한국생물공학회지 제21권 제1호
Korean J. Biotechnol. Bioeng.
Vol. 21, No. 1, 59-67(2006)

생물공정 모니터링 및 모델링을 위한 2차원 형광스펙트럼의 다변량 분석

[1,5]강 태 형 · [2,4,5]손 옥 재 · [2,4]김 춘 광 · [1,5]정 상 욱 · † [3,4,5]이 종 일
전남대학교 공과대학 [1]산업공학과, [2]물질 · 생물화공과, [3]응용화학공학부, [4]생물공정기술연구실, [5]바이오광기반기술개발사업단
(접수 : 2005. 7. 29., 게재승인 : 2006. 2. 25.)

Chemometric Analysis of 2D Fluorescence Spectra for Monitoring and Modeling of Fermentation Processes

Tae-Hyoung Kang[1,5], Ok-Jae Sohn[2,4,5], Chun-Kwang Kim[2,4], Sang-Wook Chung[1,5], and Jong Il Rhee[3,4,5]†
[1]Department of Industrial Engineering, [2]Department of Material Chemical and Biochemical Engineering,
[3]School of Applied Chemical Engineering, [4]BioProcess Technology Lab., [5]Research Center for Biophotonics,
Chonnam National University, YongBong-dong 300, GwangJu 500-757, Korea
(Received : 2005. 7. 29., Accepted : 2005. 2. 25.)

2D spectrofluorometer produces many spectral data during fermentation processes. The fluorescence spectra are analyzed using chemometric methods such as principal component analysis (PCA), principal component regression (PCR) and partial least square regression (PLS). Analysis of the spectral data by PCA results in scores and loadings that are visualized in score-loading plots and used to monitor a few fermentation processes by *S. cerevisae* and recombinant *E. coli*. Two chemometric models were established to analyze the correlation between fluorescence spectra and process variables using PCR and PLS, and PLS was found to show slightly better calibration and prediction performance than PCR.

Key Words : Chemometric method, fermentation process, process monitoring, 2D fluorescence spectra

서 론

최근, UV-Vis 또는 적외선 분광광도계와 같은 각종 분광기들이 생물공정, 식품공정 등을 모니터링하고 제어하는데 사용되고 있다(1-4). 특히, 2차원 형광분광계는 생물공정의 비침투성 (noninvasive) 모니터링을 위해 많은 주목을 받고 있으며 각종 생물공정의 모니터링 및 제어에 대한 연구는 Scheper 등의 교수가 활발히 수행하고 있다(5-7).

2차원 형광분광계는 여기 파장 (excitation wavelength)과 방출 파장 (emission wavelength)의 많은 조합을 통해 다양한 특성의 형광물질을 연속적으로 스캔할 수 있고, 얻어진 데이터는 인공 신경망 (artificial neural network) 또는 다변량 (Chemometric analysis) 분석방법 등으로 처리할 수 있다(8, 9).

인공신경망은 많은 양의 형광스펙트럼 데이터를 분류하거나 생물공정의 어떤 수학적인 관계를 확립하기 위해 많이 사용되지만(10, 11), 노드사이의 연결에 시그모이드 함수 (sigmoid function)와 같은 비선형 함수가 필요하므로 복잡한 수학적 계산이 요구된다.

분광자료의 정량적인 분석을 위해서 주성분 분석법 (Principal Component Analysis, PCA), 부분최소제곱회귀법 (Partial Least Squares regression, PLS) 그리고 주성분회귀법 (Principal Components Regression, PCR)과 같은 다변량 분석방법들이 흔히 사용되고 있다(12, 13). 분광 데이터를 분석하는데 가장 널리 사용된 방법인 PCA는 정량적인 분석을 위해서 전체 스펙트럼을 사용하며 유용한 정보를 거의 손실하지 않고 종합적으로 분석할 수 있도록 한다(14). 예를 들면, PCA는 우유의 구조적인 변화에 대한 연구에서 중요한 시스템의 특성에 관련된 정보를 추출할 수 있게 하거나(15), 올리브 오일의 여기-방출 형광 데이터의 분석이나 폐수처리공정에서 모니터링 된 형광스펙트럼의 차원을 축소하는데 적용할 수 있다(16, 17).

한편, PCR 과 PLS 방법은 스펙트럼 데이터의 분석과 모델링을 위해 사용되는데(13, 18), PCR은 주성분 분석과 선형회귀 (linear regression)의 조합이라 생각하면 될 것이다. 예를 들면, PCR은 적외선과 라만 분광분석 데이터를 사용하여 phenacetin bulk powder의 입자 크기와 같이 어떤 물질의 특성에 대한 정량적인 분석을 하는 경우에 적용되어

† Corresponding Author : School of Applied Chemical Engineering, Chonnam National University, YongBong-dong 300, GwangJu 500-757, Korea
Tel : +82-62-530-1847, Fax : +82-62-530-0846
E-mail : jirhee@chonnam.ac.kr

졌다(19, 20). 그러나 PCR을 사용하여 2차원 형광스펙트럼 데이터의 분석에 관한 연구는 거의 이루어 지지 않고 있다. 가장 많이 사용되는 다변량 분석 방법 중의 하나인 PLS는 *Claviceps purpurea*의 발효 공정에서 얻어진 2차원 형광스펙트럼 자료를 이용하여 미생물의 균체량, 단백질 및 유기산의 농도, 배기가스 속에 CO_2, O_2 등을 모델링 하는데 사용되었고, 또한 *Pseudomonas fluorescence*의 발효공정에서 숙신산 농도와 같은 공정변수를 예측하기 위한 모델로 사용되었다(9, 21). 최근 온라인으로 측정된 2차원 형광스펙트럼과 균체량의 농도에 기초한 보다 개선된 PLS 모델이 세워졌으며 *Saccharomyces cerevisiae*의 발효공정에 적용되었다(22).

한편, PCR과 PLS의 성능을 발효공정에서 얻은 형광스펙트럼 데이터를 사용하여 비교하는 연구는 지금까지 거의 수행되지 않았다. 따라서 본 연구에서는 PCR과 PLS의 공정 모델링에 관한 성능을 재조합 대장균 *E. coli*와 효모 *S.cerevisiae*의 발효공정에서 얻은 2차원 형광스펙트럼을 사용하여 비교하였다.

즉, 본 논문에서는 발효공정에서 얻어진 2차원 형광스펙트럼 데이터를 분석하기위해 PCA를 사용하여 형광스펙트럼의 차원을 축소하고 공정의 모니터링할 뿐만 아니라 PCR과 PLS로 형광스펙트럼 데이터를 평가하여 공정의 모델링에 관한 성능을 비교하고자 한다.

