A Probability-based Multivariate Statistical Algorithm for Autofluorescence Spectroscopic Identification of Oral Carcinogenesis

Chih-Yu Wang¹, Chin-Tin Chen², Chun-Pin Chiang³, Shueng-Tsong Young¹, Song-Nan Chow² and Huihua Kenny Chiang^{*1}

¹Institute of Biomedical Engineering, National Yang-Ming University, Taipei, Taiwan;

²Laser Medicine Research Center, National Taiwan University, Taipei, Taiwan and

³School of Dentistry, National Taiwan University, Taipei, Taiwan

Received 12 June 1998; accepted 20 November 1998

ABSTRACT

A probability-based multivariate statistical algorithm combining partial least-squares (PLS) and logistic regression was developed to identify the development stages of oral cancer through analysis of autofluorescence spectra of oral tissues. Tissues were taken from a 7,12-dimethylbenz[*a*]anthracene-induced hamster buccal pouch carcinogenesis model. Analyses were conducted at various excitation wavelengths, ranging from 280 nm to 400 nm in 20 nm increments, to assess classification performance at different excitations. For each excitation the PLS analysis and logistic regression were combined, on the basis of cross validation, to calculate the posterior probabilities of samples belonging to four stages of cancer development: normal tissues, hyperplasia, dysplasia and early cancers and frankly invasive cancers. Results showed that the 320 nm excitation wavelength optimally classified the cancer development stages: the accuracy rates for identifying samples at that excitation were 91.7%, 83.3%, 66.7% and 83.3% for the four respective stages. The average accuracy rate was 81.3%. These results suggest that the algorithm described in this study might be useful for the detection of human oral cancers.

INTRODUCTION

Over the past two decades, some new and effective tools for diagnosing human premalignant and malignant lesions have been developed. One promising diagnostic technique, fluorescence spectroscopy, offers the benefits of less invasiveness, real-time diagnosis and easy operation. Recent studies have shown that exogenous tumor-localized fluorescent agents, such as hematoporphyrin derivatives (HPD)[†] and profimer sodium, are quite useful in detecting cancer tissues

(1,2); however, these extrinsic fluorescent chromophores interact with cellular ultrastructures and may in turn induce some undesired side effects (3). By contrast, biological tissues contain various kinds of fluorophores, such as tryptophan, tyrosine, elastin, collagen, NADH and FAD (4,5), and cancer detection by means of the autofluorescence coming from intrinsic fluorophores can not only avoid unwanted side effects but also display the biochemical properties of the tissues. Consequently, the autofluorescence spectroscopic technique has gained prominence in recent years.

The autofluorescence spectroscopic technique has been used to identify neoplastic tissues from various organ systems, such as bronchus, colon, cervix, esophagus and head and neck (6–10). One problem, however, has been that samples used in these studies came mainly from clinical patients, making it hard to control for individual variations. Another problem with such clinical studies is the difficulty of collecting an adequate amount of tissue at the different stages of cancer development. Use of an animal model is one way to avoid these problems.

In this study, we examined the autofluorescence spectra of hamster buccal pouch tissues from a 7,12-dimethylbenz[*a*]anthracene (DMBA)-induced oral carcinogenesis animal model. The hamster buccal pouch carcinogenesis model was first described by Sally in 1954 (11) and subsequently was modified and standardized by Morris to make possible a uniform reproduction of the lesions (12). The hamster buccal pouch carcinogenesis model is a good animal system that closely resembles events in the development of precancerous lesions and squamous cell carcinoma (SCC) in the human oral cavity (13). Because SCC is the most common type of human oral cancer (14,15), the detection of early SCC is very important for oral cancer diagnosis.

