0273-2289/1559-0291 (Online)/06/135/0101/$30.00

Statistical Experimental Design for Bioprocess Modeling and Optimization Analysis

Repeated-Measures Method for Dynamic Biotechnology Process

KWANG-MIN LEE* AND DAVID F. GILMORE

Arkansas State University, Environmental Sciences Program,
State University, AR 72467,
E-mail: klee0922@yahoo.co.kr

Received August 2, 2005; Accepted February 10, 2006

Abstract

The statistical design of experiments (DOE) is a collection of predetermined settings of the process variables of interest, which provides an efficient procedure for planning experiments. Experiments on biological processes typically produce long sequences of successive observations on each experimental unit (plant, animal, bioreactor, fermenter, or flask) in response to several treatments (combination of factors). Cell culture and other biotech-related experiments used to be performed by repeated-measures method of experimental design coupled with different levels of several process factors to investigate dynamic biological process. Data collected from this design can be analyzed by several kinds of general linear model (GLM) statistical methods such as multivariate analysis of variance (MANOVA), univariate ANOVA (time split-plot analysis with randomization restriction), and analysis of orthogonal polynomial contrasts of repeated factor (linear coefficient analysis). Last, regression model was introduced to describe responses over time to the different treatments along with model residual analysis. Statistical analysis of biprocess with repeated measurements can help investigate environmental factors and effects affecting physiological and bioprocesses in analyzing and optimizing biotechnology production.

Index Entries: Statistical experimental design; repeated-measures (RM); empirical (statistical) model; dynamic bioprocess; sphericity.

*Author to whom all correspondence and reprint requests should be addressed.

Design of Experiment and Biotechnology

In most bioprocesses such as fermentation and other cell culture methods (mammalian or insect), there are no true theoretical or mathematical models that can describe the whole process with 100% certainty. Because of this limitation arising from the incredible complexity of cellular metabolism, efficient empirical approaches to explain these processes are necessary to solve research problems. These statistical or empirical methods must provide lots of data to enable a researcher to reach meaningful conclusions. However, any problem-solving approach is limited by time, money, and resources for research. Because there are limited opportunities to generate and collect data, it is critical that the data be rich in information. Given these limitations, a statistically designed experiment is one solution for obtaining the information-rich data from the process being studied *(1,2)*.

Employment of well-designed data collection methods results in successful experiments. Also, designed experiments are less time-consuming and less expensive than haphazard ones regardless of overall efficiency. Design of experiment (DOE) is a systematic approach to problem solving which is applied to data collection and analysis to obtain information-rich data. DOE is concerned with carrying out experiments under the constraints of minimum expense of time, costs, and runs. A properly designed experiment is more important than detailed statistical analysis. The primary goal of designing an experiment statistically is to obtain valid results at a minimum of effort, time, and resources *(2–4)*.

Statistical experimental design has not been widely used in the biological sciences even though it has been commonly employed in many other areas such as industrial, chemical, engineering, agricultural, medical, and food sciences. The primary reason for this is that most biological research has not been involved in many manufacturing processes. However, because genetic engineering, biomaterials, and bioprocess technologies like biodegradation and bioremediation have emerged, more scientists are getting interested in experimental designs to improve their biological processes and productions by shortening time and increasing efficiencies *(5–9)*.

Bioprocess technologies require effective problem-solving methods because they involve both adjustment of multiple parameters and complications that inhibit application of engineering principles. Additional obstacles for bioprocess research include the lack of an accurate mathematical model equation to describe the whole process, high noise levels, interactions among variables, and complex biochemical reactions. These conditions call for a good strategy to deal with such a complicated system.

Statistically designed experiments use a small set of carefully planned experiments. This method is more satisfactory and effective than other methods, such as classical one-at-a-time or mathematical methods, because it can study many variables simultaneously with a low number of observations, saving time and costs. The statistical experimental design provides a universal language with which people from different areas such as

academia, engineering, business, and industry can communicate for setting, performing, and analyzing experiments for research *(2,4)*.

Model Characteristics (Mechanistic vs Empirical)

A setup of a biological process model and design requires a broad and deep understanding of the basics of the process itself. Several state variables such as specific growth and production rate, and growth-limiting substrate and biomass concentration describe the behavior of the biological system as a whole and are what need to be optimized. There are also several parameters to be measured for explaining the process, such as maximum specific growth rate, rate constant, and yield coefficient, which vary depending on the process conditions.

