ISSN 1996-0700

Asian Journal of Biotechnology

Knowledgia
SCIENTIFIC
A Place to Publish Outstanding Research

http://knowledgiascientific.com

Asian Journal of Biotechnology 5 (1): 1-20, 2013
ISSN 1996-0700 / DOI: 10.3923/ajbkr.2013.1.20

Using Statistical Tools for Improving Bioprocesses

[1]Eutimio Gustavo Fernández Núñez, [2]Rodolfo Valdés Veliz, [1]Bruno Labate Vale da Costa, [3]Alexandre Gonçalves de Rezende and [1]Aldo Tonso

[1]Departamento de Engenharia Química, Universidade de São Paulo, São Paulo, SP, Brazil
[2]Monoclonal Antibody Department, Center for Genetic Engineering and Biotechnology (CIGB), Ave. 31 / 158 and 190, Cuabanacan, Playa, P.O. Box 6162, CP 10600, Havana, Cuba
[3]Laboratório de Imunologia Viral, Instituto Butantan São Paulo, SP, Brazil

Corresponding Author: Eutimio Gustavo Fernández Núñez, Departamento de Engenharia Química, Universidade de São Paulo, Av. Prof. Luciano Gualberto, trav. 3, 380 Butantã, 05508-900, São Paulo, SP, Brazil Tel: (55)-11-3091 2282

ABSTRACT

In this review most of statistical tools currently applied in the bioprocess area were discussed and classified. The main three categories were: fair comparison of results, mathematical modeling for little studied systems and taking advantage of large volume of data for enhance robustness and efficiency. For each statistical technique, an example from literature was commented to demonstrate its utility in bioprocess problems. Besides, others statistical tools without a wide application, at the moment, in bioprocess were also discussed, taking into consideration the near future use of them. As a conclusion, a chart was constructed for guiding researchers to select the correct statistical technique according to the specific bioprocess problem.

Key words: Artificial neural network, bioprocess, experimental design, multivariate methods

INTRODUCTION

Bioprocesses are currently used to produce a wide variety of chemicals from alcohols, organic compounds and amino acids, to antibiotics and therapeutically active recombinant proteins (Clementschitsch and Bayer, 2006). This kind of processes is characterized by the industrial application of biological pathways or reactions mediated by living cells of animal, plants and microorganisms or enzymes under controlled conditions for the biotransformation of raw material into products (Nair, 2008). Bioprocess technology is typically made up of three parts: upstream process, bioreactions and downstream processing (Nair, 2005).

In the upstream step or pretreatment, the raw material from biological or non-biological origin is first converted to a form suitable for processing. Subsequently, one or more bioreaction stages are performed. The biochemical reactors or bioreactors are the core of the bioreaction step. In this stage, the following operations are carried out: production of biomass, metabolize biosynthesis and biotransformation. Finally, the material produced in bioreactors must be further processed in the downstream section. Downstream processing consists of basically physical separations in order to purify and concentrate the product of interest (Jana, 2008).

Bioprocess goals are influenced by several parameters, being the definition of best values for them, an important and complex task for biochemical engineers. For economic reasons, large-scale established bioprocess should not be disturbed and any modification to any parameter may be not considered (Najafpour, 2006). Thus, on lab-scale is performed an intensive experimental work for understanding and optimizing the individual operation units and the process as a whole, both novel and established bioprocesses (Najafpour, 2006; Dubey and Behera, 2011). At the same time a high amount of data is stored from bioprocesses on large-scale equipped with sophisticated control, data logging and archiving systems; these could be also used to improve them (Charaniya *et al.*, 2008).

The statistical tools have demonstrated that are useful to solve this task, decreasing time and resource. The aim of this review is to expose the principal statistical techniques with applications in bioprocesses through the examples extracted from literature as well as a philosophy of work without mathematical details in order to identified what statistical method should be used as the classification of the problem under study.

DEFINING KEY TERMS

The definitions of three key statistical terms (factors, levels and responses) are necessary to facilitate the readership understanding of this text, mainly for poor trained professionals in statistical tools. This is the first step of the stair for helping bioprocess researchers to choose by themselves with a minimal statistical knowledge, the right technique for their specific problem.

Factors correspond to the independent variables of the system which we are interested in knowing as they influence the process outputs. The levels are some values in the study range of factors, when factor is a numeric variable or different categories, in the case of qualitative variable. For instance, the pH and nitrogen source could be studied for improving the yield of a fermentative process. The range of interest for factor pH (6.0-8.0), a numeric variable, could be explored at three levels: 6.0, 7.0 and 8.0. On the other hand, for nitrogen source, a qualitative variable, nitrate and urea could be two levels for this factor.

Responses are the properties of the system that are being measured and they are modified as a consequence of changes in factor values. They are also defined as dependent variables. Retaking the previous example, yield would be the response in this study case.

PROBLEM DEFINITION

Once defined the basic statistical vocabulary, we have the primary elements to understand what kind of problem we need to resolve in statistical terms. This is the first question to answer in order to choose the right statistical tool in any scientific work and as a consequence in bioprocesses too. In general, problems in bioprocesses might be statistically classified in three categories:

- Fair comparison of results
- Mathematical modeling for little-studied systems or processes
- Take advantage of large volume of data for enhance robustness and efficiency

The first one is used to determine whether or not one or more factors can affect the system outcomes (responses). The main goal is to define in qualitative way the effects of variables and levels of them on responses of a well-know system (control) or among their own responses for non-established systems. In other words, as a result of this statistical problem are identified relationships of equality, superiority or inferiority among alternatives or treatments considered (Boos and Brownie, 1995).

Statistical modeling or design of experiment methodology for little-studied systems or processes is useful when a mathematical function which connects individual factors under study and their

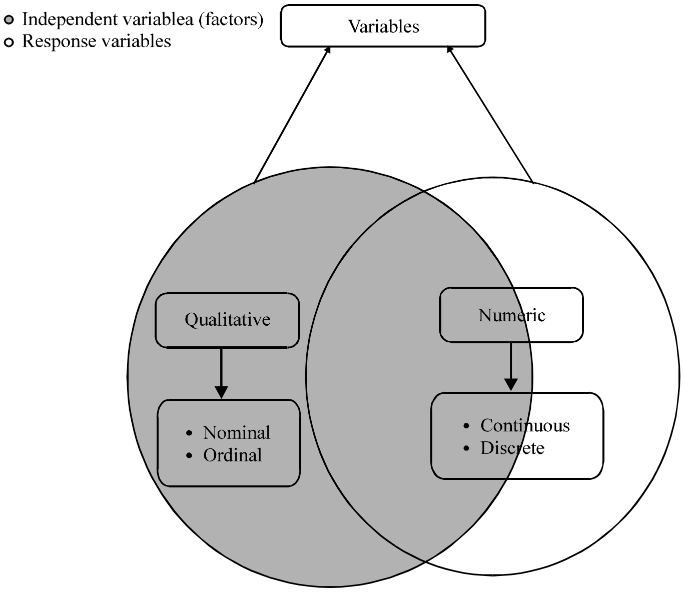


Fig. 1: Classification of variables and most frequently types of factors and response in bioprocesses

interactions with responses of the system, is desired. This is used for rapid variable screening, understanding and optimization of systems without phenomenological models associated. This general statistical tool has been boosted by Food and Drug Administration through Process Analytical Technology (PAT). PAT aims to enhance the necessary deeper understanding of the manufacturing bioprocess in the pharmaceutical field, with the purpose of adjusting manufacture online and eliminating delays in product release. However, the applications of PAT might be helpful in any other bioprocess application, beyond pharmaceutical industry (Mandenius and Brundin, 2008).

For first two types of problems, it is critical the definition of factor levels (values of independent variables) under study (Ryan, 2007). The correct definition of them can be performed by a previous revision of literature about the systems of interest or similar, through theoretical considerations or preliminary experiments. When this step is done correctly, time and resources are saved and properly results can be achieved.

Another important moment (for problems type 1 and 2) when researchers are programming their experimental work in bioprocess on lab-scale is the definition of variable types included in the system under study. The type of variable defines the appropriated statistical design and data analysis technique. Variable classification is divided in two large groups: qualitative and numeric variables (Mendenhall *et al*., 2009). Most of the problems in bioprocess include numeric variables for process responses, either continuous or discrete (e.g., yield, productivity and variables related with final product quality) and for independent variables, qualitative factors are also included, for instance: types of carbon and nitrogen sources in culture medium (Fig. 1).

