Prediction of Dynamic Plasmid Production by Recombinant *Escherichia coli* Fed-Batch Cultivations with a Generalized Regression Neural Network

T. Silva,[a,*] P. Lima,[a] M. Roxo-Rosa,[a] S. Hageman,[b] L. P. Fonseca,[b] and C. R. C. Calado[a]

[a] Faculty of Engineering, Catholic University of Portugal, Estrada Octávio Pato, 2635-631 Rio de Mouro, Portugal

[b] Institute for Biotechnology and Bioengineering, Centro de Engenharia Biológica e Química, Instituto Superior Técnico, Av Rovisco Pais, 1049-001 Lisboa, Portugal

Original scientific paper
Received: May 25, 2009
Accepted: October 12, 2009

Dedicated to the memory of Professor Dr. Valentin Koloini

A generalized regression neural network with external feedback was used to predict plasmid production in a fed-batch cultivation of recombinant *Escherichia coli*. The neural network was built out of the experimental data obtained on a few cultivations, of which the general strategy was based on an initial batch phase followed by an exponential feeding phase. The different cultivation conditions used resulted in significant differences in bacterial growth and plasmid production. The obtained model allows estimation of the experimental outputs (biomass, glucose, acetate and plasmid) based on the bioreactor starting conditions and the following on-line inputs: feeding rate, dissolved oxygen concentration and bioreactor stirring speed. Therefore, the proposed methodology presents a quick, simple and reliable way to perform on-line feedback prediction of the dynamic behaviour of the complex plasmid production process, based on simple on-line input data obtained directly from the bioreactor control unit and with few cultivation experiments for neural network learning.

Key words:
Neural network, fed-batch cultivation, plasmid production

Introduction

Plasmids are highly desirable vectors for gene therapy and DNA vaccination, as they offer multiple advantages over viral vectors, including large packaging capacity, stability without integration and reduced toxicity. Furthermore, plasmid DNA can be delivered to many different tissues, using a variety of delivery techniques currently being developed.[1] Gene therapies are a promising alternative to classical medical techniques for prevention, treatment, diagnosis or cure of genetic defects such as cystic fibrosis or acquired diseases like cancer and AIDS. Vaccines can also be developed on the basis of genes to provide immunity against infectious agents (*e.g.* malaria) and tumours, presenting several advantages over traditional vaccines as whole attenuated/killed organism or protein-based vaccines. One relevant drawback of plasmids as vectors, is the low transfection efficiency observed *in vivo* and the consequent low expression level of the target gene, that results in the requirement of large amounts of plasmids per treatment. Although considerable attention has been paid to genetic engineering of the plasmid backbone, substantially less attention has been given to the practical challenges of producing large amounts of plasmids. Therefore, the development of methodologies to control and optimize plasmid production processes constitutes a challenge for the biotechnology industry.[2]

In relation to the cultivation strategies available, the main advantage of a batch mode is simplicity, as all the nutrients, as glucose, for cell growth and plasmid production are presented at the cultivation beginning. However, in the presence of high glucose concentrations, acetate is formed aerobically by the overflow metabolism. Acetic acid production usually inhibits the cell growth and decreases the recombinant product yield. In order to optimise the glucose concentration in the bioreactor, and consequently increase the volumetric productivity, a fed-batch strategy is usually applied. In fed-batch mode, glucose is added along a certain period. However, the optimal flow rate of the feeding medium containing glucose must be determined. A higher flow rate in relation to the optimal will result in overflow metabolism, and a lower flow rate will result in low productivities.[3,4,5] Traditional models based on reaction kinetics and/or reactor kinetics, qualitative simulation of metabolic networks and complex rate law models for entire

* Corresponding author (fax: +351 214 269 800; tsilva@fe.lisboa.ucp.pt)

metabolic pathways, are usually unsuitable for fed-batch modelling and optimization.[6,7] In part, these limitations result from the complexity and high non-linearity of the biological phenomena. Furthermore, to obtain enough data to construct a suitable model framework, usually a high quantity of information of the biological reactions is needed. However, industrial cultivation with recombinant microorganisms presents aseptic requirements that result in serious limitations on on-line and off-line measurements regarding the bioprocess. There is evidence that the ability of neural networks to model any kind of function, given enough learning examples and adequate network topology, lead to models that are superior to mechanistic models in the description of bioprocesses.[7,8,9]

Generalized regression neural network (GRNN) is proposed for non-linear correlation of the process output variables with the operating variables. The advantages of the GRNN-based models are (i) unlike partial least squares (PLS) these models can efficiently and simultaneously approximate non-linear multiinput–multioutput (MIMO) relationships and, (ii) these models can be developed in a significantly shorter time in comparison with the multilayered perceptron (MLP) or radial basis function network (RBFN)-based process models, since the training of the model, which is a one-step procedure, involves fixing a value of only a single free parameter.[10]

In this work, the development and implementation of a dynamic model, based on generalized regression neural networks, of the fed-batch cultivations of recombinant *E. coli* producing the plasmid pVAX-LacZ in complex media is performed. Plasmid production is closely related to variables such as glucose, acetate and biomass concentrations. However, on-line sensors to monitor these variables are not generally used, as are the pH and dissolved oxygen concentration sensors. Furthermore, at industrial scale, and to avoid contamination, the minimum set ups of sensors should be used. Thus, the present model includes a real-time estimation of those variables, based on the bioreactor initial conditions and the few on-line data such as the feeding rate, pH, dissolved oxygen concentration and the stirring speed.

