Application of Principal Component Analysis and Self-organizing Map to the Analysis of 2D Fluorescence Spectra and the Monitoring of Fermentation Processes

Jong Il Rhee[1,2,4]*, Tae-Hyoung Kang[3,4], Kum-Il Lee[3,4], Ok-Jae Sohn[2,4], Sun-Yong Kim[2,4], and Sang-Wook Chung[3,4]

[1] School of Applied Chemical Engineering, Chonnam National University, Gwangju 500-757, Korea
[2] Laboratory of BioProcess Technology, Chonnam National University, Gwangju 500-757, Korea
[3] Department of Industrial Engineering, Chonnam National University, Gwangju 500-757, Korea
[4] Research Center for Biophotonics, Chonnam National University, Chonnam National University, Gwangju 500-757, Korea

Abstract 2D fluorescence sensors produce a great deal of spectral data during fermentation processes, which can be analyzed using a variety of statistical techniques. Principal component analysis (PCA) and a self-organizing map (SOM) were used to analyze these 2D fluorescence spectra and to extract useful information from them. PCA resulted in scores and loadings that were visualized in the score-loading plots and used to monitor various fermentation processes with recombinant *Escherichia coli* and *Saccharomyces cerevisiae*. The SOM was found to be a useful and interpretative method of classifying the entire gamut of 2D fluorescence spectra and of selecting some significant combinations of excitation and emission wavelengths. The results, including the normalized weights and variances, indicated that the SOM network is capable of being used to interpret the fermentation processes monitored by a 2D fluorescence sensor.

Keywords: fermentation processes, monitoring, principal component analysis, self-organizing map, 2D fluorescence sensor

INTRODUCTION

Fluorescence sensors have been widely used to monitor and control biotechnological processes [1-3]. In particular, 2D fluorescence sensors, which permit the simultaneous scanning of a whole range of excitation (250~650 nm) and emission (280~700 nm) wavelengths, have recently been used to monitor a variety of fermentation processes with microorganisms such as *Enterobacter aerogenes*, *Pseudomonas fluorescens*, *Escherichia coli*, and *Claviceps purpurea* [4-8]. Some significant correlations of the fluorescence spectra with process variables, *i.e.* the pH, cell mass, and the concentrations of the substrate and product, have also been found using various types of methods, such as the spectra subtraction and multivariate calibration methods [9-14], which include artificial neural networks [9] and partial least square regression analysis (PLS) [10,11].

Principal component analysis (PCA) is one of the most frequently used chemometric methods with spectroscopic data. This method allows all of the spectroscopic data to describe large amounts of data sets synthetically with the minimum loss of information [15,16]. PCA makes it possible to extract pertinent information related to the properties of the system being investigated. The investigation of the structural changes in milks [17], the analysis of the excitation-emission fluorescence matrices of olive oils [18] and the reduction of the dimensions of the fluorescence spectra used to monitor waste water treatment processes [19] are all examples of information that can be obtained using PCA. To our knowledge, this is the first study dealing with the monitoring of fermentation processes based on the analysis of 2D fluorescence spectra by PCA.

The self-organizing map (SOM), which belongs to the domain of unsupervised neural network algorithms, is considered to be a nonlinear surrogate to PCA. It has proved to be quite a simple applicative algorithm and an excellent tool for the nonlinear mapping of vectorial data in various spectrometers. The SOM has been used to cluster NMR spectra [20] and to visualize the spectroscopic data obtained using ion mobility spectrometry in a yeast fermentation process [21]. It has also been applied to the classification of chromatographic systems [22]. In our previous study the SOM technique was used to classify the 2D fluorescence spectra produced in various fermentation processes and to analyze the processes qualitatively [23].

***Corresponding author**
Tel: +82-62-530-1847 Fax: +82-62-530-0846
e-mail: jirhee@chonnam.ac.kr

Table 1. Operating conditions of the four fermentation processes

	FmPro1	FmPro2	FmPro3	FmPro4
Microorganism	recomb. *E. coli*	recomb. *E. coli*	*S. cerevisiae*	*S. cerevisiae*
Culture medium	MS8	LB	SM	SM
Process operating conditions	pH 6.2 37°C 1 vvm 450 rpm	pH 6.5 37°C 1 vvm 450 rpm	pH 5.5 30°C 1 vvm 350 rpm	pH 5.5 30°C 1 vvm 350 rpm
Addition of other components	Succinate, LA Glycine, IPTG	Succinate, LA Glycine	Glut+Gly (at 0 h), Cys (at 11 h)	Glut+Gly (at 11 h), Cys (at 11 h)

These two methods (PCA & SOM) are qualitative analysis methods for spectroscopic data. One or both of these methods have been used to correlate spectral data to some parameters. However, a comparison of the two methods in the analysis of 2D fluorescence spectra has not been performed before. Therefore, this study focused on the application of the PCA and SOM methods to the analysis of 2D fluorescence spectra and to the monitoring of fermentation processes with recombinant *E. coli* and *Saccharomyces cerevisiae* in an unsupervised manner.

MATERIALS AND METHODS

Fermentation Processes with a 2D Fluorescence Sensor

Recombinant *E. coli* BL21(DE3)pLysS (Invitrogen Co., USA) harboring the plasmid pFLS 45 with the *lac* promoter was used to produce extracellular 5-aminolevulinic acid (ALA) in a bioreactor system. For the fermentation of recombinant *E. coli*, chemically defined minimal medium [24] and LB medium were used with two precursors (succinic acid and glycine) for ALA and an inhibitor (levulinic acid, LA) of ALA dehydratase. The details of the analysis of the cell mass, ALA, substrate and organic acids, *etc.*, were described in our previous paper [24].

Yeast *S. cerevisiae* ATCC7754 (American Type Cell Collections, USA) was employed for the production of intracellular glutathione (GSH). A chemically defined medium (SM) with different glucose concentrations was used with three precursors (glutamic acid, cysteine, and glycine) for GSH [25]. The concentration of intracellular GSH was determined using the method reported by Tietze [26]. Cysteine was also analyzed using a colorimetric method based on the reaction between copper ions (II), iron ions (III), and 1.10-phenmonohydrate [27].

Four fermentation processes with recombinant *E. coli* and *S. cerevisiae* were monitored on-line with a 2D fluorescence sensor. The operating conditions of each fermentation process are listed in Table 1.