To develop a bioprocess model, we have to investigate the biophysical and biochemical mechanisms that are the essential elements of the process. Mass and energy balances should be accounted for in terms of all biological process products such as biomass, primary product, side product, and heat. These balances (stoichiometric equation) should involve chemical and molecular species, enthalpy (heat exchange), free energy, charge exchange, and reducing power. Kinetic modeling with rate expression equations is required for describing growth rate, product synthesis, and substrate consumption. A kinetic model is based on the mass balance that describes the energy exchanges in the process. Mass, heat, and gas transfer are important terms for an engineering model to improve productivity and efficiency.

This process approach produces a mechanistic model that is described by well-established theory, formulated in terms of mathematical and kinetic equations, which are based on biochemical mechanisms. To build this mechanistic model requires considerable understanding of the process and a complex quantitative description of the process. A structured model can be developed from the mechanistic model that partitions the whole process into small fractions for a detailed explanation. Although the chemical process kinetics reflect the reactor rates on a molecular level, biological process dynamics result from interactions between the living organisms and their environment which affect the biochemical and physiological reactions occurring in the bioprocess.

In the case of a fermentation process, the growth and metabolism of the microbial cells are quite complex, multistage, and highly intercorrelated biochemical processes. Thus, it is almost impossible to provide a complete description of the growth and production mechanisms. Generally, it is extremely difficult experimentally to obtain enough mechanistic information about the microbial metabolism to develop a realistic structured mechanistic kinetic model *(1)*. It is very difficult to estimate parameters, and the application of sophisticated numerical methods may easily produce physico-chemically meaningless values.

The structured models require the specification of large numbers of parameters. Even if these values are measured, other difficulties still

remain. The most serious one is the need to respecify the parameter values that are already measured for certain experimental, operational conditions *(16)*. For these reasons, structured mechanistic models are rarely used for bioprocess design, control, and optimization. They can be useful for modeling transient bioprocess systems under well-controlled limited experimental conditions *(11)*. Therefore, an alternative to this conventional model approach, which is guided by kinetics and material and energy conservation rules, is needed for biological process models. An alternative would be an empirical (statistical) model.

Modeling of complex biological systems requires simplification and assumptions at the outset. Empirical models based on simplification and assumptions require relatively little knowledge about the actual detail of the biological process being studied. The greatest disadvantage of the empirical model over the structured kinetic (mechanistic) model is that many experiments would be required to investigate relationships between factors and responses.

However, if a DOE method is used for developing an empirical model, lots of time and resources can be saved by reducing the number of experiments. This is true because DOE provides well-organized, tailored, and selected experiments. Therefore, one can overcome the most significant problem in using an empirical model by applying statistical experimental design procedures and other modified DOE methods.

Empirical models are exceptionally useful in describing processes in which the mechanisms are extremely complex and incompletely understood, which is the case with most biological processes. However, most empirical models do not involve dynamic components, unlike mechanistic models that are based on kinetics. For dynamic studies with empirical models, time-based experiments can be performed and analyzed by using modified experimental designs such as repeated-measures, time-split, time series, multivariate, or crossed experimental designs *(2,4,10–14,17)*.

Time-Based Experimental Designs

In time-based experimental designs, experimenters take observations repeatedly on the same experimental units (also called subjects). These are similar to randomized block designs with randomization limitation in the subplot (within the subject) because one cannot randomize the time order. Because these observations are taken from the same subjects, they are not independent. It is crucial to note that the important assumption in this case is not the independence of measurements but the independence of errors. Actually, these dynamic models assume a correlation between measurements. The correlations between the different measurements are not usually the same, which restrains the use of the F-test calculated as for a block. The normal F-test for blocks assumes the sphericity (compound symmetry or uniform covariance matrix) of the observations *(2,4,11–13)*.

This sphericity demonstrates that variance of all mutual differences of all possible pairs of measurements is identical. This can be tested with Mauchly's test, whose null hypothesis is sphericity. Lack of sphericity should be treated by adjusting the degrees of freedom before performing an F-test. Huynh and Felt, and Greenhouse and Geisser made an adjustment of the degree of freedom for error to correct for the deviation from compound symmetry *(12,13)*. Time series methods are more appropriate for analyzing long periods of data with more than 20 observations (collections). Multivariate experiments are ideal for analysis of ecological or environmental data, with large number of replications, collected over time or space from weather, natural niches, or pollution studies rather than laboratory results.