Nowadays, bioprocesses are well automated and they have a comprehensive data collection and archiving. These archives represent an enormous opportunity for using multivariate methods and artificial neural networks in order to enhance the robustness and efficiency of manufacturing processes (Charaniya *et al*., 2008, 2010; Albert and Kinley, 2001).

FAIR COMPARISON OF RESULTS

This kind of statistical problem could be classified according of the number of factors under consideration that could have impact in outcome parameter(s), their number corresponding levels and if levels (samples) are independent or paired. Independent samples belong to different populations; no connection exists among experimental units. On the other hand, paired samples are samples which each data point in the first sample is uniquely matched to a data point in the subsequent samples (Gerstman, 2008).

When one independent variable is being investigated, the proper experimental design to detect difference among levels is completely randomized (one-way) design (Michelson and Schofield, 1996). As a rule the sample size (N) for assessing each level or treatment is the same and N is greater or equal to three. Larger sample size is better but this will depend on cost criteria and time. These problems with one variable are subdivided according to number of levels, because they define statistical technique to analyze experimental data. If only two levels are assessed two-sample t-test or its equivalent non-parametric Mann-Whitney test could be used. Problems with three or more levels are solved by one-way Analysis of Variance (ANOVA) or the Kruskal-Wallis non-parametric test (Michelson and Schofield, 1996; Aleman *et al.*, 2007). In bioprocess experimentation, parametric test are more frequently used (Table 1). If statistical differences are detected (dispersion caused by levels is higher than dispersion caused by experimental errors or noise) the other step is for determining which levels or treatments differ from each other. For this purpose a multiple comparison procedure (multiple range tests) must be applied. Among the existing methods are: least significant difference test, Bonferroni t-statistics, Tukey Honest Significant Difference (HSD) and Duncan's multiple range test (Hinkelmann and Kempthorne, 2008). Tukey HSD test is one of the most conservative methods and produces more reliable results (Compton, 2011); this makes it one of the most used multiple range test in bioprocesses. For paired samples, paired t-test and two-way ANOVA are useful to perform fair comparisons with two and three or more levels, considering only one variable as (Rao, 2007).

In problems which are influenced by two or three factors with several levels each one, randomized complete blocks and Latin squares and related designs are used (Kowalski and Montgomery, 2011). In principle, these experimental strategies assume that interactions among variables are nonexistent. The data analysis using these statistical tools could be carried out by two, three-way ANOVA (Henderson, 2011; Taneja, 2009) (Table 1). If any factor significantly modified the response under study, a multiple range test must be performed to detect differences among levels of this factor.

Over the course of last decade, several papers in the bioprocess field have used any of these statistical techniques for fair comparison of results. For instance, a completely randomized design was performed to define the effects of three levels of one independent variable, tween 80

Table 1: Main experimental designs for fair comparison of results used in bioprocess experimentation

Experimental design	Parametric methods for data analysis	Type of sample	No. of factors	No. of levels
Completely randomized	t-student test	Independent	1	2
	Paired t-test	Paired		
	One-way ANOVA	Independent		3 or more
	Two-way ANOVA	Paired		
Randomized complete blocks	Two-way ANOVA	Independent	2	2 or more
Latin squares	Three-way ANOVA	Independent	m	n

Table 2: Latin square designs (3 factors and 3 levels) used to optimize medium composition in RB 5 decolorization by *Funalia trogii*

	X_1	X_2	X_3
Y_1	Z_1 (1)	Z_2 (2)	Z_3 (3)
Y_2	Z_3 (4)	Z_1 (5)	Z_2 (6)
Y_3	Z_2 (7)	Z_3 (8)	Z_1 (9)

X, Y and Z are independent variables (factors), for first experiment: Carbon, nitrogen and phosphate sources and for the second one: Concentration of each selected ingredient from the first Latin square, RRB 5: Reactive black 5

percentage (0, 0.1, 1.0%) on growth, lipid accumulation and fatty acid composition in *Thraustochytrium aureum* (response variables). Sample size for each assessed tween 80 percentage was 3. Data were statistically compared using one-way analysis of variance (ANOVA) and significant differences were identified by Tukey's test (Taoka *et al*., 2011).

As another example, a completely randomized block design was utilized to determine effects of two factors in pulse ultraviolet light technology (distance from the central axis of the lamp and exposition time) on major allergens mitigation. Both variables were studied at 3 levels each, 10.8, 14.6 and 18.2 cm and 2, 4 and 6 min, respectively. Comparison of level for each variable was done by the Tukey's test. It was demonstrated that reduction of the protein band intensity for peanut allergens increased with treatment time but decreased with increased distance from the pulse ultraviolet light lamp (Yang *et al*., 2011).

In one of the papers where Latin square method was used, two applications were performed in order to define the better medium composition for decolorizing of reactive black 5 (RB 5), a wastewater contaminant from textile industry, by *Funalia trogii*. Firstly, it was carried out a Latin square design to optimize three media components (carbon, nitrogen and phosphate source) at three levels of them (carbon: fructose, glycerol, starch; nitrogen: ammonium tartrate, yeast extract, peptone and phosphate: KH_2PO_4, $KHPO_4$, Na_2HPO_4). The concentration of each element resource was kept constant. Secondly, another Latin square design was performed to optimize concentrations of selected medium ingredients from first experimental plan. The variables were fructose, peptone and Na_2HPO_4 concentrations at three levels each. The response variable in both cases was decolorization percentage of RB 5. Nine combinations in both designs were done (Table 2). Authors did not consider interactions among components for reducing the experimental work and avoid inaccurate results, this justify Latin square design choice. No statistical data analysis technique was performed to define better ingredient sources and their concentrations; Latin square design was just employed for programming experimentation (Park *et al*., 2007).

STATISTICAL MODELING OF LITTLE KNOWN SYSTEMS

When we are interested in getting a rapid mathematical modeling of little known systems in bioprocess field in order to screen or optimize factors which are suspected to be important, design of experiment methodology (DoE) have been extensively used. DoE is a collection of predetermined setting of the process variables of interest which provides an efficient procedure for programming experiments (Lee and Gilmore, 2006). This general approach after experimentation allows definition of mathematical relationships between input (factors) and output (response) variables of a given system. Besides, with DoE the effects of input variables interactions can be studied as considered factors can change simultaneously and experimental biases are avoided (Mandenius and Brundin, 2008).

As a rule when it is trying to optimize a new system, firstly it is carried out experimental designs for factors screening and after that it is performed experimental plans for optimizing the most significant factors (Xu *et al.*, 2010; Arutchelvi *et al.*, 2011). This strategy reduces the required number of experiments and as consequence experimentation cost and time. Below, it will be addressed the experimental designs for screening and optimizing process variables in theory and their applications in bioprocesses.

SCREENING

Among the most used experimental designs for screening process variables in bioprocesses are: two-level factorial, fractional factorial and Plackett-Burman designs. The first design allows for the estimation of all factor effects and all interaction effects among factors. The application of two-level factorial design is limited for high number of independent variables because the experimentation could require a considerable time and financial resources (Sower, 2011). The number of experimental runs is defined by all combinations among factor levels (2^k), where k is the number of factors, 2 is the number of levels. Generally, this experimental plan is used up to 5 independent variables, where the runs number associated with this design is 32 (2^5) (Eriksson *et al.*, 2008). The response is described as a polynomial function which is defined according to the selected design and the exploration of experimental domain is peripheral (Fig. 2).

Both experimental designs for screening and optimizing variables, the actual values of levels are usually changed to a scale from -1 (minimum value) to 1 (maximum value), in order to eliminate effects of different variable ranges; facilitate the data analysis and inferences (Mills *et al.*, 2010). Besides, repetitions of experimental points included in statistical design or in the central point of the experimental domain are performed to evaluate the statistical significance of model and parameters (Brue and Howes, 2005; Santos *et al.*, 2011).

