第38卷,第12期
2018年12月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 38, No. 12, pp3790-3795
December, 2018

基于主成分分析和多元曲线分辨的蓝细菌流式荧光光谱分析方法

范贤光[1,2,3]，方晓玲[1]，王 昕[1,2,3*]，陈宇欣[1]，巫梅琴[1]，胡雪亮[1]

1. 厦门大学航空航天学院仪器与电气系，福建 厦门 361005
2. 传感技术福建省高等学校重点实验室，福建 厦门 361005
3. 厦门市光电传感技术重点实验室，福建 厦门 361005

摘 要 利用流式细胞术对细胞进行多色荧光分析时，往往获得的是由多种组分荧光光谱混合的多元荧光光谱。在对蓝细菌进行光谱流式检测时，所测得的荧光光谱同时包含了多种未知荧光光谱，且存在严重的光谱混叠。为了获得蓝细菌中的主要组分光谱及其浓度，提出主成分分析和多元曲线分辨相结合的方法，对蓝细菌的流式荧光光谱进行处理。该方法通过主成分分析获得蓝细菌的主要纯组分数量，然后利用渐进因子分析寻找各组分的起始点和终止点，并估计纯组分的初始光谱，最后利用交替最小二乘结合其纯组分光谱的单峰性和非负性，对初始估计的纯组分光谱进行迭代修正，从而得到纯组分光谱及其组分浓度。仿真和实验结果表明，该方法能够准确地估计混合光谱中纯组分的个数并对其谱峰进行拟合，进而准确地估计各个组分的浓度。该方法不但适用于蓝细菌的光谱分析，还可用于其他多元混合光谱体系的解析。

关键词 主成分分析；渐进因子分析；交替最小二乘法；蓝细菌；流式荧光光谱

中图分类号：TH79 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2018)12-3790-06

引 言

蓝细菌是一类能通过光合作用获取能量的细菌，它含有菌绿素，β-胡萝卜素和藻胆蛋白，其中藻胆蛋白能被藻类吸收是一种传递光能的天然色素。近年来，为研究蓝细菌，国内学者提出了液相色谱-质谱联用、流式细胞术以及荧光显微镜等方法[1]，但是液相色谱-质谱联用方法会破坏细胞结构，不利于活细胞分选，而荧光显微镜则不适合做大数量检测。

因此，本文利用流式细胞术对蓝细菌进行荧光标记并检测，传统流式细胞仪使用滤光片对细胞经过激光探测区所产生的光信号按不同波段进行分光，通过光电倍增管(PMT)或雪崩光电二极管(APD)对每个通道的光信号进行检测。在细胞的多色分析中，由于荧光基团发射光谱较宽，多种染料或蛋白质同时受激发导致光谱重叠，带来严重的串色问题，使得光检测器测量的信号强度和目标荧光染料的真实强度之间存在差异，因此，需要进行繁琐的荧光补偿[2-3]。但是荧光补偿不仅计算繁琐，而且容易产生补偿过度或不足，大大增加了实验的复杂程度[4]。

为了克服传统流式细胞术检测参数的限制，光谱流式细胞技术应运而生，光谱流式技术是在传统型流式细胞术的基础上，利用光栅或棱镜代替滤光片对单个细胞经过激光探测区所发出的荧光进行分光，使用阵列检测器对细胞的全光谱进行采集的新一代流式细胞检测手段，其通过荧光光谱去卷积分析，可以较轻松地实现细胞的多色同时检测[5-7]。在化学分析中，多元分辨曲线是经常用到的数据处理工具；它不需要事先知道基本化学模型，就能够建立模型来描述假设中的化学模型，从混合物演进过程的数据中解析出纯物质的性质[8-9]（如光谱曲线，pH 曲线等）。针对分析蓝细菌流式荧光光谱实验，本文利用主成分分析(principal component analysis，PCA)以及多元曲线分辨方法中渐进因子分析(evolving factor analysis，EFA)与交替最小二乘(alternating least square，ALS)相结合的方法对模拟仿真蓝细菌荧光光谱的分解进行了详细的比较和分析，并在实际蓝细菌 WH7805 实验分析中得出纯光谱组分以及组分浓度。