Previously, Balasubramanian *et al.* (16) conducted an investigation similar to ours: they analyzed the fluorescence spectra of a few extracted fluorophores from hamster buccal pouches and found that excitation at 405 nm and 420 nm gave significant differences among normal, premalignant and malignant tissues. They demonstrated that the DMBA-treated hamster buccal mucosa showed hyperplastic changes at 4–6 weeks of treatment, papillomas at 8–10 weeks, early invasive SCC at 11–13 weeks and well-differentiated SCC at 14–16 weeks (16). In the present study, we measured the

*To whom correspondence should be addressed at: Institute of Biomedical Engineering, National Yang-Ming University, 155, Section 2, Li-Nung Street, Taipei, 112, Taiwan. Fax: 886-2-28210847; e-mail: chiang@bme.ym.edu.tw

[†]Abbreviations: DMBA, dimethylbenz[*a*]anthracene; HPD, hematoporphyrin derivative; ML, maximum likelihood; MVLR, multivariate linear regression; PCA, principal components analysis; PLS, partial least squares; SCC, squamous cell carcinoma.

© 1999 American Society for Photobiology 0031-8655/99 \$5.00+0.00

Table 1. The histological assessment results of hamster buccal pouch tissues with different periods of DMBA treatment

DMBA treatment (weeks)	Histological assessment results
2–4	Hyperkeratosis and acanthosis
6	Focal mild epithelial dysplasia
8	Moderate or severe epithelial dysplasia
10	Carcinoma <i>in situ</i> , early invasive SCC
12	Papillary SCC
14–16	Frankly invasive SCC

fluorescence spectra of hamster buccal pouch samples *ex vivo* at various excitation wavelengths. In addition, to establish a diagnostic tool for detection of human oral cancers, we developed a multivariate statistical algorithm that describes the relationship between the autofluorescence spectra and oral carcinogenesis from an animal model.

Multivariate analysis methods have been used for spectrum analysis in many fields (17–19). Principal component analysis (PCA) and partial least-squares (PLS) analysis are two major methods that have been widely used (20,21). The PCA looks for the principal components that account for most variations of the spectrum, and these selected principal components are then adopted in the subsequent regression analysis. For collinear data, PCA combined with a regression model can give substantial improvement over ordinary multivariate linear regression (MVLRL) (21). It has been argued, however, that the principal components selected by using PCA may be irrelevant to the dependent variables. Hence, another multivariate method, PLS analysis, was proposed to remedy this problem. The PLS method differs from the PCA method in using the dependent variables actively during the decomposition of the independent variables (20,22). By adopting the information of the independent and dependent variables simultaneously, PLS can reduce the impact of large but irrelevant variations of the independent variables in the calibration modeling (20).

A general way for classifying samples into various categories is to choose decision lines that minimize the misclassification rates. A dichotomizing discrimination scheme, however, does not provide information regarding the uncertainty of the classification, and it is likely to have poor performance for samples located near the boundary between two categories (23,24). This situation can be improved by adopting a probability-based discriminating model—the logistic regression. The combination of the PLS analysis and the logistic regression can provide a probability-based index for identifying the sample type, and the discrimination re-

sults may be more reasonable than those obtained by the dichotomizing discrimination method.

In our previous work, we have demonstrated the group structure of cancer development stages by adopting PLS factors (25) but have difficulty in obtaining a quantitative classification for those samples, such as the determination of the threshold for separating different sample stages and calculating the probabilities of sample classification. The goal of the present study was to develop a probability-based multivariate statistical algorithm combining PLS and logistic regression to identify oral cancer development stages by analyzing the autofluorescence spectra of oral tissues taken from a DMBA-induced hamster buccal pouch carcinogenesis model. To address this objective, we measured the autofluorescence spectra of tissues from hamster buccal pouches at different cancer development stages. By combining the PLS analysis and the logistic regression on the basis of cross validation, the classification rates at different excitation wavelengths were determined, and the optimal excitation wavelength was identified. The algorithm developed may be helpful for future development of diagnostic techniques for human oral cancer.

MATERIALS AND METHODS

Sample preparation. Forty-eight 4 week-old male Syrian hamsters were used for the carcinogenesis model. The left buccal pouches of the animals were painted three times a week with a 0.5% solution of carcinogen DMBA in heavy mineral oil. The right buccal pouches were left untreated.

Six hamsters were sacrificed every 2 weeks. The left and right buccal pouches of the sacrificed animals were examined, removed, embedded in cryomatrix frozen medium, snap-frozen in liquid nitrogen within 5 min and stored at –70°C in an ultralow temperature freezer until the time of measurement. For measurement, the samples were cut into pieces 4–6 mm in diameter and 0.8–1.2 mm in thickness. After measurement, all samples were fixed in 10% neutral formalin, embedded, sectioned, hematoxylin and eosin stained and submitted for histological examination by two oral pathologists.