The mixed (crossed) experimental design that combines response surface or mixture design with process factors is also another method to analyze the time-based observations. These mixed designs can take formulation, process, and time and space effects into consideration simultaneously. They can also be applied to areas where different categorical factors should be considered together to achieve meaningful solutions. Food, metal, pharmaceutical, and other manufacturing industries as well as medical, agricultural, and bioremedial research are good examples for which the mixed design is useful.

By using these time-related experimental designs and mixed design, one can study the rate and dynamics of growth, production, and reactions associated with individual subjects, groupings, or whole populations over a period of time, in addition to the main effect of each factor and their interactions in the process. This is a new approach that combines statistical modeling with dynamic modeling to study physical mechanisms, factor effects, and process optimization for complex bioprocesses.

Repeated-Measures Designs and Bioprocess Experiments

Many bioprocess experiments produce successive observations of the same variable (response) on each of the experimental units. Repeated-measures are defined as measurements sequentially conducted in time (temporal factor) or location (spatial factor) on the same experimental unit (sampling unit or subject) without changing the treatment applied to the experimental unit. Repeated measurements are commonly employed in biochemical processes to estimate growth, measure parameters, investigate the factor effect on the process, and model and monitor the production and its process.

Repeated-measures designs usually involve experimental treatments, numeric factors such as concentration, speed, pH, and temperature, and categorical (qualitative) factors, which are applied to the experimental units. Thus, repeated-measures analysis is the investigation of the repeated-measures factors, the treatment factors, and their interactions *(18–20)*. This analysis has not been widely used by biotechnology scientists and

bioengineers because most of them are not familiar with statistical experimental design and statistical data analysis and they prefer mechanistic study (mechanism or metabolism) to an empirical one (effect or trend). In this review paper, different statistical methods were discussed that can be used to analyze polyhydroxyalkanoates (PHA) production trends and treatment effects that are obtained from repeated-measures experimental design over time.

Assumptions for Repeated-Measures ANOVA

Basic assumptions are required for the F-test to be valid. The following are three basic assumptions for normal crossed analysis of variance (ANOVA) designs.

1. Normality (normal distribution of responses)
2. Independent samples (random sampling, zero correlation)
3. Homogeneity of variance (equal variances of the responses)

Additional assumptions are needed for repeated-measures ANOVA as a result of the presence of correlations between measurements taken on the same subject at different times. Another reason for more assumptions is because of the fact that experiments with repeated measurements of various treatments are not of true split-design structures in which factor levels are randomized in two stages (1st stage also called between subject, between treatment, whole plot, whole unit, or main plot, 2nd stage also called within subject, within treatment, sub plot, sub unit, or split plot); repeated-measures involve certain repeated factors (time, order, space, location, or their combinations), which are not experimental factors whose levels can be randomly assigned to within subject.

The two typical situations present in the repeated-measures often violate the assumption of equal correlation as a result of serial correlation of successive observations and non-randomization of the repeated factors. Therefore, univariate repeated-measures ANOVA requires additional assumptions that the correlations between observations within a subject are all the same, and that the correlations between groups (treatments) are equal. These additional assumptions as shown below will make the F-test for time and for time-involved interactions of the repeated-measures valid.

1. Sphericity: assumption of compound symmetry or circularity of variance–covariance matrix.
   a. Homogeneity of within-treatment variances (constant variance).
   b. Homogeneity of the covariances among repeated measures (constant correlation).
2. Additivity: no interaction between subjects and treatment.

Repeated measures in space, data collected at different specified locations on the same subject, require basically the same assumptions as repeated measures in time.

The multivariate approach to repeated-measures ANOVA (MANOVARM) requires no assumptions about the correlation structures and is always valid. However, the multivariate approach tends to be much more conservative and less powerful than the corresponding univariate procedures when the sphericity is satisfied. The multivariate test is appropriate when the sphericity assumption does not hold. MANOVA requires more replications than measurement times for sufficient error degree of freedom for the multivariate interaction test.

Sphericity Assumption Test and Correction

The sphericity can be tested with Mauchly's test, whose null hypothesis is sphericity. Mauchly derived this test that verifies the variance-covariance matrix structure. Mauchly's test of sphericity investigates the null hypothesis that the error covariance matrix of the orthonormalized, transformed dependent variables is proportional to an identity matrix *(12)*. Lack of sphericity causes concern about the F-test because the effect of such violations is to shift the sampling distribution of F to the right; when violations are exhibited, the critical values of use (F critical) are too small. The actual critical values of need based on the correct sampling distribution are larger than those listed in the F table, which results in an F-test biased in a positive direction. Thus, the type I error rate (the α value) for F-test becomes inflated under this condition.