The 2D fluorescence sensor used in this study consisted of a spectrofluorometer (Model F-4500, Hitachi Co., Japan) and a 2-m bifurcated liquid light conductor (Lumatek GmbH, Germany) which was connected to the quartz window in the 19-mm electrode port of a stainless steel bioreactor system in conjunction with various sensors, such as pH- and dissolved oxygen (DO)-meters and an O_2-/CO_2-analyzer. A computer using homemade software was used for the configuration and control of the 2D fluorescence sensor, data acquisition, and for the direct display of the monitoring results and data saving. The measurement conditions of the 2D fluorescence sensor were as follows: scanning speed, 500 nm/sec; excitation and emission slits, 10 nm; excitation wavelength range, 250~650 nm; emission wavelength range, 280~650 nm. A scan of the whole spectrum with these parameters took 90 sec.

Principal Component Analysis (PCA)

PCA is used to reduce the dimensionality of multivariate data and to transform interdependent variables into significant and independent components [28].

The entire set of fluorescence spectra gathered during fermentation can be structured in sample numbers, *i.e.* the fermentation time and combinations of the excitation and emission wavelengths (CWLs). PCA decomposes a given spectral data matrix (**X**) as the sum of the outer product of the vectors q_a and p_a plus a residual matrix, E, using the following equation:

$$\mathbf{X} = \sum_{a=1}^{n} \mathbf{q}_a \mathbf{p}_a^T + \mathbf{E} = \mathbf{Q}\mathbf{P}^T + \mathbf{E} \qquad (1)$$

where **Q** is known as the scores matrix and contains information regarding the relationship between the samples. **P** stands for the loading matrix and includes information regarding the relationship between the variables. PCA, which uses far fewer factors than the original variables with no significant loss of information, was performed using code written in MATLAB software (vers.6.1, The MathWorks Inc., USA) [29].

Self-organizing Map (SOM)

The SOM projects high-dimensional data sets onto a space of lower dimension, while preserving the topological relationships of the input data sets. All of the fluorescence spectra collected during fermentation were projected onto a one-dimensional network consisting of the fermentation time, and the number of neurons in the output layer was equal to the number classified in the spectral data [23].

To classify the total fluorescence spectra using the SOM algorithm, the spectral data were first transformed into one-dimensional input data sets according to the CWLs. All of the spectral data so transformed were introduced into the node of the input layer in the vector form and then sent through the SOM network. Each neuron of the network computed the Euclidean distance between a weight vector and an input vector. The output neurons in the output layer are usually arranged into a two-dimensional grid. Among all of the output neurons, the best matching unit (BMU) with the minimum distance between the weight vectors and the input vectors was chosen. For the BMU and its neighborhood neurons, the weight vectors were updated by the SOM learning rule, and each spectral data point for a given class was projected onto the data sets arranged according to the fermentation time after a learning step. In this way, the topological relationships hidden in the large amounts of input data sets were classified and visualized in a class distribution card.

The optimal number of classes in a given class distribution card was determined by estimating the degree of scattering of the fluorescence intensities of all the spectral components in the corresponding class by computing the time-dependent variance of the fluorescence intensities of all of the spectral components in the class [23].

The mean variance of the fluorescence intensities for all of the spectral components of all of the classes in a distribution card usually decreases with increasing number of classes and can be used as a criteria for determining the optimal number of classes. That is, when the difference between the mean variances of two consecutive classes is less than 5%, the lower class number is selected as the optimal number of classes, by means of which the whole spectral data can be classified [23].

The SOM algorithm was also implemented using the MATLAB Neural Network Toolbox (vers.6.1, The Math Works Inc.) [29].

RESULTS AND DISCUSSION

Fermentation Processes with Recombinant *E. coli* and *S. cerevisiae*

To produce extracellular ALA using recombinant *E. coli* and intracellular GSH using *S. cerevisiae*, a number of fermentations were performed in a bioreactor with various sensors, including a 2D fluorescence sensor. A large amount of on- and off-line measurement data, including the 2D fluorescence spectra, was collected during the fermentation process.

Fig. 1 shows the characteristics of cell growth and ALA production with recombinant *E. coli* in FmPro1 and FmPro2, *i.e.* in different fermentation media with the addition of various precursors and inducers [24]. Recombinant *E. coli* usually grows faster in a complex medium (*e.g.* LB medium) than in a chemically defined medium (*e.g.* MS8 medium) [30]. In Fig. 1, the dried cell weight (DCW) and concentration of carbon dioxide (CO_2) reached their maximum values at 18.0 and 17.5 h in FmPro1 respectively, while they reached their maximum values at 8.5 and 4.7 h in FmPro2. In FmPro1, the glucose concentration reached zero at 15.0 h. The DO concentration decreased for up to 16.0 h in FmPro1, whereas it decreased for up to 4.7 h in FmPro2. The concentration of extracellular ALA increased during the fermentation in FmPro1 and reached its maximum value at the end of the fermentation, *i.e.* at 25.0 h, while it reached its maximum value at 12.5 h in FmPro2.

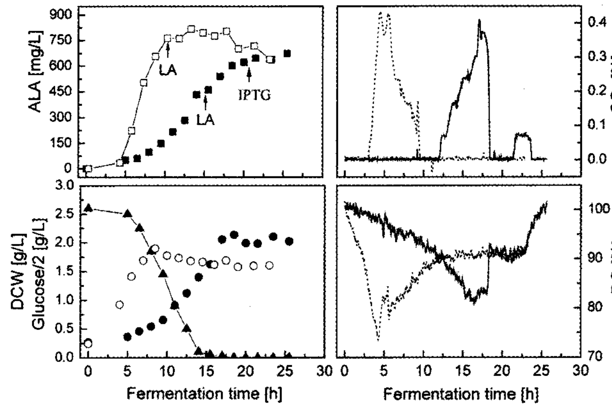

Fig. 1. On- and off-line measurement data of two fermentation processes with recombinant *E. coli*: FmPro1 data for DCW (●), ALA (■), glucose (▲), CO_2 (—), and DO (—); FmPro2 data for DCW (○), ALA (□), CO_2 (---), and DO (---).

During the fermentation of *S. cerevisiae*, cysteine was added to the processes at 11.0 h, whereas glutamic acid and glycine were added to the bioreactor at the beginning of the fermentation in FmPro3 and at 11.0 h in FmPro4, respectively [25]. The cell growth and GSH production in FmPro3 and FmPro4 are shown in Fig. 2. After 10.5 h, the DCW in FmPro3 was higher than that in FmPro4, because a higher concentration of glucose (20 g/L) was added in FmPro3 than in FmPro4 (5 g/L). The glucose concentration in FmPro3 decreased very rapidly and its concentration reached zero after 10.5 h. However, a lower concentration of cysteine (8 mM) was introduced into the FmPro3 than into the FmPro4 (16 mM). Higher amounts of intracellular GSH were produced in FmPro3 than in FmPro4, but its maximum concentration was reached at 22.0 h in both processes. The concentrations of CO_2 and O_2 in the exhaust gas of the two processes were maximal and minimal at 10.5 h, respectively, although their concentrations were quite different.

