Probabilistic classification method on multi wavelength chromatographic data for photosynthetic pigments identification

K. R. Prilianti, Y. Setiawan, Indriatmoko, M. A. S. Adhiwibawa, L. Limantara, and T. H. P. Brotosudarmo

Citation: AIP Conference Proceedings **1587**, 78 (2014); doi: 10.1063/1.4866538
View online: http://dx.doi.org/10.1063/1.4866538
View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/1587?ver=pdfcov
Published by the AIP Publishing

Probabilistic Classification Method on Multi Wavelength Chromatographic Data for Photosynthetic Pigments Identification

K.R. Prilianti[1*], Y. Setiawan[1], Indriatmoko[2], M.A.S. Adhiwibawa[2], L. Limantara[2], T.H.P. Brotosudarmo[2]

[1]*Informatics Engineering, Ma Chung University, Villa Puncak Tidar N-01, Malang- East Java*
[2]*Ma Chung Research Center for Photosynthetic Pigments (MRCPP), Villa Puncak Tidar N-01, Malang- East Java*
[*]*kestrilia.rega@machung.ac.id*

Abstract. Environmental and health problem caused by artificial colorant encourages the increasing usage of natural colorant nowadays. Natural colorant refers to the colorant that is derivate from living organism or minerals. Extensive research topic has been done to exploit these colorant, but recent data shows that only 0.5% of the wide range of plant pigments in the earth has been exhaustively used. Hence development of the pigment characterization technique is an important consideration. High-performance liquid chromatography (HPLC) is a widely used technique to separate pigments in a mixture and identify it. In former HPLC fingerprinting, pigment characterization was based on a single chromatogram from a fixed wavelength (one dimensional) and discard the information contained at other wavelength. Therefore, two dimensional fingerprints have been proposed to use more chromatographic information. Unfortunately this method leads to the data processing problem due to the size of its data matrix. The other common problem in the chromatogram analysis is the subjectivity of the researcher in recognizing the chromatogram pattern. In this research an automated analysis method of the multi wavelength chromatographic data was proposed. Principal component analysis (PCA) was used to compress the data matrix and Maximum Likelihood (ML) classification was applied to identify the chromatogram pattern of the existing pigments in a mixture. Three photosynthetic pigments were selected to show the proposed method. Those pigments are β-carotene, fucoxanthin and zeaxanthin. The result suggests that the method could well inform the existence of the pigments in a particular mixture. A simple computer application was also developed to facilitate real time analysis. Input of the application is multi wavelength chromatographic data matrix and the output is information about the existence of the three pigments.

Keywords: photosynthetic pigment, chromatogram, HPLC, PCA, maximum likelihood classifier
PACS: 82.50.-m, 82.80.Bg, 02.50.Cw, 02.70.-c

INTRODUCTION

Colorant plays important role in human life for centuries. It is widely used for many advantages such as acceptability of products like food, cosmetics and textiles [1-4]. Exploration of natural colorant is become one of the popular topic in the fields of biotechnology, especially due to various environmental problem and health hazard because of synthetic colorant usage. Generally, natural colorant obtain from living organism and can be divided as dyes and pigments. Dyes are often used for textile and food, while pigments are often used for ink, paint and cosmetics [5]. Among those pigments, the most abundant are photosynthetic pigments from plants.

The photosynthetic pigments can be separated and identified by chromatography. HPLC-DAD (high performance liquid chromatography - diode array detector) is still the most popular method [6-8].The pigments are separated by the column packing that involves various chemical and/or physical interactions between their molecules and the packing particles.Separated pigments are detected at the exit of the column by a flow-through device (detector). This detector will collect the UV light absorption data in various wavelengths in order to identify and measure the total amount of the pigment. The data visualized in the form of graph called chromatogram.

In former HPLC analysis, the researchers usually use a single chromatogram from a fixed wavelength and discard much of the information contained at other wavelength. Therefore, different approach have been proposed to utilize chromatographic data [9-10]. Rather than using a fixed wavelength, the new proposed approach involve several wavelengths in the form of two dimensional matrix (retention time x

Symposium on Biomathematics (Symomath 2013)
AIP Conf. Proc. 1587, 78-83 (2014); doi: 10.1063/1.4866538

wavelength). Unfortunately, this strategy leads to data processing problem because of the matrix size. Moreover, expert involvement is still needed to identify particular pattern in the chromatogram which is possibly lead to subjectivity problem. Hence, the study in this research is intended to develop an automation system to provide fast and accurate pigment identification.