Classification of samples. Table 1 shows the histological assessment results of hamster buccal pouch tissues after different periods of DMBA treatment. The results were consistent with the work of Balasubramanian *et al.* (16). A total of 96 samples from 48 hamsters were classified into four different categories according to their histological assessment results. Table 2 shows that these samples were classified into four important stages of oral carcinogenesis: normal tissues, hyperplasia and mild dysplasia, severe dysplasia and carcinoma *in situ* and invasive cancers.

Fluorescence measurement. A standard luminescence spectrometer (SLM Instruments, Urbana, IL, USA) was used to measure the fluorescence spectra. A monochromator with a 150 W ozone-free xenon lamp was used as the excitation light source to illuminate the sample surface at an incident angle of 37°, and the emission light was collected perpendicular to the excitation light to minimize the collection of reflected excitation light (26). The incident beam was focused on the center of the tissue surface with a spot size 4 mm

Table 2. The classification of samples based on their histological assessment results, number of samples in each category and three coded dummy variables

Classification	Cancer development stages	Sample number	Coded variables		
Category I	Normal	48	0	0	0
Category II	Hyperkeratosis, acanthosis, and mild epithelial dysplasia	18	0	0	1
Category III	Severe epithelial dysplasia, carcinoma <i>in situ</i>	12	0	1	0
Category IV	Early, papillary and frankly invasive SCC	18	1	0	0

by 1 mm. In the fluorescence measurements, excitation wavelengths ranged from 280 nm to 400 nm in 20 nm increments. To avoid collection of the first and second harmonics of the excitation light, the emission wavelength was measured at 20 nm greater than the excitation wavelength to 20 nm less than twice the excitation wavelength. Less than 10 min was required for the measurement of each sample. A blank quartz slide was used for measuring the scattered light; this measurement was then subtracted from the fluorescence spectra to eliminate the background light. Each fluorescence spectrum was then normalized by dividing the intensity at each wavelength by the integrated area under the total spectrum. A repeated emission scan was conducted at 280 nm and 300 nm excitations to monitor tissue deterioration and photobleaching. Changes in the intensities of the re-scanned spectra were less than 1%.

Cross-validation technique. Cross validation is a general and efficient method for validating the performances of an algorithm (20). A full cross validation requires splitting samples into M groups with each group consisting of only one sample. One of the M groups is kept apart for prediction purposes, and the remaining groups are used for calibration. The statistical model is then established by using the calibration set, and the prediction result is obtained from the prediction set. Next, another group is used as the prediction set, and the steps are repeated. The procedure continues until all of the M groups have been used for prediction. This "leave one out" method yields the most reasonable estimate of the predictive ability of the statistical algorithm (27). In this study, the cross-validation technique was used to evaluate performances of the algorithms for analyzing the fluorescence spectra at different excitation wavelengths.

PLS analysis. Because a multivariate statistical algorithm for numerous highly collinear variables is generally very complex and its predictive ability may be degraded, PLS analysis is used to approximate complicated multivariate fluorescence spectra with factors that are fewer but most relevant to cancer development stages. For each excitation wavelength, the fluorescence spectra are placed in a data matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & \cdots & x_{1c} \\ x_{21} & x_{22} & \cdots & \cdots & x_{2c} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{r1} & x_{r2} & \cdots & \cdots & x_{rc} \end{bmatrix} \quad (1)$$

where r represents the sample number, and c represents the number of emission wavelengths. The four cancer development stages are coded by use of three dummy variables (Table 2), constructing a dummy matrix $\mathbf{Y}$. The matrices $\mathbf{X}$ and $\mathbf{Y}$ are mean-centered and decomposed into a linear combination of a common coefficient $\mathbf{T}$ as below (22):

$$\begin{aligned} \mathbf{X} - \mathbf{1}\bar{\mathbf{X}} &\equiv \mathbf{X}_m = \mathbf{TP} + \mathbf{E} \\ &= \mathbf{t}_1\mathbf{p}_1 + \mathbf{t}_2\mathbf{p}_2 + \cdots + \mathbf{t}_a\mathbf{p}_a + \mathbf{E} \end{aligned} \quad (2)$$