When the number of factor to asses is large, two-level fractional factorial or Plackett-Burman designs are applied. Fractional factorial design is a fraction of an original two-level full factorial design for a defined number of factors. The reduction of experimental runs with these experimental plans sacrifice the non-confused effects determination of individual factors and their interactions. Therefore, they are applied when it is assumed that some interactions between factors are not significant. Two-level fractional factorial designs are classified in resolution III, IV or V, according

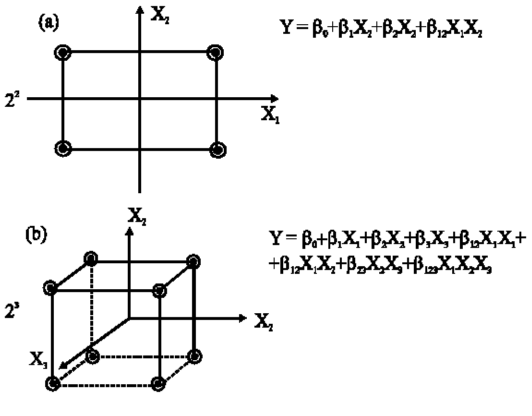

(a)
X_2
$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_{12}X_1X_2$
2^2
X_1
(b)
X_2
$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_{12}X_1X_2+$
$+\beta_{13}X_1X_3+\beta_{23}X_2X_3+\beta_{123}X_1X_2X_3$
2^3
X_2
X_3

Fig. 2(a-b): Experimental domains and runs distribution for two-level full factorial design as well as their corresponding polynomial model with (a) Two factors and (b) Three factors

Table 3: Resolutions of two-level fractional factorial design and their respective grades of confusion among interactions

Resolution	Aliasing of interactions
III	Evaluate the main effects, which are confounded with two-factor interaction
IV	Evaluate the main effects and confounded two-factor interactions (there is aliasing of the two-factor interactions)
V	Evaluate the main effects and two-factor interactions independently (there is aliasing of the two-factor with three-factor interactions)

the confusion among interactions (Table 3) (Breyfogle, 1992). In general fractional factorial designs with resolution V are widely used because rarely interactions of three-factors have significant influence on system responses.

Plackett-Burman design is another experimental plan where factors are studied at two-levels. It is also useful for screening factors, when the number of them is fairly large. This experimental strategy shows resolution III. The number of experiment is a multiple of four. Hence, designs exist for 4, 8, 12, 16, 20 and 24 etc., experiments. The number of experiments exceeds the number of factors, k, by one. Standard Plackett-Burman designs exist for 7, 11, 15, 19 and 25 factors. In cases where the number of experimental factors is less than number defined for standard design (a multiple of 4 minus 1), the final factors are dummy ones (Brereton, 2003).

After experimentation for screening variables and also for optimizing (next section), polynomial models are adjusted by matrix calculations. In order to demonstrate and improve the quality of statistical model, four tests are mainly performed:

- Goodness of fit
- Lack of fit
- Statistical significance of model coefficients
- Residues analysis

Goodness of fit test describes how well current runs can be reproduced in the proposed polynomial model. Lack of fit allows for determining whether or not errors associated to model for experimental points not included in basic design are similar to experimental errors or noise (Onsekizoglu *et al*., 2010). On the other hand, the significance of model coefficients is useful to identify the independent variables with real effects on responses, to reduce the polynomial equation and to facilitate the inferences about the system under study (Deming and Morgan, 1993). Finally, residues analysis should be performed to confirm no patterns as well as if they are normally distributed to ensure that residuals are randomly placed around the model (Harry *et al*., 2011), otherwise a mathematical transformation response values could be necessary in order to improve model fit to experimental data.

Screening experimental designs are applied in bioprocess area as first step in two-stage methodology for optimizing culture media (ingredient concentrations), process variables (temperature, time, agitation, linear flow, etc.) or a combination of both. The other remarkable applicability of this type of experimental plans is to demonstrate process robustness or to validate process in the biopharmaceutical industry (Jakobsson *et al*., 2005). In general, more than three factors are included in these two-level designs when they are used in bioreactions or downstream procedures.

As example: Plackett-Burman design was utilized during the first step for optimizing production of lipases from a newly isolated *Penicillium* sp. Six variables at two-levels were considered in 12 experiments, three repetition on the center of the experimental domain were added.

Independent variables were temperature (28-36), inoculums (25-125 ml L^{-1}), peptone (20-80 g L^{-1}), yeast extract (5-30 g L^{-1}), NaCl (5-30 g L^{-1}) and olive oil (10-30 g L^{-1}). The response variables were lipase activity after 48, 72 and 96 h of fermentation. Temperature, olive oil concentration and NaCl concentration, were the significant factors on lipase activity, with negative effects for first variables and positive for the last one. In optimization step, temperature was set at 28°C because previous results confirmed similar optimal temperatures. Olive oil and NaCl concentration were studied deeply in the optimization design according to physiological criteria (Wolski *et al.*, 2009).

Another example of screening experimental design was the application of a fraction factorial design (Resolution III) as a first stage to optimize culture media for production of phycobiliprotein by *Synechocystis* sp. PCC 6701. Seven factors (KNO_3, $NaNO_3$, $Na_2H_2PO_4$, Na_2HPO_4, $Ca(NO_3)_2$, FeEDTA, $MgSO_4$ concentrations) with two levels were studied. Addition two central point replicates were included to original design (8 runs in duplicates = 16 runs), 18 runs were performed. The response variable was specific growth rate. Nitrate and phosphate were identified as significant factors. The other factors were discarded for the optimization step (Hong and Lee, 2008).

SURFACE RESPONSE METHODOLOGY

Full factorial three-level, Central composite and Box-Behnken designs for symmetrical domains have been extensively used as surface response methods of choice in bioprocess. The central goal in these methods is to find the best combination of factors values in order to optimize the system response(s).

Full factorial three-level are used for investigating two or three factors. In the case of many factors the same problem as with a two-level design arises, the number of experiments becomes high. They can be represented in the same way as previous described for two-level designs, 3^k, k means number of factors and 3 is the number of levels (Otto, 2007; Fernandez-Nunez *et al.*, 2011). The number of experiment is 3^k. For instance, in problem with 3 factors, 27 experimental runs must be carried out (Fig. 3).

When high number of factors is required for system statistical modeling with optimizing purposes or the experimentation cost is elevated, central composite and Box-Behnken design should be used.

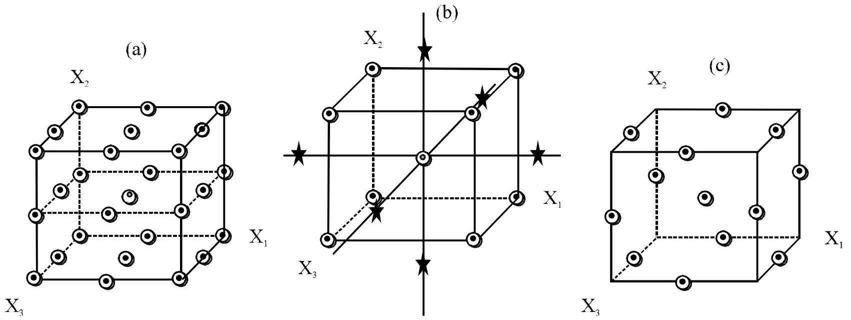


Fig. 3(a-c): Geometrical representations of experimental points arrange for (a) Full three-level, (b) Central composite and (c) Box-Behnken designs with three factors

Central composite design is generated from full factorial two-level design link to a star design. The number of run (n) is calculated by the following equation:

$$n = 2^k + 2k + n_o$$

where, n_o is the number of runs in the geometrical center of the experimental domain. For three factors, the number of experiments is 15, when repetitions number of the central point is 1 (Fig. 3). Most of the composite central designs with applications in bioprocesses are circumscribe, the distance of the star points (α) from the center can be calculated by this equation: $\alpha = 2^{k/4}$ (Anderson and Whitcomb, 2004). The experimental disadvantage of this design is the position of star point outside the hypercube, as a consequence the number of levels is higher than 3 (Fig. 3). The alternative of solution for this drawback is the Box-Behnken design.

The experimental points in Box-Behnken experimental plan lie on a hypersphere equidistant from the center point. The total number of experiment is even lower than those required in central composite designs because need few factor combinations (Otto, 2007) (Fig. 3).

