Statistical Medium Formulation and Process Modeling by Mixture Design of Experiment for Peptide Overexpression in Recombinant *Escherichia coli*

Kwang-Min Lee,[†,1] **Chang-Hoon Rhee,**[†,1]
Choong-Kyung Kang,[2] **and Jung-Hoe Kim***[,1]

*[1]Cellular Metabolic Engineering Lab.,
Korea Advanced Institute of Science and Technology,
373-1 Guseong-dong Yuseong-gu Daejeon, 305-701, South Korea,
E-mail: Klee0922@yahoo.co.kr; [2]KoBioTech Co., Ltd.,
713-12 Gojan-dong Nam-gu Inchen, 405-821, South Korea*

**Received July 25, 2005; Revised January 1, 2006;
Accepted January 26, 2006**

Abstract

The medium formulation and robust process modeling for anti-HIV peptide (T-20) production by recombinant *Escherichia coli* overexpression were studied by employing a crossed experimental design. The crossed design, a mixture design combined with process factor (induction duration), was used to find the optimal medium formulation and process time. The optimal settings for three major components (7.75 mL of NPK sources, 5.5 mL of glucose, and 11.75 mL of $MgSO_4$) characterized by %T-20 (14.45%), the proportion of peptide to the total protein, were observed in a total of 100 mL of medium inducted at an optical density of 0.67 with 0.7 m*M* isopropyl-β-D-thiogalactopyranoside) for a 3-h induction duration at shake-flask scale. These conditions were further investigated to find robust process conditions (8.2 mL of NPK sources, 5.6 mL of glucose, and 11.3 mL of $MgSO_4$, and a 3.5-h induction duration time) for T-20 production (13.9%) by applying propagation of error.

Index Entries: Anti-HIV peptide; statistical formulation; crossed design; robust process; propagation of error.

[†] Coauthors.

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

Introduction

T-20 is a synthetic peptide with 36 amino acids corresponding to a region of the transmembrane subunit of the HIV-1 envelope glycoprotein *(1–3)*. This peptide is thought to interfere with the conformational change in glycoprotein 41 (gp41) and to prevent fusion of viral and host cell membranes *(4,5)*. As a fusion inhibitor, T-20 became the most prominent of a new class of anti-HIV drug that inhibits HIV entry into host cells such as T-cells *(6,7)*.

T-20 is the most complex synthetic peptide ever chemically produced at a large scale, requiring unprecedented complexity of the manufacturing process *(8)*. Therefore, biologic production for therapeutic large chemicals such as peptides and proteins by gene-cloning technique has attracted the attention of manufacturing companies. Biologic production could be a good alternative method to a chemical one for mass production of pharmaceutical products.

Optimization of culture media and process is critical to the long-term manufacturing success of a bioproduct. To reach this goal, biotechnological research requires effective problem-solving methods because such research involves the adjustment of multiple parameters, and elucidation of the interactions among them. Additional obstacles for biotechnological research include the lack of an accurate mathematical model to describe the process in which the product is produced. These conditions call for a good strategy to deal with such a complicated system.

Statistical experimental design, also known as design of experiment (DOE) is a well-established methodology for planning and executing a predetermined setting of an informative experiment. DOE can be employed for many applications for addressing temporal, physical, space, or process-related factors. A modified application of DOE called mixture design allows researchers to find the optimal formulation for mixtures. This special DOE is increasingly used for industry and research and development. Currently, many commercially available chemical or pharmaceutical products, such as drugs, cosmetics, food, and paints, are produced and sold as mixtures that include a number of components. Owing to the large number of components, it may be difficult to optimize the properties of a mixture. In this situation, a mixture design is highly beneficial for changing the composition and properties of a mixture to get close to an optimal formulation.

When the experimental response depends on the proportions of the components, standard factorial designs are not suitable, because each factor plays an independent role in them. It is reasonable to look at the response as a function of the proportion of each factor to others, not the amount. Therefore, it is necessary to use a mixture design to account for the dependence of the response on the ratio of components when searching for optimal formulations of mixtures for which only proportions matter *(9)*.

The characteristic property of a mixture design is that the sum of all its components equals 100% ($\Sigma X_k = 1$), which means that these components

(X_k) cannot be treated independently of one another, and that their proportions must exist between 0 and 1. The requirement that the summation of all components be 100% results in some problems not present in the traditional DOE that works with an unconstrained experimental matrix. Instead, one must use a constrained experimental geometry called simplex, which is analyzed by partial least squares regression, unlike using the classic multiple linear regression method for conventional design.

The media components in a fermentation culture system are dependent on each other. This leads to the fundamental difference in the statistical experimental method applied in a media mixture from that of a culture process. In a mixture system such as culture media or product recipe, a specific relationship exists among the ingredients in that system. This indicates that no component variable can be changed without changing simultaneously any of the other component variables *(10)*.

When this mixture design is combined with the process factors, which are the experimental parameters that are not part of the actual mixture, it is called crossed design. A crossed design is a very efficient and breakthrough DOE technique that involves optimization of the formulation by mixture design and optimization of the process by other standard experimental designs such as factorial and repeated-measure designs. This comprehensive crossed mixture-process experimental design also allows one to calculate complex interactions between compositional variables (mixtures) and process factors *(9)*.

The predicted response to the selected formulation or predicted formulation for the target response may vary owing to the complexities and variations in the process of biologic production. In the case of large variations or multiple suboptimal regions around the target response, statistical confidence intervals (CIs) for the response become wide, or a multimodal response surface is produced. These wide intervals and multiple candidates for optimization suggest that further analysis should be focused on a robust process that stabilizes the process performance. The objective of robust process is to reduce variability in formulation and processing by applying propagation of error (POE) calculations for reduction of variation transmitted from controlled factors, making the production process more robust, i.e., insensitive to the levels of components (mixtures) and process factors *(9,10)*.

The goal of the current experimental design for the T-20 overexpression process was to determine the media formulation along with simultaneously determining the best process conditions. A crossed mixture-process design that combines mixture components (NPK source, glucose, and $MgSO_4$) with process factors (concentration, timing, and duration of induction) was used. The ultimate purposes of our study were to formulate the optimal culture medium, and to develop a robust process insensitive to the factor (input variable) variation.

Materials and Methods

Microorganism

A genetically modified recombinant *Escherichia coli* strain, BL21(DE3)/pET23a-G3T20, constructed at Inje University, Chooncheon, Korea was used. The strain was maintained by culturing in Luria-Bertani (LB) medium containing 5 g/L of yeast extract, 10 g/L of NaCl, and 10 g/L of tryptone with 100 mg/L of kanamycin.