Materials and methods

Bacterial strain and pre-culture

Escherichia coli DH5-α containing the plasmid pVAX-LacZ (Invitrogen, USA) was used. The stock cultures, grown on 2 % (*w/v*) Luria-broth (Sigma, UK) and 30 μg mL^{-1} kanamycin, were maintained in 40 % (*v/v*) glycerol with 10 mmol L^{-1} Tris-HCl buffer pH 8.0 at –80 °C. An aliquot of 10 μL of stock culture was inoculated into 1 L shake flask containing 300 mL with 20 g L^{-1} bactotryptone (BD), 10 g L^{-1} yeast extract (Difco, USA), 10 g L^{-1} sodium chloride (Merck, Germany) and 30 μg mL^{-1} kanamycin (Merck, Germany). The cotton-stopped flasks were incubated in an orbital shaker (Agitorb 160E, Aralab, Portugal) at 250 rpm and 37 °C for 16 h.

Cultivation

The pre-cultures were used for inoculating the culture at 10 % (*v/v*). The cultivations were performed in a 5 L bioreactor (Biostat MD, B. Braun, Germany) with a 3 L working volume, in absence of antibiotic. Cultivation was maintained at pH 7.0 by automatic control through 4 moL L^{-1} NaOH or 2 moL L^{-1} H_2SO_4 addition, and at 37 °C with a minimal dissolved oxygen concentration of 30 % of air saturation by automatic adjustment of the agitation rate by two Rushton turbines, with a constant air flow rate of 3.3 L min^{-1} during the batch phase and of 4.4 L min^{-1} during the feeding phase. The initial batch cultivation medium contained 10 g L^{-1} sodium chloride (Merck, Germany) and 10 g L^{-1} of yeast extract (Difco, USA), and different concentrations of bactotryptone (BD, UK) and D(+)-glucose (Merck, Germany) as shown in Table 1. After the

Table 1 – *Medium composition of four fed-batch cultivations, based on a batch phase followed by a feeding phase. The batch phase always contained 10 % (w/v) NaCl. The starting time of the feeding phase and the feeding rate is also indicated.*

Cultivation	Batch phase Medium composition:	Feeding phase Starting period: Medium composition:
1	Glucose: 10 g L^{-1} Yeast extract: 10 g L^{-1} Bactotryptone: 20 g L^{-1}	Started after 8 h of cultivation With: glucose: 90 g L^{-1} Yeast extract: 45 g L^{-1} Bactotryptone: 45 g L^{-1}
2	Glucose: 20 g L^{-1} Yeast extract: 10 g L^{-1} Bactotryptone: 20 g L^{-1}	Started after 21 h of cultivation With: glucose: 90 g L^{-1} Yeast extract: 45 g L^{-1} Bactotryptone: 45 g L^{-1}
3	Glucose: 8 g L^{-1} Yeast extract: 10 g L^{-1} Bactotryptone: 16 g L^{-1}	Started after 23 h of cultivation With: glucose: 53 g L^{-1} Yeast extract: 10 g L^{-1} Bactotryptone: 16 g L^{-1}
4	Glucose: 10 g L^{-1} Yeast extract: 10 g L^{-1} Bactotryptone: 20 g L^{-1}	Started after 16 h of cultivation With: glucose: 90 g L^{-1} Yeast extract: 45 g L^{-1} Bactotryptone: 45 g L^{-1}

Feeding rate (similar to all 4 cultivations):

0–30 min flowrate = 110 mL h^{-1}	30–60 min flowrate = 122 mL h^{-1}
60–90 min flowrate = 138 mL h^{-1}	90–120 min flowrate = 152 mL h^{-1}
120–150 min flowrate = 170 mL h^{-1}	150–180 min flowrate = 190 mL h^{-1}
180–210 min flowrate = 212 mL h^{-1}	210–240 min flowrate = 236 mL h^{-1}
240–270 min flowrate = 264 mL h^{-1}	270–300 min flowrate = 296 mL h^{-1}

batch-growth phase, the feeding rate was sequentially increased and started with mixtures of yeast extract, bactotryptone and D-glucose (Table 1).

Analysis of biomass, glucose, acetate and plasmid concentration

Biomass in units of dry cell weight per volume of culture medium (g L^{-1}) was determined by centrifugation of cultivation samples, washing the pellet with 0.9 % (*w/v*) sodium chloride and drying at 80 °C until constant weight.

Glucose and acetate were determined by HPLC with a L-6200 Intelligent Pump (Merck-Hitachi, UK), a L-7490 LaCrom-Ri-detector (Merck, Germany), a D-2500 Chromato-integrator (Merck-Hitachi, Germany) and with a HPLC column for culture broth monitoring from Bio-Rad at 50 °C, with H_2SO_4 at 0.6 mL min^{-1} as eluent.

