J Ind Microbiol Biotechnol (2004) 31: 469–474
DOI 10.1007/s10295-004-0171-4

ORIGINAL PAPER

Marcelo Abel Soria · Jose Luis Gonzalez Funes
Augusto Fernando Garcia

A simulation study comparing the impact of experimental error on the performance of experimental designs and artificial neural networks used for process screening

Received: 3 June 2004 / Accepted: 3 September 2004 / Published online: 6 October 2004

Abstract Many variables and their interactions can affect a biotechnological process. Testing a large number of variables and all their possible interactions is a cumbersome task and its cost can be prohibitive. Several screening strategies, with a relatively low number of experiments, can be used to find which variables have the largest impact on the process and estimate the magnitude of their effect. One approach for process screening is the use of experimental designs, among which fractional factorial and Plackett–Burman designs are frequent choices. Other screening strategies involve the use of artificial neural networks (ANNs). The advantage of ANNs is that they have fewer assumptions than experimental designs, but they render black-box models (i.e., little information can be extracted about the process mechanics). In this paper, we simulate a biotechnological process (fed-batch growth of baker's yeast) to analyze and compare the effect of random experimental errors of different magnitudes and statistical distributions on experimental designs and ANNs. Except for the situation in which the error has a normal distribution and the standard deviation is constant, it was not possible to determine a clear-cut rule for favoring one screening strategy over the other. Instead, we found that the data can be better analyzed using both strategies simultaneously.

Keywords Screening designs · Artificial neural networks · *Saccharomyces cerevisiae* · Experimental error

M. A. Soria (✉) · J. L. Gonzalez Funes · A. F. Garcia
Cátedra de Microbiología Agrícola,
Facultad de Agronomía, Universidad de Buenos Aires,
Av. San Martín 4453, 1417 Buenos Aires, Argentina
E-mail: soria@agro.uba.ar
Tel.: +54-11-45237059
Fax: +54-11-45148741

Introduction

Biotechnological processes are influenced by numerous factors and their interactions. Further complications arise from the non-linear nature and time-varying characteristics of these systems. In addition, certain measurements are affected by noise and relatively long sampling times. For these reasons, the optimization of one factor at a time, although frequently used, is not an adequate strategy to improve a process and can lead to wrong conclusions. In contrast, experimental design strategies are based on the simultaneous variation of more than one factor. Every available design specifies the factor levels to be tested, the number of repetitions and the layout of blocks. Through the statistical analysis of the results, which implies fitting the data to linear models or higher-degree polynomials, it is possible to determine quantitatively the effect of factors, or input variables, on the system dynamics. Among experimental designs, screening designs are used to study a large number of factors in a small number of experiments, with the goal of finding which factors are important and discarding the others. Other designs can be used later to analyze fewer factors with more experiments and maybe more levels, to allow fitting the data to second-degree polynomials, as is the case with response surface methodologies. Two common screening designs are fractional factorials and Plackett–Burman designs. Fractional factorials are fractions of the corresponding full factorials and Plackett–Burman designs are useful for analyze a large number of factors in relatively few runs, since they estimate only the main effects while neglecting all the interactions.

Experimental error has a large effect on the results obtained with these methods. If the data follows a linear trend, the estimates obtained using linear methods are the best possible and their variance is minimal only when the experimental error has a normal distribution (Gaussian noise) and the standard deviation (SD) is constant for any point on the experimental space.

However, these assumptions do not always hold true for biological systems; and many times the actual error distribution can be difficult to identify. Two frequent violations of these assumptions are the occurrence of a constant coefficient of variation (cv), instead of a constant standard deviation, and the presence of error with log-normal distribution. If cv is constant, SD increases with the magnitude of the observed variable. A log-normal variable can be transformed into a normal one by applying a logarithmic transformation. However, the analysis of transformed variables is difficult when more than a few output variables and their interactions are studied together, a typical situation in biotechnology [2].