Three photosynthetic pigments were used to develop the prototype of the system. Those three pigments are β-carotene, fucoxanthin and zeaxanthin. Chromatograms from well identified pigments in various samples were used to train the system. Then, by applying probabilistic method based on those patterns the system will classify each pattern that found in the new and unidentified chromatogram. Using this system the researcher could analyze the chromatogram in short time. Furthermore, this system could easily utilized to develop real time pigment identification by integrating it to HPLC software.

CHROMATOGRAPHIC DATA

Basically, chromatographic data is a two dimensional matrix (Figure 1). Each column represent the wavelength and each row represent the retention time. Each cell within the matrix represent the amount of the UV light (with corresponding wavelength) that is absorbed by the pigment which is collected at corresponding time. Each pigment will absorb light at a specific wavelength and pass through the HPLC column at a specific time. This fact will be the bases of the design of the classification method being applied.

	35204	35148	35273	35398	35523	35648	35772	35897	36022	36147
18.90133	4104	4160	4279	4300	4389	4558	4690	4646	4827	4938
18.912	4111	4189	4277	4268	4396	4561	4659	4651	4835	4906
18.92267	4120	4168	4268	4257	4388	4569	4674	4656	4852	4934
18.93333	4103	4190	4275	4278	4402	4598	4643	4653	4815	4906
18.944	4109	4199	4262	4290	4370	4566	4674	4645	4820	4918
18.95467	4126	4164	4245	4273	4376	4565	4639	4660	4827	4928
18.96533	4101	4175	4286	4275	4392	4577	4667	4679	4820	4927
18.976	4104	4200	4286	4297	4408	4587	4648	4633	4824	4952
18.98667	4119	4203	4280	4307	4422	4601	4642	4697	4827	4955
18.99733	4146	4212	4263	4330	4450	4575	4710	4697	4854	4992
19.008	4185	4211	4301	4351	4444	4620	4721	4728	4922	5016
19.01867	4236	4255	4317	4368	4504	4669	4783	4784	4990	5091
19.02933	4250	4323	4407	4423	4575	4729	4849	4862	5045	5173
19.04	4323	4378	4467	4501	4618	4836	4931	4942	5136	5259
19.05067	4411	4482	4561	4620	4741	4931	5048	5082	5272	5354
19.06133	4520	4565	4675	4750	4844	5038	5198	5196	5399	5530
19.072	4693	4739	4856	4885	4999	5256	5405	5370	5617	5757
19.08267	4855	4912	5041	5101	5206	5454	5607	5613	5883	6030
19.09333	5049	5151	5252	5313	5446	5687	5852	5896	6167	6331
19.104	5261	5385	5495	5573	5691	5959	6142	6241	6522	6696
19.11467	5515	5651	5746	5831	6013	6276	6482	6571	6886	7107
19.86133	3155	3184	3252	3253	3305	3473	3529	3477	3621	3697
19.872	3180	3162	3261	3240	3325	3504	3511	3481	3621	3679
19.88267	3165	3181	3237	3253	3299	3487	3509	3441	3623	3678
19.89333	3157	3201	3227	3249	3273	3470	3491	3446	3586	3674

FIGURE 1. Example of a chromatographic data matrix

In order to simplify the analysis, those data visualized as a line graphs called chromatogram (Figure 2) and spectrum (Figure 3). For chromatogram, plotted on the x-axis is the retention time and for spectrum, plotted on the x-axis is wavelength. For both graphs, plotted on the y-axis is the amount of the absorbed UV light. In the case of an optimal system, y is proportional to the concentration of the existing pigment. Hence, the peak shown in the graphic is used as an indicator of the particular pigment existence.