Lysis and Primary Plasmid isolation: samples from cultivation were harvested in a centrifuge (Sigma-201) at 10 000 rpm for 15 min and the pellets were frozen for later use. The bacterial pellet was resuspended in TE buffer (10 mmol L^{-1} Tris-HCl buffer pH 8.0, and 1 mmol L^{-1} ethylene-diamine tetraacetic acid). In order to obtain a similar cellular quantity in all the performed cellular lysis, the volume of TE buffer used to resuspend the cellular pellet was estimated by dividing 6 by the absorbance at 600 nm of the cellular sample before centrifugation. The dilution degree was based on a previous experiment (not published) conducted in order to determine the effect of the dilution degree on the reproducibility of the alkaline cell lyses. An equal volume of lysis buffer (200 mmol L^{-1} NaOH, 1 % SDS) was added to 400 μL of the resuspended pellet and gently mixed by inverting five times the eppendorf, and then incubated at room temperature for 5 min. An aliquot of 325 μL of cold neutralisation buffer (3.0 mol L^{-1} potassium acetate buffer pH 5.5) was added and mixed by gently inverting five times, and then incubated on ice for 15 min. Lysate was clarified by two consecutive centrifugations at 15 000 rpm, 4 °C for 30 min and 15 min, respectively, with a Sigma –1K15 centrifuge.

The plasmid concentration and its purity degree was performed by hydrophobic interaction HPLC, using a Source™ 15PHE column (Amersham Biosciences) as described in Diogo *et al.*[11]

The model

Neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs. The neural network chosen for the model was the generalized regression neural network (GRNN).[12] This type of network has a radial basis function (RBF) layer and a special linear layer.

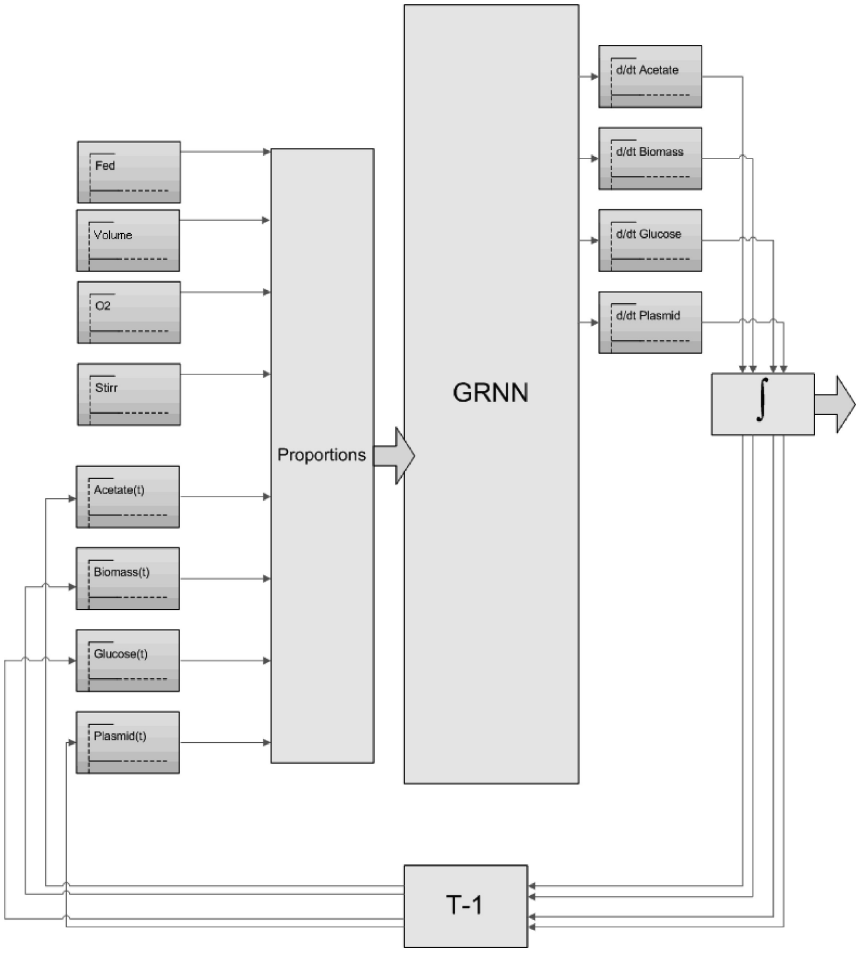
This network makes a single pass through a set of training instances and maps each instance to a neuron in the network. The parameter *S* defines the spread of the radial basis function. The first layer has as many neurons as there are training instances (input/target pairs). The first layer weights are set to the input data. The second layer weights are set to the target data, which is the desired output given the input.

GRNN training

The input/output pairs used for GRNN training were: Input: feeding rate, dissolved oxygen concentration (DO), stirring speed, biomass, glucose, acetate and plasmid at instant *t*; Output: variation of biomass, glucose, acetate and plasmid. For simulations, only initial conditions (of biomass, glucose, acetate and plasmid) and on-line data (of volume, DO, stirring speed and feeding rate) are provided to the network. Off-line inputs on training network were biomass, glucose, acetate and plasmid. Off-line inputs were not provided to the validation experiment: the network estimates this data. We used exponential feeding rate (Table 1). The on-line data (volume, DO and stirring speed) can be captured directly from the bioreactor unit control almost in real time.

In order to provide feedback predictive qualities, the models must have knowledge of previous states of the output variables, therefore these variables are on an external delay feedback loop, providing the model with previous outputs as new inputs. The externally recurrent neural network was used in a similar model[6] and was first developed by Nerrand.[13] It performed recursion over traditional Multi-Layer Perceptrons (MLP). However, MLP have a topological limitation: its number of hidden units needs to be explicitly given, so it is a parameter that is usually fine-tuned by experience. Fine-tuning network topology usually works reasonably well for systems with a limited number of variables, or whose behaviour is known to some reasonable extent: that is not the case of this problem. In order to avoid an extra parameter, with this approach we use a recurrent version of the generalized regression neural network (GRNN). GRNN determine their own number of hidden units, while performing well at non-linear noisy systems.