Therefore, if a violation has occurred, the F-test needs to be adjusted (made more conservative or stringent). To deal with such situations, the degree of deviation from sphericity should be measured to correct the F-ratio to a new critical value and to provide an adjustment to univariate tests. Two similar correction methods for adjusting the F-test in terms of degree of freedom (df) are the Huynh-Feldt (H-F) and the Greenhouse-Geisser (G-G) test (Kuehl, 2000; Oehlert, 2000). Both provide an epsilon value (e correction value) that is used to adjust the between-treatment and error df. Both tests give a value that ranges from zero to one. Higher values mean less violation (one indicates ideal situations, which show no violation of sphericity). The general rule is that when the G-G a value is 0.5 or greater, the more liberal H-F test is used and when the G-G α value is 0.5 or below, the more stringent G-G test is recommended.

The correction values should be applied to the within-subject (time) effects and their corresponding error before the F-test. To adjust for nonsphericity, the numerator and denominator df is multiplied by the appropriate α value. Mean square values (MS) change, but the value of the F-test does not. However, the df values used to calculate the *p*-value do change. The corrected df and recalculated *p*-value should be used for statistical significance when sphericity cannot be assumed (in the case of heterogeneity).

Table 1
Layout of Repeated-Measures and Results

Response: lipid (μg/mL)								
Treatment: C[a]-source (mL)	Subject: bioreactor	Repeated factor: time (h)						
		24	48	72	96	120	144	168
30	# 1	2910	5560	6870	7660	8280	8590	8860
30	# 2	3120	5260	7260	8040	8340	9150	9220
35	# 1	2830	5700	7050	9050	8440	8500	8560
35	# 2	2860	5380	7140	9050	8500	8610	8720
40	# 1	2960	5410	6860	8510	8380	8720	8290
40	# 2	3130	5510	7220	8500	8590	8700	8630
45	# 1	2540	5250	6790	8170	8760	9410	8460
45	# 2	2520	5180	6890	8630	9030	9280	8910
50	# 1	2810	5010	6360	8470	8790	9480	8460
50	# 2	2670	5130	6570	8910	9040	9680	8930

[a]Carbon.

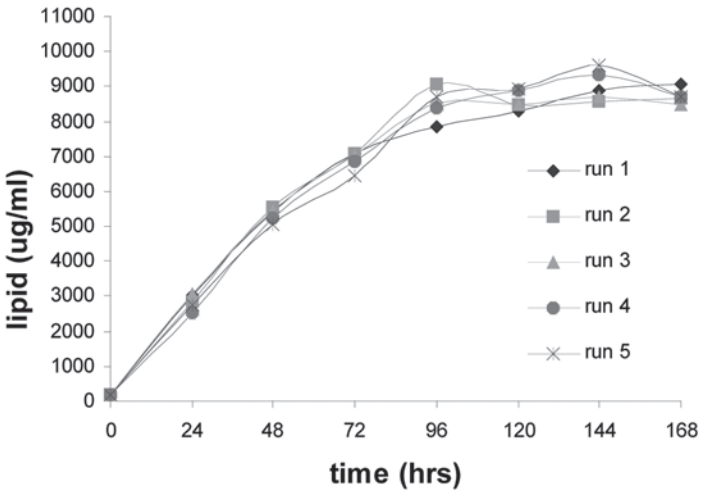

Fig. 1. Time profile of lipid production.

Analysis of Bioprocess and Bioproduction

Table 1 shows the design layout and results of repeated-measures experiment with five concentration levels (treatments) and seven successive measurements (sample collections) in two replicates. Figure 1 shows the time profile of responses (lipid production at each collection). This graph reveals dynamic variation of lipid production when experiments were performed under the various treatments over time.

Table 2
Mauchly's Test of Sphericity

Response: lipid (μg/mL)							
Within Subjects Effect	Mauchly's W	Approx χ^2	df	Sig.	Epsilon (ε) Greenhouse-Geisser	Huynh-Feldt	Lower-bound
DAY	0	.	20	.	0.42	1.00	0.17