Culture Media

Seed cultures were prepared by growing in a 250-mL shake flask containing 50 mL of LB medium supplemented with 100 μg/mL of kanamycin at 37°C for 12 h in a rotary shaker at 250 rpm. Fermentation assay media for T-20 production were prepared by mixing concentrated stock solutions in the 250-mL shake flask and adding distilled water to make a total volume of 50 mL. Assay media were varied with treatments obtained from the experimental design matrix (Table 1). Major stock components of the assay media used in varying treatments (formulations) were as follows: glucose (20%), 200 g of glucose; NPK sources (1 *M*), 136 g of KH_2PO_4, 132 g of $(NH)_2HPO_4$, 19.2 g of citric acid; $MgSO_4$ (0.1 *M*), 26.4 g of $MgSO_4 \cdot 7H_2O$; trace element (100X, per liter of 5 *M* HCl), 10 g of $FeSO_4 \cdot 7H_2O$, 2 g of $CaCl_2 \cdot H_2O$, 2.2 g of $ZnSO_4 \cdot 7H_2O$, 0.5 g of $MnSO_4 \cdot 4H_2O$, 1 g of $CuSO_4 \cdot 5H_2O$, 0.1 g of $(NH_4)_6Mo_7O_{24} \cdot 4H_2O$, 0.02 g of $Na_2B_4O_7 \cdot 10H_2O$; isopropyl-β-D-thiogalactopyranoside (IPTG) (0.1 *M*), 23.83 g of IPTG. All media were sterilized at 121°C for 15 min and cooled to room temperature prior to use.

Expression of Recombinant Peptide T-20

Batch fermentation for T-20 expression was carried out in 250-mL shake flasks containing 50 mL of assay media supplemented with 100 μg/mL of kanamycin. Fermentation was initiated by inoculating 1 mL of seed culture cultured in LB for 12 h into the flasks variously formulated by the design. The assay cultures were grown at 37°C in a shaker at 250 rpm. For the induction of T-20 gene expression, IPTG was simultaneously added at an optical density (OD) of 0.6 for the center point at 600 nm (OD variations according to the formulations) at a final concentration of 0.7 m*M*. Samples were collected at regular 2-h intervals and analyzed immediately after being drawn from the cultures.

Experimental Design

In mixture design, the experimental area forms a specific geometric triangular layout called a simplex (Fig. 5), which is defined as a shape with one more vertex than the number of dimensions. In this simplex, the vertices (apices) represent pure component blends (only one component), and center edges (binary blends) that can provide estimates of second-order

Table 1
Design Layout and Results for Crossed Design

Experiment	ID	Mix 1: NPK (mL) actual	Mix 2: glucose (mL) actual	Mix 3: $MgSO_4$ (mL) actual	Process duration (h) actual	Mix 1: NPK (mL) pseudo	Mix 2: glucose (mL) pseudo	Mix 3: $MgSO_4$ (mL) pseudo	Mix 1: NPK (mL) real	Mix 2: glucose (mL) real	Mix 3: $MgSO_4$ (mL) real	Process duration (h) coded	Resp. 1 OD (abs.)	Resp. 2 T-20 (%)
1	1	5.50	5.50	14.00	2.00	0.00	0.00	1.00	0.22	0.22	0.56	–1.00	1.30	10.91
2	2	10.00	5.50	9.50	2.00	1.00	0.00	0.00	0.40	0.22	0.38	–1.00	1.16	9.36
3	3	5.50	10.00	9.50	2.00	0.00	1.00	0.00	0.22	0.4	0.38	–1.00	1.28	10.35
4	4	7.75	5.50	11.75	2.00	0.50	0.00	0.50	0.31	0.22	0.47	–1.00	1.25	10.32
5	5	5.50	7.75	11.75	2.00	0.00	0.50	0.50	0.22	0.31	0.47	–1.00	1.28	10.24
6	6	7.75	7.75	9.50	2.00	0.50	0.50	0.00	0.31	0.31	0.38	–1.00	1.12	9.24
7	7	7.00	7.00	11.00	2.00	0.33	0.33	0.33	0.28	0.28	0.44	–1.00	1.14	9.20
8	8	7.00	7.00	11.00	2.00	0.33	0.33	0.33	0.28	0.28	0.44	–1.00	1.16	8.64
9	1	5.50	5.50	14.00	3.00	0.00	0.00	1.00	0.22	0.22	0.56	–0.33	1.33	12.51
10	2	10.00	5.50	9.50	3.00	1.00	0.00	0.00	0.40	0.22	0.38	–0.33	1.30	12.71
11	3	5.50	10.00	9.50	3.00	0.00	1.00	0.00	0.22	0.4	0.38	–0.33	1.43	13.73
12	4	7.75	5.50	11.75	3.00	0.50	0.00	0.50	0.31	0.22	0.47	–0.33	1.34	14.45
13	5	5.50	7.75	11.75	3.00	0.00	0.50	0.50	0.22	0.31	0.47	–0.33	1.31	12.84
14	6	7.75	7.75	9.50	3.00	0.50	0.50	0.00	0.31	0.31	0.38	–0.33	1.28	13.13
15	7	7.00	7.00	11.00	3.00	0.33	0.33	0.33	0.28	0.28	0.44	–0.33	1.35	12.81
16	8	7.00	7.00	11.00	3.00	0.33	0.33	0.33	0.28	0.28	0.44	–0.33	1.32	11.99
17	1	5.50	5.50	14.00	4.00	0.00	0.00	1.00	0.22	0.22	0.56	0.33	1.43	11.93
18	2	10.00	5.50	9.50	4.00	1.00	0.00	0.00	0.40	0.22	0.38	0.33	1.51	13.01
19	3	5.50	10.00	9.50	4.00	0.00	1.00	0.00	0.22	0.4	0.38	0.33	1.53	13.24
20	4	7.75	5.50	11.75	4.00	0.50	0.00	0.50	0.31	0.22	0.47	0.33	1.49	13.05
21	5	5.50	7.75	11.75	4.00	0.00	0.50	0.50	0.22	0.31	0.47	0.33	1.27	12.24
22	6	7.75	7.75	9.50	4.00	0.50	0.50	0.00	0.31	0.31	0.38	0.33	1.46	13.32
23	7	7.00	7.00	11.00	4.00	0.33	0.33	0.33	0.28	0.28	0.44	0.33	1.40	13.02
24	8	7.00	7.00	11.00	4.00	0.33	0.33	0.33	0.28	0.28	0.44	0.33	1.43	12.54
25	1	5.50	5.50	14.00	5.00	0.00	0.00	1.00	0.22	0.22	0.56	1.00	1.69	10.27
26	2	10.00	5.50	9.50	5.00	1.00	0.00	0.00	0.40	0.22	0.38	1.00	1.71	11.25
27	3	5.50	10.00	9.50	5.00	0.00	1.00	0.00	0.22	0.4	0.38	1.00	1.79	10.29
28	4	7.75	5.50	11.75	5.00	0.50	0.00	0.50	0.31	0.22	0.47	1.00	1.74	10.80
29	5	5.50	7.75	11.75	5.00	0.00	0.50	0.50	0.22	0.31	0.47	1.00	1.66	10.29
30	6	7.75	7.75	9.50	5.00	0.50	0.50	0.00	0.31	0.31	0.38	1.00	1.64	11.00
31	7	7.00	7.00	11.00	5.00	0.33	0.33	0.33	0.28	0.28	0.44	1.00	1.57	9.95
32	8	7.00	7.00	11.00	5.00	0.33	0.33	0.33	0.28	0.28	0.44	1.00	1.58	8.88

