Bioprocess Biosyst Eng (2008) 31:41–46
DOI 10.1007/s00449-007-0161-9

ORIGINAL PAPER

Product formation kinetics in genetically modified *E. coli* bacteria: inclusion body formation

Stefan Gnoth · Marco Jenzsch · Rimvydas Simutis · Andreas Lübbert

Received: 8 May 2007 / Accepted: 7 September 2007 / Published online: 10 October 2007

Abstract A data-driven model is presented that can serve two important purposes. First, the specific growth rate and the specific product formation rate are determined as a function of time and thus the dependency of the specific product formation rate from the specific biomass growth rate. The results appear in form of trained artificial neural networks from which concrete values can easily be computed. The second purpose is using these results for online estimation of current values for the most important state variables of the fermentation process. One only needs online data of the total carbon dioxide production rate (tCPR) produced and an initial value *x* of the biomass, i.e., the size of the inoculum, for model evaluation. Hence, given the inoculum size and online values of tCPR, the model can directly be employed as a softsensor for the actual value of the biomass, the product mass as well as the specific biomass growth rate and the specific product formation rate. In this paper the method is applied to fermentation experiments on the laboratory scale with an *E. coli* strain producing a recombinant protein that appears in form of inclusion bodies within the cells' cytoplasm.

Keywords Fermentation process · Protein formation kinetics · Neural networks · Soft-sensing · Online-predicion

S. Gnoth · M. Jenzsch · A. Lübbert (✉)
Center of Bioprocess Engineering,
Martin-Luther-University Halle-Wittenberg,
06120 Halle/Saale, Germany
e-mail: andreas.luebbert@biochemtech.uni-halle.de

R. Simutis
Institute of Automation and Control Systems,
Kaunas University of Technology, 3028 Kaunas, Lithuania

Introduction

E. coli is the most important host cell system for recombinant protein production systems if the desired products do not need posttranslational modifications to obtain efficacy [9]. In many practical cases, the heterologous products appear in form of inclusion bodies within these bacterial cells. Then, several downstream processing steps including cell disruption followed by solubilization and refolding are necessary before clinical efficacy of the protein is achieved.

In order to obtain a high product titer in the fermenter, the process operational procedure must be optimized for high cell density, i.e., high biomass concentration X, and, at the same time high specific product formation rates π. The latter can only be adjusted to their optimal values if the relationship between π and the variables that can be adjusted during the cultivation process is known. Such relationships are not well investigated, and thus, only very rough estimates can be found in literature. Usually it is simply assumed that the specific product formation rate is in a fixed stoichiometric relationship to the specific biomass growth rate $\mu: \ \pi = Y_{p_x}\mu$, or the specific substrate consumption rate. Only a few groups developed more complex kinetic expressions for product formation (e.g., [4]).

Strict mechanistic approaches are extremely difficult to quantify as the anabolic metabolism of the cells is rather complex and not yet completely understood. Anyway, for process control purposes, it is straightforward to look for correlations, preferably of variables that are online accessible during industrial fermentation runs. Such data driven approaches can be very well performing provided the right variables are chosen and the relations are identified using many data sets from the single process under consideration.

The approach used here is based on artificial neural networks that are trained on an extended set of data records. These networks are known to depict very good mapping properties for complicated nonlinear relationships (e.g., [2]).

Data driven approach to product formation kinetics

A schematic view of the data-driven approach used here is shown in Fig. 1 [1]. It is based on two simple feedforward artificial neural networks (ANNs). The first one determines the specific biomass growth rate μ from online measured carbon dioxide production rate (tCPR) data as well as t_{ai}, the time after induction. The time signal t_{ai} is zero before induction and increases continuously thereafter. Acting as a switch this information first signals the induction point to the network model and then delivers the current process time axis after induction. The specific growth rate μ determined by the identified inputs is used in an ordinary differential equation, i.e., a simple process model, to determine the biomass x, which is then fed back onto the input layer of the ANN. Thus, the model is a hybrid one, where the kinetics, represented by ANNs are combined with dynamic mass balances [5].