재료 및 방법

2차원 형광분광계의 발효 시스템

미생물 발효시스템은 7.5 L 스테인리스 반응기 (KoBiotech Co., Korea)와 pH, DO 센서 (Mettler-Toledo Co., USA), O_2/CO_2 가스분석기 (Lokas Co., Korea) 그리고 온도, 교반속도, 거품제거 및 pH 제어기와 같은 제어 시스템들로 구성되어 있다. 온라인 데이터의 수집과 저장은 LabVIEW (vers. 6.1, National Instruments Co., USA) 프로그램을 이용하였다. 2차원 형광분광계 (Model F-4500, Hitachi Co., Japan)는 두 가닥으로 된 2 m 길이의 액체광학전도관 (Lumatek GmbH, Germany)을 스테인리스 생물반응기 표면에 설치된 직경 19 mm 석영창에 직접 연결하여 사용하였다. 형광분광계의 제어와 데이터 수집, 모니터링 결과의 표현 및 데이터의 저장을 위한 프로그램은 직접 작성되었다. 또한 측정조건은 스캔속도 30000 nm/min, 여기파장과 방출파장간격 10 nm, 여기파장범위 250-650 nm, 방출파장범위 280-650 nm로 설정되었다. 위의 측정조건에서 전 파장스캔은 1.5분이 소요되었다(24).

발효공정

5-aminolevulinic acid (ALA) 생산을 위해 플라스미드 pRLS45를 재조합 대장균 *E.coli* BL21(DE3) pLysS (Invitrogen Co., USA)에 삽입하여 사용하였다. 화학적으로 정의된 최소배지(MS8)(25)와 LB 복합배지는 전구체 (숙신산, 글라이신)와 ALAD의 저해제 (levulinic acid)가 첨가되어 사용되었다.

효모 *S.cerevisiae* ATCC7754 (American Type Cell Collections, USA)는 세포내 glutathione (GSH)의 생산을 위해 사용되었다. 발효 배지는 글루코오스, 염, 미량 원소, 비타민 그리고 글루탐산, 시스테인, 글라이신으로 구성된 합성배지(SM)이다(26). 각각의 발효공정에 대한 실험조건은 Table 1에 나타나 있다.

Table 1. Operating conditions of 4 fermentation processes

	FermPro1	FermPro2	FermPro3	FermPro4
Microorganism	recom. *E.coli*	recom. *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 0h), Cys (at 11 h)	Glut+Gly (at 11h), Cys (at 11 h)

Off-line 분석

재조합 대장균 *E.coli*을 발효하는 동안 시료들은 오프라인에서 얻으며 균체량, ALA, 포도당의 농도와 유기산 등이 결정된다. 균체량, ALA, 유기산 등의 분석에 대한 자세한 사항은 우리의 이전 논문(25)을 참조하면 될 것이다. 세포 내 GSH의 분석을 위해서 균들은 초음파 분쇄기로 파쇄하였다. 효모 발효에서 GSH의 농도는 글루타치온 환원효소와 5.5-dithiobis (2-nitrobenzoic acid)에 의한 효소 반응을 이용해 Tietze의 방법으로 결정된다(27). L-Cysteine은 cupper ion(II), ferrous ion(III) 그리고 1.10-phenmonohydrate 들간의 반응에 기초한 유색반응으로 분석된다(26).

다변량 분석 방법

주성분 분석 (PCA)

발효공정으로부터 얻어진 2차원 형광 스펙트럼은 발효시간과 여기 및 방출 파장 조합의 곱에 해당하는 많은 데이터를 가지고 있다. 선형 패턴인식 기술이라 할 수 있는 PCA는 다변량 데이터를 소수 (n)개의 주성분 (PC)으로 차원을 축소하여 분석하는 것이다. 즉, PCA는 주어진 형광분광 데이터 행렬(X)를 벡터 q_a와 p_a 외적의 합과 잔차 행렬 (E, error matrix)와의 합으로 분해한다.

$$X = \sum_{a=1}^{n} q_a p_a^T + E = QP^T + E \quad (1)$$

여기서 Q는 데이터들 간에 고려해야 할 정보를 가지고 있는 주성분 점수행렬 (score matrix)이고, P는 변수들 간에 고려해야 할 정보를 가지고 있는 적재행렬 (loading matrix)이다. PCA에서 첫번째 주성분 (q1 & p1 쌍 또는 PC1)은 데이터에서 가장 큰 분산을 가지고 있으며 두 번째 주성분은 그 다음 큰 분산을 갖는다. 순차적으로 그 다음 주성분은 그 다음 큰 분산을 가지게 된다. 따라서 PCA는 형광데이터를 중요한 정보의 손실 없이 원래의 데이터 변수보다 적은 소수개의 변수로 해석할 수 있게 한다(14).

주성분 회귀 (PCR)

PCR은 형광 데이터 (X)와는 직접적으로 관련은 없지만 PC들과는 관련이 있다. 즉, PC들은 서로 직교하므로 다중 선형회귀 (multiple linear regression, MLR)와 같은 다변량 회귀에서 공선성 (colinearity) 문제를 피할 수 있다(29). 따라서 PCR에서 출력 데이터 행렬(y)에 대한 회귀 등식은 다음과 같이 MLR과 PC의 조합으로 표현될 수 있다.

$$y = X\beta + E = XPP^T\beta + E = QB_{PCR} + E \quad (2)$$

여기서 Q는 XP와 같고 $B_{PCR}(=P^T\beta)$는 회귀계수 행렬을 나타낸다.

부분최소제곱 회귀 (PLS)

PLS는 형광 데이터와 공정 변수들의 내적 관계가 선형이라는 것을 이용하여 형광데이터의 가장 큰 분산을 갖는 요인을 찾기 위해 적용될 수 있다. 즉, 형광 데이터 행렬 (X)와 출력 데이터 행렬 (y)는 다음과 같이 분해된다.

$$X = QP^T + E = \sum_{k=1}^{A} q_k p_k^T + E \quad (3)$$

$$y = TU^T + F = \sum_{k=1}^{A} t_k u_k^T + F \quad (4)$$

여기서 T, U는 y를 주성분으로 분해할 때 사용한 주성분 점수와 적재행렬이고 F는 잔차행렬이다. X와 y의 관계를 살펴볼 수 있는 PLS의 내적 관계는 다음과 같다.

$$T = BQ + H \quad (5)$$

여기서 B는 단위행렬이고 H는 잔차행렬이므로 PLS에서 회귀등식을 다음과 같이 표기할 수 있다.

$$y = XB_{PLS} + F \quad (6)$$

위식에서 회귀계수 (B_{PLS})는 X와 y의 관계를 표현할 수 있는 적재행렬인 가중치와 관련이 있고 PLS에 대한 보다 깊은 설명 및 방법들은 다른 문헌 등을 참조하면 될 것이다(30).

위에서 언급된 다변량 분석방법은 상용 소프트웨어인 MatLab 6.2 (The MathWorks, Inc., Natick, USA)를 사용하여 계산하였다(31).