With this model, product quantities are used to calculate proportions between each of the products which are used as neural network inputs along with feeding rate, volume, DO and stirring speed. Outputs are the derivative of product concentrations. The outputs are then integrated and fed back with a



F i g . 1 – *Generalized regression neural network with external feedback applied*

time delay, as described in Fig. 1. Proportions layer is a linear layer that produces one output for every combination of two inputs. Those outputs are the proportion of each input in the sum of both. To every quantity a small value (0.1) is added to avoid prediction distortions caused by a division of a near zero value.

The neural network was trained with only three cultivations, and only a small number of inputs were provided for it to learn the system dynamics: reactor initial conditions, evolution over experiment time of the off-line concentration values of biomass, glucose, acetate, and plasmid, evolution of total volume over time and the feedback control provided during the experiment: dissolved oxygen concentration and stirring speed. Latter values were obtained on-line from bioreactor control unit; the other values were obtained by *a posteriori* sample analysis. Samples were acquired at irregular intervals of time. Due to this sampling irregularity and noisy nature of data, cultivation was divided into sections and a second-degree polynomial curve was fitted to each product quantity for each of the sections. The following four sections were defined: (i) On the batch phase, the glucose consumption phase; (ii) On the batch phase, the acetate consumption phase, that started just after all glucose was consumed and ended just before the starting of the feeding phase; (iii) The feeding phase; (iv) After ending the feeding phase and until the end of cultivation (Fig. 2). For on-line data, linear interpolation was a reasonable option as data was acquired in very short lime intervals.

On-line data was collected for each cultivation periodically, but with a different number of samples. Cultivations C1, C2, C3 and C4 were sampled, respectively, 12, 15, 18 and 19 times. On-line data samples were linearly interpolated in order to have inputs (to feed the network) with an exact 30-minute periodicity. Therefore, as C1, C2, C3, and C4 took 11.5 h, 20.5 h, 27.5 h and 29.5 h, respectively, the number of inputs was 23, 41, 55 and 59.

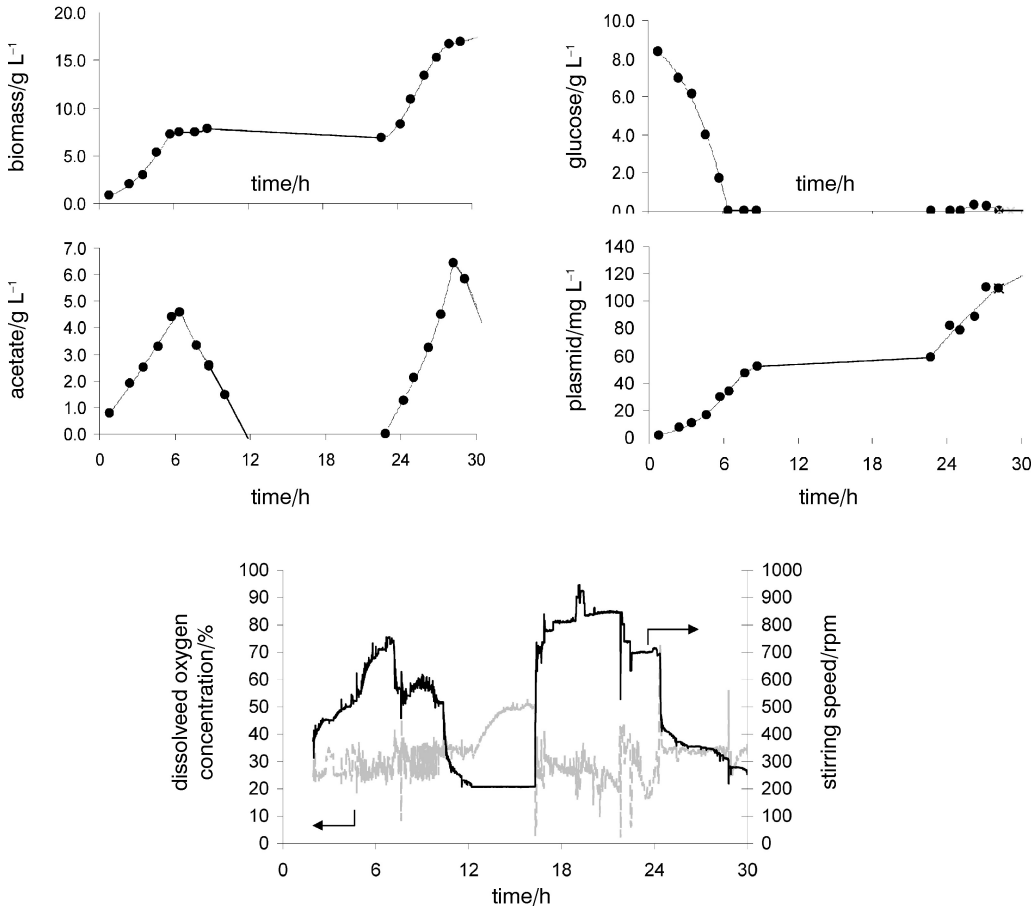
When testing an experiment, GRNN was fed with the following data: i) initial system conditions (mass values of biomass, glucose, acetate, and plasmid); ii) on-line data over time: feeding rate, dissolved oxygen concentration, and stirring speed.