Table 3
Tests of Within-Subjects (Time) Effects

Response: lipid (μg/mL)						
Source	Condition	Sum of squares	df	Mean square	F	Sig.
Time	Sphericity Assumed	318059294	6.0	53009882	2297.50	0.000
	Greenhouse-Geisser	318059294	2.5	125845730	2297.50	0.000
	Huynh-Feldt	318059294	6.0	53009882	2297.50	0.000
	Lower-bound	318059294	1.0	318059294	2297.50	0.000
Time C[a]-source	Sphericity Assumed	5123291	24.0	213470	9.25	0.000
	Greenhouse-Geisser	5123291	10.1	506780	9.25	0.000
	Huynh-Feldt	5123291	24.0	213470	9.25	0.000
	Lower-bound	5123291	4.0	1280823	9.25	0.016
Error (time)	Sphericity Assumed	692186	30.0	23073		
	Greenhouse-Geisser	692186	12.6	54775		
	Huynh-Feldt	692186	30.0	23073		
	Lower-bound	692186	5.0	138437		

[a]Carbon.

Univariate Repeated-Measures ANOVA

According to Table 2, G-G condition should be used to analyze this experiment because of , less than 0.5, which also means there is violation of sphericity. Thus, df correction under the G-G condition is done by multiplying df for each source by the correction value of 0.42. Modified df for time, time*ice cream, and error of within-subject becomes 2.5, 10.1, and 12.6, respectively, which are smaller than df under the sphericity assumption, time (6), time*ice cream (24), and error (30). This reduction of df resulting in decreased power is a penalty to pay for the analysis of univariate repeated-measures under the nonsphericity condition.

Tables 3 and 4 display univariate tests for the within-subjects factors and between-subjects factors, respectively. Even when the conservative df are used, time and time*ice cream effects show statistical significance. Therefore, these conclusions are valid. However, ice cream by itself turned out to be a non-significant factor as indicated by high *p*-value of 0.9276 in

Table 4
Tests of Between-Subjects (sugar) Effects

Response: lipid (μg/mL)					
Source	Sum of squares	df	Mean square	F	Sig.
Intercept	3557870036	1	3557870036	32679.13	0.0000
C[a]-source	87328.57143	4	21832.14286	0.20	0.9276
Error (subject)	544364.2857	5	108872.8571		

[a]Carbon.

Table 4. It is concluded that there are significant differences between different time levels and significant time*ice cream interactions at the 5% significant level.

Polynomial Contrast of Time Trend (Analysis of Contrasts)

It is useful to analyze contrasts among the within-subjects variables to study the levels of the within-subjects factors as shown in Table 5. This set of contrasts does not affect the results of the univariate and multivariate analyses. The repeated-measures analysis of contrast can be considered as a series of analyses of variance, each on a different linear combination of the responses across time. It is useful only for comparing different levels of the within-subjects factor (repeated factor = time).

The first column indicates the effect (variation source) being tested. The second column displays the type of contrast being used. If the significance level is less than 0.05, then the contrast is significant at the 5% significant level. In time, linear and quadratic contrast show two most significant effect. Quadratic contrast gives the most significant effect in time*ice cream. As seen in the error term, quadratic contrast shows the smallest mean square of error (MSE). Thus, quadratic contrast would be the best linear combination to demonstrate responses trend over time.

Regression Model for Repeated-Measures Analysis

Regression and Prediction Equation

An alternative method of repeated-measures analysis can be performed by fitting a polynomial regression of lipid (PHA) production on fermentation time. The coefficients calculated from least square method are used for fitting a model and making graphs. Table 6A introduces the selected model for lipid production by the stepwise regression reduction method. The standard error of the regression is the estimated standard deviation associated with the regression coefficient estimate (Table 6B). The 95% confidence interval gives the estimated range in which the true coefficient can be found. The VIF (variance inflation factor) represents how much the variance of that model coefficient increases from the lack of orthogonality in the design. If a coefficient is orthogonal to the rest of the model term, then its VIF is 1 as shown in these model terms (Table 6B).

Table 5
Tests of Within-Subjects (Time) Contrasts

Response: lipid (μg/mL)						
Source	Time	Sum of squares	df	Mean square	F	Sig.
Time	Linear	253974603	1	253974603	6383.05	0.0000
	Quadratic	62152000	1	62152000	13438.62	0.0000
	Cubic	387207	1	387207	29.76	0.0028
	Order 4	29149	1	29149	0.88	0.3918
	Order 5	7620	1	7620	0.21	0.6641
	Order 6	1508715	1	1508715	126.24	0.0001
Time C[a]-source	Linear	1530284	4	382571	9.62	0.0144
	Quadratic	1013639	4	253410	54.79	0.0003
	Cubic	1361377	4	340344	26.16	0.0015
	Order 4	520994	4	130249	3.92	0.0832
	Order 5	65359	4	16340	0.46	0.7668
	Order 6	631639	4	157910	13.21	0.0072
Error (time)	Linear	198945	5	39789		
	Quadratic	23124	5	4625		
	Cubic	65050	5	13010		
	Order 4	166058	5	33212		
	Order 5	179255	5	35851		
	Order 6	59754	5	11951		

[a]Carbon.