1 方 法

流式细胞仪测到的光谱矩阵 D 满足

收稿日期：2017-10-19，修订日期：2018-02-13

基金项目：国家自然科学基金项目(21503171)，国家重大科研仪器研制项目(21627811)资助

作者简介：范贤光，1980 年生，厦门大学航空航天学院仪器与电气系副教授 e-mail：fanxg@xmu.edu.cn

* 通讯联系人 e-mail：xinwang@xmu.edu.cn

$$\boldsymbol{D} = \boldsymbol{C}\boldsymbol{S}^{\mathrm{T}} + \boldsymbol{E} \tag{1}$$

式(1)中，光谱矩阵 $\boldsymbol{D}$ 和误差矩阵 $\boldsymbol{E}$ 的维数为 $N_C \times N_S$，浓度矩阵 $\boldsymbol{C}$ 的维数为 $N_C \times N$，纯光谱 $\boldsymbol{S}^{\mathrm{T}}$ 矩阵的维数为 $N \times N_S$。其中 N_C 是光谱曲线数(即所测细胞数)，N_S 是波长点数，N 是组分数。

本文利用主成分分析、渐进因子分析得到蓝细菌纯光谱的组分数以及纯光谱的初始估计，将其作为交替最小二乘法的初始值，结合光谱的单峰性及非负性对蓝细菌荧光光谱数据进行迭代处理，得出蓝细菌含有的荧光组分并估计每种荧光组分的浓度。

1.1　主成分分析

主成分分析通过计算光谱数据 X_{mn} 协方差矩阵 $X^{\mathrm{T}}X$，求得协方差矩阵的特征向量和特征值，从而选择成分组成模式矢量，得到光谱数据的降维数据。

计算步骤如下：

(1) 对原始光谱矩阵 X_{mn}(m 为行数，n 为列数)进行奇异值分解，则根据式(2)计算协方差矩阵的特征值 λ 与特征向量 $\boldsymbol{Q}$；

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \tag{2}$$

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}\boldsymbol{\Sigma}^{2}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}\boldsymbol{\Sigma}^{2}\boldsymbol{V}^{-1} \tag{3}$$

则协方差矩阵的特征值 λ 及特征向量 $\boldsymbol{Q}$ 为 $\lambda = \Sigma^2$，$\boldsymbol{Q} = \boldsymbol{V}$；

(2) 计算主成分贡献率：$\dfrac{\lambda_K}{\sum_{i=1}^{t}\lambda_i}(k = 1, 2, \cdots, t;)$；

(3) 计算主成分累计贡献率并确定组分数：累计贡献率为 $\dfrac{\sum_{j=1}^{K}\lambda_j}{\sum_{i=1}^{t}\lambda_i}(k = 1, 2, \cdots, t)$，当累计贡献率达到 95% 以上时，所选取的 K 个特征值即可描述初始变量的绝大部分信息。

1.2　EFA 渐进因子分析

EFA 渐进因子分析在化学定量分析中是一个很重要的分析方法，其基本思想是按照洗脱过程的进展，利用主成份分析渐进对二维数据矩阵进行分析，逐步扩展到整个数据阵，跟踪 $\boldsymbol{D}$ 子阵的秩的变化。

计算步骤如下：

(1) 计算正向 f 列矩阵 $\boldsymbol{D}_f(f = 2, 3, \cdots, n)$ 以及反向 b 列矩阵 $\boldsymbol{D}_b(b = 2, 3, \cdots, n)$ 的特征值；