The estimate of μ from this ANN is then used as an input to the second ANN computing the specific product formation rate π. A further input to this second ANN is again the time after induction, t_{ai}. Additionally, the third input is the specific protein concentration $p_x = p/x$ (where x is total biomass, and p the total product mass). p is obtained from x and π by solving another simple mass balance equation shown in the Fig. 1.

As there are no direct measurements available for μ and π, the networks must be trained using offline-measured biomass x and product mass p data. This can be done with the sensitivity equation technique discussed by Simutis and Lübbert [7] and Gnoth et al. [1].

Experimental

Experiments were performed with genetically modified *E. coli* bacteria that are able to produce the commercially interesting gastric inhibitory polypeptide GIP [3]. The desired product in the process reported about here is in form of inclusion bodies. All experiments used *E. coli* BL21(DE3) as the host cell. The target protein was coded on the plasmid pET 28a and expressed under the control of the T7 promoter after induction with isopropyl-thiogalactopyranosid (1 mM IPTG). The strain was resistant against kanamycin. The product appears as inclusion bodies within the cytoplasm. The particular strain used did not produce notable amounts of acetate (data not shown) under the cultivation conditions adjusted in the experiments reported.

All the experiments were performed within BIOSTAT C 15-L-bioreactor (BBI Sartorius) operated at maximal 8-L volume. The fermenter was equipped with three standard six-blade Rushton turbines that could be run up to 1,400 rpm. The aeration rate could be increased up to 24 sLpm. Aeration rate and then stirrer speed were increased one after the other in order to keep the dissolved oxygen concentration at 25% saturation.

The fermentations were operated in the fed-batch mode immediately after inoculation. The initial volume was 5-L. Temperature and pH were adjusted to 35 °C and 7, respectively. The main C- and energy source, glucose, was fed at a concentration of 300 and 600 g/kg. For more details about the medium the reader is conferred to Jenzsch et al. [3].

All fermentations were started during night by automatic transfer of the inoculum from a refrigerator into the reactor. Substrate feeding was started in an open loop fashion with predefined exponential profiles. When, after some cultivation time, the signal to noise ratio of the offgas data reached a predefined level, closed loop control was started with the total biomass or the total carbon dioxide produced as control variables. Additionally experiments were performed under unlimited conditions.

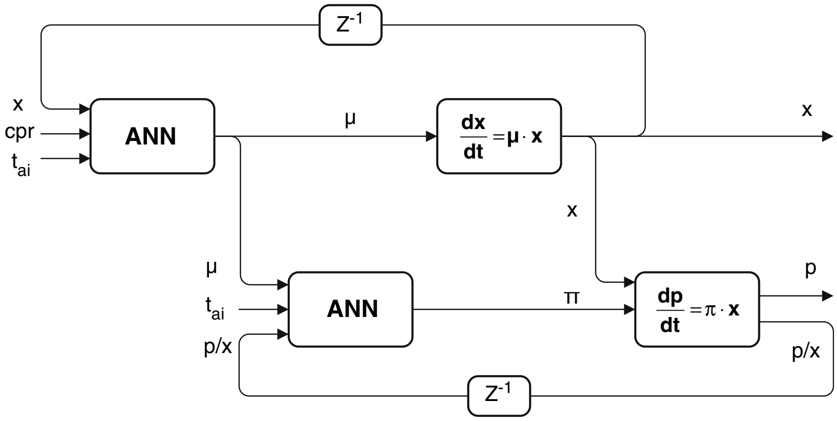


Fig. 1 Scheme of the ANN-based approach of deriving the $\pi(\mu)$ relationship as well as biomass x and product mass p