A different, black-box strategy is the use of artificial neural networks (ANNs). ANNs are trained with sets of inputs and outputs and they "learn" how to reproduce an output from the input. If adequately trained, ANNs can yield outputs similar to the real world from different sets of inputs. Some of the reasons for their appeal are that they can model non-linear models and they require no previous knowledge of the system dynamics. Their main drawback is their black-box nature. Even when the ANN can satisfactorily simulate the system it is modeling, little can be said about the underlying relationships between variables. Another problem is over-training the ANN, which happens when the network is very good at predicting only the outputs from the training data set and performs poorly with other inputs. However, there are several procedures available to limit over-training [6, 13].

Fed-batch culture is the method of choice for the industrial production of baker's yeast, *Saccharomyces cerevisiae*. This method of production ensures that most of the sugar available in the culture medium is channeled through the oxidative metabolism with minimum production of ethanol and high conversion of carbon source into biomass. This microorganism is an excellent model system for simulation and theoretical studies because its genetics and metabolism are well known and there are many theoretical models describing its growth in bioreactors [10, 11].

The goal of this paper is to compare experimental screening designs analyzed with standard linear regression models and ANNs for the optimization of a simulated fed-batch process for biomass production of *S. cerevisiae* under varying levels of experimental noise with different underlying statistical distributions.

Materials and methods

Simulation model

Simulated experiments were generated using the model developed by Pham et al. [11]. This model simulates aerobic fed-batch fermentations of *S. cerevisiae* based on a kinetic model of overflow metabolism and allows variations in a large number of parameters and factors. For this study, seven factors were selected: initial dry cell weight (X_0), initial glucose concentration (S_0), initial ethanol concentration (E_0), initial volume (V_0), initial feed rate (F_0), glucose concentration in feed (S_f) and specific feed rate (SFR). For every factor, one high level and one low level were chosen and simulations were then run with all 128 (2^7) possible combinations of levels, plus an additional simulation with all factors set at the middle point.

The biomass concentration after 12 h of growth was the output variable analyzed. The levels tested for each factor and their coding are shown in Table 1. It was assumed that the fermentor had a total volume large enough to accommodate the final volume of any experiment.

Statistical analysis and random numbers generation

R ver. 1.7.1 statistical software (R Foundation for Statistical Computing, Vienna, Austria) was used for data analysis. Random numbers with a uniform distribution were generated using the random number generator of Microsoft Excel. The randomness of these data was analyzed with the test described by Banks et al. [1] and proved satisfactory for the short sequences required. Random numbers with normal or log-normal distributions were obtained from the uniformly distributed series applying inverse probability functions. Three screening designs design were used: a 12-experiment Plackett–Burman design and fractional factorial designs for 2^{7-4} and 2^{7-3} experiments [2]. Different random error terms were added to every one of the 129 biomass values simulated. Four different distributions were selected for error distribution: normal with constant SD, log-normal with constant SD, normal with constant cv and log-normal with constant cv. Several values of SD and cv

Table 1 Factors (variables) tested in this study and range of values

Factor	Low value (code −1)	High value (code +1)	Center value(code 0)	Factor name (units)
X_0	0.5	2	1.25	Initial dry cell weight (g l^{-1})
S_0	0.005	0.1	0.0525	Initial glucose concentration (g l^{-1})
E_0	0	1	0.5	Initial ethanol concentration (g l^{-1})
V_0	4	7	5.5	Initial volume (l)
F_0	4	10	7	Initial feed rate (g h^{-1})
S_f	60	120	90	Glucose concentration in feed (g l^{-1})
SFR	0.15	0.35	0.25	Specific feed rate (h^{-1})

were tested (Table 3). Two data sets were generated for every combination of design, distribution of the error and magnitude of SD or cv.