The fluorescence spectra usually depend on such factors as the fermentation medium and process operating parameters. Figs. 3A and B show two of the fluorescence spectra recorded during the fermentation process, *i.e.* the fluorescence spectra of FmPro1 at 10.0 and 17.0 h. The change in the fluorescence intensity was also visualized by subtracting one fluorescence spectrum (t = 10 h) from the other (t = 17 h). There was a large difference in the fluorescence intensities ($FS_{17h} - FS_{10h}$) in the regions of NAD(P)H (360 nm (ex)/440 nm (em), which is used

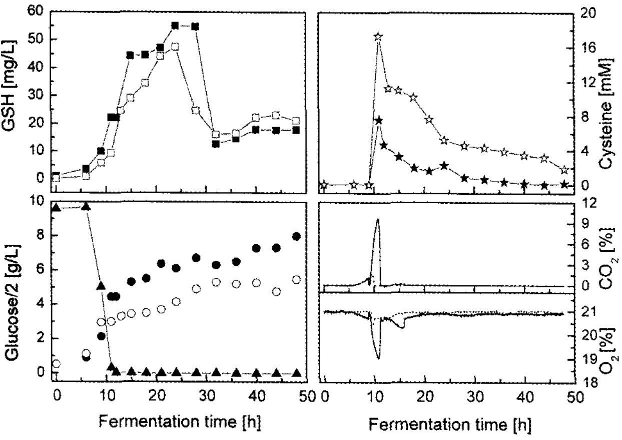

Fig. 2. On- and off-line measurement data of two fermentation processes with *S. cerevisiae*: FmPro3 data for DCW (●), GSH (■), glucose (▲), CO_2 (—), O_2 (—), and cysteine (ξ); FmPro4 data for DCW (○), GSH (□), CO_2 (---), O_2 (---), and cysteine (ψ).

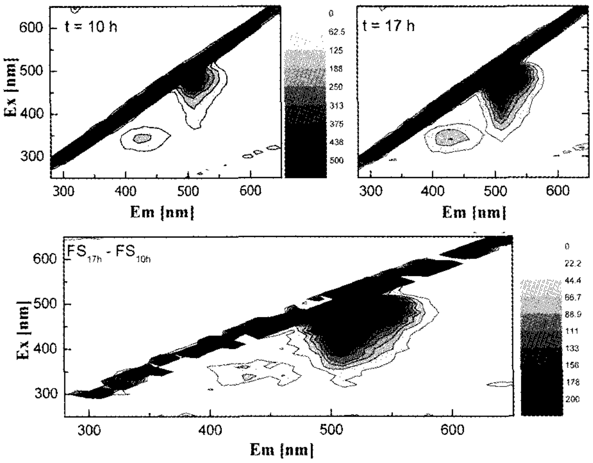

Fig. 3. Fluorescence spectra at 10.0 h (a) and at 17.0 h (b) and difference Difference in fluorescence spectra between 10.0 and 17.0 h in the case of FmPro1 (c).

as an indicator of the metabolic activity) and of EGFP (470 nm (ex)/520 nm (em), which is used as an indicator of the ALA production) between 10.0 and 17.0 h, because the cells were in the exponential growth phase and produced ALA. However, no distinct differences in the fluorescence intensities were visualized in the other fluorescence regions, so some available spectra could not be ascertained from large amounts of spectral data by the spectra subtraction technique [9,23]. Therefore, it is necessary to apply a holistic approach such as PCA and SOM to the interpretation of the overall fluorescence spectra.

Analysis of 2D Fluorescence Spectra by PCA and SOM

The large volume of 2D fluorescence spectra produced during fermentation either can be reduced in dimension by PCA or clustered into several classes by the SOM method. The score and loading plots produced by the PCA, as well as the class distribution card produced by the SOM, do not only help to understand the relationship between each fluorescence spectrum and the cellular states, but also provide some qualitative information on the fermentation process.

PCA

After filtering out some of the light scattering data, the spectral data can be reduced and used as the column of the fluorescence spectral matrix (**X**) in Eq. (1). The whole spectral matrix can then be decomposed into the score and loading data matrices. The number of columns in the fluorescence spectral matrix, *i.e.* 378 CWLs in this study, can affect the computation time of the data matrix and the amount of information. Therefore, a total of 189 CWLs, which were selected simply by using a scan interval of 20 nm, was employed to calculate the variance which can be captured by each PC. In the case of FmPro1, PC1, PC2, PC3, PC4, and PC5 captured 58.73, 4.97, 1.94, 0.84, and 0.79% of the total variances in the entire fluorescence spectra, respectively.

Figs. 4A and B show the score and loading plots of the PCA for FmPro1, respectively. The score plot contains 426 score data concerning the fermentation time and can be used to interpret the tendency of the fermentation process. For example, in the score plots of PC1 and PC2, the increase in the scores of PC1 and PC2 at the beginning of the fermentation represents the start of cell growth. The small increase and saturation in the scores of PC2 between 6.0 h (PC1 = –10.353; PC2 = 1.5907) and 15.0 h (PC1 = 9.8175; PC2 = 4.5761) may result from either the exponential growth of the cells or the fast consumption of the substrate.

The loading data are normalized between +1 and –1 and presented in the form of a one dimensional loading spectrum, *i.e.* the loading value vs. CWL selected using a scan interval of 20 nm. The loading values for the PCs indicate the importance of each CWL in the fermentation process. That is, those CWLs having a high loading value (*e.g.* over ± 0.09) show a large change in the fluorescence intensity as a function of the fermentation time, so that monitoring the process using the 2D fluorescence sensor based on these CWLs would provide more information on the cellular behavior. Many CWLs for PC1 have loading values of over 0.09, especially in the spectral regions of EGFP (470 nm (ex)/520 nm (em)) and NAD(P)H (360 nm (ex)/440 nm (em)), as shown in Fig. 4C, and provide some useful information on the cell growth and ALA production. Some high positive loading values are also observed for PC2 in the spectral regions of the proteins (270~290 nm (ex)/370~390 nm (em)), which provides us with some information on the change in the cellular proteins. PC3 has high positive loading values in the spectral region of amino acids such as tryptophan (250~270 nm (ex)/290~310 nm (em)) and negative loading values in the region of the proteins (290~310 nm (ex)/370~390 nm (em)) (Fig. 4C) [8]. In the bot-