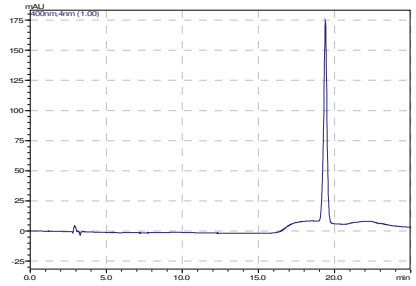

FIGURE 2. Example of a chromatogram from sample containing zeaxhantin, reading for absorption at 400 nm

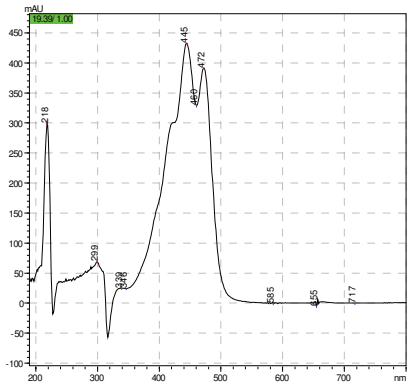

FIGURE 3. Example of a spectrum from sample containing zeaxhantin, reading after 19.43 min running

Commonly, different pigment will produce different spectrum. This difference allows the identity confirmation of the existing pigment in a sample. However, some spectrum have small differences and cannot be absolute confirmations by themselves. Therefore, both retention time and spectrum is used to determine a probability of identifying a pigment in a sample.

MATERIALS AND METHOD

Three of the most common photosynthetic pigments from carotenoid family (β-carotene, fucoxanthin and zeaxanthin) were used as the object of the classification. The unique spectrum of those pigments is depict in Figure 4.

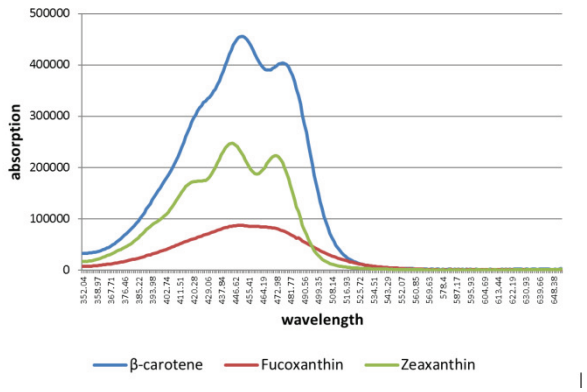

FIGURE 4. Spectrums of carotenoid pigments

Forty eight chromatographic data matrix were collected to develop the classification method. Among them 45 matrix was obtain from samples that contain single pigment (15 for each pigment) and the rest are from the sample that contain mixture of the three pigments. Each sample was run for 80 min and the absorption of UV light was recorded at 189 nm – 800 nm wavelength. Therefore, the dimension of each matrix is 7502x493.Those matrices was then divided into 2 dataset, 39 for training set and 9 for test set. The training set will be used to provide probability distribution information of each pigment chromatographic data. In general, development of the classification system consist of three main procedure which are data preprocessing, model building and validation.

Data Preprocessing

Prior to the classification process, those data should be well prepared such that the process could run faster. Data reduction technique was applied here to reduce the dimension of the chromatogram matrix while keeping the important information contained on it. The reduction process was done sequentially. Begin with the reduction of the row (time retention) and then followed by the reduction of the column (wavelength).

Row reduction was done by select 50 row with the lowest SNR (signal to noise ratio) using following equation [11]:

$$SNR = \frac{\sigma}{\mu} \quad (1)$$

The aim of this step was to assure that only good spectrums were used to identify the pigment existence. Good spectrums are those which are low in noise. This concept is important regarding that the unique pattern of the spectrum could be vanishes because of the noise and lead to misidentification of the pigment being examine. Moreover, the selection was applied in particular range of retention time based on the fact that particular pigment will pass through the HPLC column at a specific time. The specific retention time for the three pigments is shown in Table 1 [12].