Low and high values of factors were coded as −1 and +1, respectively, prior to statistical and ANN analysis. This coding yields experimental designs that are orthogonal and in consequence the estimated coefficients have minimum variance [2]. It was empirically shown that the Bayesian regularization technique for ANN training also works better with −1, +1 coded variables [4].

Neural networks

ANNs were built using the Neural Network Toolbox included in MATLAB ver. 6.0. (The MathWorks, Mass., USA). The architecture consisted of a feed-forward network with three layers: one input layer with seven inputs, one hidden layer with three neurons and one output layer with one output. The transfer function of the neurons in the hidden layer was the hyperbolic tangent sigmoid transfer function (tansig) and the neuron in the output layer had a linear transfer function. The Bayesian regularization back-propagation (trainbr) method was used for training. With this method, it is possible to choose the number of neurons in the hidden layer that ensures good predictions while minimizing the risk of over-fitting [4, 8].

Results

Data simulation and analysis

Growth curves of *S. cerevisiae* were simulated using the model for aerobic fed-batch cultures developed by Pham et al. [11]. Seven factors were varied and 129 simulations were run as described in the Materials and methods. The biomass concentration after 12 h of growth was the output variable analyzed throughout this study.

A summary review of the 129 simulated biomass data showed that the lowest value of final biomass was 5.20 g l^{-1} and was obtained in two experiments with factors set at (in coded values): $X_0 = -1$, $S_0 = -1$, $V_0 = +1$, $F_0 = -1$, $S_f = -1$, SFR = −1. E_0 was −1 in one experiment and +1 in the other. The highest biomass value was 20.18 g l^{-1} and was observed in two experiments, differing only in the E_0 setting. The settings for the other factors were: $X_0 = +1$, $S_0 = +1$, $V_0 = -1$, $F_0 = +1$, $S_f = +1$, SFR = +1. The biomass values had a 3.9-fold variation (20.18:5.20) across the experimental space, which ensured that a wide range of physiological conditions were covered.

The combination of factors and levels used to generate the biomass data, excluding the central point, corresponded to a full factorial experiment of 2^7. These data were analyzed by analysis of variance (ANOVA) and the six-factor and seven-factor interactions were pooled and used as the error term. The results are shown in Table 2. The initial glucose (S_0) and ethanol (E_0) concentrations were not significant ($P > 0.05$), while all other factors were highly significant ($P < 0.0001$), but with different effects: V_0 had a large negative effect, while S_f, SFR, F_0 and X_0 had decreasing positive effects. Also, some two-factor interactions were significant at $P < 0.01$, involving either S_f or SFR: $S_f \times X_0$, $S_f \times V_0$, $S_f \times F_0$ and SFR × V_0, SFR × F_0, SFR × S_f.

Table 2 Full factorial analysis of the whole data set (128 experiments). The mean squared error (*MSE*) is an estimator of the magnitude of the effect on the overall variability. + Positive effect, − negative effect, *NS* not significant

Factor	*P*-value	MSE
X_0	0.000001*	59.419^{+}
S_0	0.363525	0.292^{NS}
E_0	0.577316	0.106^{NS}
V_0	0.000000*	429.240^{-}
F_0	0.000000*	124.899^{+}
S_f	0.000000*	1046.927^{+}
SFR	0.000000*	131.990^{+}

*P-values are highly significant ($P < 0.05$)

Generation of simulated data sets with different error structures

From the pool of 129 simulations, biomass data corresponding to the layouts of the screening designs (Plackett–Burman, fractional factorial) were extracted. Then, random errors with different magnitudes and distributions were added to the biomass values to generate duplicate "experimental" data sets.

The different "experimental" biomass data sets were fitted using linear regression. As the error term for the regression models, we used the SD of three replicated central points or the pooled effects of interaction factors that did not include the main factors. The *t*-test was used to assess the statistical significance of factors. Table 3 shows the average number of significant factors detected under every condition tested. As expected, the ability to detect significant factors decreases with increasing SD or cv values and data with a normally distributed error gave better predictions than log-normal data. Although only one combination (normally distributed errors with constant SD) fulfills the requirements of linear models, it was possible to detect significant factors with the other error structures. The use of the pooled effects of interaction terms for error estimation showed an enhanced sensitivity, compared with central points.