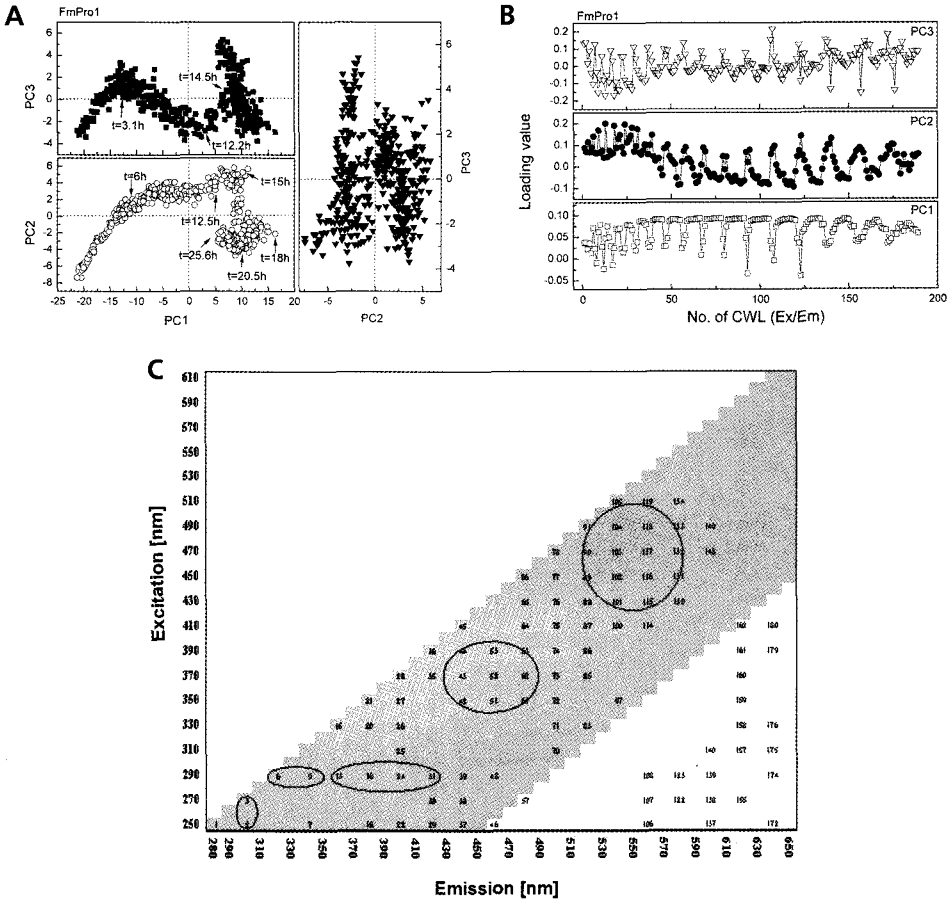

Fig. 4. (A) Score plots, (B) loading plots, and (C) CWLs (77 CWLs with loading values over +0.09) of PC1, PC2, and PC3 for FmPro1.

tom half of the fluorescence spectrum in Fig. 4C, PC2 and PC3 have a few CWLs with high loading values of over +0.09, due to the influence of the scattered light caused by the monochrometer of the 2D fluorescence spectroscopy. These CWLs should be subtracted when analyzing the fluorescence spectra. The CWL whose loading value for the PCs is very low or zero, provides little information on the behavior of a process, so must not be selected to monitor a fermentation process on-line. As a result, the fluorescence spectra collected during fermentation may be effectively analyzed by the PCA using 77 of the 189 CWLs selected simply based on a scan interval of 20 nm.

SOM

In FmPro3, the whole range of fluorescence spectra were collected during the fermentation with *S. cerevisiae* and clustered into several classes. After calculating the mean variance, the optimal number of classes was found to be 5 [23]. Fig. 5 shows the class distribution card for the 5 classes, including the central regions as well as the time courses of the normalized weights and the variances of each class. The CWLs within 20% of the largest elements in each class are represented as the central regions of each class. That is, the high values of the central regions showed areas where large changes in the time-dependent spectral data can be predicted. In Fig. 5A, the distribution card of the 5 classes had 27 central regions for class 1, 19 for class 2, 25 for class 3, 10 for class 4, and 17 for class 5.

The trends for the normalized weights may give some qualitative information about the cellular metabolism and characteristics of the fermentation process, such as the substrate consumption, the production of metabolites and enzymes, or the introduction of some components into the bioreactor [31]. In Fig. 5B, the time course of the normalized weight of class 1 in FmPro3 may represent the change in the concentration of cysteine due to its addition at 11.0 h, when compared with the on- and off-line data in Fig. 2. The change in the variances according to time, *i.e.* the degree of scattering of the fluorescence intensity of a class with regard to the fermentation time, also provides some information about the fermentation process. The change in the variances of class 3 is also similar to that of the DCW in FmPro3 in Fig. 2.

Monitoring the Fermentation Processes

The analysis of 2D fluorescence spectra by PCA and

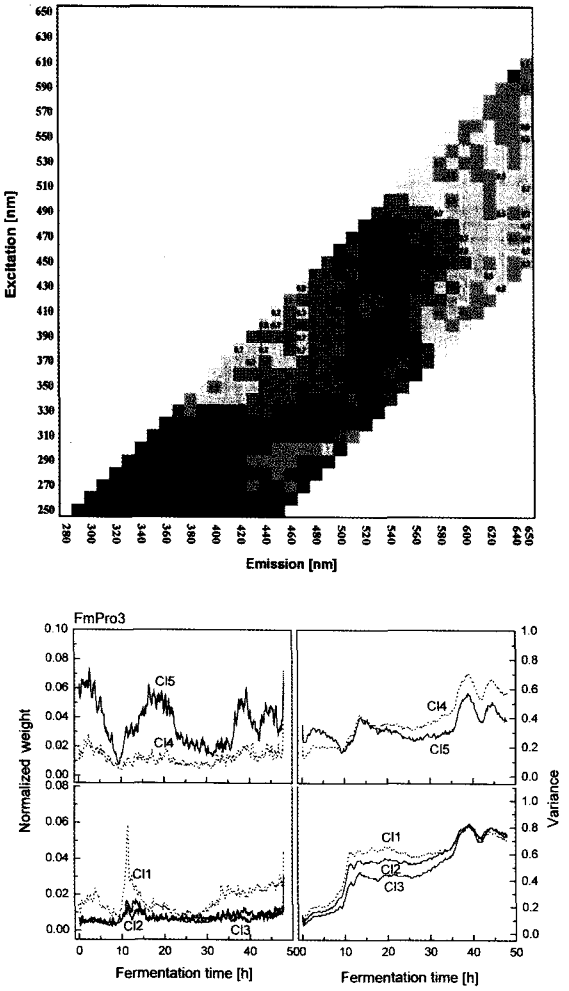

Fig. 5. (A) Class distribution card of the 5 classes including the 98 central regions: (■) class 1; (■) class 2. ( ) class 3; (▨) class 4; (■) class 5, and (B) time courses of the normalized weights and the variances of each class (Cl) for FmPro3.