Quadratic polynomials are adjusted for data provided by these experimental plans, for this reason quadratic terms for each factor are included in equation describing response(s) (Anderson and Whitcomb, 2004). This allows the local maximum and minimum detection.

Several works using three levels factorial design for two or three factor in bioprocess have been published. As example a 3^2 full factorial design was conducted to locate the optimum concentrations of yeast extract (2.5-10 g L^{-1}) and peptone (2.5-10 g L^{-1}) for the maximum cell growth of *Bacillus fusiformis* CICC 20463, used to transform 4´-demethyl epipodophyllotoxin (DMEP) into 4´-demethyl epipodophyllic acid (DMEPA), at the same DMEP consumption and DMEPA accumulation were considered as response variables (y) too. Three y variables were modeled by quadratic polynomial, the number of experimental points was 10, one central point was added to 9 initial runs for a 3^2 design. Each model was statistically analyzed and non-significant coefficients were eliminated. As example: the DMEP consumption was defined by the following Eq. 1. Then, each combination for maximum cell growth, DMEP consumption and DMEPA accumulation were defined (Tang *et al.*, 2010):

$$Y_{DMEP} = 23.64X_1 - 1.83X^2_1 \tag{1}$$

where, Y_{DMEP} is the DMEP consumption response and X_1 is the yeast extract concentration.

Central composite designs are frequently used in bioprocess problems too. Exemplifying, this type of experimental plan was applied to model the purification of α-amylase from the cultivation of *Bacillus subtilis* in a polyethylene glycol-citrate aqueous two-phase system. The PEG3350, citrate and sodium chloride concentrations were selected as variables to evaluate partition coefficients of α-amylase, total protein, purification factor and α-amylase yield. Value of α for this problem (k = 3) was 1.68 ($2^{k/4} = 2^{3/4}$). The experimental plan was composted by 8 runs associated to the factorial design, 6 runs associated to the star points and 5 repetitions of the central point of the experimental domain. At the end of the work, the optimal values for three factors determined by the statistical models were confirmed experimentally (Zhi *et al.*, 2005). This is an important step when the results on small scales based on empiric models are going to be used on large-scale in order to decrease risks in scale-up decisions.

In general the Box-Behnken designs are suggested in bioprocess field when experiments are costly or there is not much time for a large number of experimental runs. For instance, Box-Behnken design was used to optimize the levels of four factors (maltose, beef extract, $MgSO_4$ concentrations and incubation time) for chitosanase production by *Bacillus* sp. RKY3 (Wee *et al*., 2009).

MIXTURE DESIGN

Mixtures designs constitute a category within the experimental designs, specifically fitted to problems where factors are proportions or fractions (mass, volumetric and molar). They have found applications in different areas of knowledge. However, the number of papers with this statistical tool for optimizing pharmaceutical formulations is significant (Gabrielsson *et al*., 2002). In bioprocess, mixture designs have been also utilized for optimizing proportion of components of culture media, mainly in bioremediation processes (Prakasham *et al*., 2009) and they could be also utilized to seek best composition of liquid mixtures for metabolites extractions or chromatographic mobile phases in downstream steps. It is prudent to clarify that mixtures design cannot be applied for optimizing culture media if components are expressed in concentrations units, in these cases experimental designs with process variables (screening and surface response) are suitable.

The factors in this type of problems are not totally independent like those discussed for problems with process variables. Fractions (x_i) must sum to unity and then the effective number of variables is the total number minus 1 (Cornell, 2002). This has implications in the domains under investigations. The geometrical domains in mixture problem are different to those for process variable problem with the same number of factors, for this reason the experimental point arrange discussed in previous sections cannot be used in mixture problems. For instance, mixtures with three components are represented as equilateral triangle whereas a problem with three process variables defines a cube (Fig. 4) (Fernandez *et al*., 2008a).

Problems with mixtures are divided in two main categories: without constrains and with constrains. The first one encompasses mixtures in which all components (factors) are studied in

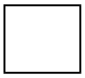
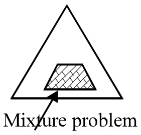
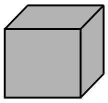
	Geometrical Domains	
No. of factors	Mixture	Process variables
2		
3	 Mixture problem with constrains	

Fig. 4: Geometrical domains related to two or three factors for problems with mixture and process variable. Arrow indicates a mixture problem with constrains, describing no similar experimental area respect to original domain without constrains

0-1 range. They are defined by symmetric geometrical areas (Fig. 4). Nevertheless, mixtures with ingredients constrains ($a\leq x_i\leq b$; $a\neq 0$ or $b\neq 1$) could generate irregular experimental areas (Fig. 4), as a consequence different statistical design must be applied in each case. Simplex lattice designs and simplex centroid designs are utilized in problems without constrains whereas D-optimal designs are the experimental plans of choice for modeling response in mixtures studies with constrains describing irregular areas (Fernandez *et al.*, 2008a). When formulation with component constrains lead to an experimental area being analogous to original ($0\leq x_i\leq 1$), simplex lattice and simplex centroid designs can be used.

The mathematical models for responses in mixtures could be linear, quadratic, special cubic and cubic. The linear models are usually employed for screening mixture ingredients and the others are used to find the best combination of factors which satisfy the response(s) goal (Cornell, 2002). The decision among quadratic, special cubic and cubic model is taken balancing the level of precision for describing response and the experimental time and cost.

In bioprocess, non-constrains problems have been the most discussed. As example, an augmented simplex-centroid design for three components was used to improve glutaminase production in solid-state fermentation by isolated *Bacillus* sp. RSP-GLU. Wheat bran, Bengal gram and Palm seed fiber were de ingredients under study and the response variable was glutaminase activity. A quadratic model described satisfactorily glutaminase activity. Two combinations of factors were identified in order to maximize glutaminase production, 100% of Bengal gram husk or a mixture of Bengal gram husk and wheat bran (66:34%) (Sathish *et al.*, 2008).

Sometime, bioprocess professional need to define a polynomial equation for describing a response as a function of ingredients proportions and process variables. In this case, crossed experimental design is recommended (Lee and Gilmore, 2006). On the other hand, there exist problems where a mixture inside another mixture should be optimized; in this situation crossed mixture design must be applied (Delaroza and Scarminio, 2008). Recently, the last statistical tool was used for improving a recombinant protein in continuous culture (Didier *et al.*, 2009).

In near future, mixture design could be more utilized in bioprocess because there are many challenges involving optimization of ingredient proportions like definition of culture media and composition of liquid phase in bioseparations.

UNIFORM DESIGN

Uniform designs were created by Fang at the end of 70's to solve an industrial problem where 6 factors with at least 12 levels each should be considered. Nevertheless, the experimental runs could not exceed 50 because of cost limitations. Using, the classical experimental designs for polynomial modeling of systems described in previous sections, the number of runs generated by these techniques would have been impossible for carrying out. Then, this problem was overcome using Uniform design with only 31 experiments and each factor with 31 levels (Liang *et al.*, 2001). This experimental plan produces uniform scattering of the design points over the experimental domain (Fang and Chan, 2006). If point distribution in screening and surface response experimental designs inside experiment domain is observed, it will be noticed that most of the points are placed on the periphery of experimental domain. Therefore, the sample to be experimented for generating inferences could be not representative. Classical experimental designs show uniform scattering of designs point in their dimensions but not in the geometrical area defined by factor and their levels. One of the most important advantages of Uniform design over traditional experimental design is that, even when factors number and their corresponding levels are large, the experiment can be performed in a relatively small number of runs (Fang and Chan, 2006).

Most of experimental designs are based on model assumptions. However, experimental plans insensitive to this assumption are desired. In other words, changes in the underlying distribution or model should cause small change in the performance of the design (Liang *et al.*, 2001). Uniform designs do not have a mathematical model associated, thus it is particular suitable for studying systems with an unknown underlying model. Artificial neural networks and multiple linear regressions could be used to model non-linear and linear systems, respectively, after experimentation and data collection with uniform designs.

For programming experimentation (point arrange) in common problems by means of uniform design, softwares and tables (http://www.math.hkbu.edu.hk/UniformDesign/) are available (Sun *et al.*, 2010; Zhou *et al.*, 2011). The choice of particular experimental [U_n (q^s)] is defined by number of runs (n), number of factors (s) and number of levels (q). Both process and mixture variables can be studied through this kind of experimental design (Fang and Chan, 2006).