Model building and testing

Linear models were built from every "experimental" data set, considering only the significant factors deter-

Table 3 Average number of significant factors detected by 2^{7-4} (*Frac8*), 2^{7-3} (*Frac16*) fractional factorials and a 12-experiment Plackett–Burman (*PB*) linear model. Factors were considered significant when a *t*-test on its effect had $P \leq 0.05$. Experimental error had normal or log-normal distributions with constant SD or constant cv; and the columns labeled SD or cv show their values. The error term for estimating the effect significance was either the SD of three central points (*C.p.*) or the pooled effects of interaction terms that did not include the main factors (*Int.*)

Normal distribution							Log-normal distribution						
Constant SD				Constant cv			Constant SD				Constant cv		
SD	Design	C.p.	Int.	cv	Design	C.p.	Int.	SD	Design	C.p.	Int.	cv	Design	C.p.	Int.
0.20	Frac8	5.5	–	0.05	Frac8	4.5	–	0.05	Frac8	5.5	–	0.05	Frac8	2.5	–
	Frac16	2.0	6.0		Frac16	2.0	5.5		Frac16	2.0	5.5		Frac16	1.0	4.5
	PB	4.5	6.0		PB	4.0	5.0		PB	5.0	4.5		PB	5.0	5.5
0.40	Frac8	4.5	–	0.10	Frac8	2.0	–	0.10	Frac8	2.5	–	0.10	Frac8	0	–
	Frac16	2.0	5.0		Frac16	2.0	3.5		Frac16	4.0	3.0		Frac16	0	0
	PB	4.5	5.5		PB	3.0	1.0		PB	4.5	3.0		PB	1.0	0.5
0.80	Frac8	4.0	–	0.20	Fac8	0.5	–	0.20	Frac8	1.0	–	0.20	Frac8	1.5	–
	Frac16	1.5	4.5		Frac16	1.0	1.5		Frac16	1.0	2.0		Frac16	0	2.5
	PB	3.5	2.5		PB	2.5	1.0		PB	0.5	0		PB	0	0
1.20	Frac8	2.5	–	0.40	Frac8	0	–	0.40	Frac8	0	–	0.40	Frac8	3.5	–
	Frac16	3.0	3.5		Frac16	0	0		Frac16	1.0	0		Frac16	0	3.5
	PB	2.5	2.5		PB	0.5	1.0		PB	0	0		PB	0.5	0
								0.80	Frac8	5.5	–				
									Frac16	0	5.5				
									PB	0	0				

mined in the previous section. Additional data sets were generated for validation (validation data set); and they had the same error structure as the corresponding model data set but the biomass values were sampled at different points over the experimental space. ANNs were also built from the "experimental" data sets. All the ANNs had the same architecture and were trained with the same methods. Additional validation data sets were also generated.

To validate the models, predicted values were estimated from the models built before at the same factor levels used for generating the validation data sets. Predicted values were compared with the validation data sets using the mean squared error (MSE). MSE values were lower for 2^{7-3} fractional factorials than for 2^{7-4} experiments. The average MSE values for Plackett-Burman and 2^{7-3} fractional factorial designs were slightly lower for models built using interaction terms instead of central points for error estimation, although the differences were not significant. For ANNs, MSE values decreased when the number of data points used for training increased. So, the lowest MSE was obtained with networks trained with 16 points distributed according to the layout of the 2^{7-3} fractional factorial. Figure 1 shows the MSE values obtained with the models that performed better: linear models from a 2^{7-3} fractional factorial and a 12-experiment Plackett–Burman design and a model from an ANN trained with 16 data points.