SOM helps to monitor a fermentation process qualitativeely on-line. Herein, four fermentation processes are monitored based on the score plots of the PCs, as well as the normalized weights and variances of the classes in a given class distribution card in an unsupervised manner.

Fermentation Process 1 (FmPro1)

For FmPro1, the score plots of the PCs produced by the PCA shown in Fig. 4A and the normalized weight and variances of the classes produced by the SOM shown in Fig. 6 can be compared with the on- and off-line measurement data shown in Fig. 1.

In the score plot of PC1 and PC2 in Fig. 4A, the scores (426 data points) of PC1 increased with increasing fermentation time until 18.0 h and decreased thereafter until the end of the fermentation process. However, the scores of PC2 increased from the beginning of the fermentation process up to 15.0 h, decreased from 15.0 to 20.5 h and then increased till the end of the fermentation process. The increase in the scores of PC1 and PC2 at the beginning of the fermentation process were found to represent the start of cell growth, when they were compared with the on- and off-line data in Fig. 1. The small increase and saturation in the scores of PC2 between 6.0 and 15.0 h resulted from the exponential growth of the cells or the fast consumption of the substrate. The increase in the scores of PC2 between 12.5 and 15.0 h possibly reflected the maximum cell growth rate and the start of CO_2 accumulation. The complete consumption of the substrate was also reflected in the sharp decrease in the scores of PC2 after 15.0 h. The decrease in the scores of PC1 between 18.0 and 25.6 h represented the stationary phase of cell growth, and the change in the scores of PC2 at 20.5 h showed that some change in the CO_2 concentration occurred due to the addition of IPTG [32].

The scores of PC1 and PC3 increased from the beginning of the fermentation to 3.1 h, which represented the lag phase. The decrease in the scores of PC3 from 3.1 to 12.2 h represented the exponential growth of the cells. The increase in the scores of PC3 between 12.2 and 14.5 h also reflected the attainment of the maximum cell growth rate and the start of CO_2 accumulation. After 15.1 h, a complex phenomenon was observed in the score data of PC1 and PC3, due to various factors, including the consumption of the substrate, the addition of LA and IPTG.

PC2 and PC3 captured only 6.91% of the total variances in the entire fluorescence spectra and therefore would be expected to explain very little of the characteristics of the fermentation process. The scores of PC3 in the score plot of PC2 and PC3 oscillated, so that it is difficult to explain the process based on the measurement data in Fig. 1.

The class distribution card of the 5 classes had 34 central regions for class 1, 14 for class 2, 21 for class 3, 13 for class 4, and 16 for class 5. The time courses of the normalized weights and the variances of some of the classes in Fig. 6 were compared with the on- and off-line measurement data in Fig. 1.

In Fig. 6, the time courses of the normalized weights of classes 3 and 4 for which the spectral components agreed well with the spectral region of NAD(P)H (data not shown [18]), might be associated with the dried cell weight (DCW), whereas the normalized weight of class 2 might be correlated with the courses of the ALA concentrations in Fig. 1. The steep increase in the variance of class 1 after 20.0 h may result from the production of various metabolites, such as acetic acid, or large amounts of foam, and the addition of antifoam agents to the bioreactor.

Fermentation Process 2 (FmPro2)

In the case of FmPro2, PC1, PC2, and PC3 captured 35.64, 25.43, and 2.27% of the total variances in the entire fluorescence spectra, respectively. The score plots of PC1, PC2, and PC3 in Fig. 7A can also be compared with the on- and off-line measurement data shown in Fig. 1.

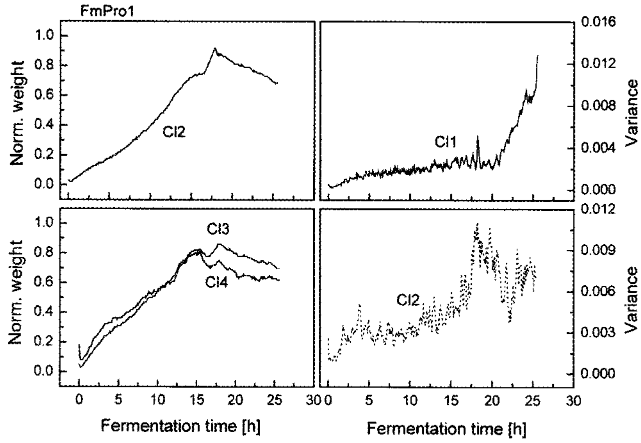

Fig. 6. Time courses of the normalized weights and the variances of some of the classes (Cl) for FmPro1.

In the score plots of PC1 and PC2, the increase in the scores of PC1 from the beginning of the fermentation until 8.5 h represented the increase in DCW, *i.e.* cell growth. Between 10.0 and 13.0 h the scores of PC2 decreased, whereas those of PC1 remained almost constant. These trends represented the maximal concentration of ALA in Fig. 1. The slow decrease in the scores of PC1 after 13.0 h indicated the slow degradation of ALA to porphobilinogen (PBG).

The meaning of the score values of each PC in the score plots of PC1 and PC3 after 5.6 h is not clear, due to the complexity of cell metabolism, which includes cell death and biosynthesis of the various amino acids. However, the fluctuation of the scores of PC3 between 4.7 and 5.6 h represented the two maximal peaks of the CO_2 concentrations in the exponential growth phase of the cells. The score values of PC2 and PC3 also reflected some trends in the cell growth (DO, CO_2 *etc.*). That is, the steep decrease of DO between 0.1 and 4.7 h corresponded to the sharp increase in the scores of PC3, and the decrease in the CO_2 concentrations after 5.6 h agreed well with the decrease in the scores of PC3.

The cellular metabolism and biosynthesis of ALA might be well explained through the change in the score values of each PC in the score plots, if more on- and off-line measurement data were to be obtained.

The entire range of fluorescence spectra in FmPro2 could also be classified into 8 classes. There were 13 central regions for class 1, 14 for class 2, 5 for class 3, 12 for class 4, 5 for class 5, 18 for class 6, 11 for class 7, and 20 for class 8.