PCR과 PLS의 성능비교

다변량 회귀모델의 성능은 실험 데이터와 모델에 의해 추정된 데이터간의 평균차이를 추정하는 것으로 평가할 수 있다. 즉, 어느 공정에서 온라인 및 오프라인에서 얻은 데이터에 대해 외삽과 내삽의 과정을 거친 후 70%는 학습 데이터 (training data)로 30%는 보정 데이터 (calibration data)로 무작위로 선택하였다. 즉, 합리적인 데이터 선택을 위해 처음 10개에서 학습 데이터 7개와 보정 데이터 3개를 선택하고, 다음 10개에서 7개와 3개를 가각 선택하는 등 순차, 반복적으로 학습데이터와 보정데이터를 선택하였다.

PCR과 PLS의 보정 모델이 얼마나 적합한지를 알아보기 위해 3가지 기준값들을 계산하였다. 먼저, RMSEC (root mean square error of calibration)는 다음과 같이 정의된다.

$$\mathrm{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_{i,\mathrm{cal}} - y_i)^2}{n}} \quad (7)$$

여기서 $\hat{y}_{i,cal}$은 i번째 샘플에 대해서 모델로 보정된 값이고 y_i는 공정에서 측정한 i번째 샘플의 값이며 n은 보정하는데 사용한 샘플수이다.

새로운 데이터에 대한 모델의 예측력은 아래에 정의된 PRESS (predicted residual error sum-of-squares)와 RMSEP (root mean square error of prediction)로 계산할 수 있다.

$$\mathrm{PRESS} = \sum_{i=1}^{m}(\hat{y}_{i,\mathrm{pred}} - y_i)^2 \quad (8)$$

$$\mathrm{RMSEP} = \sqrt{\frac{PRESS}{m}}. \quad (9)$$

여기서 $\hat{y}_{i,pred}$는 모델을 형성하는데 포함되지 않은 i번째 샘플에 대해 모델로 예측된 값을 나타내고 m은 예측 데이터의 수를 나타낸다.

결과 및 고찰

PCA를 이용한 공정 모니터링

PCA는 발효공정에서 얻어진 많은 양의 형광 데이터의 규모를 줄여 공정 변화에 대하여 중요한 정보를 가지고 있는 공정변수들 또는 요인들의 조합을 찾는데 도움을 준다. 본 연구에서는 2차원 형광스펙트럼을 PCA의 주성분 점수와 적재행렬 데이터로 분석하였다.

여기 및 방출 파장의 최적조합수의 결정

형광분광계를 여기 파장은 250-650 nm, 방출 파장은 280-650 nm의 범위에서 10 nm 간격으로 어떤 발효공정을 50시간 동안 5분 간격으로 측정하도록 하면, 1588 (여기 및 방출파장의 조합수, CWL) x 600 (스캔 수)개의 전체 형광 데이터를 얻을 수 있다(24). 공정을 분석하는데 큰 영향을 미치지 않는 데이터를 제거하고 나면, 여기 및 방출파장의 조합수 (CWL)는 493개로 감소되며 수식(1)에서 형광스펙트럼의 행렬로 사용할 수 있다. 전체 형광스펙트럼, X [493 600]는 주성분 점수행렬 (score matrix)과 적재행렬 (loading matrix)로 분해하여 분석할 수 있다. 그러나 CWL은 공정을 분석하는데 중요한 영향을 미치므로 최적조합

수를 결정할 필요가 있어 본 연구에서는 SOM 분류(24)에 기초하여 493개의 20%인 98개로 줄인 경우, 스캔 간격을 20 nm로 하여 493개 중 25%인 124개를 얻은 경우, 그리고 형광세기의 표준편차 (SD)를 계산하여 7 이상이 되는 경우만을 선택한 경우 (예; 493개 중 247개)로 나누어 주성분 분석을 하는데 사용하였다. Table 2는 FermPro1과 FermPro2에 대하여 위의 세 종류의 CWL을 적용했을 때 얻어진 주성분 분석의 분산 값이다. 첫번째 주성분 (PC1)이 가장 큰 분산을 갖고 있음을 알 수 있는데 이 중에서도 SOM 분류에 기초한 여기 및 방출 파장의 조합수를 사용했을 때 큰 분산을 얻을 수 있는 것을 알 수 있다. 따라서 본 연구에서는 SOM 분류에 기초하여 여기 및 방출파장의 조합수를 얻은 후 PCA에 적용하여 공정을 분석하고 모델을 형성하였다.

Table 2. Variance values captured by the PCA with three different CWL for the ALA concentration in FermPro1 and FermPro2

		PC1	PC2	PC3	PC4	PC5
FermPro1	SOM	76.2	4.4	1.2	1.0	0.9
	20nm Scan	58.7	5.0	1.0	0.9	0.8
	SD(>7)	73.6	6.1	1.1	0.7	0.4
FermPro2	SOM	45.8	33.0	1.5	0.7	0.7
	20nm Scan	35.6	25.4	0.9	0.9	0.8
	SD(>7)	43.1	28.3	1.8	0.9	0.4

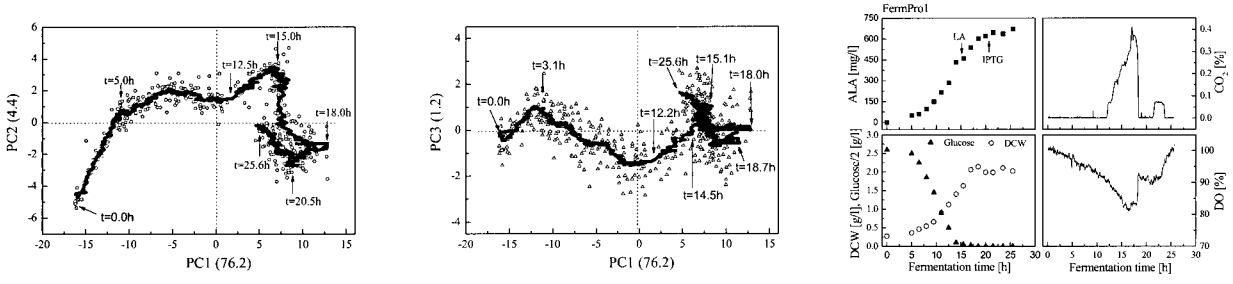
주성분의 점수 산점도 (Score plots)

PCA에 의해 얻어진 주성분의 점수 데이터는 2개의 주성분 (PC1 vs PC2 또는 PC1 vs PC3)이 축이 되는 2차원 좌표상에 나타낼 수 있다. 좌표상에 나타난 주성분 점수 산점도는 2차원 형광스펙트럼의 차원을 축소시킬 뿐 아니라 공정의 흐름에 대한 정보를 제공해 준다.