effects are located on the sides of the simplex. One can augment the design by adding the overall centroid points that represent three component blends containing equal amounts of all three parts, and by adding axial check blends that indicate the interior points between the centroid and each vertex containing two-thirds of each respective component and one-sixth of each of the other two *(10–13)*.

The factor treatments and experimental settings were performed by experimental design. The Design Expert (version 6.11; Stat-Easy, Minneapolis, MN) was used to build and analyze the experimental designs. To optimize the composition of the assay medium and the fermentation process for T-20 production, a dynamic crossed statistical design was created prior to performing experiments. On the basis of the results of previous experimental designs (data not shown), significant medium factors were selected and varied according to the design layout obtained from Design Expert. The design layout is based on the factor range constraint and number and type of design point.

Table 1 shows the design matrix in terms of coded factor levels such as for mixture components from 0 to 1 for lowest concentration to highest one, respectively, and for process factors from –1 to +1 for lowest to highest levels, respectively. In mixture design, coding to the pseudoscale calculated from the intermediate stage, called real coded values, makes statistical computations much easier than actual values. To calculate pseudoscale, first actual data need to be converted into real values, where components are expressed as fractions defined as real coded = actual/total of actuals. For pseudocoded values, each real component is recalculated to be a fraction of the active part of the mixture to express the presence of individual constraints. Pseudoscales are defined as follows *(9,12)*:

$$\text{Pseudocoded} = (\text{real coded} - L_i)/(1 - L)$$

in which L_i is the lower constraint in real coded value, and L is the sum of lower constraints in real value.

Actual values and ranges for mixture are as follows:

Mixture\constraint	Lower limit (mL)	Higher limit (mL)
A (NPK)	5.5	10
B (glucose)	5.5	10
C ($MgSO_4$)	9.5	14
Active medium (A + B + C = 25)		
Basal medium (water + trace = 75)[a]		

[a] Total 100 mL of assay medium for each treatment.

The major hurdle with mixture data is that the components sum to a constant, which gives rise to exact colinearity among the x variables *(11)*. A special method should be used to avoid this colinearity. The most widely employed approach is the use of so-called Scheffe canonical polynomials, which eliminate the intercept and square terms from the basic models in order to use ordinary least square method for fitting *(10)*. Proper models by Scheffe method for mixture designs including three components consist of linear, quadratic, and special cubic models as follows:

Linear model:

$$y = b_1X_1 + b_2X_2 + b_3X_3 + e$$

Quadratic model:

$$y = b_1X_1 + b_2X_2 + b_3X_3 + b_{12}X_1X_2 + b_{13}X_1X_3 + b_{23}X_2X_3 + e$$

Special cubic model:

$$y = b_1X_1 + b_2X_2 + b_3X_3 + b_{12}X_1X_2 + b_{13}X_1X_3 + b_{23}X_2X_3 + b_{123}X_1X_2X_3 + e$$

The mixed mathematical model for mixture combined with process factors is obtained by multiplying two models for each part as

$$y = f(x)\ \mathrm{H}\ g(z) + e$$

in which y is the response, $f(x)$ indicates a mixture model (Scheffe), $g(z)$ is an ordinary model for the process factors, and e is a random error left out of the fitted model.

The best mathematical model was selected according to the comparisons of representative statistical parameters such as standard deviation (SD), coefficient of variation (CV), coefficient of determination (R^2), adjusted R^2, and lack of fit test.

Analytical Methods

Cell growth was monitored by measuring the OD at 600 nm. For the qualitative analysis of T-20, sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) was performed on a 15% (w/v) gel system. Gels were stained using the Coomassie brilliant blue R method for protein *(14)*. The quantitative analysis of peptide T-20 was done by measuring image profiles of scanned gel with ImageJ 1.29x (National Institutes of Health, Bethesda, MD). All measurements for OD were duplicated, and two center points were used to estimate experimental errors and variations among T-20 production.

Results and Discussion

Analysis of Variance, Regression, and Prediction Equation

Table 1 presents the DOE matrix and the results of the crossed design. Figures 1 and 2 present cell growth and T-20 production time sequences, respectively. Figure 3 demonstrates that different concentrations of protein are expressed under the different media culture conditions. Figure 4 provides an image profile of the scanned SDS-PAGE gel for T-20 quantification. The area indicated by the arrow represents the proportion of T-20 peptide to the whole protein.

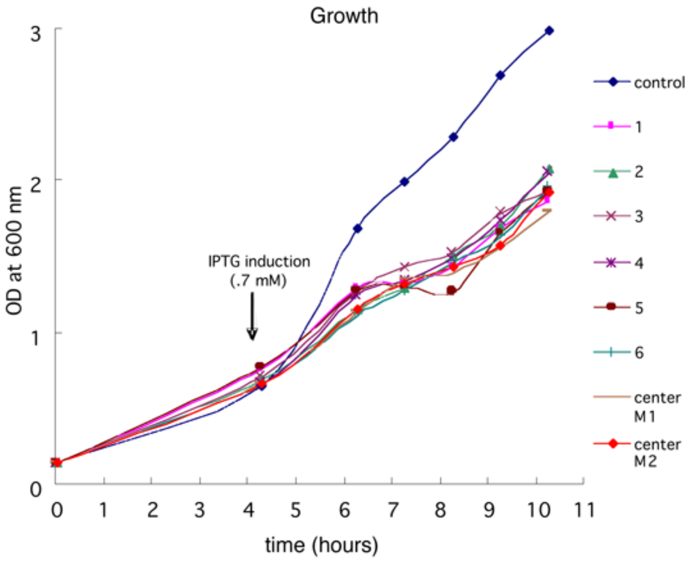

Fig. 1. Time courses of cell growth from different formulations for crossed design.