The concentrations of ALA in Fig. 1 could be associated with the time courses of the normalized weights of class 8 shown in Fig. 7B, for which some of the spectral components lie in the fluorescence region of the EGFP (470 nm (ex)/520 nm (em)). The time courses of the normalized weights of class 7 corresponded to the changes in DCW in Fig. 1. However, class 7 in the class distribution card (Figure not shown) does not belong to the fluorescence region of NAD(P)H (360 nm (ex)/440 nm (em)), and may be due to the degradation and bio-synthesis of some amino acids in the LB-complex medium. The change in the variance of class 4, which lies in the region of NAD(P)H, was associated with the time course of CO_2 shown in Fig. 1.

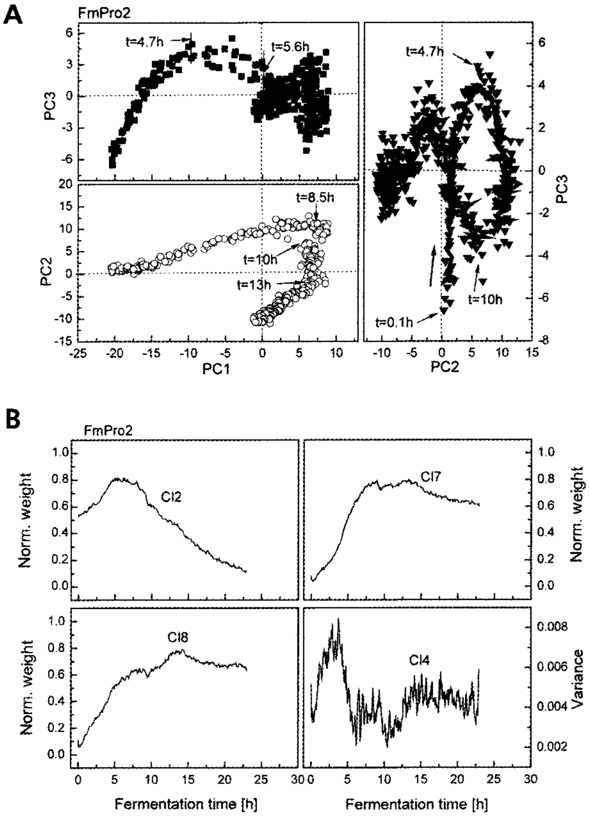

Fig. 7. (A) Score plots of PC1, PC2, and PC3 and (B) time courses of the normalized weights and the variances of a few classes (Cl) for FmPro2.

Fermentation Process 3 (FmPro3) and 4 (FmPro4)

The score plots of the PCs in FmPro3 and FmPro4 are shown in Fig. 8. The cell growth, as well as the difference in the time of addition of the two amino acids (glutamic acid and glycine) to the process, can be interpreted by comparing the score plots of the two processes. Table 2 presents the total variances captured by the 5 PCs in both processes.

From the score plots shown in Fig. 8A, the trends in the score values of the PCs at the beginning of fermentation and at 11.0 h in FmPro3 were different from those in FmPro4. Glutamic acid and glycine were added to FmPro3 at the beginning of fermentation, whereas they were added to FmPro4 at 11.0 h. This difference can be observed in the change of the scores of PC1 and PC2 in the score plots. The score values of PC1 and PC3 reflected the addition of cysteine to both processes at 11.0 h. In the score plots of PC1 and PC3, the increase in the score values of PC3 starting from 11.0 h might also result from the change from an oxidative to an oxidoreductive metabolism [8], for example the conversion of glucose to

Table 2. Total variances captured by 5 PCs in FmPro3 and FmPro4 (%)

	PC1	PC2	PC3	PC4	PC5
FmPro3	41.87	7.17	2.72	1.59	0.92
FmPro4	41.11	4.49	3.39	1.03	0.92

ethanol. The increase in the scores of PC3 starting from 35.0 h may represent the production of other metabolites or the degradation of GSH within the cells, as shown in Fig. 2.

In FmPro3, the difference in the mean variances between classes 5 and 6 was less than 5%, while the difference in the mean variances between classes 7 and 8 in FmPro4 was less than 5%. Therefore, the fluorescence spectra gathered in FmPro3 and FmPro4 were classified into 5 and 7 classes, respectively. Class 1 (290~340 nm (ex)/330~490 nm (em)) and class 3 (340~410 & 370~590 nm (ex)/380~460 & 560~650 nm (em)) in FmPro3 have 52 central regions which capture more than 50% of the useful information about the process. However, in FmPro4, classes 5 and 6 within the wavelength range of 340~560 nm (ex) and 400~640 nm (em) contain about 43% of the biological and environmental information concerning the process.

The difference in the addition time of the two amino acids can also be interpreted by comparing the normalized weights of some of the classes in the class distribution cards of FmPro3 and FmPro4. In Fig. 8B, the normalized weights of class 1 in FmPro3 and class 5 in FmPro4 correspond well to the respective cell mass concentrations. The fluorescence region of NAD(P)H lay in the regions of class 1 in FmPro3 and class 5 in FmPro4. The sudden change in the normalized weights of class 5 in FmPro4 at 11.0 h might also correspond to the addition of the two amino acids. That is, the sudden decrease may result from some metabolic change within the cells due to the simultaneous addition of three amino acids to FmPro4. The concentrations of cysteine added to FmPro3 and FmPro4 at 11 h can also be interpreted by analyzing the time courses of the variances in class 1 in FmPro3 and class 3 in FmPro4.

In most previous studies, the large amounts of fluorescence spectra collected with a 2D fluorescence sensor were analyzed by multivariate regression methods such as partial least square regression [10-14,33]. For example, a multivariate calibration method applying parallel factor analysis (PARAFAC) was employed to quantify the anti-inflammatory materials in biological samples, based on 2D fluorescence spectral data [33]. The process variables, such as the cell mass and protein production were well correlated with the fluorescence spectra in the fermentation processes of *E. coli* and *S. cerevisiae* [13,14]. A backpropagation neural network (BPNN) was also introduced to interpret the highly complex fluorescence maps obtained from complex bioprocesses [9] and to predict the fermentation process of recombinant *E. coli* [29]. Although the multivariate calibration method and supervised neural networks have good accuracy and high prediction capability, they require extensive numerical computations when dealing with the entire range of the 2D fluorescence spectra.