(2) 将正、反向过程所得的特征值作为渐进变量(波长)的函数在同一图上作图，确认浓度窗口，获取起始位置和终止位置；

(3) 绘制特征值估计曲线，作为纯组分初始估计。

1.3　ALS 交替最小二乘法

交替最小二乘法利用 EFA 近似估计纯光谱矩阵对测量数据进行反复迭代运算，直到混合光谱的均方误差 MSE 满足 $\mathrm{MSE} \leqslant \zeta$ 或者迭代次数达到设定值 M，从而求得纯光谱和组分浓度。其中均方误差 MSE 定义为

$$\mathrm{MSE} = \sqrt{\frac{\sum_{i=1}^{N_C}\sum_{j=1}^{N_S}E_{ij}^2}{N_C \times N_S}} \tag{4}$$

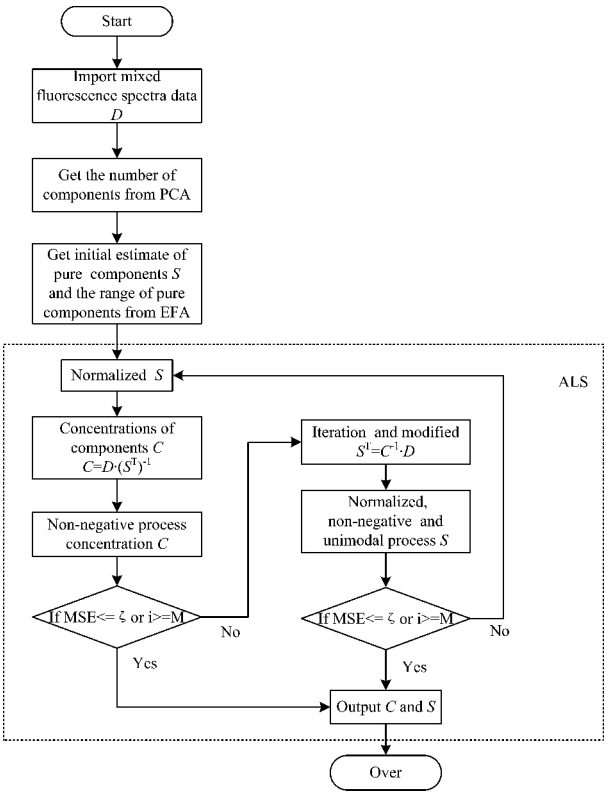


图 1　蓝细菌流式荧光光谱分析方法算法流程图

Fig. 1　Algorithm flowchart of flow cytometer data analysis

计算步骤如下：

(1) 将纯光谱估计 $\boldsymbol{S}^{\mathrm{T}}$ 进行非负、单峰性、归一化处理，并对其起始与终止位置进行牛顿三次插值修正后代入式(5)中；

$$\boldsymbol{C} = \boldsymbol{D}(\boldsymbol{S}^{\mathrm{T}})^{-1} \tag{5}$$

(2) 根据式(6)估计混合光谱 $\hat{\boldsymbol{D}}$；

$$\hat{\boldsymbol{D}} = \boldsymbol{C}\boldsymbol{S}^{\mathrm{T}} \tag{6}$$

(3) 根据式(7)计算混合光谱误差矩阵 $\boldsymbol{E}$

$$\boldsymbol{E} = \boldsymbol{D} - \hat{\boldsymbol{D}} \tag{7}$$

(4) 若均方误差小于等于迭代阈值或者迭代次数大于设定值 M，则判断 $\boldsymbol{S}^{\mathrm{T}}$ 为混合光谱的纯组分，$\boldsymbol{C}$ 为组分浓度并结束运算，否则继续运算；

(5) 将得到的浓度矩阵 $\boldsymbol{C}$ 代入式(8)中，返回步骤(1)。

$$\boldsymbol{S}^{\mathrm{T}} = \boldsymbol{C}^{-1}\boldsymbol{D} \tag{8}$$

2　仿真分析

利用已知六种荧光染料的发射光谱数据(FITC，QD545，PE，QD605，AF610-PE，PerCP 等)，根据式(1)模拟仿真产生 1 000 个细胞的光谱数据 D。其中，纯组分光谱的波长范围在 450～750 nm，归一化最大谱峰强度为 100。