Fig. 1은 발효공정 FermPro1에 대한 2개의 주성분 점수 산점도와 온라인과 오프라인 측정 데이터를 산점한 그림을 보여주고 있다. 각각의 주성분 점수 산점도는 공정 시간에 대해 426개의 점수 데이터를 가지고 있으며 FermPro1의 발효공정의 경향을 이해하는데 사용될 수 있다. Fig. 1(a)에서 PC1의 주성분 점수 데이터는 발효시간이 18.0시간이 될 때까지 계속 증가한다. 반면 PC2의 주성분 점수 데이터는 발효 시작부터 15시간이 될 때까지 증가하다가 15~20.5시간에는 감소하고, 발효공정이 끝날 때 다시 증가하는 경향을 보인다. 이러한 PC1과 PC2의 주성분 점수 데이터의 경향이 Fig. 1(c)의 온라인과 오프라인 측정 데이터와 비교되어 공정을 분석하는데 이용되었다. 즉, 발효공정 초기에 PC1과 PC2의 주성분 점수 데이터의 증가는 미생물 성장의 시작을 나타낸다고 볼 수 있다. 그리고 5.0~12.5시간에서 PC2의 주성분 점수 데이터의 작은 증가는 균체의 지수적 증가 또는 기질의 빠른 소비 때문 일 것으로 생각된다. 12.5~15시간에서 PC2의 주성분 점수 데이터의 증가는 최대 균체 성장률과 CO_2가 누적되는 시점을 나타낸다. 또한 기질이 완전히 소비되는 15.0~18.0시간에서는 PC2의 주성분 점수 데이터의 급격한 감소를 나타낸다. 18.0~25.6시간에서 PC1의 주성분 점수 데이터의 감소는 균체 성장의 정지기를 나타내며 20.5시간에 PC2의 주성분 점수 데이터에 변화는IPTG 첨가로 인한 CO_2 농도에 변화 때문일 것이다. 공정 FermPro1에 대하여 PC1과 PC2의 분산은 형광스펙트럼 데이터의 전체 분산 중 80.6%를 차지하며 발효공정에 관련된 중요한 특성을 설명하는데 도움을 준다. Fig. 1(b)에서 PC1과 PC3의 주성분 점수 산점도는 FermPro1 공정에 대한 좀 더 많은 정보를 제공한다. PC1과 PC3의 주성분 점수 데이터는 세포 적응기를 나타내는 0.0~3.1시간까지는 증가하는 경향을 보이다가 3.1~12.2시간까지 PC3의 주성분 점수 데이터의 감소는 균체의 지수적 성장을 나타낸다. 또한, 12.2~14.5시간까지 PC3의 주성분 점수 데이터의 증가는 최대 균체 성장률과 CO_2가 누적되는 시점을 나타낸다. 15.1시간 이후, PC1과 PC3의 복잡한 현상들은 기질의 소비, LA와 IPTG의 주입 등의 이유 때문에 나타나는 것으로 볼 수 있다.

발효공정 FermPro2에 대하여 PC1과 PC2의 분산은 형광스펙트럼 데이터의 전체 분산중 78.8%를 차지하는데, PC1과 PC2의 주성분 점수 산점도를 DCW, DO와 함께 그림 2에 나타내었다. 발효초기, 0.0~7.4시간까지 PC1과 PC2의 주성분 점수 데이터의 증가는 DCW의 증가, 즉 균체 성장을 나타내는데, DO 농도가 최소가 되는 4.5시간에 PC2의 주성분점수 데이터의 변화는 아주 적었다. 배양배지에 LA가 첨가되는 14.4시간에 주성분 점수 데이터는 작은 변화를 나타내며 7.4시간부터 시작되는 PC1의 주성분 점수 데이터의 느린 증가는 DCW의 감소를 나타낸다고 할 수 있다.

효모의 발효공정, FermPro3과 FermPro4에서 두개의 주성분 점수 산점도를 Fig. 3에 나타냈다. 두 공정에 세 개의 아미노산(L-cysteine, L-glutamic acid, L-glycine)이 주입되었는데 L-glutamic acid와 L-glycine의 주입시간이 두 공정에서



(a) (b) (c)

Figure 1. Score plots and on- and off-line measurement data for FermPro1.

각기 다르다. FermPro3 에서 두개의 아미노산은 발효 초기에 주입되었지만, FermPro4 에서는 11시간에 주입되었다. 두 아미노산 주입시간의 차이를 두 공정의 주성분 점수 산점도의 비교를 통해 공정을 해석할 수 있다. Fig. 3(a)에서 FermPro4에 대한 PC2의 주성분 점수 데이터는 공정에 3개의 아미노산이 주입된 11시간에 상당히 높게 증가한다. 한편, Fig. 3(b)에서 FermPro3의 초기에 PC3의 주성분 점수 데이터의 두드러진 변화는 없지만 FermPro4에서는 발효초기 (0.0~4시간)에 증가함을 볼 수 있다. 이와 같이, 주성분 점수 산점도에서 점수 크기의 변화는 공정 시간에 따른 형광세기 변화가 얼마나 큰가를 나타내므로 공정의 불안정성을 찾아내거나 다변량 회귀 모델을 개발할 때 잡음을 제거하는데 사용될 수 있다.

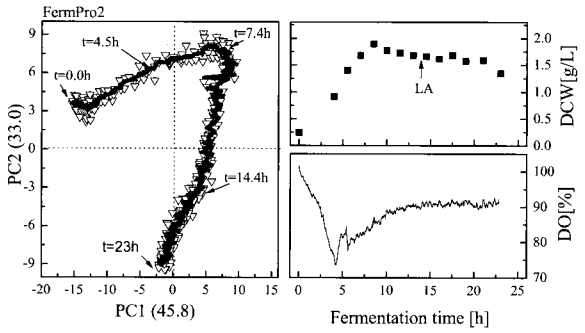
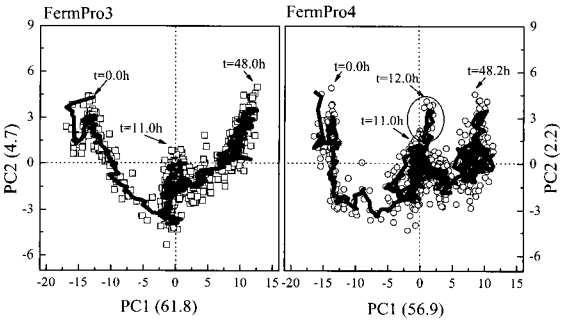


Figure 2. Score plot with DCW and DO measurement data for FermPro2.

(a)



(b)

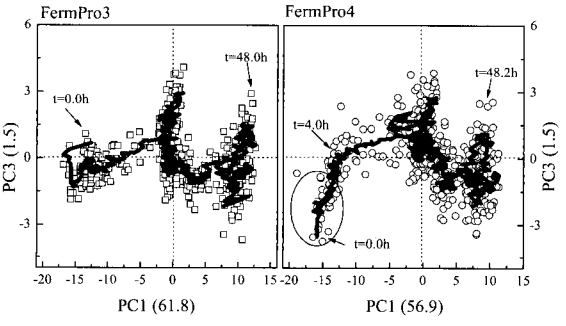


Figure 3. Two score plots for FermPro3 and FermPro4.