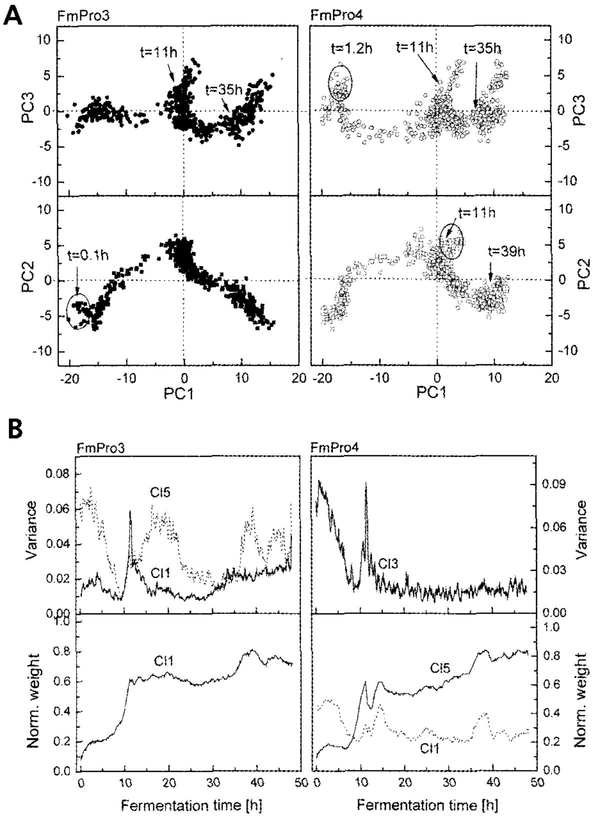

Fig. 8. (A) Score plots of PC1, PC2, and PC3 and (B) time courses of the normalized weights and the variances of some of the classes (Cl) for FmPro3 and FmPro4.

Therefore, in this study the 2D fluorescence spectra gathered during the fermentation process were analyzed with the PCA and SOM techniques. The analysis of the fluorescence spectra using the PCA and SOM methods helped to describe the fermentation process rapidly and qualitatively.

The PCA method was used to reduce the dimensionality of the fluorescence spectra and to find the relationship between the process parameters and PCs. In this work, the total score values of PC1 and PC2 for the 4 fermentation processes were about 50~60%, so that the use of a single score plot was insufficient to interpret the process. The low score values of the PCs may result from the use of 185 CWL selected based on a 20 nm scan, of which some of the CWLs had no significant variances regarding the fermentation time. The loading data provided us with some information, such as how large was the variance of each CWL, *i.e.* which CWLs had a substantial effect on the description of each PC. The larger the value of a particular CWL, the larger the variance of the fluorescence spectra, *i.e.* the more information that could be obtained on this particular CWL. Therefore, if some significant CWLs are selected among all of the CWLs of a fluorescence spectrum and used to analyze the spectra of a process, a higher percentage of the total variances can be captured, with the result that the score plots of the PCs will be able to interpret the fermentation process and be used to evaluate the process instability.

The SOM made it possible to classify a large amount of fluorescence spectra into a few useful classes. A large amount of the spectral data collected during the fermentation could be classified into a few classes (*e.g.* 2, 5, or 8 classes), and some important spectral components, *i.e.* CWLs, could be extracted from the whole range of spectral data (*e.g.* 98 among the 493 spectral components [23]). These spectral components could be used as the input data sets for the analysis of the whole fluorescence spectra by the PCA, as well as for the modeling of the fermentation process by a supervised neural network algorithm [29]. In the case where 98 CWLs were extracted from the SOM algorithm and used to analyze the whole range of fluorescence spectra of FmPro1, PC1, and PC2 captured 80.6% of the total variances, *i.e.* 16.9% more of the total variances than that obtained using a scan interval of 20 nm.

CONCLUSION

This study addressed the application of the PCA and SOM methods to the analysis of the whole range of fluorescence spectra obtained during the monitoring of fermentation processes. During the fermentations of recombinant *E. coli* and *S. cerevisiae*, large amounts of 2D fluorescence spectral data were collected, and the data were analyzed by the PCA and SOM methods.

The score plots of the PCs were used successfully to interpret the tendency of the fermentation processes. Some significant combinations of excitation and emission wavelengths could be selected from the whole fluorescence spectra based on the loading data. The SOM was used to classify the entire range of fluorescence spectra and to describe the relationships between certain parameters and variables in the fermentation processes phenomenologically.

The meaningful CWLs extracted from the entire range of spectral data by the PCA and SOM methods represent an important step forward in the modeling of biological processes. That is, the 77 CWLs selected from the loading data of the PCA or the 98 CWLs extracted by the SOM network in this study could be utilized to model the process by supervised neural networks or multivariate regression methods.

Acknowledgments This work was supported by grant No. R01-2002-000-00027-0 from the Basic Research Program of the Korea Science & Engineering Foundation (KOSEF) and grant No. RTI04-03-03 from the Regional Technology Innovation Program of the Ministry of Commerce, Industry and Energy (MOCIE).

ABBREVIATIONS AND NOTATIONS

ALA	5-Aminolevulinic acid
BMU	Best matching unit
BPNN	Backpropagation neural network
CWL	Combinations of the excitation and emission wavelength
DCW	Dry cell weight
DO	Dissolved oxygen concentration (%)
E	Residual matrix
EGFP	Enhanced green fluorescent protein
FmPro	Fermentation process
GSH	Glutathione
IPTG	Isopropylthiogalactoside
LA	Levulinic acid
LB	Luria bertini medium
NMR	Nuclear magnetic resonance
P	Loading matrix
PBG	Porphobilinogen
PC	Principal component
PCA	Principal component analysis
Q	Scores matrix
PLS	Partial least square regression analysis
SM	Chemically defined medium
SOM	Self-organizing map
X	Given spectral matrix

REFERENCES

[1] Sonnleitner, B. (2000) Instrumentation of biotechnological processes. pp. 1-64. In: K. Schugerl (ed.). *Advances in Biochemical Engineering and Biotechnology.* Springer, Berlin, Germany.

[2] Harms, P., Y. Kostov, and G. Rao (2002) Bioprocess monitoring. *Curr. Opin. Biotechnol.* 13: 124-127.

[3] Hantelmann, K., M. Kollecker, D. Hull, B. Hitzmann, and T. Scheper (2006) Two-dimensional fluorescence spectroscopy: a novel approach for controlling fed-batch cultivations. *J. Biotechnol.* 121: 410-417.

[4] Schügerl, K., C. Lindemann, S. Marose, and T. Scheper (1998) Two-dimensional fluorescence spectroscopy for on-line bioprocess monitoring. pp. 1-27. *Course Material for the Bioprocess Engineering Course.* Supertar, Island of Brac, Croatia.

[5] Mukherjee, J., C. Lindermann, and T. Scheper (1999) Fluorescence monitoring during cultivation of *Enterobacter aerogenes* at different oxygen levels. *Appl. Microbiol. Biotechnol.* 52: 489-494.

[6] Boehl, D., D. Solle, B. Hitzmann, and T. Scheper (2003) Chemometric modelling with two-dimensional fluorescence data for *Claviceps purpurea* bioprocess characterization. *J. Biotechnol.* 105: 179-188.