当仿真的荧光光谱只有 3 种组分时，采用本算法得到的纯光谱图与理想仿真纯光谱基本重合，谱线形状也基本重合，识别的谱峰位置准确，能够得到准确的光谱波长起始点和终止点，从而得出较为准确的组分浓度。图 2 为三种组分

的仿真光谱及其处理结果。

为研究本方法对混合光谱解析的适用性，逐渐增加仿真光谱中的纯组分数(6 种，10 种，12 种)。由于纯光谱并不都满足 EFA 的“先进先出”原则，所以需要选择合适的纯组分起始和终止位置，对纯光谱进行浓度修正；结合图 3，图 4 及表 1 可以看出，虽然总体上都能够拟合出较好的混合光谱波形，但是随着纯组分数的增多，对纯组分光谱的估计会出现一定的偏差。该偏差一方面取决于纯组分的数目，另一方面还取决于纯组分的重合程度。从表 1 中可以看出，组分数由 6 增加到 8 和 10 时，其拟合效果下降较多。因此，在 300 nm 的范围内，当纯组分光谱数量小于 6 时，得到的纯光谱及浓度矩阵还是较为准确可靠的。

3 实验与结果讨论

本文实验仪器是厦门大学化学化工学院自主开发的光谱流式细胞仪，其能有效同时测量纳米颗粒和生物分子的荧光和散射光，实现颗粒计数和获取颗粒的生物性质及其体积、浓度特性[10-11]；实验对象是蓝藻细菌 WH7805。

由于实验设备和实验环境的影响使得测到的原始光谱曲线存在毛刺，显然噪声的存在会影响算法的结果分析，所以本文用滑动平均值法对原始数据进行了预处理，处理后的结

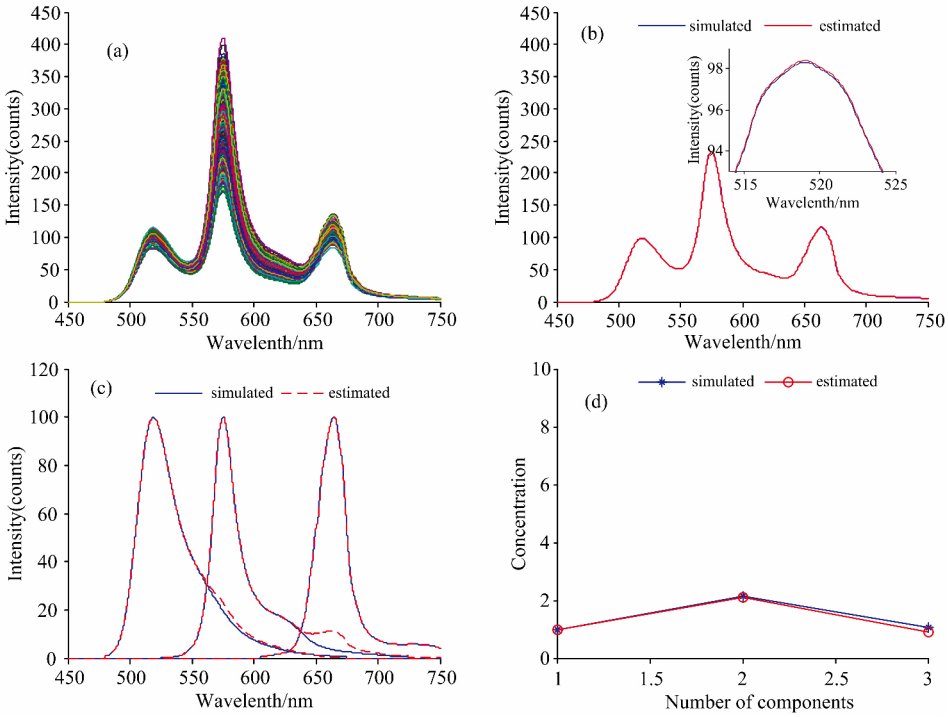
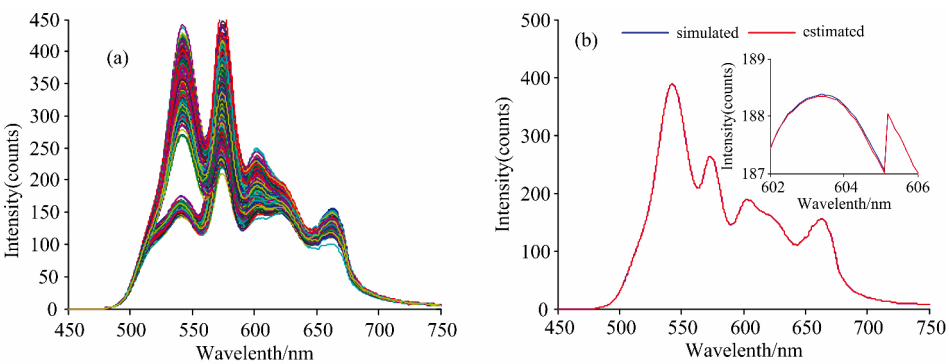