Table 6
Regression Model for Lipid Production

Response: lipid (μg/mL)					
A. Stepwise Regression with α to Enter = 0.100, α to Exit = 0.100					
Forced Terms : Intercept					
Terms added	Coefficient estimate	t for H_0 Coeff = 0	Prob > |t|	R-Squared	MSE
B	2857.18	11.00	<0.0001	0.79	1049929
B^2	–2448.11	–16.69	<0.0001	0.98	111614
AB	218.14	1.89	0.0677	0.98	103275
Hierarchical Terms Added After StepWise Regression					
A					

B. Regression analysis						
Factor	Coefficient estimate	DF	Standard error	95% CI Low	95% CI High	VIF
Intercept	8217.33	1	84.19	8045.39	8389.28	
A: C-source (mL)	25.86	1	77.95	–133.33	185.05	1
B: Time (h)	2857.18	1	82.68	2688.33	3026.03	1
B^2	–2448.11	1	143.20	–2740.56	–2155.65	1
AB	218.14	1	116.92	–20.64	456.93	1

The purpose of the prediction equation is to fit the data to the model for prediction or optimization. The final regression functions for lipid production in terms of coded factors, shown below, were used for making a statistical model.

Final Equation in Terms of Coded Factors:

$$\begin{aligned}\text{lipid } (\mu g/mL) = {} & 8217.3333 \\ & + 25.857143 * A \text{ (carbon source)} \\ & + 2857.1786 * B \text{ (time)} \\ & - 2448.1071 * B^2 \\ & + 218.14286 * A * B\end{aligned}$$

Residual Analysis: Assessment of Model Assumptions

Before accepting any model, the adequacy of the adopted model should be checked by the appropriate statistical method. The model assumptions shown below are on the error terms (e_i's), which are mimicked by the residuals, the difference between the observed value of the response variable and the value predicted by the regression function.

Assumptions on the Error Terms: $e_i \sim N(0, \sigma^2)$
with Independent & Identical Distribution

1. $E(e_i) = 0$ (zero mean)
2. Var $(e_i) = \sigma^2$ (constant variance)
3. Independent e_i (no correlation among variables)
4. Normality of e_i distribution (random variable)

The major diagnostic method is residual analysis as shown in Fig. 2, providing diagnostics for residual behavior. There are several residual plots to test the model assumptions. The primary analysis is to examine a normal probability plot of the studentized residuals, that is, the number of standard deviations of the actual values from their respective predicted values (Fig. 2A). The normal probability plot is employed to determine whether the residuals follow a normal distribution, which is the most important assumption for statistical modeling and model adequacy checking.

The next analysis is to look at the residuals plotted vs the predicted responses (Fig. 2B). There should be no systemic pattern in the plot, and the points should fall within a horizontal band centered at zero. Departure from this may suggest a violation of the constant variance assumption. The size of the studentized residual should be independent of its predictive value, which means that the spread should be about the same across all levels of the predicted values. Residual versus run order (number) graphs reveal any time-based effects or sequential component (Fig. 2C).

Actual vs predicted displays the real response data plotted against the predicted responses (Fig. 2D). Points above or below the diagonal line mean areas of over or under prediction. Residual versus factors graphs also need to be examined to see if there might be variance changes (dispersion effect)