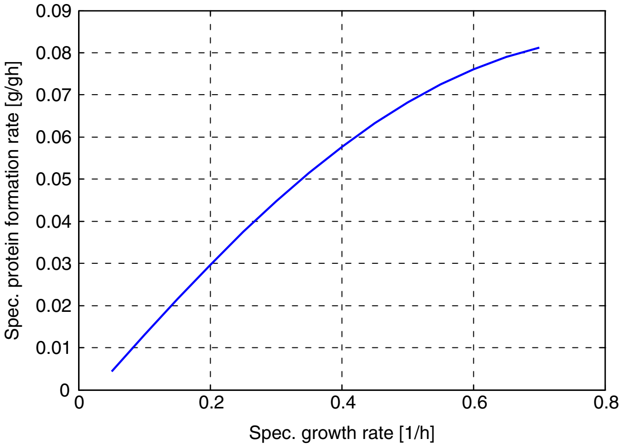


Fig. 2 $\pi(\mu)$ relationship derived from the network system depicted in Fig. 1 for fixed t_{ai} = 1 (h) and p/x = 0.01 (g/g)

CO_2 in the vent line was measured with MAIHAK's Unor 610, O_2 with MAIHAK's Oxor 610. The total ammonia consumption during pH control was recorded by means of a balance beneath the base reservoir. These three quantities were measured online.

Biomass concentrations were measured offline via optical density at 600 nm with a Shimadzu photo-spectrometer (UV-2102PC). Glucose was determined enzymatically with a YSI 2700 Select Bioanalyzer. The product was measured with SDS PAGE after separation of the inclusion bodies and their solubilization.

Results

Forty-nine data sets from the *E. coli* fermentations described above were used for training and validation (cross-validation procedure) of the hybrid model depicted in Fig. 1. These fermentations were performed under very different conditions. Some of them were controlled to small specific growth rates in the order μ = 0.1 (1/h). Others were run in an unlimited way with respect to the substrate concentration *S*. Furthermore, some runs were controlled to fairly high specific growth rates in the beginning of the product formation phase.

Figure 2 depicts the result of the training of the network system illustrated in Fig. 1 with respect to the simple $\pi(\mu)$ relationship.

In order to assure that the data-driven model is truly mapping the biomass growth and product formation kinetics, the model solutions were compared to the corresponding experimental data. For this purpose, the cross-validation technique was employed. The typical results shown in Fig. 3 are from experiments, the data of which were not used during the network training.

Note that the model depicted in Fig. 1 only needs the online available tCPR signal, the initial biomass *x* as well as the time t_{ai} at which the culture is induced. The actual values of total biomass *x* and total product concentration *p* then appear as model outputs together with the specific growth and product formation rates at each time where a new tCPR value becomes available. Hence, the method can