Discussion

The smallest data layout (2^{7-4} experiments) analyzed with linear models could detect several significant factors with low values of SD or cv, but the predictive capacity was rather poor, i.e., MSEs were large for both regression and ANN models. These designs can be easily augmented to 2^{7-3} designs, as we did here, taking the smaller design as one half of the larger. However, if one smaller design indicates that some factors are not significant, it is not recommendable to augment that design. Instead, a new analysis should be applied only to the significant factors. The performance of Plackett–Burman designs was lower but close to that of 2^{7-3} factorials, while requiring four experiments less. In cases where the cost or difficulty of experiments is an issue, the use of Plackett–Burman designs can be a satisfactory tool, although they cannot be augmented and their interactions are difficult to interpret. The best results were obtained with the 2^{7-3} experiments analyzed with linear models and ANNs. Adding central points to the designs neither increased the capacity to detect significant factors nor improved the predictive value of models. However, even if they are not included in the design, it is worth running central points because they are a valuable estimation of the dispersion at the center of the layout.

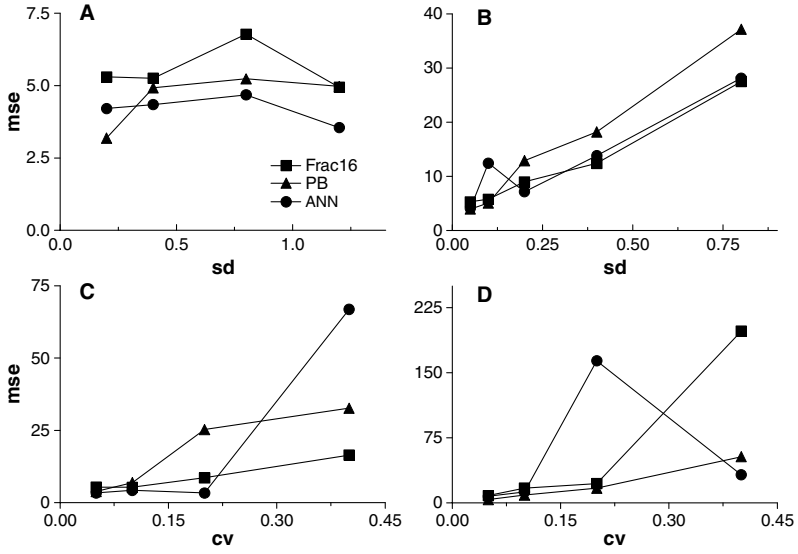

Fig. 1 MSE (*mse*) for the actual versus predicted values of models built from simulated data with different levels of SD (*sd*) and cv and four different distributions of experimental error: **a** normal distribution with constant SD, **b** log-normal distribution with constant SD, **c** normal distribution with constant cv, **d** log-normal distribution with constant cv. The predicted values were obtained using the regression coefficients calculated from a 2^{7-3} fractional factorial design (*Frac16*), a 12-experiment Plackett–Burman design and a simulation with a feed-forward ANN.

ANOVA of the full set of simulated data showed that five factors and six interactions were significant (Table 2). Either SFR or F_0 were always present in the interactions; and the non-significant factors were E_0 and S_0. These results are in agreement with what is expected of a fed-batch process, in which most of the control lies in the feeding stage. The linear models sometimes detected six significant factors, one more than the five actually significant. This can be due to the distortion introduced when fitting a highly interacting and at some regions non-linear process to linear models that disregards those characteristics. Even with these simplifications, the models still allowed a good insight into system dynamics.

For the ANN models, the lowest MSE values were obtained with the largest data sets: 16 data points (2^{7-3} fractional factorial). This is in agreement with other published results which show that predictions improve when increasing the size of the training set [6, 7]. The Bayesian regularization technique was used for training because it takes all the available data for training, while other learning algorithms require splitting the data into one subset for training and another for testing in order to get good fit without over-fitting [4, 8]. Since the data sets tested were quite small, it would not have been recommended to reduce their size further.