주성분의 적재 산점도 (Loading plots)

주성분의 적재 데이터는 여기 및 방출 파장 조합의 분산이 얼마나 큰 가에 대한 정보를 제공한다. 예를 들면, 적재 성분은 각각의 주성분을 묘사하는데 가장 큰 영향을 미치기 때문에 적재 성분의 크다는 것은 형광스펙트럼의 분산이 커진다는 것을 의미한다. 여기서 적재 데이터는 +1 과 1 사이로 표준화 하였고 1차원의 적재 스펙트럼으로 표현했다.

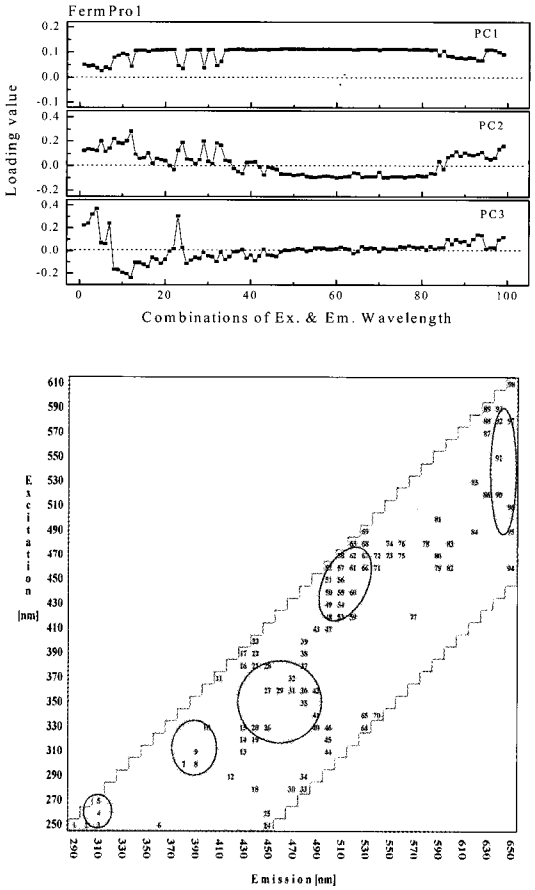


Figure 4. Loading plots of PC1, PC2 and PC3 for FermPro1, and 98 combinations of excitation and emission wavelengths for each loading component. The number in the classification map indicates the combination of excitation and emission wavelength for each loading component.

Fig. 4는 공정 FermPro1에 대한 PC1, PC2, PC3의 적재 산점도와 SOM 분류를 통해서 얻은 98개의 여기 및 방출 파장의 조합을 나타내고 있다. 각각의 PC에 대한 적재값은 발효공정의 특성에 의존적이며 여기 및 방출 파장의 조합들이 공정을 온라인 모니터링하기 위해 얼마나 중요한가를 보여준다. PC1에서 많은 적재 성분들이 EGFP (480 nm (ex)/520 nm (em))과 NADH (340 nm (ex)/440 nm (em)) 의 스펙트럼 영역에서 0.1의 적재값을 갖는데 이는 PC1이 균체 성장과 ALA 생산에 있어 유용한 정보를 갖고 있다는 것을 의미한다. 또한, PC2에서 양 (positive)의 적재값들은 세포내 단백질 (270-290 nm (ex)/370-390 nm (em))의 형광 파장 범위에서 얻어진다. PC3은 트립토판(250-270 nm

(ex)/290-310 nm (em))과 같은 아미노산의 형광 파장 범위에서 양(positive)의 높은 적재값을 갖고, 단백질(270-290 nm (ex)/370-390 nm (em))의 범위에서는 음 (negative)의 적재값을 갖는다(7). 한편, 각각의 주성분에 대한 적재값이 "0"이 되는 여기 및 방출 파장의 조합은 공정을 온라인 모니터링 하는 데에는 무의미한 형광영역이다. 이와 같이 주성분의 점수와 적재 데이터를 사용한 2차원 형광스펙트럼의 PCA 분석은 형광 스펙트럼과 균체 상태의 관계를 이해하는데 도움을 주며 발효공정에 대한 정성적인 정보를 제공한다.

PCR과 PLS에 의한 공정 모델링

발효공정에서 얻어진 2차원 형광스펙트럼은 각종 공정변수들과 함께 공정에 대한 다변량 분석 모델을 수립하기 위해서 이용될 수 있다. 발효 시간에 따른 형광 스펙트럼 데이터는 보정 모델을 위한 학습 데이터와 예측 모델을 위한 타당성 데이터로 나눠질 수 있다. 예를 들면, FermPro1 공정에서 전체 데이터의 70%인 298개 데이터는 학습 데이터로 사용하고 나머지 30%가 되는 128개 데이터는 타당성 데이터로 사용되었다.

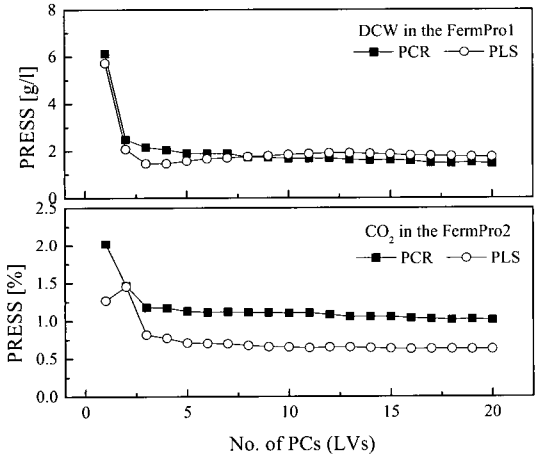


Figure 5. PRESS as number of PCs in the PCR model (LVs in the PLS model) for DCW in FermPro1 and for CO_2 concentrations in FermPro2.

PCR에서 PC 또는 PLS에서 LV의 최적 개수 결정

PCR에서 PC 또는 PLS에서 LV의 개수를 결정하는 것은 다변량 분석 모델의 설정에 커다란 영향을 미친다. PC나 LV의 수가 너무 적으면 공정변수가 적절히 모델화 되지 못하고 너무 많으면 모델의 융통성이 감소할 것이다. 따라서 PCR에서 PC 또는 PLS에서 LV의 최적 개수를 결정하는 것은 매우 중요하며 다변량 분석 모델로부터 유도된 추정 값과 실제 측정 데이터와의 상관성에 따라 결정된다. 여기서 PRESS를 상관관계를 나타낼 수 있는 척도로 사용하였다. Fig. 5에서 발효공정 FermPro1에서 DCW에 대한 PC와 LV의 최적 개수 그리고 FermPro2에서 CO_2 농도에 대한 최적의 PC와 LV의 개수를 PRESS값을 계산한 후 나타내었다. DCW에 대한 PRESS값을 보면 PCR 모델에서는 5개일 때가 최소였고, PLS 모델에서는 3개일 때가 최소였다. 그러나 CO_2 농도에 대해서는 PCR 모델에서는 3개, PLS 모델에서는 5개일 때가 최소였다. 따라서 3개 또는 5개의 PC 나 LV들이 PCR 과 PLS 보정 모델을 세우기 위해 적합하다. 그러나 두 가지 경우에 대하여 3개나 5개의 PC (LV)의 PRESS값의 변화는 5% 이하이므로 본 연구에서는 3개의 PC 또는 LV를 사용하여 공정을 모델링하였다.