TABLE 1. Retention time of the pigments

Pigment	RT (min)
β-carotene	58 – 64
Fucoxanthin	8 – 11
Zeaxanthin	18 – 20

Column reduction was done by applying selection and sampling to the column. Selection was done in order to obtain the relevant wavelengths. Relevant wavelength is particular range of wavelength that theoretically have maximum absorption for particular pigment. In this research the relevant wavelength for the three pigments is 350 nm - 653 nm. Sampling was done by using only even column for the next calculation. This technique was proven will not diminish the unique pattern of the spectrum. Now the size of each chromatogram matrix is become much more smaller (50x246). The next step is transforms those matrix into a 1x12300 vector in order to prepare the training set.

Training Set Preparation

Training set for each pigment is a 13x12300 matrix obtained from the stacked chromatogram vector which is the representation of the whole samples spectrum of each pigment. The compression of the training set was done by applying the most popular statistical method for dimensionality reduction of a large data set which is Karhunen Loeve method, also called principal component analysis (PCA) [13]. PCA provide interpretable overview of the main information in a matrix by extracting and displaying the existing systematic variation [14]. The information in the original variables (refer to the column of the matrix) is compressed into a smaller number of uncorrelated variables called principal components (PCs) using following equation:

$$X = CS^T + E \quad (2)$$

Where X is the final compressed matrix (this matrix will be used in the classification process), C is the scores matrix,S^T is loading matrix transpose and E is an error matrix. At the end of the process, the training set (X) will be a 13x12 matrix.

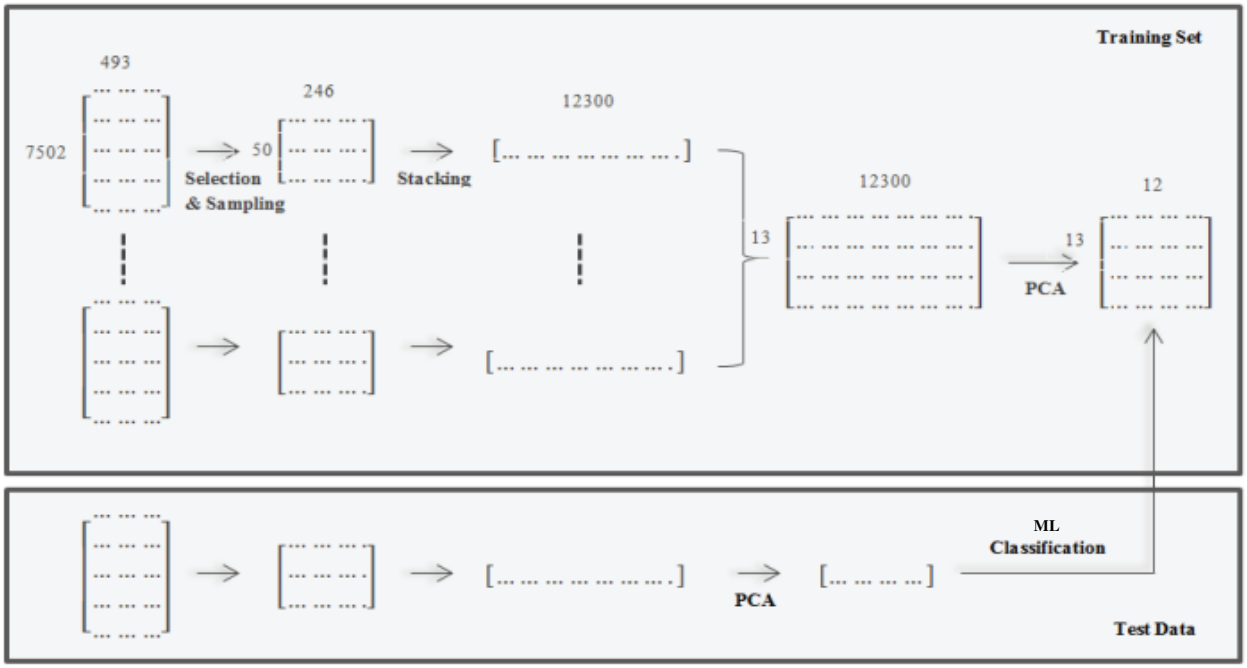

FIGURE 5. Data preprocessing procedure

Test Set Preparation

Test set consist of 9 vectors, the size of each vector is1x12 obtained from the multiplication of the score matrix and the transposed loading matrix. Among the vectors, 6 of them contain pattern of a single pigment (2 for each pigment) and the rest contain patterns from 2 pigments. Each vector will be an input for the classification process. Figure 5 depict the data preprocessing procedure and relationship between training and test data.