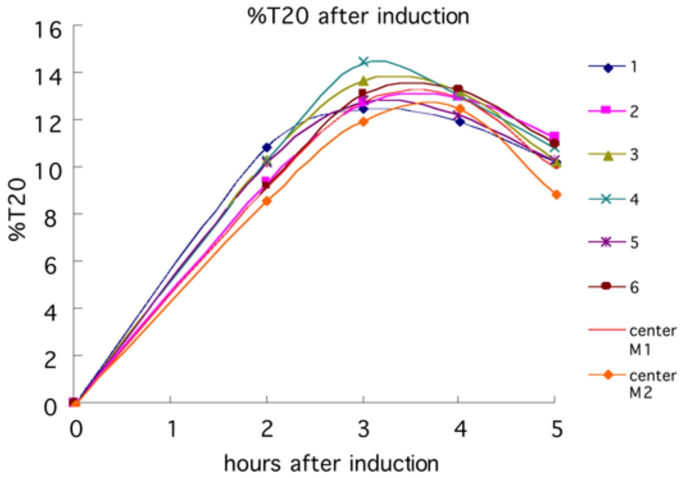

Fig. 2. Time courses of %T-20 from different formulations for crossed design.

As seen from Tables 2 and 3, the selected models for crossed mixture-process designs are quite complex owing to crossing of the mixture model terms with the process model ones. The high R^2 (0.928) indicates a good explanation of the variability by the selected model for %T-20, as shown in the analysis of variance (ANOVA) presented in Table 2. Low (5.6%) CV also shows the high degree of precision and accuracy of this model. Therefore, the crossed model (mixtures [A, B, and C] H process [D]) appears to be a reliable model for %T-20 from the crossed design.

As shown in Table 2, the linear mixture of only AC ($p = 0.091$) turned out to be significant among three mixtures (AB, AC, and BC) at the

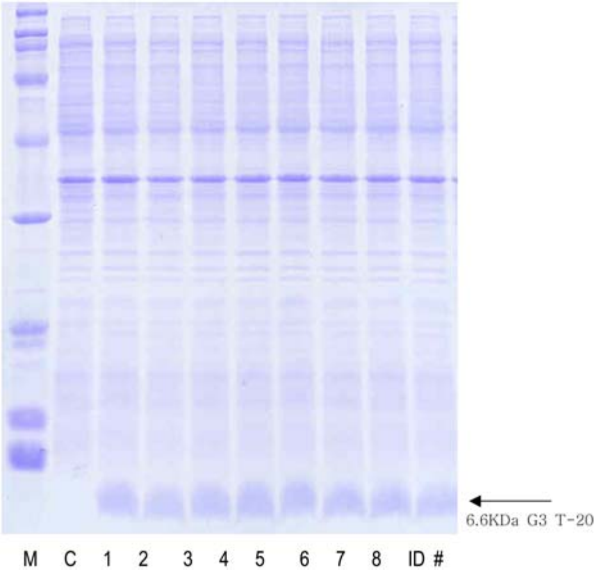

Fig. 3. SDS-PAGE analysis for total cell protein containing T-20 at 4 h after induction. Lane M, marker; lane C, control without induction; lanes 1–6, design ID; lanes 7 and 8, center points (ID 0) of crossed design.

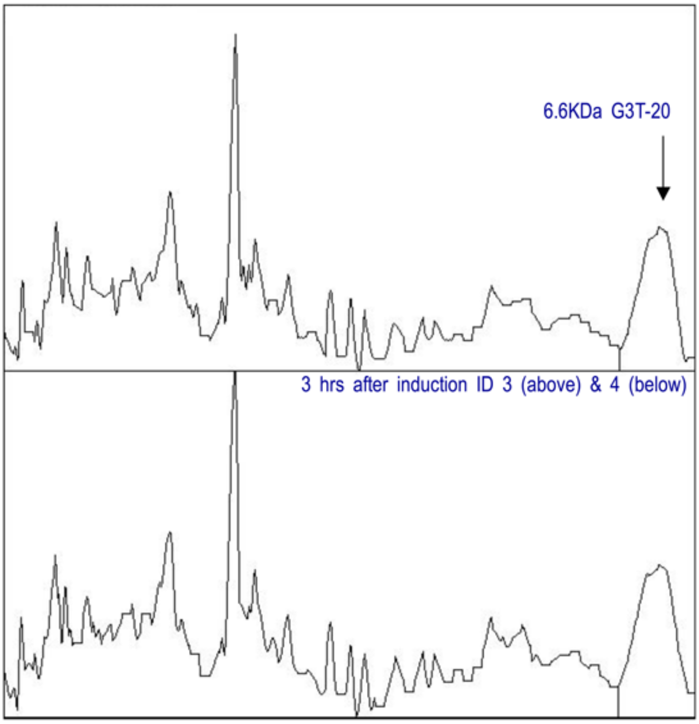

Fig. 4. T-20 quantification by ImageJ. The ratio of the area indicated by the arrow to the whole one was calculated for %T-20.

Table 2
ANOVA for %T-20 From Crossed Design[a]

Source	SS	df	MS	F value	p value Prob > F
Model	74.237	17.000	4.367	10.554	<0.0001
Linear mixture	49.276	2.000	24.638	59.548	<0.0001
AB	0.066	1.000	0.066	0.160	0.695
AC	1.360	1.000	1.360	3.286	0.091
AD	1.817	1.000	1.817	4.392	0.055
BC	0.526	1.000	0.526	1.272	0.278
BD	0.023	1.000	0.023	0.055	0.818
CD	0.314	1.000	0.314	0.760	0.398
AD^2	6.379	1.000	6.379	15.417	0.002
BD^2	9.864	1.000	9.864	23.840	0.000
CD^2	2.547	1.000	2.547	6.155	0.026
ABD	0.318	1.000	0.318	0.769	0.395
ACD	0.124	1.000	0.124	0.299	0.593
BCD	0.049	1.000	0.049	0.118	0.736
ABD^2	0.171	1.000	0.171	0.414	0.530
ACD^2	1.396	1.000	1.396	3.375	0.088
BCD^2	0.006	1.000	0.006	0.015	0.904
Residual	5.793	14.000	0.414		
Lack of fit	4.621	10.000	0.462	1.578	0.350
Pure error	1.172	4.000	0.293		
Corrected total	80.030	31.000			
SD	0.643			R^2	0.928
Mean	11.485			Adj R^2	0.840
CV	5.601			Adeq precision	9.801

[a] SS, sum of squares; df, degrees of freedom; MS, mean square; SD, standard deviation; CV, coefficient of variation; Adj, adjusted; Adeq, adequate.