图 2 三种组分混合光谱解析结果

(a)：三种组分 1 000 个仿真混合光谱图；(b)：三种组分第 561 个细胞的荧光光谱对比图；(c)：三种组分的纯光谱对比图；(d)：三种组分第 561 个细胞的组分浓度估计对比图

Fig. 2 Estimated results of the mixed spectra with three components

(a): Three components 1 000 cells simulated mix fluorescence spectra; (b): Comparison between the 561st cell fluorescence spectra and signal after processed with three components; (c): Comparison between the simulated pure components and estimated ones from three components; (d): Comparison of the 561st cell concentration estimation with three components



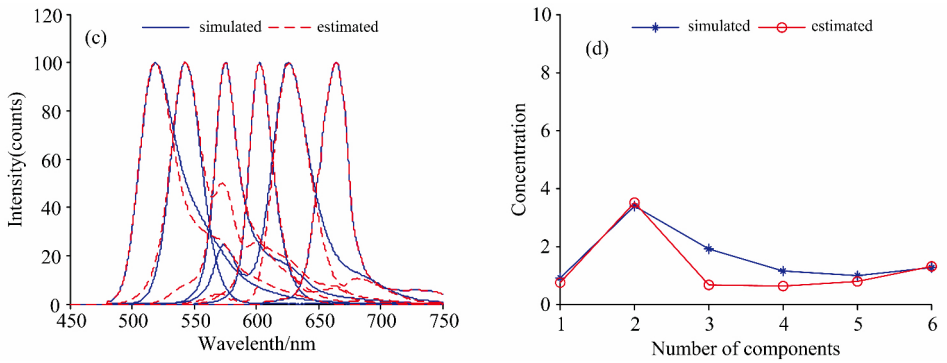


图 3　六种组分混合光谱解析结果

(a)：六种组分 1 000 个仿真混合光谱图；(b)：六种组分第 561 个细胞的荧光光谱对比图；
(c)：六种组分的纯光谱对比图；(d)：六种组分第 561 个细胞的组分浓度估计对比图

Fig. 3　Estimated results of the mixed spectra with six components

(a)：Six components 1 000 cells simulated mix fluorescence spectra；(b)：Comparison between the 561st cell fluorescence spectra and signal after processed with six components；(c)：Comparison between the simulated pure components and estimated ones with six components；(d)：Comparison of the 561st cell concentration estimation with six components