Uniform design is very attractive for statistical system modeling in bioprocess where the experimentation is very expensive. In general, it has been used for optimizing concentration of culture medium components and process variables related to bioreactions step (Hua and Xu, 2011; Wei *et al.*, 2009; Xu *et al.*, 2006).

As example: A uniform design was used to optimize medium composition in order to improve the ethanol tolerance of self-flocculating yeast. Seven component were studied at 6 levels each in experimental plan with 12 runs [U_{12} (7^6)]. The ingredients were vitamins (x base), $(NH_4)_2SO_4$(g L^{-1}), K_2HPO_4 (g L^{-1}), $MgSO_4.7H_2O$ (g L^{-1}), $CaCl_2.2H_2O$ (g L^{-1}), $ZnSO_4.7H_2O$ (mg L^{-1}), $CoCl_2.6H_2O$ (mg L^{-1}) and the response variables were viability and ethanol concentration. The statistical model was adjusted by means a linear-regression method. The optimization of medium component led to 90.2% of cell viability which demonstrated noteworthy improvement of ethanol tolerance of the self-flocculating yeast (Xue *et al.*, 2008).

MULTIPLE RESPONSE OPTIMIZATIONS

Finding optimal values of factors is a major goal in statistical modeling of little known systems, in general research and therefore in bioprocess field too. However, in most of the cases systems need to be evaluated by more than one response variable (Hendriks *et al.*, 1992). Sometimes models for response differ greatly and then optimal values could be quite different for each particular response. Besides, outcome parameters could have dissimilar positive criteria on the system, for instance: in group of response variable, one of them would need to be maximized and another could need minimized or set in a range to guarantee the best system performance comprehensively. Thus, professionals in bioprocess research and development area must be trained in multiresponse optimization. The main issue related to this kind of optimization is to make a selection out of a set of factor values that result in a compromise solution of a multicriteria problem (Hendriks *et al.*, 1992).

A large number of statistical techniques with this purpose are now available. These include: overlay plots, pareto optimality, utility functions, desirability functions and methods based on the standardized Euclidian distance between the predicted value of each response and the optimum one that was obtained individually (Hendriks *et al.*, 1992; Sivakumar *et al.*, 2007; Takayama *et al.*, 2003). Among them, the most employed in bioprocesses is the desirability function.

Desirability function includes all the response variables into a single function in order to consider all of them simultaneously. This method requires minimum and maximum acceptable values for each outcome variable. The individual response can be normalized to the desirability functions (d_k, k = 1, 2, 3,..., n) with values inside the interval [0,1] using the distance between minimum and maximum acceptable values. The d_i functions are then combined to obtain a global desirability function D which should be maximized choosing the best conditions of the designed variables. D is calculated by the following equation (Takayama *et al.*, 2003; Moreira *et al.*, 2007):

$$D = (d_1^{r1} \times d_2^{r2} \times d_3^{r3} \times \ldots. \times d_n^{m})^{1/(r1+r2+r3 \cdots m)}$$

where, r_i is the relative importance assigned to each response.

The application of this method for multiresponse optimization is performed mainly in bioprocess for defining best values of factors associated to bioreaction step (Moreira *et al.*, 2007; Liu and Tang, 2010). For instance: Desirability function was used to find the best combination of four medium component concentrations (sucrose, yeast extract, peptone, Mg^{2+}) in order to maximize three response variables: dry cell weight, extracellular and intra cellular polysaccharides. The relative importance for three outcome variables was the same. The optimal values were 73, 11, 8 g L^{-1} and 46 mM of sucrose, yeast extract, peptone and Mg^{2+}, respectively. Under these conditions the predicted values for dry cell weight, extracellular and intra cellular polysaccharides were 24.50, 4.10 and 3.20 g L^{-1}, respectively (Liu and Tang, 2010).

HOW TO TAKE ADVANTAGE LARGE VOLUME OF DATA

The statistical techniques included in categories: fair comparison of results and statistical modeling of little known systems are applied almost always on lab scale, for further application on large-scale, scale-up criteria should be considered. However, nowadays modern bioprocess plants are equipped with proper automation systems for capturing and recording material input, process output, control action as well as physical parameters (Charaniya *et al.*, 2008; Alford, 2006). These data might provide the cause of process response variables fluctuations. Besides, industrial bioprocess data allows for improving process robustness and efficiency (Charaniya *et al.*, 2008).

To identify novel and useful relations and patterns that associate process factors with different process response, a methodology involving four interactive steps is commonly used. These steps are: data processing, feature selection and/or dimensionality reduction, data mining and expert analysis for interpretation of the results. Multivariate methods and artificial neuronal networks have been extensively used in dimensionality reduction and data mining steps (Charaniya *et al.*, 2008, 2010).

Among multivariate methods for improving bioprocesses from production plant data, Principal Component Analysis (PCA) and Partial Least Squares (PLS) are the most used (Charaniya *et al.*, 2010).

Specifically, PCA is applied to get a quick overview and to detect deviations from the desired bioprocess behavior. Its main purpose is to reduce the dimension of the space of the inter-correlated variables to a lower dimensional representation space with less correlated variables. The PCA is

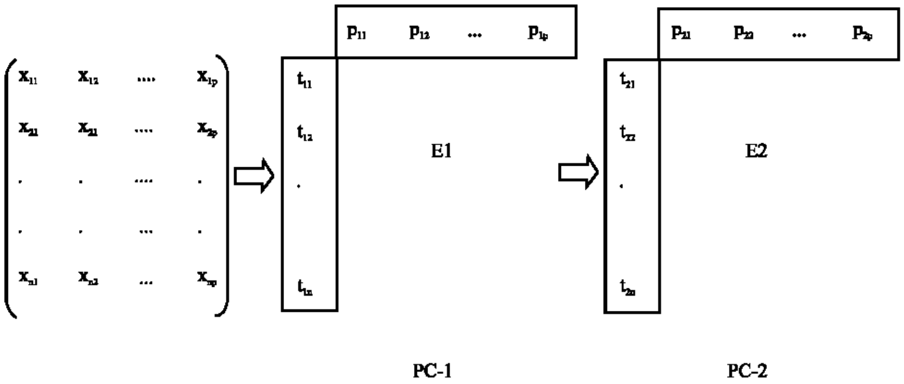


Fig. 5: Principal component analysis: representation of matrix X decomposition in principal components (PC), E: Matrix containing residuals

applied to bi-dimensional data structures (matrix-X) defined by p columns (number of variables) and n rows (number of observations). Matrix-X is discomposed into a number of principal components that maximize explained variance in data on each successive component under the constraint of being orthogonal to the previous principal components (Nucci *et al.*, 2010). A bilinear model is the result of this latent variable projection method, a product of scores T and the loadings P matrices (Fig. 5). Two main graphics are generated from PCA, score and loading plots. Each observations gets a score value on each principal component, then observations can be represented in scores plots. This graph reveals clusters, trends and outliers in data. Similarly, variables are represented in loading plots, allowing to detect correlations among variables and to interpret patterns observed in the score plot (Rajalahti and Kvalheim, 2011). For instance: PCA was applied to identify key variables from cell culture data related to a fed-batch culture of mouse hybridoma by reducing their dimensionality. For each observation was defined: cell density, concentration of amino acids, major carbon sources and by-products. PCA identified three major clusters characterized by a particular amino acids consumption/production rates (Selvarasu *et al.*, 2010).

Regression method based on PLS in bioprocess has been used to identify predictive correlations between output parameters (Y) and factors (X) as well as detect process abnormalities (Charaniya *et al.*, 2008). PLS takes PCA one step further, as it deals with two matrixes (X and Y). Each observation is represented in both spaces X and Y. In PLS, firstly a PCA is carried out for both the descriptive (X) and the response (Y) variables. Then, the best correlation between X and Y is calculated using least square technique. Therefore, the resulting PLS model is not always the best description of X and Y but rather of the relation between them. Summarizing, the emphasis in PLS is both on the correlation between X and Y and a good description of X and Y (Gabrielsson *et al.*, 2002). As a bioprocess example: a PLS regression was used to calibrate models in order to predict volatile fatty acid concentrations from near infrared spectra obtained in biogas test plants. These Volatile Fatty Acids (VFA) should be well controlled to guarantee a stable biogas production. The model displayed acceptable to very good prediction performances for total VFA as well as for three other essential individual acids based on test set validations. This PLS application was useful for improving the bioreaction control (Holm-Nielsen and Esbensen, 2011).