Classification Method

Maximum likelihood (ML) classification was applied as the classification method. It is a supervised learning based on bayes theorem [15]. The probability of an input data with feature vector *X* (obtain from the compressed chromatogram matrix) belong to pigment *i*, is given by:

$$P(i|X) = \frac{P(X|i).P(i)}{P(X)} \qquad (2)$$

Where *P(X*\|i*)* is likelihood function, *P(i)* is probability that pigment *i* occur and *P(X)* is probability that ω is observed. ML assumes that the distribution of the data within a given class *i* is a multivariate Gaussian distribution as follow:

$$f(X) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-(X-\mu)'\Sigma^{-1}(X-\mu)/2} \qquad (3)$$

Therefore, the likelihood function become:

$$\ln P(i|X) = -\frac{1}{2}(X-\mu_i)'\Sigma_i^{-1}(X-\mu_i) - \frac{N}{2}\ln(2\pi) + -\frac{1}{2}\ln(|\Sigma_i|) \qquad (4)$$

Each input data is assigned to the class (pigment) with the highest likelihood or labeled as "unknown" if the probability value are below a particular threshold. An input data is assigned to pigment *i* using the rule:

$$If\ P(i|X) > P(j|X)\ for\ all\ j \neq i \qquad (5)$$

Figure 6 provide the scheme of the ML method and the general ML procedures applied in this research are as follows:

1. Determine the number of the pigment type being identified
2. Estimate the mean vector and covariance matrix for each pigment type from the training set
3. Every input data is classified into one of the pigment type or labeled as unknown.

Input data is obtained from a chromatographic data matrix which is preprocessed based on specific retention time due to identification of particular pigment. Thus, in this research one chromatographic

data matrix will be preprocessed three times respectively.

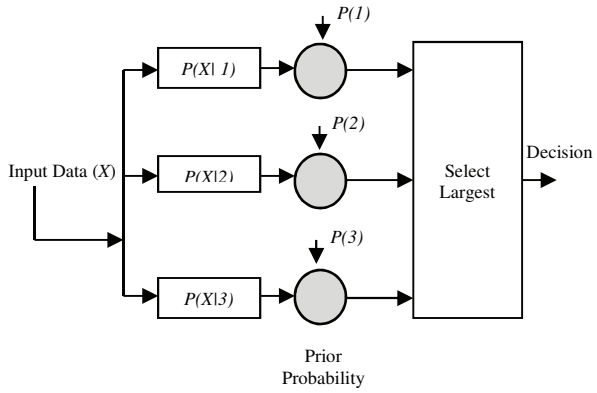

FIGURE 6. The scheme of ML method

Computer Application

To facilitate rapid identification of a pigment from a chromatogram data matrix, a prototype of identification software was developed. Figure 7 depict the user interface of the software.

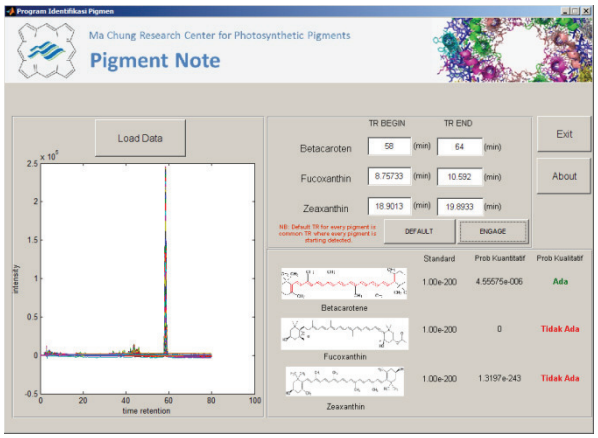

FIGURE 7. User interface of the software

User inputted the chromatographic data matrix by clicking the Load Data button. The input file format is .csv. Default retention time as the basis of identification process for each pigment is as shown in Table 1, but user could specify different retention time. Therefore, user could extent the pattern matching process of the spectrum regarding to the condition of the laboratory experiment. Output of the software is existence status of each pigment.