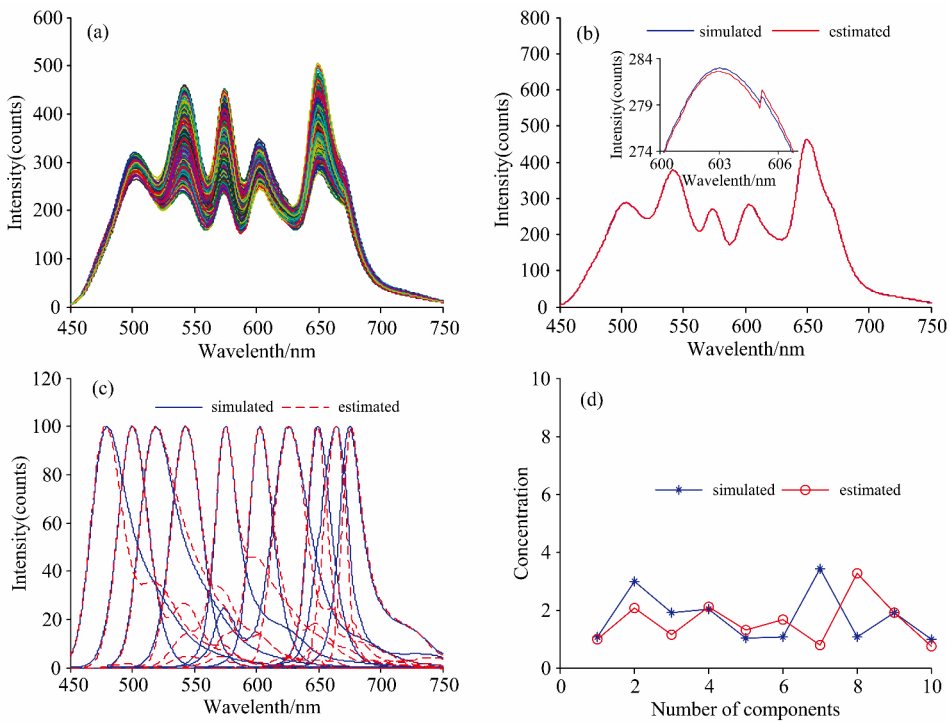


图 4　10 种组分混合光谱解析结果

(a)：10 种组分 1 000 个仿真混合光谱图；(b)：10 种组分第 561 个细胞的荧光光谱对比图；
(c)：10 种组分的纯光谱对比图；(d)：10 种组分第 561 个细胞的组分浓度估计

Fig. 4　Estimated result of the mixed spectra with ten components

(a)：Ten components 1 000 cells simulated mix fluorescence spectra；(b)：Comparison between the 561st cell fluorescence spectra and signal after processed withten components；(c)：Comparison between the simulated pure components and estimated ones with ten components；(d)：Comparison of the 561st cellconcentration estimation with ten components

果如图 5(a)所示。根据目前的研究结果，已知蓝细菌 WH7805 含有 APC 和 PE 两种组分。首先假设组分完全未知，对实验数据进行处理；然后在已知两种组分的条件下，再次对实验数据进行分析，并比较两种情况下的处理结果。

根据主成分分析，当选取组分数为 6 时达到特征值累计贡献率为 99.96%，因此选择 6 种组分纯光谱来拟合混合荧光光谱。在纯组分完全未知的前提下，得到如图 5(b)的纯组分估计。图5(d)和(c)为第561个细胞的拟合光谱及其浓度

表 1 不同数目纯组分估计结果的均方误差

Tabel 1 MSE of estimated results for different numbers of components

Number of components	MSE
3	1.279 8
4	3.408 9
6	5.976 4
8	11.039 3
10	32.946 2
12	33.237 8

估计。

在已知两种组分前提下，将图 5(b)中峰位置与 APC 和 PE 最接近的组分 1 和 6 替换为 APC 和 PE 的纯光谱，处理得到如图 5(e)的纯光谱，获得的拟合光谱及其浓度如图 5(g)和(f)所示。