PCR 과 PLS의 비교

생물 공정내 어떤 변수들에 대하여 PCR 과 PLS은 보정능력 (calibration power)을 제공할 수 있다. 즉, PCR과 PLS의 보정모델을 공정에서 얻어진 2차원 형광스펙트럼과 온라인, 오프라인 수집된 측정 데이터를 사용하여 수립하고 RMSEC라는 척도를 사용하여 수립된 보정모델이 공정 변수 데이터에 대하여 얼마나 적합한지를 평가한다. Table 3에는 두개의 공정 FermPro1과 FermPro2에서 각공정 변수들의 RMSEC 값을 3개의 PC와 LV를 적용하여 PCR과 PLS 모델에서 계산하여 나타냈다. Table 3에서 PCR 모델의 RMSEC 값이 PLS 모델의 RMSEC 값보다 약 30% 정도가 더 큰 것으로 보아 PLS 모델이 PCR 모델보다 더 좋은 보정능력을 가진 것으로 생각된다. 또한 Table 3에서 공정변수들의 최대값에 대한 상대오차의 백분율도 계산하였는데, 예를 들면 FermPro1에서 DCW의 상대오차는 PCR 모델에 대하여 4.992% (100 x 0.1066/2.135)이다. 즉, RMSEC 값과 상대오차가 작을수록 다변량 분석 모델이 실제 측정 데이터를 잘 추정하는 것으로 생각되며 모델이 새로운 데이터를 예측하는데 적용될 수 있다는 것을 의미한다. 한편, CO_2와 ALA의 상대오차는 비교적 높음을 알 수 있고 ALA 생합성의 대사적 계산을 위해서는 숙신산와 글라이신을 포함하는 복잡한 모델이 필요하다.

한편, PC나 LV의 수를 5이상으로 하였을 때에도 두 다변량 분석 모델의 RMSEC 값은 크게 감소하지 않았다.

Table 3. RMSEC values and relative maximum error of some process parameters in FermPro1 and FermPro2 by the PCR and PLS

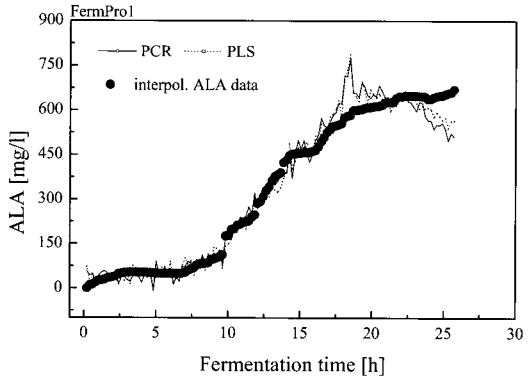
		PCR		PLS	
		RMSEC	Rel. error (%)	RMSEC	Rel. error (%)
FermPro1	DCW	0.1066 [g/l]	4.992	0.0811 [g/l]	3.798
	ALA	45.69 [mg/l]	6.788	34.88 [mg/l]	5.182
	CO_2	0.0843 [%]	20.56	0.0648 [%]	15.81
	DO	3.195 [%]	3.195	2.490 [%]	2.490
FermPro2	DCW	0.0661 [g/l]	3.492	0.0521 [g/l]	2.752
	ALA	85.90 [mg/l]	11.10	69.59 [mg/l]	8.995
	CO_2	0.0659 [%]	15.18	0.0572 [%]	13.82
	DO	2.821 [%]	2.821	2.592 [%]	2.529

다변량 모델을 이용하여 새로운 데이터를 예측하는데 보정모델이 얼마나 타당한지를 알아보기 위해 보정모델 형성에 사용되지 않은 30% 데이터를 이용하였다. Fig. 6은 FermPro1에서 ALA 와 FermPro2에서 DO의 내삽된 측정 데이터를 PCR과 PLS의 보정모델로 추정된 데이터들과 함께 나타내었다. Fig.에서 ALA에 대한 모델추정 데이터와 실제 측정 데이터와는 18시간 이후부터 차이를 나타내기 시작

하지만 DO의 경우에는 발효 초기에 현저한 차이를 보이고 있다. ALA 예측을 위한 128개의 형광 데이터에 대한 RMSEP 값은 62.51 mg/l (PCR)와 52.17 mg/l (PLS)이고 DO의 예측을 위한 115개의 형광 데이터에 대한 RMSEP 값은 5.891% (PCR)와 5.246% (PLS)이다. 그리고 ALA에 대한 상관계수 (correlation coefficients, R^2)는 0.979 (PCR)과 0.984 (PLS)이며 DO에 대해서는 0.845 (PCR)과 0.884 (PLS)이다.

이와 같이 다변량 분석 모델과 측정 데이터간의 밀접한 상관성 및 RMSEP 값은 2차원 형광스펙트럼을 이용한 다변량 분석 모델이 새로운 측정 데이터에 대한 훌륭한 예측 기법으로 사용될 수 있다는 것을 나타낸다.

(a)



(b)

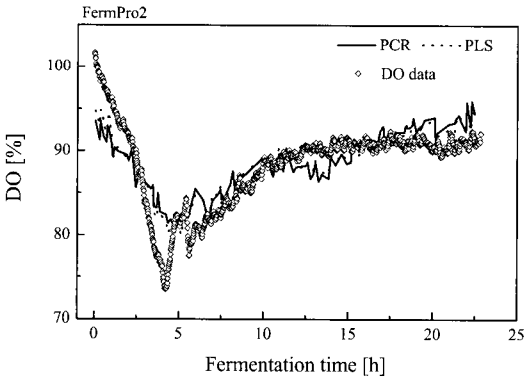


Figure 6. Comparison of the calibration model of the PCR and PLS with interpolated measurement data of ALA in FermPro1 and DO in FermPro2.

한편, 어떤 한 공정에서 개발된 PCR과 PLS 모델의 예측 능력은 또 다른 공정의 예측을 위해 적용할 수 있다. 즉, 공정 FermPro3에서 DCW의 학습 데이터를 사용하여 개발된 PCR과 PLS 모델은 비슷한 공정 FermPro4에서 DCW의 경향을 예측하는데 사용할 수 있다(두 공정간에는 전구체의 주입시간의 차이만 있음).