10% significance level, which means that blending components A and C (at the center edge) produces higher %T-20 values than would be expected just by averaging the %T-20 production of the pure blends (at vertices). This is an example of synergistic effects as shown in the positive coefficient estimate (4.135) in Table 3, as opposed to the insignificant antagonistic effects of the negative coefficient estimate of AB (–0.914) and BC (–2.572). The crossed factors (mixture and process) of AD ($p = 0.055$), AD^2 ($p = 0.002$), BD^2 ($p < 0.0001$), CD^2 ($p = 0.026$), and ACD^2 ($p = 0.088$) also showed significant effects on T-20 production at the 10% significance level.

Table 3 introduces regression analysis for the selected model for %T-20. Coefficient estimates are equal to one-half the factorial effects in orthogonal designs. The standard error of the regression is the estimated SD associated with the regression coefficient estimate. The 95% CI implies the estimated range in which the true coefficient could be found 95% of the time. The purpose of the prediction equation obtained from the regression analysis is to fit the data to the model for prediction or optimization. The final

Table 3
Regression Analysis for %T-20 From Crossed Design[a]

Component	Coefficient estimate	df	SE	95% CI: low	95% CI: high
A-NPK	13.241	1.000	0.512	12.143	14.339
B-glucose	13.946	1.000	0.512	12.847	15.044
C-$MgSO_4$	12.489	1.000	0.512	11.391	13.587
AB	–0.914	1.000	2.281	–5.806	3.979
AC	4.135	1.000	2.281	–0.757	9.027
AD	0.899	1.000	0.429	–0.021	1.820
BC	–2.572	1.000	2.281	–7.464	2.320
BD	–0.100	1.000	0.429	–1.021	0.820
CD	–0.374	1.000	0.429	–1.295	0.546
AD^2	–2.826	1.000	0.720	–4.369	–1.282
BD^2	–3.514	1.000	0.720	–5.057	–1.970
CD^2	–1.786	1.000	0.720	–3.329	–0.242
ABD	1.677	1.000	1.912	–2.423	5.777
ACD	–1.046	1.000	1.912	–5.146	3.054
BCD	0.657	1.000	1.912	–3.444	4.757
ABD^2	–2.063	1.000	3.206	–8.940	4.813
ACD^2	–5.890	1.000	3.206	–12.766	0.987
BCD^2	–0.394	1.000	3.206	–7.270	6.482

[a] df, degrees of freedom; SE, standard error; CI, confidence interval.

regression functions for %T-20 in terms of pseudocomponents and coded factors, shown next, were used to make various statistical models.

The final equation in terms of pseudomixtures and coded process was as follows:

$$\begin{aligned}
\text{T-20 (\%)} = {} & 13.2 \cdot A \\
& 13.9 \cdot B \\
& 12.5 \cdot C \\
& -0.914 \cdot A \cdot B \\
& 4.13 \cdot A \cdot C \\
& 0.899 \cdot A \cdot D \\
& -2.57 \cdot B \cdot C \\
& -0.100 \cdot B \cdot D \\
& -0.374 \cdot C \cdot D \\
& -2.83 \cdot A \cdot D^2 \\
& -3.51 \cdot B \cdot D^2 \\
& -1.79 \cdot C \cdot D^2 \\
& 1.68 \cdot A \cdot B \cdot D \\
& -1.05 \cdot A \cdot C \cdot D \\
& 0.657 \cdot B \cdot C \cdot D \\
& -2.06 \cdot A \cdot B \cdot D^2 \\
& -5.89 \cdot A \cdot C \cdot D^2 \\
& 0.394 \cdot B \cdot C \cdot D^2
\end{aligned}$$

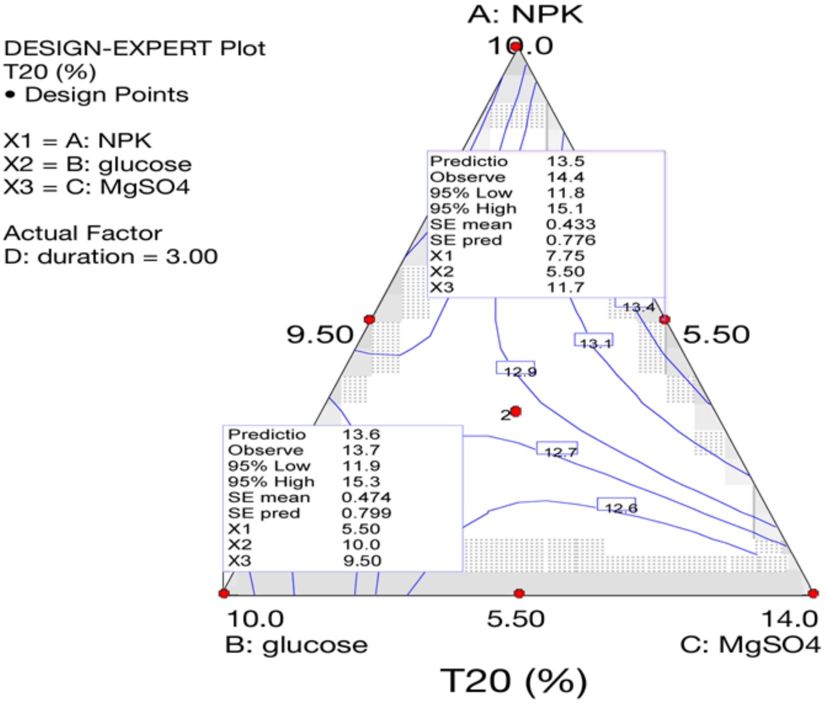

Fig. 5. Isocontour plot of crossed design for %T-20 production after 3 h of induction.

Modeling

The predictive model used to generate a response surface graph and a contour plot requires equations for mixture components representing the blending properties, equations for describing linear and quadratic effects of the process, and equations for representing interactions between process factors and the mixture. Since there are significant interaction effects between mixture and process factors, the response surface graphs and contour plots change as the process conditions (induction duration) are varied (graphs not shown). Figures 5 and 6 show an isocontour plot and a three-dimensional model of %T-20 production, respectively. The toggled (flagged) point in the isoresponse contour plot in Fig. 5 represents a tentative optimal point for the corresponding response at a specific duration time (3 h after induction), and dots on the plot indicate design points created by the crossed design.