We favoured mass prediction instead of concentration prediction. GRNN makes prediction, at each step t, for the mass of biomass, glucose, acetate, and plasmid at time $t + 1$, which corresponds to 30 minutes later. These values were used to self-feed the GRNN with the necessary inputs at the next learning iteration and (using the acquired on-line data) make the next prediction.

Results and discussion

Fed-batch cultivations

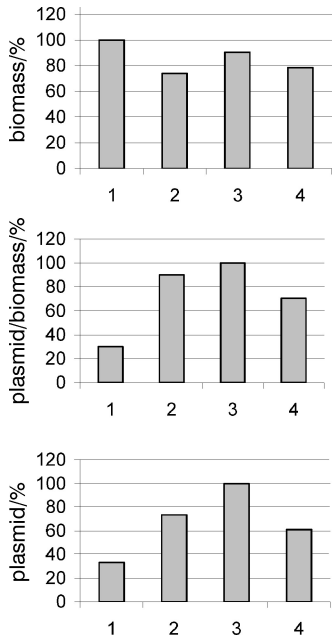
Fed-batch cultivations used for neural network learning and model validation were performed following a similar general strategy based on an initial batch phase followed by an exponential feeding phase. To establish different performances on the fed-batch cultivations, the four cultivations differ in relation to the glucose concentration on the batch phase or in the feeding medium, the ratio between the C-source and N-source used and on the feeding start period (Table 1). For example, cultivation 2



F i g . 2 – *Fed-batch cultivation 3*

was conducted with two-times more glucose concentration on the batch phase than cultivation 4. Cultivation 3 presented a totally different medium composition concerning the C-source/N-source ratio in relation to the other three cultivations. In cultivations 2 to 4, the feeding phase started only after all the acetate formed during the batch phase was consumed, contrary to cultivation 1. During the batch phase, the high glucose concentrations result in acetate formation (Fig. 2), that may inhibit both the cell growth and the plasmid production. For that reason, it is relevant that the acetate formed during the batch phase is consumed before the feeding phase started. With the present cultivation conditions, a high range of plasmid productions per biomass and plasmid final productions was observed (Fig. 3).

The present expression system, in a high-density cultivation conducted in non-defined and non-selective media, presented a complex behaviour. For example, in cultivation 1 the feeding started before the acetate produced during the batch phase was metabolised, contrary to cultivation 4. Contrary to the expected, cultivation 4 presented a lower biomass concentration than cultivation 1, but as expected a higher plasmid production per bio-



F i g . 3 – *Biomass, plasmid and plasmid produced per biomass obtained at the end of cultivations 1 to 4 (as defined in Table 1), in relation to the maximum value obtained in all cultivations. The maximum values obtained for each variable (equivalent to the graph 100 %) was for biomass concentration, plasmid produced per biomass and plasmid concentrations 19.2 g* L^{-1}*; 7.0 mg* g^{-1} *and 121 mg* L^{-1}*, respectively.*

mass than cultivation 1. One possible explanation is that in cultivation 4, when the feeding phase was started in an environment with low acetate concentration, a 2.4-fold higher plasmid production per biomass was obtained in relation to cultivation 1. However, a higher plasmid content will submit cells in cultivation 4 to a higher consumption of energy and metabolic precursors (as nucleotide bases for plasmid replication) and as consequent in 28 % lower biomass concentration in relation to cultivation 1 (Fig. 3).

An increased production of plasmid per biomass was observed in the following order: cultivation, 1, 4, 2 and 3. While the biomass concentration obtained at the end did not follow a similar increase or decreased order. If it is considered that plasmid maintenance and multiplication represents a metabolic burden to the cell, it is expectable that, in two different cultivations, the one presenting higher plasmid concentration per biomass will present a lower biomass concentration. However, a decreased biomass concentration was observed in the following order: cultivation 1, 3, 4 and 2, different from the observed relation of increased plasmid productions per biomass.

In recombinant systems, plasmid stability is a very important factor. High plasmid content per cell, which results in an extra-cellular metabolic burden, usually represents lower biomass concentrations and specific growth rates in relation to cells lacking plasmids. In complex rich media, without a selection marker (in this work the antibiotic kanamycin), the selective advantage of the cells that lost part or all the plasmid could result in a cellular population with a higher content of cells lacking plasmid population, and consequently cultivation productivity decrease. These factors are exacerbated in high-density cultivations, as in our case where dry cell weights as high as 19 g L^{-1} were achieved, using complex media and without selection pressure. Therefore, a dynamic and complex interaction, as the one represented in Fig. 2 usually occurs. If a kinetic model is to be defined, usually numerous experiments to define kinetics constants are required. Furthermore, some kinetics constants are highly difficult to determine. An example of an important parameter is the acetate threshold determination. It is well known that a primary barrier to total productivity in recombinant *E. coli* cultivation is the acetic acid production, which can inhibit growth, and decrease the recombinant product yield.[14] Therefore, a common methodology to optimize fed-batch cultivations is to avoid acetate synthesis. However, in rich cultivation media, containing for example yeast extract and tryptone as in the present case, the detection of acetate threshold tracking is highly difficult.[15]