For qualitative analysis, the chromatogram of the sample being analyzes was also displayed in the user interface. As soon as the software read the input file, the chromatogram will be created.

RESULT AND DISCUSSION

The performance of the method was analyzed using the test set. ML prediction was compared to the information of the actual existing pigment in the sample. The result is shown in Table 2. For all samples, ML performs very good result when classify the carotenoid pigments. The method could well inform the existence of both single and multiple pigment. Thus, it is proven that matrix compression method using PCA could well encapsulated the unique pattern information of the carotenoid spectrum.

TABLE 2. Performance evaluation result

Sample ID.	Existing Pigment (Actual)			ML Prediction		
	B	***F***	***Z***	***B***	***F***	***Z***
1	√	-	-	√	-	-
2	√	-	-	√	-	-
3	-	√	-	-	√	-
4	-	√	-	-	√	-
5	-	-	√	-	-	√
6	-	-	√	-	-	√
7	√	√	-	√	√	-
8	√	-	√	√	-	√
9	-	-	√	-	-	√

B = β-carotene, *F* = Focoxanthin, *Z*=Zeaxanthin

CONCLUSION

The proposed method to automate chromatogram data analysis shows promising result to be developed into more sophisticated computer application. The method could recognize more photosynthetic pigments by simply update the training set database with various chromatogram data matrix from other photosynthetic pigments.

FUTURE WORK

Since samples used in this research is well prepared such that all experiment variable is controlled, further performance confirmation of the method should be done. Complex chromatographic profile from a plant extraction mixture will be a good material to study the weaknesses of the proposed method.

For future software development, prototype of the computer application is being developed into a real

time pigment identification system by integrating it with HPLC software.

REFERENCES

1. Aberoumand, World J Dairy and Food Sci. **6**, 71 (2011).
2. D.Cardon, "Natural dyes today: why?" in *International Symposium on Natural Dyes Proceedings*, 2006, pp. 4-8.
3. K.D.Casselman, S.J. Kadolph,"Horticultural Hues - Natural dyes from plants",*Chronica Horticulturae,* 2010, Vol. 50(2), pp. 19-24.
4. P. Chattopadhyay, S.Chatterjee and S.K. Sen, Afri.J. Biotech. **7**, 2972 (2008).
5. M. Visalakshi and M. Jawaharlal, Journal of Agriculture and Allied Sciences. **2**, 42 (2013).
6. L. Schluter, T.L. Lauridsen, G. Krogh and T. Jorgensen, Freshwater Biology. **51**, 1474 (2006).
7. V. Brotas and M.R. Plante-Cuny, "The Use of HPLC Pigment Analysis to Study Microphytobenthos Communities" i*n ActaOecologica*, *Proceedings of the Plankton Symposium*, Elsevier, 2003, Vo.24(1), pp.S109-S115.
8. J.G. Lashbrooke, P.R. Young, A.E. Strever, C. Stander and M.A. Vivier, Australian Journal of Grape and Wine Research. **16**, 349 (2010).
9. J. Ricardo Lucio-Gutierrez, J. Coello and S. Maspoch, Analytica Chimica Acta. **710**, 40 (2012).
10. J. Ricardo Lucio-Gutierrez, A. Garza-Juarez, J. Coello, S. Maspoch, M.L. Salazar-Cavazos, R. Salazar-Aranda and N.W. de Torres, Journal of Chromatography A. **1235**, 68 (2012).
11. A. Felinger, *Data Analysis and Signal Processing in Chromatography*, Elsevier, 1998.
12. G. Britton,*Carotenoids:Handbook,*BirkhauserVerlag : Basel, 2004.
13. M. Kantardzic, *Data Mining: Concept, Models, Methods and Algorithms*, IEEE Press, Wiley, 2001.
14. Y.Chen, M.Xie, Y.Yan, S.Zhu, S.Nie, C.Li, Y.Wang, X.Gong, Anal.Chim.Acta. **618**, 121 (2008).
15. G.J. Miao, M.A. Clements, *Digital Signal Processing and Statistical Classification*, Artech House Inc., 2002.