level is obtained

after 3 hours of induction beginning with an OD of 0.75, while the initial sequence needs more than 6 hours.
Thus, the modeling presented in this work, in order to compute the growing and production parameters of induced *E. coli* recombinant cultures, suggest a new data scheme, that globalize the data coming from several runs of the experimental process. In particular, for the growth parameters, a nonlinear estimation problem is lead to a linear least squares estimation. For recombinant gene expression two kind of parameters are considered: The 'static' parameters of the model, which can be seen as constraints for the fitting method, and the 'dynamic' parameters, which are specific of each sequenced sub-culture. Therefore, for this kind of experimental process, of gene expression in recombinant bacterial cultures, it is not recommended to use data from induced culture sequences in an independent way, since it leads to low precision results. Our proposal is to batch all the data referred to the same experiment in which various aliquots of a single initial culture are induced sequentially, and to run only one time a constrained least squares method, designed specifically for this experiment. By this way the errors of the estimated parameters are significatively reduced.

References

Barlow, R. 1989. In: Statistics, a guide to the use of statistical methods in the physical sciences. John Wiley and Sons Ltd., Chichester.

Corchero, J.L.; Vila, P.; Cubarsi, R.; Villaverde, A. 1994. Production of thermally induced recombinant proteins relative to cell biomass is influenced by cell density in *Escherichia coli* batch cultures. Biotechnology Letters, 16: 777-782.

Cubarsi, R.; Corchero, J.L.; Vila, P.; R.; Villaverde, A. 1998. Numerical techniques and mathematical modelling for CI857-controlled gene expression and cell growth in recombinant *E.* coli. IMA Journal of Mathematics Applied in Medicine and Biology, 15, 257-278.

Flickinger, M.C.; Rouse, M.P. 1993. Sustaining protein synthesis in the absence of rapid cell division: an investigation of plasmid-encode protein expression in *Escherichia coli* during very slow growth. Biotechnology Progres, 9, 555-572.

Hirsch, M.W.; Smale, S. 1974. In: Differential Equations, Dynamical Systems, and Linear Algebra. Academic Press, Inc., London.

Stuart, A.; Ord, J. K. 1991. In: Kendall's advanced theory of statistics: Classical inference and relationship (Vol. 2). Edward Arnold, London.

Villaverde, A; Benito, A.; Viaplana, E.; Cubarsi, R. 1993. Fine regulation of CI857-controlled gene expression in continuous culture of recombinant *Escherichia coli* by temperature. Appl. Environ. Microbiol., 59, 3485-3487.

time (min)	0	60	120	180	240	300
OD	0.05	0.27	1.04	2.12	2.85	-
β-gal	0	750	4520	12125	22090	-
OD	0.13	0.56	1.60	2.30	2.94	-
β-gal	0	1500	9280	15655	17430	-
OD	0.30	0.90	2.06	2.73	3.03	-
β-gal	0	2706	8999	14279	17318	-
OD	0.50	1.38	2.34	3.12	-	-
β-gal	0	4880	9750	13520	-	-
OD	0.88	2.07	2.70	3.00	-	-
β-gal	0	5560	8610	10950	-	-
OD	1.66	2.01	2.27	2.32	2.37	2.39
β-gal	0	1466	3106	3606	4160	4710
OD	1.92	2.21	2.38	2.56	-	-
β-gal	0	1490	2820	3335	-	-
OD	2.11	2.27	2.34	2.37	-	-
β-gal	0	823	1656	2017	-	-

Table 4: Actual data for the eight runs of experiment **E42.**

time (min)	0	30	60	90	120	150	180	210	240
OD	0.04	0.07	0.13	0.25	0.38	0.60	0.99	1.47	-
β-gal	0	6	42	102	225	468	835	1920	-
OD	0.06	0.10	0.18	0.31	0.49	0.86	1.49	1.90	2.4
β-gal	0	0	54	66	333	503	953	2248	2675
OD	0.08	0.15	0.27	0.42	0.76	1.27	1.80	2.35	2.45
β-gal	0	25	69	128	247	588	1830	1935	2340
OD	0.11	0.19	0.32	0.61	0.88	1.50	2.18	2.50	2.81
β-gal	0	0	23	228	462	940	2110	2150	3370
OD	0.16	0.28	0.44	0.62	1.03	1.80	2.38	2.57	2.79
β-gal	0	40	122	195	545	1543	2355	2720	3215
OD	0.22	0.37	0.52	0.72	1.24	2.23	2.26	2.74	-
β-gal	0	0	196	408	748	2013	2210	3013	-
OD	0.30	0.44	0.63	0.98	2.00	2.25	2.55	2.85	-
β-gal	0	56	140	445	1623	2180	2520	2705	-
OD	0.39	0.54	0.88	1.68	2.12	2.41	2.74	3.44	-
β-gal	0	64	320	1265	1838	1985	2390	3655	-
OD	0.48	0.68	1.23	1.95	2.29	2.73	3.04	-	-
β-gal	0	125	609	1545	2018	2575	3280	-	-
OD	0.60	1.05	1.73	2.17	2.55	2.91	2.92	-	-
β-gal	0	196	675	1657	2330	2465	2635	-	-
OD	0.88	1.16	2.05	2.52	2.69	2.91	3.33	3.46	-
β-gal	0	49	1021	1843	2398	2865	3620	3910	-
OD	1.03	1.82	2.44	2.57	2.83	3.05	3.36	3.71	-
β-gal	0	730	1389	1692	2298	3000	3165	3215	-

Table 5: Actual data for the twelve runs of experiment **E40.**