The unique characteristic of this experimental design and modeling is that it is a combination of response surface method (RSM) with repeated measurement design, which is able to show statistical effects and the dynamic nature of the process simultaneously from the single design of experiment. The RSM combined with repeated measures is a novel experimental design consisting of multistage response surfaces explained by active medium component factors along with a process factor (induction duration).

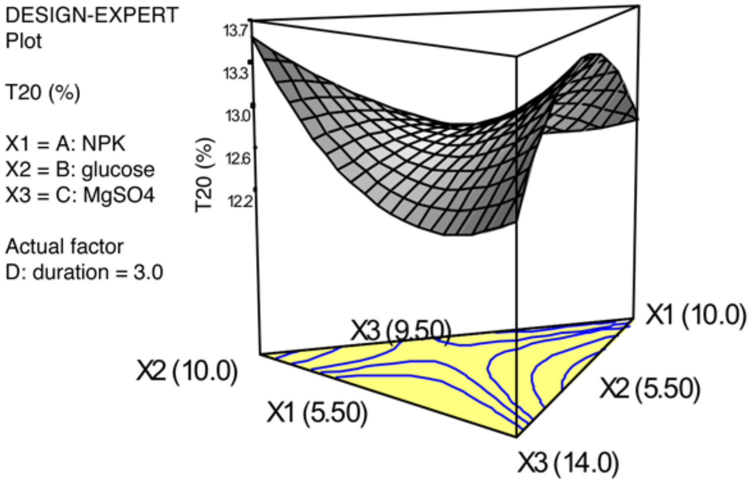

Fig. 6. Three-dimensional model for %T-20 production of crossed design after 3 h of induction.

Table 4
Induction-Timing (OD 0.66–0.76) Variation Effects on %T-20[a]

Source	df	Seq SS	Adj SS	Adj MS	*F*	*p* value	Covariate with
Induct OD	1	0.014	0.943	0.943	0.21	0.652	NPK
Induct OD	1	0.014	1.439	1.439	0.32	0.578	Glucose
Induct OD	1	0.014	1.208	1.208	0.27	0.610	$MgSO_4$
Induct OD	1	0.014	0.014	0.014	0.02	0.889	Duration

[a] df, degrees of freedom; Seq SS, sequential sum of squares; Adj SS, adjusted sum of squares; Adj MS, adjusted mean square.

Analysis of Induction Timing

One of the most important factors in determining efficient recombinant protein overexpression is the timing of induction. Because the medium compositions varied as different mixture combinations, the cultural OD at IPTG induction was different among treatments (OD range of 0.66–0.76). Thus, the effects of induction timing on T-20 production should be investigated to determine how they influence T-20 production. The statistical method used for determining the effects of induction timing variation on %T-20 was a general linear model taking induction timing as a covariate. Table 4 shows that induction timing had insignificant effects on T-20 production as it was covariated with NPK, glucose, $MgSO_4$, and induction duration ($p = 0.652$, $p = 0.578$, $p = 0.610$, and $p = 0.889$, respectively) within an OD range of 0.66–0.76. This implies that the OD variation at the induction timing of IPTG dose not affect T-20 production and further analysis.

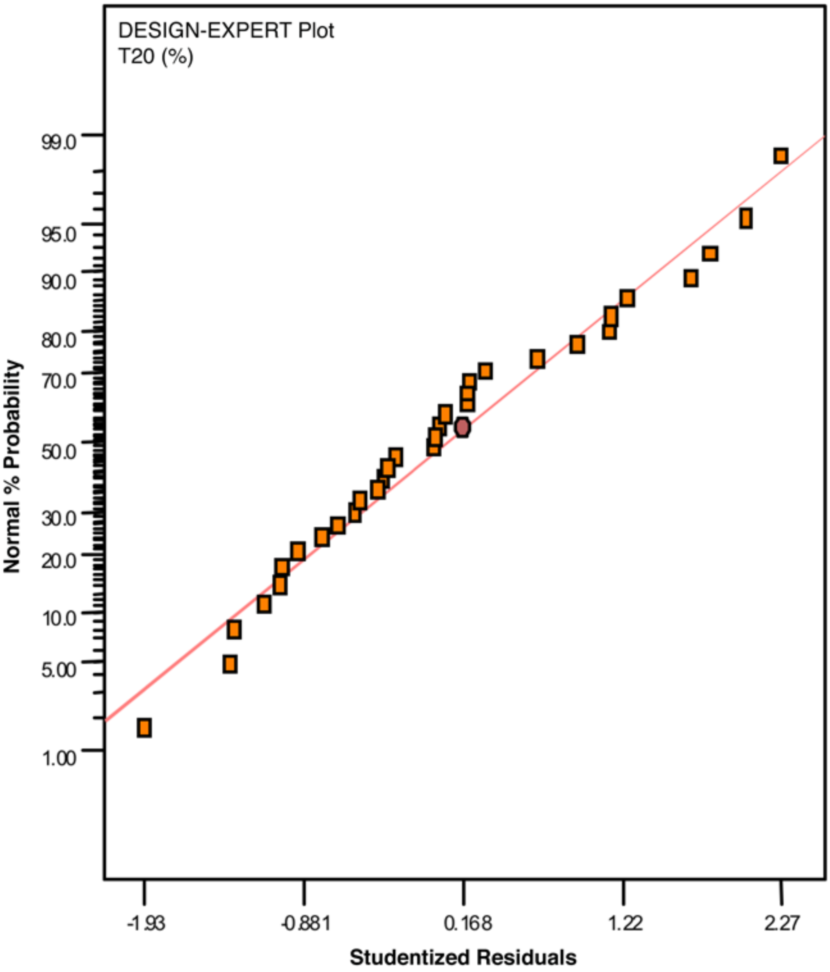

Fig. 7. Residual analysis of %T-20 production of crossed design for model verification.

Model Diagnostics

Before accepting any model, the adequacy of the adopted model should be checked by the appropriate statistical method. The major diagnostic method is residual (observed minus predicted) analysis, as shown in Fig. 7, providing diagnostics for residual behavior. There are several residuals graphs to test the model assumptions. The primary analysis is to examine a normal probability plot of the studentized residuals, i.e., the number of SDs of the actual values from their respective predicted values. 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. No significant violations of the model assumptions were found in this residual. Therefore, the modeling could be used for further studies such as formulation, optimization, or process simulation without any bias.