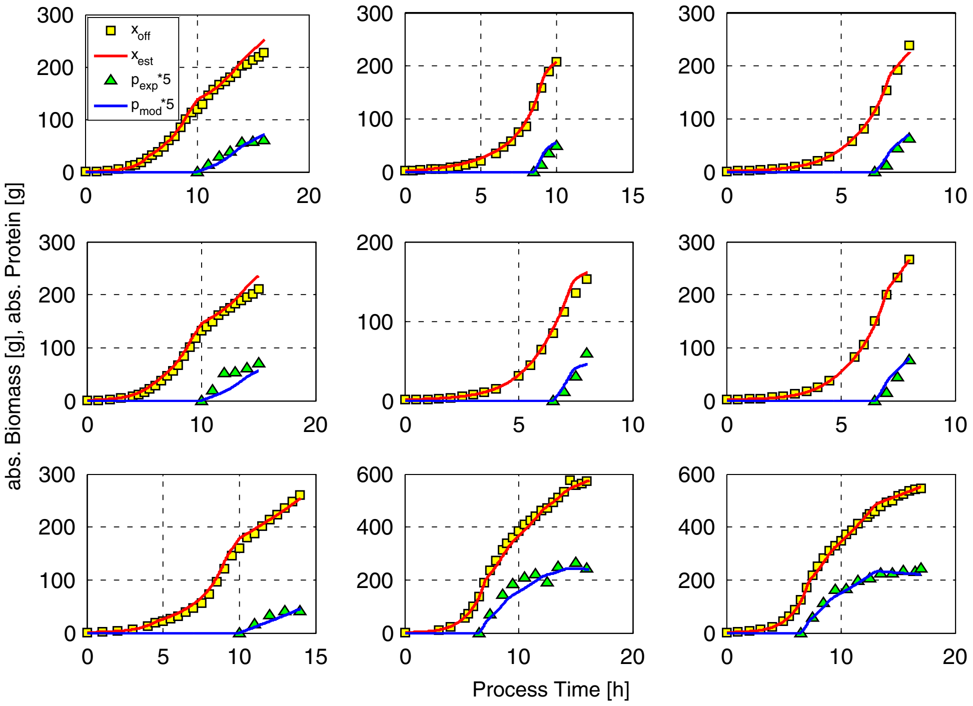


Fig. 3 Typical example for simulations using the data-based kinetics. The biomass and product mass profiles are shown together with the corresponding offline measured data (*symbols*). These data were not used during network training, thus, the comparison in the plot can be considered as a model's validation procedure

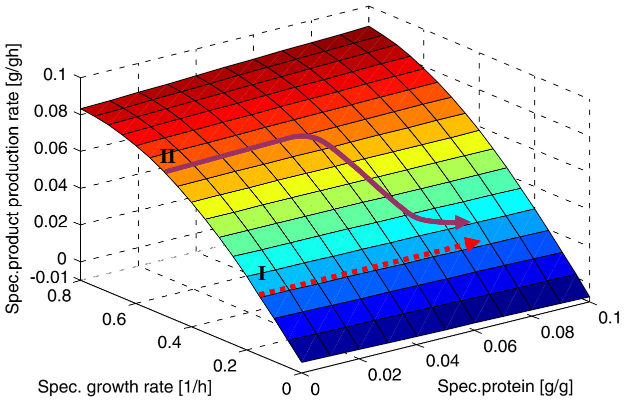


Fig. 4 Specific product formation rate as a function of μ, the specific biomass growth rate and p_x, the specific protein concentration. In general, two types of $\mu - \pi - p/x$ dependencies were obtained. Type I (*dashed line*) shows constant relationship if the fermentations run under limited conditions below the critical specific growth-rate. Type II depicts the relationship under maximum growth conditions, i.e., the increasing protein-load on the cells reduces the achievable maximum growth-rate after induction. *Arrows* indicate the process evolution with increasing process time

be used as a soft sensor for x, p, μ, and π. In Fig. 3 the lines depict the outputs of this soft sensor for x and p, the symbols show the corresponding offline measurement values which, in the case of the protein data, are available only days after the fermentation had been finished. It is worthwhile to mention, that the online measured variable tCPR and the information t_{ai} about the induction state are sufficient for an online adaptation of the ANN-based model to the current state of the process, particularly in the product formation phase. With this online information, biomass and product mass as well as the corresponding specific formation rates μ and π can quite accurately be estimated. The full lines in Fig. 3 corresponding to the biomass and the product mass shown confirm this.

Heterologous protein formation is usually accompanied by a metabolic load of the cells. In order to quantify this, the influence of the specific protein concentration p_x, i.e., the protein load of the cells, on π was examined as well. The resulting three-dimensional graph is depicted in Fig. 4.

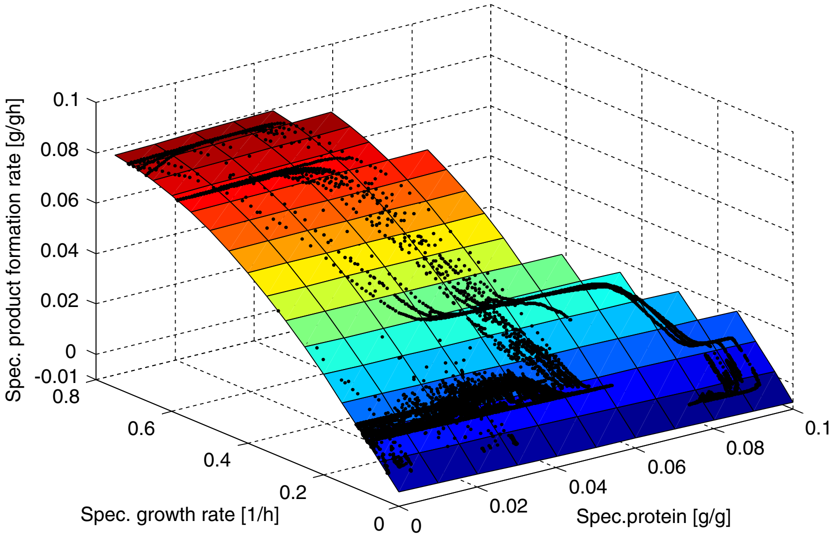
The specific product concentration p_x is seen not influencing the $\pi(\mu)$ relationship significantly. At a given μ, π is only slightly decreasing with the accumulation of the inclusion body protein within the cytoplasm of the cells. However, there is a significant influence of p_x, the protein load of the cells, on the specific growth rate μ, which is referred to in literature as a metabolic burden.