分析图 5(d)和图 5(g)，可以看出拟合的混合光谱还是能比较好的模拟实际蓝细菌光谱，谱线平滑，识别的谱峰位置准确。但是由于实际蓝细菌组分的复杂度比仿真光谱大得多，利用 EFA 得到的纯组分只是估计主要几种纯组分，因此拟合光谱与实际光谱的形状不能完全的重合。另外，无论是否已知两种纯组分，其对混合光谱的拟合效果接近，但是对于纯组分光谱和浓度的估计则略有不同。这说明，对于相同的混合光谱，存在多组可能的解析结果，因此两种组分已知的解析结果相对更可靠。

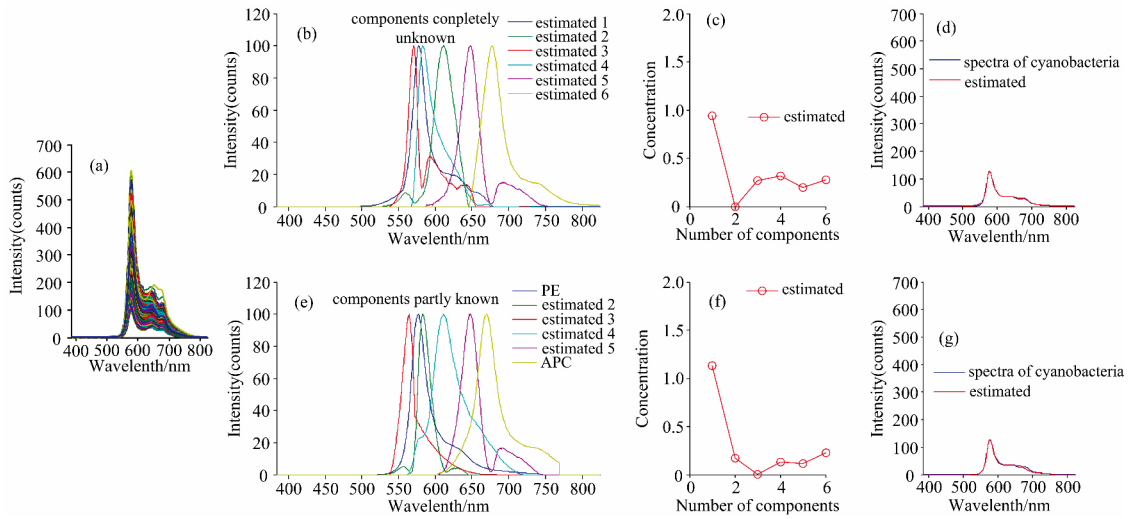


图 5 蓝细菌 WH7805 荧光光谱解析结果

(a)：蓝细菌 WH7805 荧光光谱；(b)：在组分完全未知情况下估计的纯组分光谱；(c)：在组分完全未知情况下第 561 个蓝细菌的纯光谱组分浓度；(d)：在组分完全未知情况下第 561 个蓝细菌混合荧光光谱对比图；(e)：在组分部分已知情况下估计的纯组分光谱；(f)：在组分部分已知情况下第 561 个蓝细菌的纯光谱组分浓度；(g)：在组分部分已知情况下第 561 个蓝细菌混合荧光光谱对比图

Fig. 5 Estimated results of cyanobacteria WH7805 mixed spectra

(a): WH7805 cells mix fluorescence spectra; (b): The estimated result of pure components with components completely unknown; (c): Concentration estimation of the 561st cell with components completely unknown; (d): Comparison between the 561st cell fluorescence spectra and estimated signal with components completely unknown; (e): The estimated result of pure components with components partly known; (f): Concentration estimation of the 561st cell with components partly known; (g): Comparison between the 561st cell fluorescence spectra and estimated signal with components partly known