GRNN training and validation results

In order to test the prediction ability of the model, we used cross-validation. For that purpose, we used 3 cultivations for learning and one for testing, and then changed the learning set (and consequently the testing experiment). We used 3 different learning-dataset combinations, always using the remaining one for testing: dataset D1={C1,C2,C3}; dataset D2={C1,C3,C4}; D3={C2,C3,C4}. Cultivation 3 was always included for learning and never for testing, because this experiment was made under much different conditions in comparison with the others – see Table 1, and was deliberately introduced to enforce learning diversity and therefore generalization power, so it would make no sense to use it for testing. Learning included both determining an optimal spread while fitting the neural network to the experimental data, using that spread. This could have been done by using just 2 experiments for optimizing neural network weights and the remaining one for adapting spread parameter. Unfortunately, learning examples provided by just 2 experiments usually lead to very poor covering of the search space, and using all 3 experiments for fitting weights was found to be strongly desirable. We decided to make a brute-force approach, using all possible spread values (with a 0.01 step) from 0.10 to 2.00, for updating the neural network. We then determined the spread that leads to the best coefficient of determination R^2 (16), for the 3-experiment learning data set. Although the spread was not updated with an independent experiment, we sacrificed learning generalization power in favour of reasonable network optimization. Nevertheless, prediction power was later tested with the independent testing experiment, by measuring the coefficient of model prediction Q^2 (this measure was obtained in the same way as R^2, but applied to test experiment) for each learning-and-testing combination.

We used Matlab Software for model design, training and testing. Coefficient of determination R^2 was obtained for learning datasets D1, D2 and D3, under a spread ranging from 0.10 to 2.00. Table 2 shows spread results from 0.10 to 0.30; remaining results were excluded for clarity. For each learning dataset, obtained GRNN weights and optimal spread parameter were used to test model prediction (Q^2) with the remaining cultivation. Therefore, D1 was tested with C4; D2 was tested with C2, and D3 was tested with C1. GRNN simulations are shown in Fig. 4, and Q^2 values are shown in Table 3. It is usual to quantify the variables as biomass, glucose, acetate and plasmid in units of concentration, as represented in Fig. 2. However, to compare different fed-batch, as represented in Fig. 4, the quantities of the variables in mass instead of concentra-

Table 2 – *Coefficient of determination (R^2) for each dataset, for each value of Spread (we represent only the most interesting range, from 0.10 to 0.30). For each dataset, R^2 is measured for estimating biomass, glucose, acetate, and plasmid. We optimize spread by considering the value that leads to greater R^2 average. Spread was set to 0.17 for D1; 0.16 for D2 and 0.28 for D3.*