Another predictive approach used to analyze bioprocess data is Artificial Neural Networks (ANN). Especially, it has been applied to predict the output of a fermentation process as a nonlinear function of the process inputs. This artificial intelligence approach can be used together with optimization methods to identify the optimal factor values to maximize the desire process response (Charaniya *et al*., 2008). The inputs to the model are process measurements, then data are weighted individually or in groups, after that combined using a nonlinear activation function at a node referred to as a neuron. A process ANN model is often composed of an input layer, an output layer and one or more hidden layers of nodes (Komives and Parker, 2003). There are many types of ANN, the most often used ANN is a fully connected, supervised network with backpropagation learning rule. This type of ANN is excellent at prediction and classification tasks (Agatonovic-Kustrin and Beresford, 2000). A concrete example of ANN application in bioprocess for taking advantage large volume of data was the satisfactory dynamic modeling of biochemical oxygen demand, chemical oxygen demand, suspended solid and total nitrogen removal in a Waste Water Treatment Plant (WWTP) using a data set collected from a full-scale WWTP (Lee *et al*., 2011).

COMBINING STATISTICAL MODELING OF LITTLE KNOWN SYSTEMS AND MULTIVARIATE METHODS

To finish this review, it is appropriated to note that a useful statistical tool with wide applications in pharmaceutical industry, it has not found the same application in bioprocess, the multivariate designs. They are a combination of classical methods for modeling of little known systems and multivariate methods (PCA and PLS). In pharmaceutical formulation, this experimental strategy has been used to optimize type of ingredients and their concentrations (Fernandez *et al*., 2010). Ingredient types for the same purpose were characterized by several descriptive variables and then PCA was applied to turn this categorical variable into numeric variables (scores) (Gabrielsson *et al*., 2002). Subsequently, a higher number of ingredients could be considered, even for new ones non evaluated, the developed models could estimate the responses associated to the system under study. This application would be interesting for culture media optimization where most of the time different resources of nutrients are assessed.

Another application of the multivariate design is related to multi-stages processes which has not frequently used in bioprocesses. In general, most of the time just one system unit is optimized and it is not considered the system as a whole. Multivariate design allows for considering the previous step by means of latent variables generate from PCA and integrate them to the own variables of the system unit being optimize (Bergman *et al*., 1998). This would be very useful to connect bioreaction and downstream steps. Specially for connecting bioreaction with the first downstream operation unit.

INFERENCE

In this review, the statistical tools were classified so that professionals with low statistics expertise involved in bioprocess research can choose the correct technique for a similar investigation problem. Figure 6 is a graphical guide for this purpose. Examples from literature were also commented in order to help the understanding of theoretical issues of each statistical technique. Besides, some statistical tools without wide application at the moment in bioprocess were also discussed, taking into consideration the near future use of them.

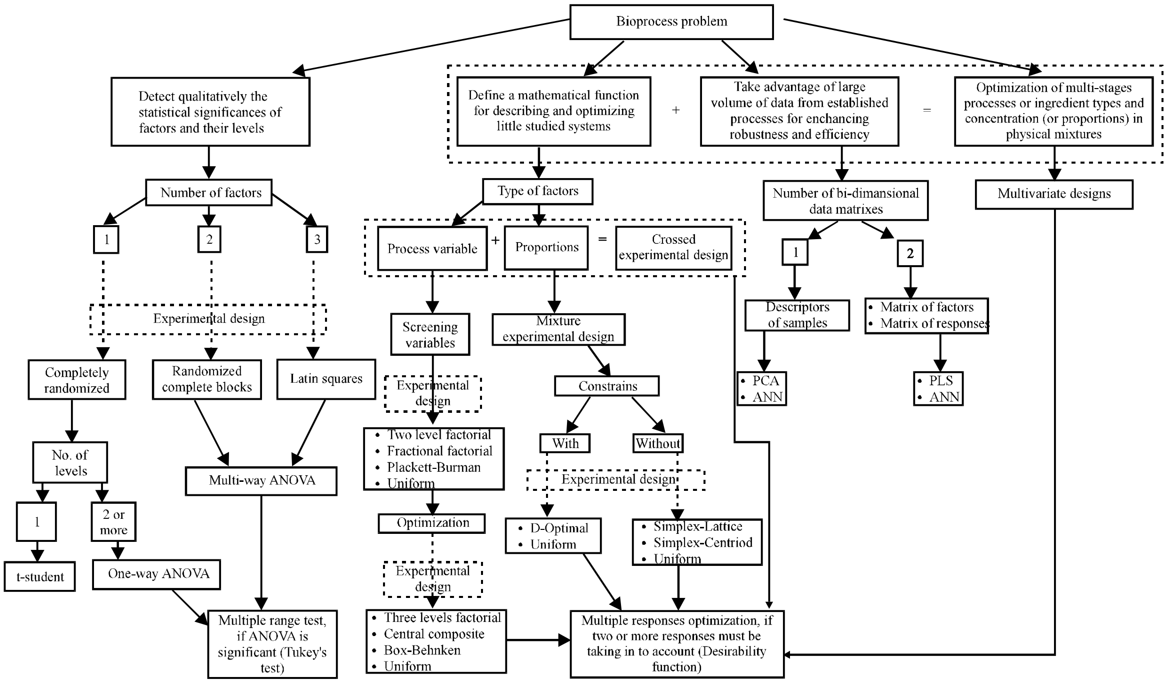

Bioprocess problem
Detect qualitatively the statistical significances of factors and their levels
Define a mathematical function for describing and optimizing little studied systems
+
Take advantage of large volume of data from established processes for enchancing robustness and efficiency
=
Optimization of multi-stages processes or ingredient types and concentration (or proportions) in physical mixtures
Number of factors
1
2
3
Experimental design
Completely randomized
Randomized complete blocks
Latin squares
No. of levels
1
2 or more
t-student
One-way ANOVA
Multi-way ANOVA
Multiple range test, if ANOVA is significant (Tukey's test)
Type of factors
Process variable
+
Proportions
=
Crossed experimental design
Screening variables
Experimental design
• Two level factorial
• Fractional factorial
• Plackett-Burman
• Uniform
Optimization
Experimental design
• Three levels factorial
• Central composite
• Box-Behnken
• Uniform
Mixture experimental design
Constrains
With
Without
Experimental design
• D-Optimal
• Uniform
• Simplex-Lattice
• Simplex-Centriod
• Uniform
Multiple responses optimization, if two or more responses must be taking in to account (Desirability function)
Number of bi-dimansional data matrixes
1
2
Descriptors of samples
• Matrix of factors
• Matrix of responses
• PCA
• ANN
• PLS
• ANN
Multivariate designs

Fig. 6: Guide for choosing statistical tools in bioprocess problems

ACKNOWLEDGMENTS

First author gratefully acknowledges his wife and daughter, Relma and Giovanna as well as Ph.D. students: Bruno Oliva Oishi and Hector Aguilar Vitorino, for the inspiration to write this study.

REFERENCES

Agatonovic-Kustrin, S. and R. Beresford, 2000. Basic concepts of Artificial Neural Network (ANN) modeling and its application in pharmaceutical research. J. Pharm. Biomed. Anal., 22: 717-727.

Albert, S. and R.D. Kinley, 2001. Multivariate statistical monitoring of batch processes: An industrial case study of fermentation supervision. Trends Biotechnol., 19: 53-62.

Aleman, M.D.R., E. Noa, A. Tamayo, M. Dubed and S. Padilla *et al.*, 2007. Downstream processing: A revalidation study of viral clearance in the purification of monoclonal antibody CB.Hep-1. BioPharm Int., 20: 46-55.

Alford, J.S., 2006. Bioprocess control: Advances and challenges. Comput. Chem. Eng., 30: 1464-1475.

Anderson, M.J. and P.J. Whitcomb, 2004. RSM Simplified: Optimizing Processes Using Response Surface Methods for Design of Experiments. Productivity Press, USA., ISBN-13: 9781563272974, Pages: 292.