Table 5
Numerical Optimization for %T-20 From Crossed Design and Constraints

Name	Goal	Lower limit	Upper limit	Lower weight	Upper weight	Importance
NPK (mL)	Is in range	5.5	10	1	1	3
Glucose (mL)	Is in range	5.5	10	1	1	3
$MgSO_4$ (mL)	Is in range	9.5	14	1	1	3
Duration (h)	Is in range	2	5	1	1	3
OD	Is in range	1.12	1.79	1	1	3
T-20 (%)	Maximize	8.64	14.45	1	1	5

Table 6
Selected Formulations From Numerical Optimization

NPK (mL)	Glucose (mL)	$MgSO_4$ (mL)	Duration (h)	OD	T-20 (%)	Desirability
5.50	10.0	9.50	3.48	1.46	13.9	0.914
8.17	5.50	11.3	3.53	1.43	13.9	0.912
8.14	5.50	11.4	3.51	1.43	13.9	0.912

Optimization and Formulation

Numerical optimization can be represented by a general nonlinear algorithm with constraints applied to the main objective function, which is a desirability function. In numerical optimization, the desired goals (constraints) for each response and factor are selected along with the weight and importance that can be assigned to each goal. A weight for each goal can adjust the shape of the desirability function, and the importance of each goal can determine the relative importance to other goals. The goals are combined into an overall desirability function, which is an objective function of optimization, with its value ranging from 0 (beyond the goal limits) to 1 (matching the exact goal) *(15)*.

Numerical optimization looks for a point that maximizes the desirability function. All goals that are obtained from responses and factors become combined into one desirability function. There may be several maxima owing to curvature of the response surface and their combination in the desirability function. Tables 5 and 6 introduce constraints for the response and factors, and suggested optimal mixtures based on the desirability function, respectively.

Point prediction is used to make predictions for responses at any factor combination (identity 4 at 3-h induction in Table 7). As shown in Table 8, SE Mean stands for the standard error of the prediction of an average, and 95% CI means the 95% CI for the true mean. SE Pred represents the standard error of the prediction of an individual observation, and 95% PI is the 95% prediction interval for the true value of an individual

Table 7
Point Optimization for Crossed Design

Component	Name	Level	Low level	High level	SD
A	NPK (mL)	7.75	5.5	10	1.5
B	Glucose (mL)	5.5	5.5	10	1.5
C	$MgSO_4$ (mlL	11.75	9.5	14	1.5
D	Duration (h)	3.00	2	5	1
Total = 25					

Table 8
Prediction and Statistical Interval for Crossed Design

Response	Prediction	SE mean	95% CI: low	95% CI: high	SE Pred	95% PI: low	95% PI: high
OD	1.351	0.032	1.281	1.420	0.058	1.226	1.475
T-20 (%)	13.479	0.433	12.549	14.408	0.776	11.815	15.142

SE, Standard error; CI, confidence interval; PI, prediction interval.

observation. The value for 95% CI is always narrower than that for 95% PI, owing to its association with the mean rather than with single observations. These values explain what to expect for an individual verification or confirmation experiment.

The final optimal formulation for T-20 production was calculated based on the mixture-process crossed model. From the optimization method, three suggested optimal formulations according to desirability were selected for further robustness and formulation studies, as shown in Table 6.

POE and Robust Process

The purposes of the robust process is the same as that of six sigma: to find the most stable conditions of product quality and quantity and, thus, find the most efficient process range *(12,16,17)*. Such a desirable process region can be obtained by analyzing the response surface or more accurately by using a mathematical method (calculus) to minimize POE (response variation) from varying factors.

The POE method makes the production process robust (insensitive) to variations in input factors. This method requires calculations of partial derivatives to find broad, flat areas on the response surface and to generate POE plots that show how that error is transmitted to the response, as shown in Figs. 8 and 9. Then, a search is conducted for conditions that minimize the transmitted variation from media components. The lower the POE is, the more robust the process is, and less error (variation) from factors is transmitted to the response. However, POE will be varied to the measured response to differing degrees only when the response surface is nonlinear (involves curvature). Thus, POE is available only for second-order or higher models.

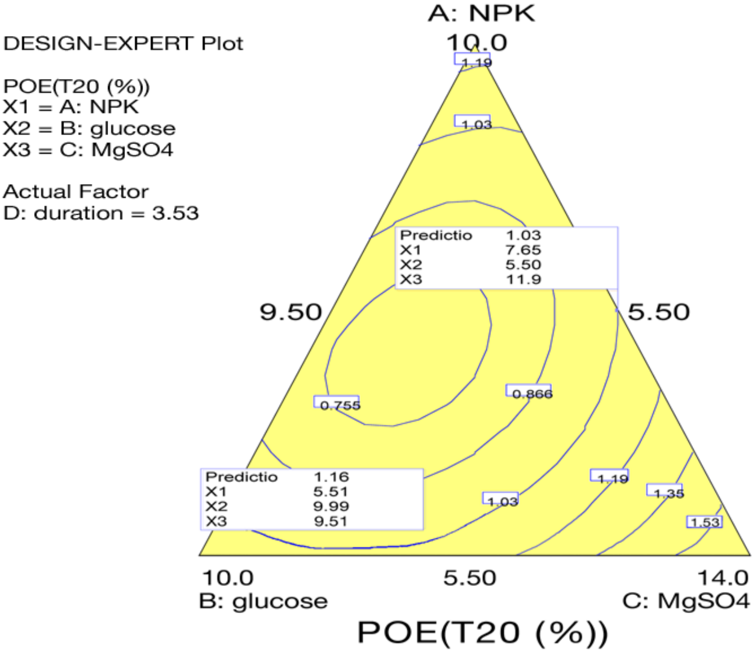

Fig. 8. Isocontour plot of POE model from crossed design for %T-20 production.

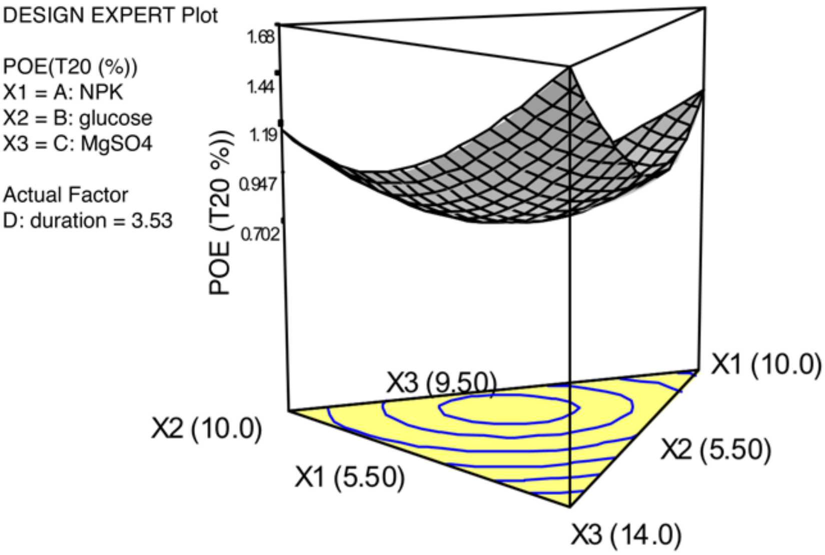

Fig. 9. Three-dimensional POE model from crossed design for %T-20 production.