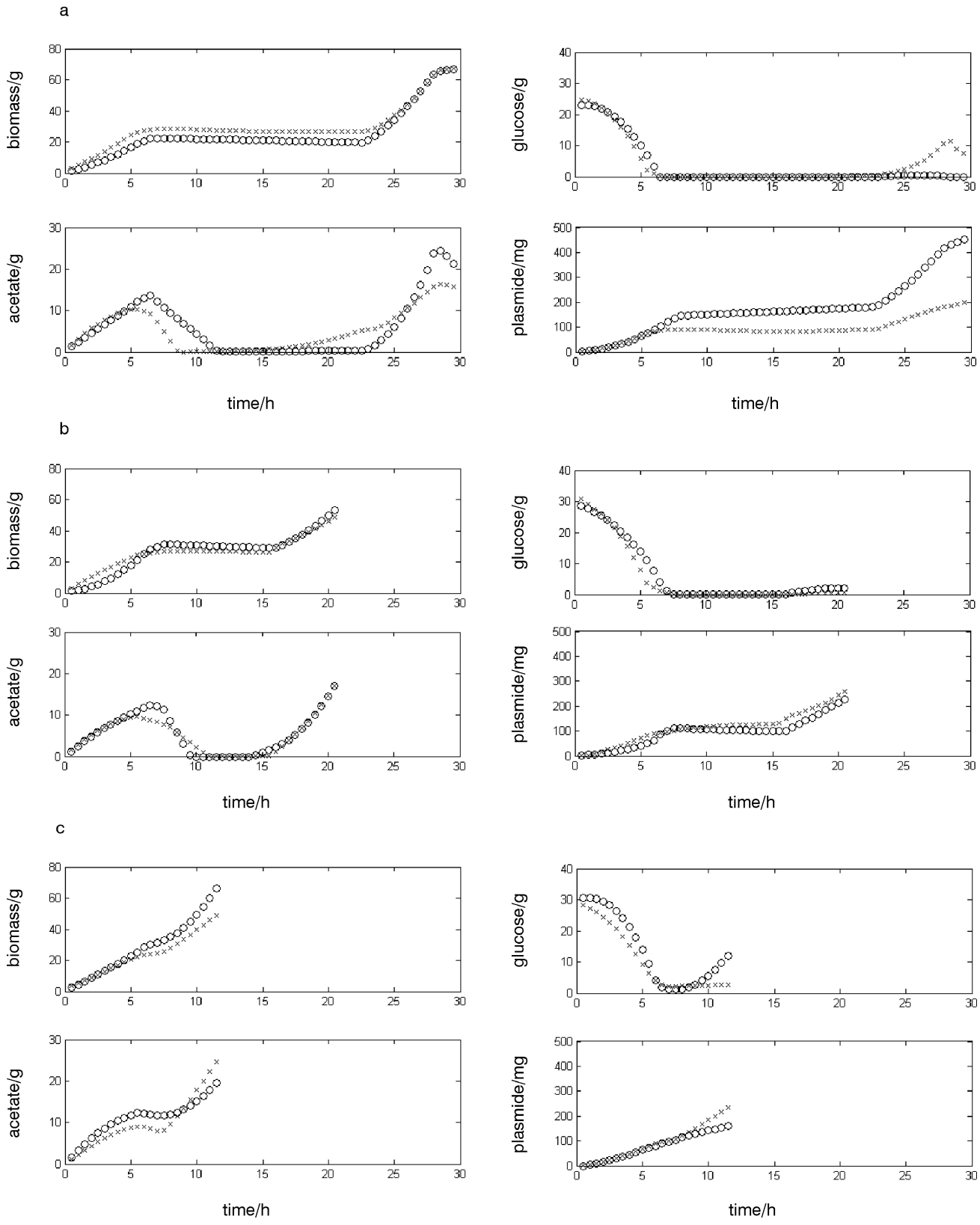
Spr.	Dataset D1 = {C1,C2,C3}					Dataset D2 = {C1,C3,C4}					Dataset D3 = {C2,C3,C4}				
	bio.	gluc.	acet.	plas.	aver.	bio.	gluc.	acet.	plas.	aver.	biom.	gluc.	acet.	plas.	aver.
0.10	0.949	0.951	0.904	0.991	0.949	0.983	0.960	0.927	1.000	0.968	0.851	0.865	0.541	0.922	0.795
0.11	0.954	0.955	0.915	0.992	0.954	0.986	0.960	0.926	0.999	0.968	0.846	0.870	0.504	0.928	0.787
0.12	0.961	0.959	0.926	0.992	0.960	0.989	0.961	0.930	0.999	0.970	0.860	0.875	0.630	0.917	0.821
0.13	0.968	0.963	0.934	0.993	0.964	0.990	0.961	0.935	0.999	0.971	0.863	0.879	0.681	0.908	0.833
0.14	0.976	0.966	0.939	0.993	0.968	0.990	0.960	0.939	0.999	0.972	0.854	0.880	0.695	0.898	0.832
0.15	0.982	0.968	0.940	0.993	0.971	0.990	0.959	0.943	0.999	0.973	0.828	0.878	0.684	0.882	0.818
0.16	0.987	0.971	0.939	0.993	0.972	0.990	0.957	0.945	0.998	0.973	0.791	0.878	0.676	0.882	0.806
0.17	0.991	0.973	0.936	0.992	0.973	0.989	0.956	0.946	0.998	0.972	0.779	0.876	0.677	0.845	0.795
0.18	0.994	0.975	0.932	0.991	0.973	0.988	0.956	0.945	0.997	0.972	0.775	0.878	0.688	0.855	0.799
0.19	0.997	0.976	0.926	0.990	0.972	0.987	0.956	0.943	0.997	0.971	0.780	0.881	0.703	0.844	0.802
0.20	0.998	0.977	0.920	0.988	0.971	0.986	0.958	0.940	0.996	0.970	0.785	0.884	0.717	0.837	0.806
0.21	0.998	0.978	0.913	0.987	0.969	0.984	0.960	0.936	0.994	0.969	0.793	0.888	0.731	0.822	0.808
0.22	0.997	0.978	0.905	0.985	0.966	0.983	0.962	0.932	0.992	0.967	0.797	0.893	0.747	0.831	0.817
0.23	0.995	0.977	0.897	0.983	0.963	0.980	0.965	0.927	0.990	0.965	0.806	0.901	0.768	0.829	0.826
0.24	0.992	0.975	0.889	0.981	0.959	0.977	0.967	0.921	0.987	0.963	0.820	0.912	0.797	0.825	0.839
0.25	0.988	0.973	0.881	0.979	0.955	0.973	0.970	0.915	0.983	0.960	0.834	0.924	0.826	0.817	0.850
0.26	0.984	0.970	0.872	0.977	0.951	0.968	0.971	0.908	0.979	0.957	0.846	0.934	0.846	0.812	0.860
0.27	0.978	0.967	0.863	0.976	0.946	0.961	0.972	0.902	0.974	0.952	0.855	0.941	0.859	0.797	0.863
0.28	0.973	0.962	0.853	0.974	0.940	0.954	0.972	0.894	0.968	0.947	0.862	0.946	0.865	0.788	0.865
0.29	0.966	0.957	0.841	0.973	0.934	0.946	0.970	0.886	0.962	0.941	0.868	0.950	0.868	0.767	0.863
0.30	0.959	0.952	0.829	0.971	0.928	0.937	0.967	0.877	0.956	0.934	0.873	0.952	0.869	0.750	0.861

tion should be used. Indeed, in a fed-batch the final concentration of variables depends on the degree of dilution of the feeding medium. For example, for two fed-batches based on the same strategy, the one with a more diluted feeding medium will present the same biomass and plasmid quantities in mass, but not the same concentrations. Fig. 4 presents the performance of the network by comparing predicted and real values over time, at all cross-validation tests, for plasmid, acetate, biomass and glucose. It shows masses instead of concentrations, because GRNN was designed to estimate variable masses.