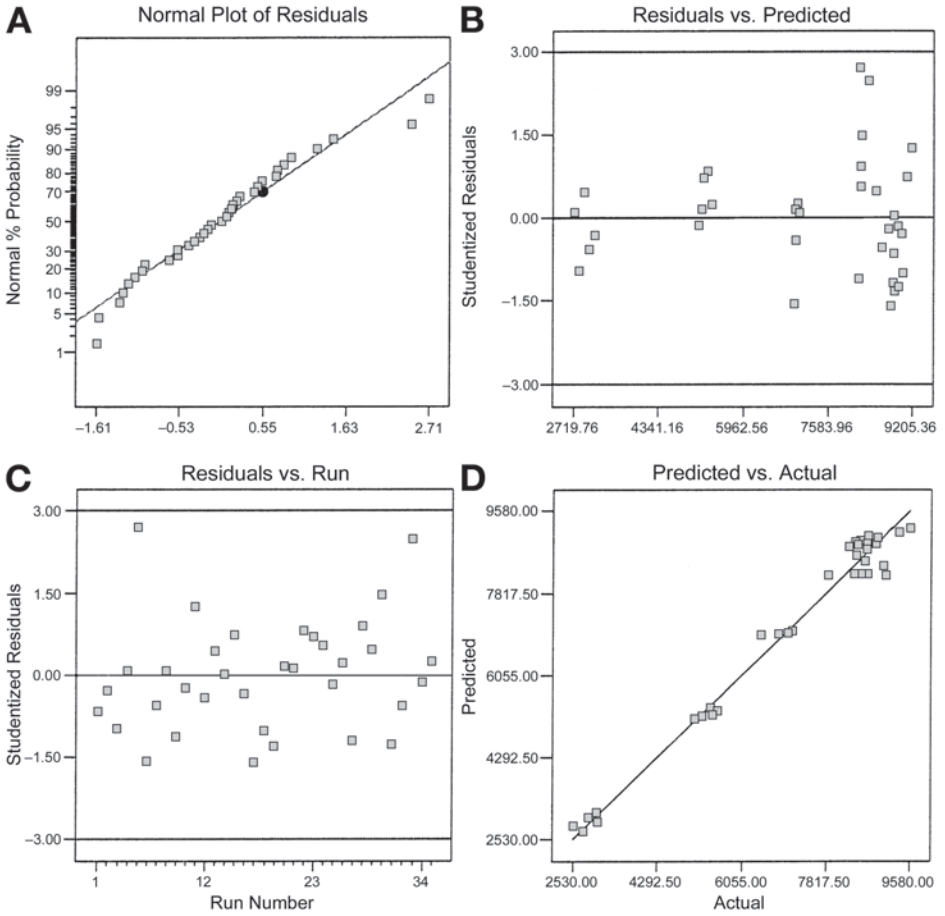

Fig. 2. Residual test of regression model for lipid production: **A**, normality; **B**, residual vs predicted response; **C**, residual vs experimental order; **D**, predicted vs actual value.

with any factor or level of factor. There were no significant violations of the model assumptions found in this residual analysis as shown in Fig. 2. So the selected model can be used for further studies such as modeling, graph, and optimization without any bias.

Multivariate Analysis

As a result of insufficient residual df for the multivariate interaction test (more repeated measurements [seven times] than replications [two subjects per treatment]), it is not possible to compute multivariate statistics for the repeated factor main effect even if it is possible to analyze the individual contrasts of the repeated factor as shown in Table 5. The major disadvantage of multivariate analysis is lack of power because of calculating a large number of parameters, $t(t–1)/2$ where t = number of repeated measurements. This is a critical consideration when t is large and n is small,

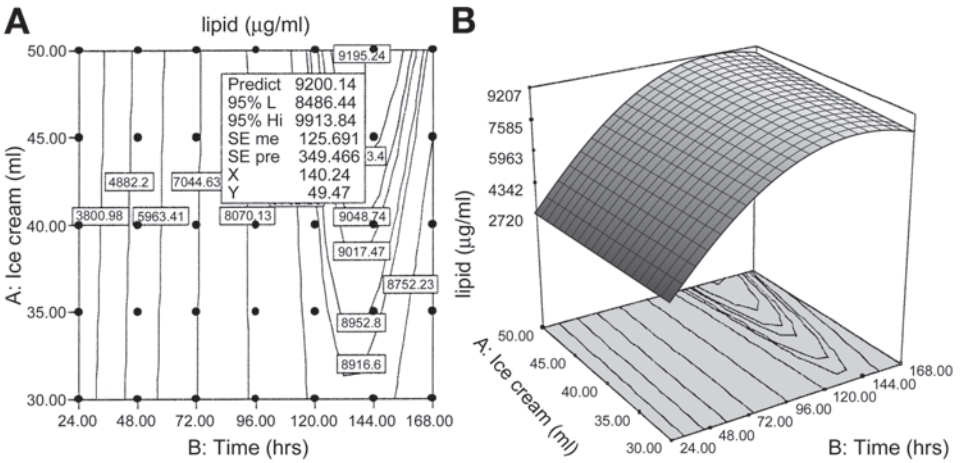

Fig. 3. Iso-contour plot **(A)** and three-dimensional graph **(B)** for lipid production.

a typical situation to long-term fermentation process or microbial culturing with frequent sample collection. In addition, one missing measurement influences the total analysis. However, no statistical assumption is required for the multivariate test.