This application of the POE method to the process modeling is necessary when searching for compositions of mixtures that minimize variation in the response, a formulation robust to variation in the input values. Use of the POE method begins with construction of response surface mod-

Table 9
Numerical Optimization for %T-20 From Crossed Design With POE and Constraints

Name	Goal	Lower limit	Upper limit	Lower weight	Upper weight	Importance
NPK (mL)	Is in range	5.5	10	1	1	3
Glucose (mL)	Is in range	5.5	10	1	1	3
$MgSO_4$ (mL)	Is in range	9.5	14	1	1	3
Duration (h)	Is in range	2	5	1	1	3
OD	Is in range	1.12	1.79	1	1	3
POE (OD)	Is in range	0.075	0.35	1	1	3
T-20 (%)	Maximize	8.64	14.45	1	1	5
POE (T-20 [%])	Minimize	0.65	5.56	1	1	5

Table 10
Selected Formulations With POE

NPK (mL)	Glucose (mL)	$MgSO_4$ (mL)	Duration (h)	OD	POE (OD)	T-20 (%)	POE (T-20 [%])	Desirability
8.17	5.50	11.3	3.53	1.43	0.193	13.9	0.960	0.955
7.93	5.50	11.6	3.55	1.43	0.193	13.9	0.987	0.954
5.65	9.85	9.50	3.50	1.46	0.220	13.9	1.10	0.951
5.54	9.96	9.50	3.46	1.46	0.220	13.9	1.14	0.950

els, and information about the SD of factors should be prepared. Then, a POE model of the factor variation transmitted to the selected response is generated, as shown in Figs. 8 and 9. Ultimately, optimal factor settings can be detected that get the selected response on target with minimal variation by employing numerical optimization, with setting the goal for POE to minimize, as shown in Tables 9–12. The final optimal formulation obtained from the POE method is not identical to the one from the ordinary optimization method displayed in Tables 5–8. In addition, the confidence and prediction intervals of both formulae are inconsistent with each other. Based on the POE analysis, the conditions of 8.17 mL of NPK, 5.5 mL of glucose, and 11.3 mL of $MgSO_4$ are recommended for optimal formulation owing to the lowest POE value for %T-20 production among other formulae. It appears that a higher ratio of NPK to $MgSO_4$ in the mixture than that of optimal formulation is required to have a robust process of lower POE. It is logical that NPK mainly consists of buffer components as introduced in Materials and Methods.

A further verification experiment with this optimal formulation was not performed because this formulation was targeted for the robust process under the suggested optimal conditions and no higher production of T-20 was expected from further experimental design. Our study demonstrates the applicability of statistical theories to the optimization and formulation of recombinant overexpression. Further studies for enhancing T-20 pro-

Table 11
Point Optimization for Crossed Design With POE

Component	Name	Level	Low level	High level	SD
A	NPK (mL)	8.2	5.5	10	1.5
B	Glucose (mL)	5.5	5.5	10	1.5
C	$MgSO_4$ (mL)	11.3	9.5	14	1.5
D	Duration (h)	3.53	2	5	1
	Total (mL)	25			

Table 12
Prediction and Statistical Interval for Crossed Design

Response	Prediction	SE mean	95% CI: low	95% CI: high	SE Pred	95% PI: low	95% PI: high
OD	1.432	0.034	1.359	1.505	0.059	1.306	1.558
POE (OD)	0.193						
T-20 (%)	13.934	0.456	12.956	14.912	0.788	12.243	15.625
POE (T-20 [%])	0.982						

SE, standard error; CI, confidence interval; PI, prediction interval.

duction should be focused on fermentation kinetics in bioreactors and metabolic and genetic engineering under optimized conditions obtained from the present research.

Acknowledgment

This research was supported by a grant from the Korea Science and Engineering Foundation (KOSEF).

References

1. Richman, D. (1998), *Nat. Med.* **4(11),** 1232–1233.
2. Rimsky, L. T., Shugars, D. C., and Matthews, T. (1998), *J. Virol.* **72,** 986–992.
3. Wei, X., Ghosh, S. K., Taylor, M. E., et al. (1995), *Nature* **373,** 117–122.
4. Chan, D. C., Fass, D., Berger, J. M., and Kim, P. S. (1997), *Cell* **89,** 263–273.
5. Kilby, J. M., Hopkins, S., Venetta, T. M., et al. (1998), *Nat. Med.* **4(11),** 1302–1307.
6. Rice, W. G., Supko, J. G., Malspeis, L., et al. (1995), *Science* **270,** 1194–1197.
7. Vanot, G., Valerie, D., Guilhem, M. C., Phan-Tan-Luu, R., Comeau, L. C. (2002), *Appl. Microbiol. Biotechnol*. **60,** 417–419.
8. www.fuzeon.com (accessed July 7, 2004).
9. Lee, K.-M. and Gilmore, D. F. (2005), *Process Biochem*. **40,** 226–249.
10. Hu, R. (1999), *Food Product Design*, Technomic, Lancaster, PA.
11. Cornell, J. A. (2002), *Experiments with Mixtures*, 3rd ed., John Wiley & Sons, New York.
12. Myers, R. H. and Montgomery, D. C. (2002), *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 2nd ed., John Wiley & Sons, New York.
13. Montgomery, D. C. (2004), *Design and Analysis of Experiments*, 6th ed., John Wiley & Sons, New York.

14. Bollag, D. M., Rozycki, M. D., and Edelstein, S. J. (1996), *Protein Methods,* 2nd ed., John Wiley & Sons, New York.

15. Lee, K.-M. and Gilmore, D. F. (2005), *Appl. Biochem. Biotechnol*. **133,** 113–148.

16. Montgomery, D. C. (1997), *Introduction to Statistical Quality Control,* 3rd ed., John Wiley & Sons, New York.

17. Mitra, A. (1998), *Fundamentals of Quality Control and Improvement,* 2nd ed., Prentice Hall, Upper Saddle River, NJ.