Model prediction coefficient Q^2, for all cross-validation tests, is shown at Table 3. We observe very good behaviour at D2 validation and an overall good network prediction (Q^2 above 0.7) at all validations except in Acetate prediction at D3 validation and Plasmid prediction at D1 validation. Besides obvious difficulties due to the scarce information provided by only 3 cultivation sets of train-

Table 3 – *Q^2 obtained for each dataset, in respect to each variable. Dataset D1 was validated with C4; dataset D2 was validated with C2; dataset D3 was validated with C1.*

	D1 → C4	D2 → C2	D3 → C1
Biomass	0.865	0.900	0.840
Glucose	0.803	0.953	0.846
Acetate	0.735	0.922	0.564
Plasmid	0.045	0.799	0.727



F i g . 4 – *GRNN model cross-validation. Comparison between estimated and experimentally observed masses of plasmid, acetate, biomass and glucose for: a) cultivation test C4 after learning dataset D1; b) cultivation test C2 after learning dataset D2; c) cultivation test C1 after learning dataset D3. O: experimental data; X: estimated data.*

ing data, the higher analytical error associated with plasmid prediction is most probably associated with the low reproducible procedure of the alkaline cell lyses method used to extract plasmid.

Conclusions

With the present GRNN model, it was possible to predict the dynamic behaviour of high-density cultivations of recombinant *E. coli,* conducted in complex and non-selective medium. GRNN learning was based on a few numbers of complex cultivations, on the bioreactor initial conditions and few on-line variables: the feed-rate, percentage of dissolved oxygen concentration and the bioreactor stirring speed. The use of on-line measurement variables that are routinely analyzed in the bioreactor will simplify the control process as it avoids the use of other measurement probes, reducing further interferences with fluid circulation and higher costs. Oxygen- and stirring-speed-based feedback controllers present several advantages due to their sensitivity, versatility and on-line availability. Therefore, the present GRNN may be used to detect real on-line changes in the system, having the potential to be implemented as a simple and inexpensive on-line control of plasmid production process. The present process is in accordance with the Process Analytical Technology guidance that stimulates pharmaceutical companies to develop global methods for on-line process monitoring, to ensure a pre-defined final product quality. This model presents the advantages to be based on few cultivations for learning and uses the general probes present in a regular stirrer-tank bioreactor, operating *in situ* and is able to generate on-line information on multiple key bioprocess variables.

ACKNOWLEDGMENTS

Work supported by a PTDC/Bio/69242/2006 research grant, by Fundação para a Ciência e Tecnologia, Portugal.

List of symbols

GRNN – Generalized Regression Neural Network

DO – Dissolved Oxygen Concentration

MIMO – MultiInput–MultiOutput

MLP – MultiLayered Perceptron

RBFN – Radial Basis Function Network

S – Spread of the radial basis function

RBF – Radial Basis Function

References

1. *Stoll, S. M., Calos, M. P.,* Curr. Opin. Mol. Ther. **4** (2002) 299.
2. *Prather, K. J., Sagar, S., Murphy, J., Chartrain, M.,* Enzyme Microb. Technol. **33** (2003) 865.
3. *Calado, C. R. C., Almeida, C., Cabral, J. M. S., Fonseca, L. P.,* J. Biosc. Bioeng. **96** (2003) 141.
4. *Calado, C. R. C., Ferreira, B. S., Fonseca, M. R., Cabral, J. M. S., Fonseca, L. P.,* J. Biotechnol. **109** (2004) 147.
5. *Calado, C. R. C., Mannesse, M., Egmond, M., Cabral, J. M. S., Fonseca, L. P.,* Biotechnol. Bioeng. **78** (2002) 692.
6. *Zelic, B., Bolf, N., Vasic-Racki, D.,* Bioprocess Biosyst. Eng. **29** (2006) 39.
7. *Kiviharju, K., Salonen, K., Leisola, M., Eerikaien, T. E.,* J. Biotechnol. **10** (2006) 365.
8. *Harada, L. H., Costa, A. C., Maciel Filho, R.,* Appl. Biochem. Biotechnol. **98-100** (2002) 1009.
9. *Kovarova-Kovar, K., Gehlen, S., Kunze, A., Keller, T., Daniken, R. V., Kolb, M., van Loon, A. P.,* J. Biotechnol. **79** (2000) 39.
10. *Kulkarni, S. G., Chaudhary, A. K., Nandi, S., Tambe, S. S., Kulkarni, B. D.,* Biochemical Eng. Journal **18** (2004) 193.
11. *Diogo, M. M., Queiroz, J. A., Prazeres, D. M. F.,* J. Chromatog. A **1006** (2003) 137.
12. *Specht, D. F.,* IEEE Trans. Neural Networks **2** (6) (1991) 568.
13. *Nerrand, O., Roussel-Ragot, P., Urbani, D., Personnaz, L., Dreyfus, G.,* IEEE Transactions Neural Networks **5** (2) (1994) 178.
14. *Johnston, W. A., Stewart, M., Lee, P., Cooney, M. J.,* Biotechnol. Bioeng. **84** (2003) 314.
15. *Whiffin, V. S., Cooney, M. J., Cord-Ruwisch, R.,* Biotechnol. Bioeng. **85** (2004) 422.
16. *Alman, D. H., Ningfang, L.,* Color Res. & Application **27** (2) (2001) 122.