[7] Tartakovsky, B., M. Scheintuch, J.-M. Hilmer, and T. Scheper (1996) Application of scanning fluorometry for monitoring of a fermentation process. *Biotechnol. Prog.* 12: 126-131.

[8] Marose, S., C. Lindemann, and T. Scheper (1998) Two-dimensional fluorescence spectroscopy: a new tool for on-line bioprocess monitoring. *Biotechnol. Prog.* 14: 63-74.

[9] Wolf, G., J. S. Almeida, C. Pinheiro, V. Correia, C. Rodrigues, M. A. M. Reis, and J. G. Crespo (2001) Two-dimensional fluorometry coupled with artificial neural

networks: a novel method for on-line monitoring of complex biological processes. *Biotechnol. Bioeng.* 72: 297-306.

[10] Skibsted, E., C. Lindemann, C. Roca, and L. Olsson (2001) On-line bioprocess monitoring with a multi-wavelength fluorescence sensor using multivariate calibration. *J. Biotechnol.* 88: 47-57.

[11] Cimander, C. and C. F. Mandenius (2002) Online monitoring of a bioprocess based on a multi-analyser system and multivariate statistical process modelling. *J. Chem. Technol. Biotechnol.* 77: 1157-1168.

[12] Hisiger, S. and M. Jolicoeur (2005) A multiwavelength fluorescence probe: is one probe capable for on-line monitoring of recombinant protein production and biomass activity? *J. Biotechnol.* 117: 325-336.

[13] Eliasson Lantz, A., P. Jorgensen, E. Poulsen, C. Lindemann, and L. Olsson (2006) Determination of cell mass and polymyxin using multi-wavelength fluorescence. *J. Biotechnol.* 121: 544-554.

[14] Haack, M. B., A. Eliasson, and L. Olsson (2004) On-line cell mass monitoring of *Saccharomyces cerevisiae* cultivations by multi-wavelength fluorescence. *J. Biotechnol.* 114: 199-208.

[15] Jolliffe, I. T. (1986) *Principal Component Analysis.* Springer, New York, NY, USA.

[16] Bro, R. (2003) Multivariate calibration. What is in chemometrics for the analytical chemist? *Anal. Chim. Acta* 500: 185-194.

[17] Dufour, E. and A. Riaublanc (1997) Potentiality of spectroscopic methods for the characterization of dairy products. I. Front-face fluorescence study of raw, heated and homogenized milks. *Lait* 77: 657-670.

[18] Guimet, F, J. Ferre, R. Boque, and F. X. Rius (2004) Application of unfold principal component analysis and parallel factor analysis to the extrapolatory analysis of olive oils by means of excitation-emission matrix fluorescence spectroscopy. *Anal. Chim. Acta* 515: 75-85.

[19] Tartakovsky, B., L. A. Lishman, and R. L. Legge (1996) Application of multi-wavelength fluorometry for monitoring wastewater treatment process dynamics. *Water Res.* 30: 2941-2948.

[20] Dow, L. K., S. Kalelkar, and E. R. Dow (2004) Self-organizing maps for the analysis of NMR spectra. *BioSilico* 2: 157-163.

[21] Kolehmainen, M., P. Ronkko, and O. Raatikainen (2003) Monitoring of yeast fermentation by ion mobility spectrometry measurement and data visualization with self-organizing maps. *Anal. Chim. Acta* 484: 93-100.

[22] Debeljak, Z., M. Strapac, and M. Medic-Saric (2001) Application of self-organizing maps for the classification of chromatographic systems and prediction of values of chromatographic quantities. *J. Chromatogr. A* 925: 31-40.

[23] Rhee, J. I., K.-I. Lee, C.-K. Kim, Y.-S. Yim, S.-W. Chung, J. Wei, and K.-H. Bellgardt (2005) Classification of two-dimensional fluorescence spectra using self-organizing maps. *Biochem. Eng. J.* 22: 135-144.

[24] Chung, S.-Y., K.-H. Seo, and J. I. Rhee (2005) Influence of culture conditions on the production of extra-cellular 5-aminolevulinic acid (ALA) by recombinant *E. coli*. *Process Biochem.* 40: 385-394.

[25] Shimizu, H., K. Araki, S. Shioya, and K.-I. Suga (1991) Optimal production of glutathione by controlling the specific growth rate of yeast in fed-batch culture. *Biotechnol. Bioeng.* 38: 196-205.

[26] Tietze, F. (1969) Enzymic method for quantitative determination of nanogram amounts of total and oxidized glutathione: applications to mammalian blood and other tissues. *Anal. Biochem.* 27: 502-522.

[27] Teshima, N., H. Katsumate, M. Kurihara, T. Sakai, and T. Kawashima (1999) Flow-injection determination of copper (II) based on its catalysis on the redox reaction of cysteine with iron (III) in the presence of 1,10-phenanthroline. *Talanta* 50: 41-47.

[28] Geladi, P., B. Sthson, J. Nystrom, T. Lillhinga, T. Lestander, and J. Burger (2004) Chemometrics in Spectroscopy. *Spectrochim. Acta Part B* 59: 1347-1357.

[29] Lee, K.-I., Y.-S. Yim, S.-W. Chung, J. Wei, and J. I. Rhee (2005) Application of artificial neural networks to the analysis of two-dimensional fluorescence spectra in recombinant *E. coli* fermentation processes. *J. Chem. Technol. Biotechnol.* 80: 1036-1045.

[30] Kim, J. E., E. J. Kim, W. J. Rhee, and T. H. Park (2005) Enhanced production of recombinant protein in *Escherichia coli* using silkworm hemolymph. *Biotechnol. Bioprocess Eng.* 10: 353-356.

[31] Rhee, J. I., A. Ritzka, and T. Scheper (2004) On-line monitoring and control of substrate concentrations in biological processes by flow injection analysis systems. *Biotechnol. Bioprocess Eng.* 9: 156-165.

[32] Hur, W. and Y.-K. Chung (2005) On-line monitoring of IPTG induction for recombinant protein production using an automatic pH control signal. *Biotechnol. Bioprocess Eng.* 10: 304-308.

[33] Munoz de la Pena, A., N. Mora Diez, D. B. Gil, A. C. Olivieri, and G. M. Escandar (2006) Simultaneous determination of flufenamic and meclofenamic acids in human urine samples by second-order multivariate parallel factor analysis (PARAFAC) calibration of micellar-enhanced excitation-emission fluorescence data. *Anal. Chim. Acta* 569: 250-259.

[Received July 12, 2006; accepted September 25, 2006]