Arutchelvi, J., C. Joseph and M. Doble, 2011. Process optimization for the production of rhamnolipid and formation of biofilm by *Pseudomonas aeruginosa* CPCL on polypropylene. Biochem. Eng. J., 56: 37-45.

Bergman, R., M.E. Johansson, T. Lundstedt, E. Seifert and J. Aberg, 1998. Optimization of a granulation and tabletting process by sequential design and multivariate analysis. Chemometr. Intell. Lab., 44: 271-286.

Boos, D.D. and C. Brownie, 1995. ANOVA and rank tests when the number of treatments is large. Stat. Probab. Lett., 23: 183-191.

Brereton, R.G., 2003. Chemometrics Data Analysis for the Laboratory and Chemical Plant. John Wiley and Sons Ltd., USA., ISBN-13: 9780470845745, Pages: 504.

Breyfogle, F.W., 1992. Statistical Methods for Testing, Development and Manufacturing. John Wiley and Sons Inc., USA., ISBN-13: 9780471540359, Pages: 516.

Brue, G. and R. Howes, 2005. The McGraw-Hill 36-Hour Course Six Sigma. The McGraw Hill Companies, USA., ISBN-13: 9780071430081, Pages: 304.

Charaniya, S., H. Le, H. Rangwala,, K. Mills, K. Johnson, G. Karypis and W.S. Hu, 2010. Mining manufacturing data for discovery of high productivity process characteristics. J. Biotechnol., 147: 186-197.

Charaniya, S., W.S. Hu and G. Karypis, 2008. Mining bioprocess data: Opportunities and challenges. Trends Biotechnol., 26: 690-699.

Clementschitsch, F. and K. Bayer, 2006. Improvement of bioprocess monitoring: Development of novel concepts. Microb. Cell Factories, Vol. 5. 10.1186/1475-2859-5-19

Compton, M.E., 2011. Elements *In vitro* Research. In: Plant Tissue Culture, Development and Biotechnology, Trigiano, R.N. and D.J. Gray (Eds.). CRC Press/Taylor and Francis Group, USA., pp: 57-74.

Cornell, J.A., 2002. Experiments with Mixtures: Designs, Models and the Analysis of Mixture Data. 3rd Edn., John Wiley and Sons Ltd., New York, USA., ISBN-13: 9780471393672, Pages: 649.

Delaroza, F. and I.S. Scarminio, 2008. Mixture design optimization of extraction and mobile phase media for fingerprint analysis of *Bauhinia variegata* L. J. Sep. Sci., 31: 1034-1041.

Deming, S.N. and S.L. Morgan, 1993. Experimental Design: A Chemometric Approach. 2nd Edn., Elsevier Science Publishers, USA., ISBN-13: 9780444891112, Pages: 437.

Didier, C., G. Forno, M. Etcheverrigaray, R. Kratje and H. Goicoechea, 2009. Novel chemometric strategy based on the application of artificial neural networks to crossed mixture design for the improvement of recombinant protein production incontinuous culture. Anal. Chim. Acta, 650: 167-174.

Dubey, K.K. and B.K. Behera, 2011. Statistical optimization of process variables for the production of an anticancer drug (colchicine derivatives) through fermentation: At scale-up level. New Biotechnol., 28: 79-85.

Eriksson, L., E. Johansson, N. Kettaneh-Wold, C. Wikstrom and S. Wold, 2008. Design of Experiments: Principles and Applications. 3rd Edn., Umetrics Academy, USA., ISBN-13: 9789197373043, Pages: 459.

Fang, K.T. and L.Y. Chan, 2006. Uniform Design and its Industrial Applications. In: Springer Hanbook Engineering Statistics, Pham, H. (Ed.). Springer-Verlag, USA., pp: 228-248.

Fernandez, E.G., M. Fernandez, R.T. Oliveira, B. Bermudez and I. Perez *et al*., 2008a. Disenos de experimentos en tecnologia y control de los medicamentos (Design of experiments in technology and medication management). Lat. Am. J. Pharm., 27: 286-296.

Fernandez, E.G., R. Valdes, J.A. Montero, A. Figueroa and T.A. Alvarez *et al*., 2008b. Application of the partial least square technique to identify critical variables in the immunosorbent manufacturing. Chormatographia, 68: 375-380.

Fernandez, E.G., M. Fernandez, H.M. Hoang, I. Perez and D. Guerra *et al*., 2010. A multivariate strategy for tablet manufacturing optimization. Latin Am. J. Pharm., 29: 1336-1344.

Fernandez-Nunez, E.G., B.L. Faintuch, R. Teodoro, D.P. Wiecek and N.G. da Silva *et al.*, 2011. Parameters optimization defined by statistical analysis for cysteine-dextran radiolabeling with technetium tricarbonyl core. Applied Radiat. Isot., 69: 663-669.

Gabrielsson, J., N.O. Lindberg and T. Lundstedt, 2002. Multivariate methods in pharmaceutical applications. J. Chemom., 16: 141-160.

Gerstman, B.B., 2008. Basic Biostatistics: Statistics for Public Health Practice. Jones and Bartlett Publishers Inc., USA., ISBN-13: 9780763735807, Pages: 557.

Harry, M.J., P.S. Mann, O.C. De Hodgins, R.L. Hulbert and C.J. Lacke, 2011. Practitioner's Guide to Statistics and Lean Six Sigma for Process Improvements. John Wiley and Sons Inc., USA., ISBN-13: 9781118210215, Pages: 800.

Henderson, G.B., 2011. Six Sigma Quality Improvement with Minitab. 2nd Edn., John Wiley and Sons Ltd., USA., ISBN-13: 9781119976189, Pages: 528.

Hendriks, M.M.W.B., J.H. de Boer, A.K. Smilde and D.A. Doornbos, 1992. Multicriteria decision making. Chemom. Intell. Lab. Syst., 16: 175-191.

Hinkelmann, K. and O. Kempthorne, 2008. Design and Analysis of Experiments Volume 1: Introduction to Experimental Design. 2nd Edn., John Wiley and Sons, USA., ISBN-13: 9780471727569, Pages: 631.

Holm-Nielsen, J.B. and K.H. Esbensen, 2011. Monitoring of biogas test plants: A process analytical technology approach. J. Chemom., 25: 357-365.

Hong, S.J. and C.G. Lee, 2008. Statistical optimization of culture media for production of phycobiliprotein by *Synechocystis* sp. PCC 6701. Biotechnol. Bioprocess. Eng., 13: 491-498.

Hua, D. and P. Xu, 2011. Recent advances in biotechnological production of 2-phenylethanol. Biotechnol. Adv., 29: 654-660.

Jakobsson, N., M. Degerman and B. Nilsson, 2005. Optimisation and robustness analysis of a hydrophobic interaction chromatography step. J. Chromatogr. A., 1099: 157-166.

Jana, A.K., 2008. Chemical Process Modeling and Computer Simulation. Prentice Hall of India Private Limited, New Delhi, India, ISBN-13: 9788120331969, Pages: 273.

Komives, C. and R.S. Parker, 2003. Bioreactor state estimation and control. Curr. Opin. Biotechnol., 14: 468-474.

Kowalski, S.M. and D.C. Montgomery, 2011. Design and Analysis of Experiments. 7th Edn., John Wiley and Sons Inc., USA.

Lee, J.W., C. Suh, Y.S.T. Hong and H.S. Shin, 2011. Sequential modelling of a full-scale wastewater treatment plant using an artificial neural network. Bioprocess. Biosyst. Eng., 34: 963-973.

Lee, K.M. and D.F. Gilmore, 2006. Statistical experimental design for bioprocess modeling and optimization analysis: Repeated-measures method for dynamic biotechnology process. Applied Biochem. Biotechnol., 135: 101-115.

Liang, Y.Z., K.T. Fang and Q.S. Xu, 2001. Uniform design and its applications in chemistry and chemical engineering. Chemom. Intell. Lab. Syst., 58: 43-57.

Liu, R.S. and Y.J. Tang, 2010. *Tuber melanosporum* fermentation medium optimization by Plackett-Burman design coupled with Draper-Lin small composite design and desirability function. Bioresour. Technol., 101: 3139-3146.