Normally and log-normally distributed data with either constant SD or constant cv could be reliably analyzed with conventional fractional factorials or ANNs when the SD for the three central points was lower than 2. There was no clear pattern to favor experimental designs or neural networks; and increasing cv or SD increased MSE values for all conditions tested, except for normal distributions with constant SD. A convenient strategy is using the data layouts of the 2^{7-3} experiments, building linear regression and ANN models and then comparing the predictions of both techniques. Since the analytical procedures are very different for each method, when the predictions are similar they can be reliably accepted. In contrast, those regions of the experimental space yielding grossly different predictions need further research. This concept can be extended to other situations with a different number of factors. The data layout from an appropriate experimental design should be selected, the experiments performed and the data fitted to linear, or polynomial, models and used for ANN training. Then, predicted values for the experimental data and for unexplored, intermediate values should be calculated with both methods. Using MSE or correlation coefficients and graphical methods, it is possible to make comparisons between the predictions of both models and between predicted and observed values. If the agreement between values from the two models is poor, the design should be augmented or replicates should be added. If replicates are added, the value to be analyzed is their average.

Linear models can reveal significant factors, the sign of their effect and make predictions, while ANN models can be used for optimization and for making predictions. Nagata and Chu [9] combined ANNs with genetic algorithms to search the experimental space to find maximum points. In contrast, 2^{n-k} fractional factorial designs have to be augmented to perform the search for maximum or minimum points. Screening experiments have been used to improve medium composition [3, 5, 12]. This study focused on different model parameters and not only culture medium design, but it could be easily applied to such a situation.

References

1. Banks J, Carson JS, Nelson BL (1996) Discrete-event system simulation, 2nd edn. Prentice Hall, Upper Saddle River
2. Box GEP, Hunter GH, Hunter JS (1979) Statistics for experimenters. An introduction to design, data analysis and model building. Wiley, New York

3. Castro PM, Hayter PM, Ison AP, Bull AT (1992) Application of a statistical design to the optimization of culture medium for recombinant interferon-gamma production by Chinese hamster ovary cells. Appl Microbiol Biotechnol 38:84–90
4. Foresee DF, Hagan MT (1997) Gauss–Newton approximation to Bayesian learning. Proc Int Joint Conf Neural Networks 1997
5. Kalil SJ, Maugeri F, Rodrigues MI (2000) Response surface analysis and simulation as a tool for bioprocess design and optimization. Process Biochem 35:539–550
6. Kennedy M, Krouse D (1999) Strategies for improving fermentation medium performance: a review. J Ind Microbiol Biotechnol 23:456–475
7. Kennedy MJ, Prapulla SG, Thakur MS (1992) Designing fermentation media: a comparison of neural networks to factorial design. Biotechnol Tech 6:293–298
8. MacKay D (1992) Bayesian interpolation. Neural Comput 4:415–447
9. Nagata Y, Chu KH (2003) Optimization of a fermentation medium using neural networks and genetic algorithms. Biotechnol Lett 25:1837–1842
10. Ostergaard S, Olsson L, Nielsen J (2000) Metabolic engineering of *Saccharomyces cerevisiae*. Microbiol Mol Biol Rev 64:34–50
11. Pham HT, Larsson G, Enfors SO (1998) Growth and energy metabolism in aerobic fed-batch cultures of *Saccharomyces cerevisiae*: simulation and model verification. Biotechnol Bioeng 60:474–482
12. Silveira RG, Kakizono T, Takemoto S, Nishio N, Nagai S (1991) Medium optimization by an orthogonal design for the growth of *Methanosarcina barkeri*. J Ferm Bioeng 72:20–25
13. Zupan J, Gasteiger J (1993) Neural networks for chemists. An introduction. VCH, Weinheim