To further clarify this, the π values, obtained from experimental data were plotted into the $\pi(\mu,p_x)$-surface depicted in Fig. 4. Figure 5 shows the data points for all 49 experiments. The values practically remain on the surface. They clearly show which part of the (μ p_x)-space has been explored during the fermentations performed. With higher p_x values smaller and smaller μ values were obtained, even when the culture does not run in a substrate-limited way. Thus, only a part of the surface depicted in the model (Fig. 4) is accessible during the process.

The monotonic increase of the specific product formation rate with the specific biomass growth rate μ leads to the consequence that the biomass growth rate must be kept as high as possible in order to obtain a maximal product titer at the end of the cultivation.

These results were successfully tested in some spot checks. For this purpose additional validation experiments were performed. In the first one the culture was grown at its maximal specific growth rate after induction, in the second one, a small specific growth rate which was

Fig. 5 Comparison of the $\pi(\mu,p_x)$ relationship derived from the ANN-based model with π values, obtained from experimental data (*symbols*). The data on the plane show the size of the design space, i.e., the about the range of values that are possible at all in the (μp_x)-space



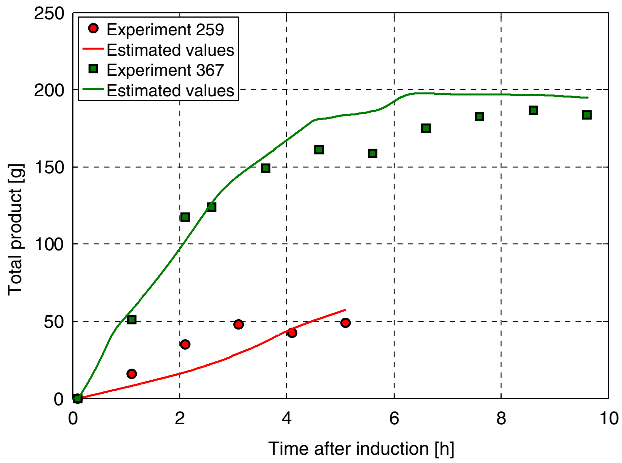


Fig. 6 Validation tests of the results depicted in Fig. 2 for two cases: one cultivation (Exp.367) was operated with substrate concentration well above values where substrate limitation occurs and one (Exp.259) was operated under substrate limitation. Total biomass was similar in both experiments by induction time. The *curves* present the total product obtained with data-driven model and the points are experimental values

suboptimal for the expression of inclusion body proteins was chosen. Both resulting protein formation patterns are depicted in Fig. 6.

The product mass profiles p are very different and show the behavior expected from the monotonically increasing π with increasing μ.

Discussion

The most important point to note for the protein investigated here, which is packed in form of inclusion bodies within the cells, is, that the $\pi(\mu)$ relationship is a simple monotonic function of essentially the specific biomass growth rate μ only. The metabolic load of the cell, resulting from the product accumulation within its cytoplasm and characterized by the specific protein concentration p_x, influences the maximal specific biomass growth $\mu_{\max}$, but not directly the specific product formation rate π. In other words, the specific product formation rate π is only dependent on the specific growth rate μ as assumed by many researches in bioprocess engineering (e.g., [6, 8, 10]. However, μ cannot be freely adjusted as its maximally attainable value $\mu_{\max}$ decreases with the specific product concentration p_x. Thus, in this particular system, the influence of the cell's internal protein concentration or accumulation on the cell's protein formation performance π is an indirect one.