Mandenius, C.F. and A. Brundin, 2008. Bioprocess optimization using design-of-experiments methodology. Biotechnol. Prog., 24: 1191-1203.

Mendenhall, W., R.J. Beaver and B.M. Beaver, 2009. Introduction to Probability and Statistics. 13th Edn., Brooks/Cole, USA.

Michelson, S. and T. Schofield, 1996. The Biostatistics Cookbook: The Most User-Friendly Guide for Bio/Medical Scientist. Springer, New York, USA., ISBN-13: 9780792341055, Pages: 176.

Mills, K.L., J.J. Filiben, D.Y. Cho, E. Schwartz and D. Genin, 2010. Study of proposed internet congestion control mechanisms. National Institute of Standards and Technology Special Publication No. 500-282. http://www.nist.gov/itl/antd/upload/P1-SP-500-282-Cover-Pages.pdf

Moreira, G.A., G.A. Micheloud, A.J. Beccaria and H.C. Goicoechea, 2007. Optimization of the *Bacillus thuringiensis* var. *kurstaki* HD-δ-endotoxins production by using experimental mixture design and artificial neural networks. Biochem. Eng. J., 35: 48-55.

Nair, A.J., 2005. Basics of Biotechnology. Laxmi Publications (P) Ltd., India, ISBN-13: 9788170086123, Pages: 302.

Nair, A.J., 2008. Principles of Biotechnology. Laxmi Publications (P) Ltd., India, ISBN-13: 9788131800621, Pages: 886.

Najafpour, G., 2006. Biochemical Engineering and Biotechnology. Elsevier, USA., ISBN-13: 9780444528452, Pages: 421.

Nucci, E.R., A.J.G. Cruz and R.C. Giordano, 2010. Monitoring bioreactors using principal component analysis: Production of penicillin G acylase as a case study. Bioproc. Biosyst. Eng., 33: 557-564.

Onsekizoglu, P., K.S. Bahceci and J. Acar, 2010. The use of factorial design for modeling membrane distillation. J. Membr. Sci., 349: 225-230.

Otto, M., 2007. Chemometrics-Statistics and Computer Application in Analytical Chemistry. 2nd Edn., Wiley-VCH, Weinheim, Germany, Pages: 328..

Park, C., J.S. Lim, Y. Lee, B. Lee, S.W. Kim, J. Lee and S. Kim, 2007. Optimization and morphology for decolorization of reactive black 5 by *Funalia trogii*. Enzyme. Microb. Technol., 40: 1758-1764.

Prakasham, R.S., T. Sathish, P. Brahmaiah, C. Subba Rao, R. Sreenivas Rao and P.J. Hobbs, 2009. Biohydrogen production from renewable agri-waste blend: Optimization using mixer design. Int. J. Hydrogen Energ., 34: 6143-6148.

Rajalahti, T. and O.M. Kvalheim, 2011. Multivariate data analysis in pharmaceutics: A tutorial review. Int. J. Pharm., 417: 280-290.

Rao, K.V., 2007. Distribution Free Statistical Tests of Inference. In: Biostatistics, A manual of statistical methods for use in health, nutrition and Anthropology, Rao, K.V., (Ed.). Jaypee Brothers Medical Publishers (P) Ltd., New Delhi, India, pp: 631-657..

Ryan, T.P., 2007. Modern Experimental Design. John Wiley and Sons, New York, USA., Pages: 593.

Santos, V.C., F.A. Hasmann, A. Converti and A. Jr. Pessoa, 2011. Liquid-liquid extraction by mixed micellar systems: A new approach for clavulanic acid recovery from fermented broth. Biochem. Eng. J., 56: 75-83.

Sathish, T., G.S. Lakshmi, Ch.S. Rao, P. Brahmaiah and R.S. Prakasham, 2008. Mixture design as first step for improved glutaminase production in solid-state fermentation by isolated Bacillus. Lett. Appl. Microbiol., 47: 256-262.

Selvarasu, S., Y. Kim do, I.A. Karimi and D.Y. Lee, 2010. Combined data preprocessing and multivariate statistical analysis characterizes fed-batch culture of mouse hybridoma cells for rational medium design. J. Biotechnol., 150: 94-100.

Sivakumar, T., R. Manavalan, C. Muralidharan and K. Valliappan, 2007. Multi-criteria decision making approach and experimental design as chemometric tools to optimize HPLC separation of domperidone and pantoprazole. J. Pharmaceut. Biomed. Anal., 43: 1842-1848.

Sower, V.E., 2011. Essentials of Quality With Cases and Experimental Exercises. John Wiley and Sons, New York, USA., Pages: 416.

Sun, L., C. Wang, C. Ma and L. Shi, 2010. Optimization of renewal regime for improvement of polysaccharides production from *Porphyridium cruentum* by uniform design. Bioproc. Biosyst. Eng., 33: 309-315.

Takayama, K., M. Fujikawa, Y. Obata and M. Morishita, 2003. Neural network based optimization of drug formulations. Adv. Drug. Delivery Rev., 55: 1217-1231.

Taneja, H.C., 2009. Statistical Methods for Engineering and Sciences. I.K. International Publishing House Pvt, Ltd., East Bangalore, India, ISBN: 9789380026664, Pages: 368.

Tang, Y.J., X.L. Xu and J.J. Zhong, 2010. A novel biotransformation process of 40-demethylepipodophyllotoxin to 4-demethylepipodophyllic acid by Bacillus fusiformis CICC 20463, Part II: process optimization. Bioproc. Biosyst. Eng., 33: 237-246.

Taoka, Y., N. Nagano, Y. Okita, H. Izumida, S. Sugimoto and M. Hayashi, 2011. Effect of Tween 80 on the growth, lipid accumulation and fatty acid composition of *Thraustochytrium aureum* ATCC 34304. J. Biosci. Bioeng., 111: 420-424.

Wee, Y.J., L.V.A. Reddy, K.C. Chung and H.W. Ryu, 2009. Optimization of chitosanase production from *Bacillus* sp. RKY3 using statistical experimental designs. J. Chem. Technol. Biotechnol., 84: 1356-1363.

Wei, G., X. Yang, T. Gan, W. Zhou, J. Lin and D. Wei, 2009. High cell density fermentation of *Gluconobacter oxydans* DSM 2003 for glycolic acid production. J. Ind. Microbiol. Biotechnol., 36: 1029-1034.

Wolski, E., E. Rigo, M. Di Luccio, J.V. Oliveira, D. de Oliveira and H. Treichel, 2009. Production and partial characterization of lipases from a newly isolated *Penicillium* sp. Using experimental design. Lett. Appl. Microbiol., 49: 60-66.

Xu, C.P., J. Sinha, J.T. Bae, S.W. Kim and J.W. Yun, 2006. Optimization of physical parameters for exo-biopolymer production in submerged mycelial cultures of two entomopathogenic fungi *Paecilomyces japonica* and *Paecilomyces tenuip*. Lett. Appl. Microbiol., 42: 501-506.

Xu, L.J, Y.S. Liu, L.G. Zhou and J.Y. Wu, 2010. Optimization of a liquid medium for beauvericin production in fusarium redolens Dzf2 mycelial culture. Biotechnol. Bioprocess Eng., 15: 460-466.

Xue, Ch., X.Q. Zhao, W.J. Yuan and F.W. Bai, 2008. Improving ethanol tolerance of a self-flocculating yeast by optimization of medium composition. World J. Microbiol. Biotechnol., 24: 2257-2261.

Yang, W.W., N.R. Mwakatage, R. Goodrich-Schneider, K. Krishnamurthy and T.M. Rababah, 2011. Mitigation of major peanut allergens by pulsed ultraviolet light. Food Bioprocess Technol., 10.1007/s11947-011-0615-6

Zhi, W., J. Song, F. Ouyang and J. Bi, 2005. Application of response surface methodology to the modeling of α-amylase purification byaqueoustwo-phase systems. J. Biotechnol., 118: 157-165.

Zhou, Y., S. Lakshminarayanan and R. Srinivasan, 2011. Optimization of image processing parameters for large sets of in-process video microscopy images acquired from batch crystallization processes: Integration of uniform design and simplex search. Chemometr. Intell. Lab. Syst., 107: 290-302.