4 结 论

基于主成分分析、渐进因子分析及交替最小二乘法对蓝细菌流式荧光光谱进行解析，该方法先利用主成分分析得到蓝细菌的纯组分数量；再通过渐进因子分析得到各组分的起始位置和终止位置并估计纯光谱的初始波形；然后利用交替最小二乘对纯光谱和浓度矩阵进行迭代估计，并在处理过程中加入了非负性、归一化、单峰性等约束条件；最后，通过设定阈值获得较为准确的纯光谱和浓度矩阵。在仿真实验中，通过对不同组分数的混合光谱进行处理，结果表明本文方法能够得到较为准确的谱峰位置、纯组分光谱以及浓度估计。在实验中，将本方法应用于蓝细菌流式荧光光谱处理，能够较好地估计纯组分光谱及其组分浓度，获得较为理想的拟合光谱，可供蓝细菌后续生化研究使用。

References

[1] Hyka P, Lickova S, Přibyl P, et al. Biotechnology Advances, 2013, 31(1): 2.
[2] Nolan J P, Condello D. Current Protocols in Cytometry/Editorial Board, J. Paul Robinson, Managing Editor, 2013, Chapter 1(467): Unit1. 27.
[3] Novo D, Grégori G, Rajwa B. Cytometry Part A, 2013, 83A(5): 508.
[4] Byrd T, Carr K D, Norman J C, et al. Cytometry Part A, 2015, 87A(11): 1038.
[5] Futamura K, Sekino M, Hata A, et al. Cytometry Part A, 2015, 87A(9): 830.
[6] Grégori G, Patsekin V, Rajwa B, et al. Cytometry Part A, 2012, 81A(1): 35.
[7] Grégori G, Rajwa B, Patsekin V, et al. Current Topics in Microbiology & Immunology, 2014, 377: 191.
[8] Hormozi-Nezhad M R, Jalali-Heravi M, Kafrashi F. Journal of Chemometrics, 2013, 27(10): 353.
[9] ZHANG Qiu-lan, NI Yong-nian(张秋兰, 倪永年). Journal of Analytical Science(分析科学学报), 2012, 28(1): 6.
[10] Yang L, Zhu S, Hang W, et al. Analytical Chemistry, 2009, 81(7): 2555.
[11] ZHU Shao-bin, WANG Shuo, YANG Ling-ling, et al. SCIENCE CHINA Chemistry, 2011, 54(8): 1244.

Analytical Method of Cyanobacteria Flow Fluorescence Spectrum Based on Principal Component Analysis and Multivariate Curve Resolution

FAN Xian-guang[1,2,3], FANG Xiao-ling[1], WANG Xin[1,2,3*], CHEN Yu-xin[1], WU Mei-qin[1], HU Xue-liang[1]

1. Department of Instrumental and Electrical Engineering, School of Aerospace Engineering, Xiamen University, Xiamen 361005, China
2. Fujian Key Laboratory of Universities and Colleges for Transducer Technology, Xiamen 361005, China
3. Xiamen Key Laboratory of Optoelectronic Transducer Technology, Xiamen 361005, China

Abstract When flow cytometry is used to analyze the polychromatic fluorescence of cells, multiple fluorescence spectra were often obtained, mixed with multicomponent fluorescence spectra. In this paper, the fluorescence spectra of cyanobacteria including many unknown fluorescence spectra were detected by flow cytometer with serious spectral overlap. In order to extract the main components and their concentrations from cyanobacteria spectra, a method of principal component analysis combined with multivariate curve resolution was used to process the fluorescence spectra of cyanobacteria. At first, the number of main components of cyanobacteria was given by principal component analysis, and then Evolving Factor Analysis was adopted to find the starting and end position of each component and to estimate the initial spectrum of pure components, finally Alternating Least Square combined with the pure components spectral unimodality and non-negativity was used to correct the initial estimation of pure components and concentrations. In the simulation and experiment, it was proved that the method could accurately estimate the number of pure components in the mixed spectra and fit the spectral peaks, and then accurately estimate the concentration of each component. This method can not only be applied in the spectral analysis of cyanobacteria, but also used for other multiple spectral mixture analysis.

Keywords PCA; EFA; ALS; Cyanobacteria; Flow fluorescence spectrum

(Received Oct. 19, 2017; accepted Feb. 13, 2018)

* Corresponding author