Modeling

Figure 3 show models with contour and three-dimensional plots of lipid production. Toggled (flagged) points in the iso-response contour plot represent tentative optimal settings for the lipid production, and dots on the plot indicate design points created by repeated-measurers design (Fig. 3A). The predictive model was used to generate a response surface graph that contains equations for describing linear and quadratic effects of carbon source and time, and equations for representing interactions between them (Fig. 3B). From the optimization method, optimal formulation (50.00 mL of sugar and 140 h of fermentation time) with predicted value of 9200 µg/mL were selected for further studies.

Conclusions

Selecting an appropriate statistical approach for the analysis of repeated-measures depends on the structure of data, covariance matrix, the assumptions, sample sizes, and missing data. To summarize the method for the RM, if equal correlations are present in RM, univariate ANOVA is the best approach. This approach is valid only when compound symmetry is satisfied, which is a special case of sphericity of uniform covariance matrix. As shown in most cases of RM, adjusted univariate tests are needed when sphericity is not met. Multivariate analysis needs in-depth statistical knowledge to analyze data and interpret results. A linear regression method is a simple and efficient statistical analysis to provide information about

treatment response over time or location. In this paper, the relationship between repeated factor and treatment was demonstrated through several statistical approaches. Statistical analysis of biprocess with repeated measurements can help investigate environmental factors and effects affecting physiological and biochemical processes in analyzing and optimizing biotechnology based production.

References

1. Haaland, P. D. (1989) *Experimental Design in Biotechnology*, Marcel Dekker, INC., New York, NY.
2. Montgomery, D. C. (2004) *Design and Analysis of Experiments*, *6th ed*., John Wiley & Sons, INC. New York, NY.
3. Cornell, J. A. (2002) *Experiments with Mixtures*, *3rd ed*., John Wiley & Sons, INC., New York. NY.
4. Myers, R. H. and Montgomery, D. C. (2002) *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, *2nd ed*., John Wiley & Sons, INC., New York, NY.
5. Krishna, C. and Nokes, S. E. (2001) *J. Ind. Microbiol. Biotechnol.* **26,** 161–170.
6. Lee, K.-M. and Gilmore, D. F. (2005) *Process Biochem.* **40,** 226–249.
7. Narang, S., Sahai, V., and Bisaria, V. (2001) *J. Biosci. Bioeng.* **91(4),** 425–427.
8. Saxena, S. and Saxena, R. K. (2004) *Biotechnol. Appl. Biochem.* **39,** 99–106.
9. Sundaran, B., Rao, U. B., and Boopathy, R. (2001) *J. Biosci. Bioeng.* **91(2),** 123–128.
10. Dunn, I. J., Heinzle, E., Ingham, J., and Prenosil, J. E. (1992) *Biological Reaction Engineering*, VCH, New York, NY.
11. Dean, A. and Voss, D. (1999) *Design and Analysis of Experiments*, Springer-Verlag, New York, NY.
12. Kuehl, R. O. (2000) *Design of Experiments: Statistical Principle of Research Design and Analysis*, *2nd ed*., Brooks/Cole, Pacific Crove, CA.
13. Oehlert, G. W. (2000) *A First Course in Design and Analysis of Experiments*, W. H. Freeman and Company, New York, NY.
14. Wu, C. F. J. and Hamada, M. (2000) *Experiments: Planning, Analysis, and Parameter Design Optimization*, John Wiley & Sons, INC., New York, NY.
15. Volesky, B. and Votruba, J. (1992) *Modelling and Optimization of Fermentation Process*, Elsevier Publisher B. V., Amsterdam, The Netherlands.
16. Baughman, D. R. and Liu, Y. A. (1995) *Neural Networks in Bioprocessing and Chemical Engineering*, Academic Press, San Diego, CA.
17. Chen, C., Heineken, K., Szeto, D., Ryll, T., Chamow, S., and Chung, J. D. (2003) *Biotechnol. Prog*. **19,** 52–57.
18. Lee, K. M. and Gilmore, D. F. (2006) *Appl. Biochem. Biotechnol.* **133(2),** 113–148.
19. Lee, K. M., Rhee, C. H., Kang, C. K., and Kim, J. H. (2006a) *Appl. Biochem. Biotechnol.*, In production.
20. Lee, K. M., Rhee, C. H., Kang, C. K., and Kim, J. H. (2006b) *Appl. Biochem. Biotechnol.*, In production.