Fig. 7을 보면 공정 FermPro3에 의해 개발된 PCR과 PLS 모델에 의해 추정된 DCW 데이터를 FermPro3과 FermPro4의 오프라인 DCW 데이터와 비교하였다. FermPro3에서 DCW의 RMSEC 값은 PCR에 대해서는 0.5848 g/L이고 PLS에 대해서는 0.5048 g/L이다. 여기서 두 모델에 사용된 3개의 PC는 전체 분산의 68.0% (PCR) 과 60.6% (PLS)를 차지하고 있다. 두 모델의 FermPro4에서 DCW 측정 데이터의 예측능력을 비교하기 위해 RMSEP 값을 계산하였는데 PCR에 대해서는 0.7241 g/l, PLS에 대해서는 0.6872 g/l이었다. 또한, 오프라인에서 측정한 최대의 DCW 값에 대한 상대오차는 13.35% (PCR)와 12.67% (PLS)이었다.

이와 같이 RMSEP와 상대오차의 높은 값은 두 공정에서 두 전구체의 주입시간이 다르기 때문인 것으로 생각된다. FermPro3의 초기에 두개의 아미노산 (글루탐산, 글라이신)이 주입되었고, FermPro4에는 11시간에 주입되었다. 각각의 공정에서 11시간에 L-cysteine이 주입된 이후 글루타치온이 균체 내에서 합성되는데, 이러한 생합성은 균체량과 대사물질의 형성에 영향을 끼쳐 모델 예측값과 실제 측정값에서 차이를 나타낸다. 그러나 글루타치온의 생합성이 일어나지 않았던 11시간까지는 FermPro4에서 DCW의 모델 데이터는 오프라인 결과와 잘 일치함을 볼 수 있다.

다변량 분석 모델을 개발하거나 개발된 모델로 다른 생물공정의 성능을 예측하기 위해 어떤 실험을 반복 수행할 수 있다. 본 연구에서 PCR과 PLS 모델은 단지 한 공정에 한번씩 만 실험하여 생성되었지만 이러한 모델은 유사 공정에도 적용할 수 있다는 능력을 보여주었다.

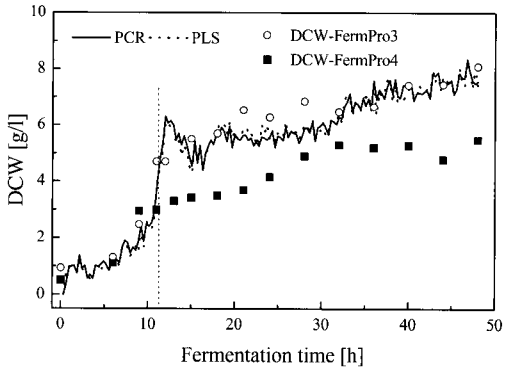


Figure 7. Comparison of the model data of DCW by the PCR and PLS in FermPro3 with off-line measurement data in FermPro3 and FermPro4.

많은 형광스펙트럼 중 적절한 스펙트럼을 추출하는 방법 (subtraction method)은 관심있는 성분 (예: biogenic fluorophor)이나 공정변수 (예: dissolved oxygen 농도)와 2차원 형광스펙트럼 데이터와의 상관관계를 알아보기 위해 사용될 수 있다 (6, 7, 33). 그러나 스캔된 전 범위의 스펙트럼 데이터를 이용하는 데는 한계가 있다. 따라서 생물공정으로부터 얻어진 많은 형광 데이터를 인공신경망 (ANN) 기술을 이용하여 분석, 처리하기도 하였다(8, 32). 그러나 ANN은 높은 정확도와 예측 능력을 보여주었지만 복잡한 수리적 계산을 요구하고 어려움이 있다.

본 연구에서 PCA 방법은 형광스펙트럼의 차원을 축소하

고 공정변수와 주성분들 간의 관계를 규명하기 위해 사용되었다. 또한, PCA의 주성분 점수 산점도로 재조합 대장균 *E.coli*와 효모 *S.cerevisiae* 의 발효공정 특성에 대한 정성적인 분석을 하였다(24).

Tartakovsky 등은 *E.coli*와 *S.cerevisiae* 등을 사용한 각종 발효공정에서 공정 변수들을 2차원 형광데이터와 함께 단계적 다중선형 회귀분석으로 설명하였다(5). 한편, *Pseudomonas fluorescences*를 발효하는 공정에서 배출된 CO_2 와 O_2 그리고 발효 배지에서 숙신산과 단백질을 관찰하기 위해 형광 스펙트럼에 기초한 PLS 모델이 사용되었다(21). PLS와 비선형 PCR 그리고 단계적 회귀모델의 성능이 CCD 어레이 (array) 광섬유 형광분광계로 혐기적 소화과정에서 수집된 형광스펙트럼을 사용하는 공정에서 비교되었고, 그 중에서 PLS 모델이 우수하다는 결과를 얻었다(34). 그러나 폐수처리 공정에서 형광 신호는 빈약하여 비선형 PCR의 성능을 저하시키기 때문에 비선형 PCR 모델의 학습을 위해 많은 양의 데이터를 얻을 수 있는 2차원 형광분광계를 이용하여 개선하기도 한다. 한편, PCA와 ANN를 조합한 모델이 *Alcaligenes eutrophus* 의 회분배양에서 형광스펙트럼으로부터 균체량과 기질의 농도를 예측하기 위해서 사용되었고 선형 PLS 모델보다 더 정확한 예측능력을 보여주었다(35).

본 연구에서 다변량 분석 모델 (PCR와 PLS)을 재조합 대장균 *E.coli*와 효모 *S.cerevisiae*의 발효공정에서 얻어진 2차원 형광스펙트럼의 분석을 위해 이용하였다. 그리고 공정변수에 대한 두 모델의 보정과 예측능력이 비교되었다. 3개의 PC (LV)를 선택하여 PCR과 PLS 보정 모델을 형성하였을 때 PLS 모델이 PCR 모델보다 성능이 우수함을 알 수 있었다. 그러나 본 연구에서 사용한 두 모델들이 모두 합리적이고 타당한 이유는 전체 데이터의 70%가 되는 많은 양의 학습데이터와 SOM 분류에 기초한 여기 및 방출 파장의 최적 조합 수를 잘 선택했기 때문이라고 볼 수 있다. 따라서 이러한 모델들은 유사한 다른 생물공정의 성능을 예측하는데 사용될 수 있다.

요 약

본 연구에서는 2차원 형광스펙트럼의 PCA분석을 통하여 발효 공정을 모니터링하고 PCR과 PLS과 같은 다변량 분석 기법을 이용하여 공정을 모델링하였다. 재조합 대장균 *E. coli* 와 효모 *S.cerevisiae*의 발효 공정 중에 얻어진 많은 양의 2차원 형광스펙트럼 자료는 우선 PCA를 통해 축소된다. 그리고 PCA에서 주성분점수와 적재 산점도는 발효 공정의 정성적 경향을 묘사하기 위해 사용되었다. 또한, PCR과 PLS는 2차원 형광스펙트럼의 분석을 위해 사용되었으며 PLS모델이 보정과 예측 능력에서 PCR모델보다 조금 더 우수한 성능을 나타냈다. 따라서 2차원 형광스펙트럼 자료를 이용하여 생물공정을 모델링 하고자 할 때는 PCR 방법보다는 PLS 방법을 사용하는 것이 유리할 것이다.