The $\pi(\mu)$-relationship of the strain investigated here is qualitatively different from the one of other *E. coli* systems, e.g., one where the product appears in a soluble active form. For a strain expressing the soluble green fluorescence protein (GFP), Gnoth et al. [1] found that the $\pi(\mu)$ relationship depicts a maximum at a rather low specific biomass growth rate of about $\mu = 0.14$ (1/h). Other strains possibly show further forms of the $\pi(\mu)$ relationship. Hence, it is straightforward to look for a well-performing technique that allows determining the product kinetics without the assumption of unproven models or constraints on the cell metabolism. Such a method is presented in this paper. It is not restricted to a special product and works without any assumption about kinetic parameters. In so far it is suitable for any kind of expressing protein (e.g., soluble/insoluble).

As the approach proposed here is a purely data-driven approach, relatively many data records are required to train the networks and to validate the results. In the beginning of the developments with a new biological system, having a few data records only, the prediction might not be sufficiently good. This could be considered a disadvantage of the proposed method. However, this approach has the advantage of excellent learning abilities. After each cultivation the networks can automatically be retrained using the extended database. In this way the software learns without much additional efforts of the plant personnel.

Experiments performed in much different ways as compared to the records used for network training, will also not be predicted sufficiently well. However, after adding the data records to the database, the automatic learning will quickly lead to a better model performance. Thus, in the following runs the model will be sufficiently accurate for state estimation.

Providing many data records might be a problem in small laboratories, but it is definitely not a problem in industrial production environments. The number of data records necessary to obtain reliable results is dependent on the quality of the data, particularly on the accuracy of the product concentration values. Typically, data from about ten experiments are needed. Again, this is not so much a problem in industrial environments, where the laboratories are usually very experienced in measuring the concentrations of their particular product.

One of the main advantages of the method summarized in Fig. 1 is that the evaluation of the identified model is extremely quick. It can thus perfectly be used for online model supported process monitoring and control purposes. One example is its use as a software sensor for x and p as well as for μ and π. For these quantities no physical sensors are available.

Acknowledgments The financial support from the "Ministry of Cultural Affairs" of Sachsen-Anhalt, Germany, is gratefully acknowledged.

References

1. Gnoth S, Jenzsch M, Simutis R, Lübbert A (2006) Product formation kinetics in a recombinant protein production process, CAB-2007, IFAC. Elsevier, Amsterdam (accepted)
2. Haykin S (1999) Neural networks: a comprehensive foundation, 2nd edn. Prentice Hall, Upper Saddle River
3. Jenzsch M, Gnoth S, Beck M, Kleinschmidt M, Simutis R, Lübbert A (2006) Open loop control of the biomass concentration within the growth phase of recombinant protein production processes. J Biotechnol 127:84–94
4. Levisauskas D, Galvanauskas V, Henrich S, Wilhelm K, Volk N, Lübbert A (2003) Model-based optimization of viral capsid protein production in fed-batch culture of recombinant *Escherichia coli*. Bioprocess Biosyst Eng 25:255–262
5. Schubert J, Simutis R, Dors M, Havlik I, Lübbert A (1994) Bioprocess optimization and control: application of hybrid modelling. J Biotechnol 35:51–68
6. Shioya S (1992) Optimization and control in fed-batch bioreactors. Adv Biochem Eng Biotechnol 46:1
7. Simutis R, Lübbert A (1997) Exploratory analysis of bioprocesses using artificial neural network-based methods. Biotechnol Prog 13(4):479–487
8. Soons ZITA, Voogt JA, van Straten G, van Boxtel AJB (2006) Constant specific growth rate in fed-batch cultivation of *Bordetella pertussis* using adaptive control. J Biotechnol 125:252–268
9. Walsh G (2006) Biopharmaceutical benchmarks 2006. Nat Biotechnol 24(7):769–776
10. Yoon SK, Kang WK, Park TH (1994) Fed-batch operation of recombinant *Escherichia coli* containing Trp promoter with controlled specific growth rate. Biotechnol Bioeng 43:995–999