감 사

본 연구는 한국과학재단 (KOSEF)의 특정기초과제연구 (과제번호R01-2002-000- 00027-0)과 산업자원부 지방기술혁신사업 (RTI04-03-03) 지원에 의해 이루어졌으며, 이에 감사드립니다.

REFERENCES

1. Langergraber G., Fleischmann N., and F. Hofstaedter (2003), A multivariate calibration procedure for UV/VIS spectrometric quantification of organic matter and nitrate in wastewater, *Water Sci. Tech.* **47**(2), 63-71.
2. Givens D. I. and E. R. Deaville (1999), The current and future role of near infrared reflectance spectroscopy in animal nutrition, *J. Agr. Res.* **50**(7), 1131-1145.
3. Vaidyanathan, S., S. White, L. Harvey, and B. McNeil (2003), Influence of morphology on the near-infrared spectra of mycelial biomass and its implications in bioprocess monitoring, *Biotech. Bioeng.* **82**(6), 715-724.
4. Climander, C. and C. F. Mandenius (2002), Online monitoring of a bioprocess based on a multi-analyzer system and multivariate statistical process modeling, *J. Chem. Tech. Biotech.* **77**, 1157-1168.
5. Tartakovsky, B., M. Scheintuch, J. M. Hilmer, and T. Scheper (1996), Application of scanning fluorometery for monitoring of a fermentation process, *Biotech. Progr.* **12**, 126-131.
6. Mukherjee, J., C. Lindermann, and T. Scheper (1999), Fluorescence monitoring during cultivation of Enterobacter aerogenes at different oxygen levels, *Appl. Microbiol. Biotech.* **52**, 489-494.
7. Marose, S., C. Lindemann, and T. Scheper (1998), Two-dimensional fluorescence spectroscopy: A new tool for on-line bioprocess monitoring, *Biotech. Prog.* **14**, 63-74.
8. Wolf, G., J. S. Almeida, C. Pinheiro, V. Correia, C. Rodrigues, MAM. 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, *Biotech. Bioeng.* **72**, 297-306.
9. Boehl, D., D. Solle, B. Hitzmann, and T. Scheper (2003), Chemometric modeling with two-dimensional fluorescence data for Claviceps purpurea bioprocess characterization, *J. Biotech.* **105**, 179-188.
10. Basheer, I. A. and M. Hajmeer (2000), Artificial neural networks: fundamentals, computing, design, and application, *J. Microbiol. Meth.* **43**, 3-31.
11. Bhat, N. V. and T. J. (1992), Determining model structure for neural network stripping, *Comp. Chem. Eng.* **16**, 271-281.
12. Bo, R. (2003), Multivariate calibration. What is in chemometrics for the analytical chemist? *Anal. Chim. Acta.* **500**, 185-194.
13. Geladi P., B. Sthson, J. Nystrom, T. Lillhinga, T. Lestander, and J. Burger (2004), Chemometrics in Spectroscopy, *Spectrochim. Acta. Part B* **59**, 1347-1357.
14. Jolliffe, I. T. (1986), Principal component analysis, New York, Springer.
15. 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, *Le Lait* **77**(6), 657-670.
16. 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.

17. 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.
18. Karim. M. N., D. Hodge, and L. Simon (2003), Data-based modeling and analysis of bioprocesses: some real experiences, *Biotech. Prog.* **19**, 1591-1605.
19. Cooper, J. B. (1999), Chemometric analysis of Raman spectroscopic data for process control applications, *Chemomet Intell. Lab. Sys.* **46**, 231-247.
20. Otsuka, M. (2004), Comapartive particle size determination of phenacetin bulk powder by using Kubelka-Munk theory and principal component regression analyis based on near-infrared spectroscopy, *Powder Tech.* **141**, 244-250.
21. Skibsted, E., C. Lindemann, C. Roca, and L. Olsson (2001), On-line bioprocess monitoring with a multi-wavelength fluorescence sensor using multivariate calibration, *J. Biotech.* **88**, 47-57.
22 Haack, M. B., A. Eliasson, and L. Olsson (2004), On-line cell mass monitoring of Saccharomyces cerevisiaecultivations by multi-wavelength fluorescence, *J. Biotech.* **114**, 199-208.
23. Wentzell, P. D. and L. V. Montoto (2003), Comparison of principal components regression and partial least squares regression through generic simulations of complex mixturtes, *Chem. Intell. Lab. Sys.* **65**, 257-279.
24. Rhee, J. I., Lee K.-I., Kim C.-K., Yim Y.-S., Chung S-W., Wei, J., and K.-H. Bellgardt (2005), Classification of two-dimensional fluorescence spectra using self-organizing maps, *Biochem. Eng. J.* **22**, 135-144.
25. Chung, S. Y., Seo K. H., and Rhee J. I. (2005), Influence of culture conditions on the production of extra-cellular 5-aminolevulinic acid (ALA) by recombinant E. coli, *Proc. Biochem.* **40**. 385-394.
26. 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, *Biotech. Bioeng.* **38**, 196-205.
27. Tietze, F. (1969), Enzymic method for quantitative determination of nanogram amount of total and oxidized glutathione, *Anal. Biochem.* **27**, 502-522.
28. 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.
29. Liu, R. X., J. Kuang, Q. Gong, and X. L. Hou (2003), Principal component regression analysis with SPSS, *Comp. Meth. Prog. Biomed.* **71**, 141-147.
30. Geladi, P. and B. R. Kowalski (1986), Partial least-squares regression: tutorial, *Anal. Chim. Acta.* **185**, 1-17.
31. Maltlab manual, vers. 6.1, The Mathworks, Inc., USA, 2002.
32. Lee, K. I., Yim Y. S., Chung S. W., Wei J., and Rhee J. I. (2006), Application of artificial neural networks to the analysis of 2D fluorescence spectra in recombinant E.coli fermentation processes, *J. Chem. Tech. Biotech.* in print.
33. Lindemann, C., S. Marose, H. O. Nielson, and T. Scheper (1998), 2-Dimensional fluorescence spectroscopy for on-line bioprocess monitoring, *Sens. Actuat. B* **51**, 271-277.
34. Morel, M., K. Santamaria, M. Perrier, S. R. Guiot, and B. Tartakovsky (2004), Application of multi-wavelength fluorometery for on-line monitoring of an anerobic digestion process, *Water Res.* **38**, 3287-3296.
35. Hegedorn, A., R. L. Legge, and H. Budman (2003), Evaluation of spectrofluorometry as a tool for estimation in fed-batch fermentations, *Biotech. Bioeng.* **83**(1), 104-111.