$$\begin{aligned} \mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}} &\equiv \mathbf{Y}_m = \mathbf{UQ} + \mathbf{F} \\ &= \mathbf{u}_1\mathbf{q}_1 + \mathbf{u}_2\mathbf{q}_2 + \cdots + \mathbf{u}_a\mathbf{q}_a + \mathbf{F} \end{aligned} \quad (3)$$

$$\mathbf{U} = \mathbf{TB} + \mathbf{D} \quad (4)$$

where matrices $\mathbf{T}$ and $\mathbf{U}$ are referred to as PLS scores of $\mathbf{X}_m$ and $\mathbf{Y}_m$, respectively; the subscript a represents the PLS factor number. First pairs of PLS factors are extracted from $\mathbf{X}_m$ and $\mathbf{Y}_m$, and the additional pairs of PLS factors can then be calculated from residuals $\mathbf{E}$ and $\mathbf{F}$ sequentially. Thus, the PLS factors are determined to describe $\mathbf{X}_m$ as accurately as possible and simultaneously to correlate with $\mathbf{Y}_m$ (22). The row vectors in matrix $\mathbf{P}$ are all normalized to unit, and the column vectors in $\mathbf{T}$ are orthogonal to each other.

The cross-validation technique is usually used in PLS analysis for determining the optimal number of PLS factors. The criterion for calculating the optimal number of PLS factor is to get a minimal value of the objective function $(\mathbf{Y}_m - \mathbf{UQ})^T(\mathbf{Y}_m - \mathbf{UQ})$ or $\mathbf{F}^T\mathbf{F}$. That is, the number of PLS factors are chosen such that the mean-square error computed from the actual and predicted values in the $\mathbf{Y}_m$ matrix is minimized.

Combination of PLS analysis and the logistic regression. The

classification algorithm developed by combining the PLS analysis and the logistic regression was based on cross validation. The PLS factor scores were obtained by using the calibration set. Let $\mathbf{s}$ denote the vector of variables comprising these PLS scores. The samples of hamsters were classified into four categories (denoted by D_1 , D_2 , D_3 and D_4) according to the cancer development stages. Let the conditional probability distribution function of $\mathbf{s}$ in the four categories be $\Pr(\mathbf{s}|D_1)$, $\Pr(\mathbf{s}|D_2)$, $\dots$, $\Pr(\mathbf{s}|D_4)$, respectively. The likelihood ratio of the variable vector $\mathbf{s}$ for categories D_i and D_1 can be written as

$$L_{i1}(\mathbf{s}) = \frac{\Pr(\mathbf{s}|D_i)}{\Pr(\mathbf{s}|D_1)} \quad i = 1 \sim 4 \quad (5)$$

where D_1 is the reference category that contains the normal samples.

We then proceed by assuming that the likelihood ratio of $\mathbf{s}$ for categories D_i and D_1 has the simple exponential form

$$L_{i1}(\mathbf{s}) = \exp(\mathbf{a}_{0i}^* + \mathbf{a}_i^T\mathbf{s}). \quad (6)$$

Based on Bayes' theorem, the posterior probability for a given sample with variable vector $\mathbf{s}$ lying in category D_i is expressed as (28)

$$\Pr(D_i|\mathbf{s}) = \frac{L_{i1}(\mathbf{s})}{\sum_{j=1}^4 L_{j1}(\mathbf{s})} = \frac{\exp(\mathbf{a}_{0i}^* + \mathbf{a}_i^T\mathbf{s})}{\sum_{j=1}^4 \exp(\mathbf{a}_{0j}^* + \mathbf{a}_j^T\mathbf{s})} \quad (7)$$

Next we calculate the intercepts $\tilde{\mathbf{a}}_{0i}^*$ and the coefficients vectors $\tilde{\mathbf{a}}_i$ ($i = 1, 2, 3, 4$) by using maximum likelihood (ML) estimations. Let $\mathbf{s}_{ij}$ denotes variable vector of the i th sample in j th category. Albert and Anderson indicated that for n samples with n_1 belonging to D_1 , n_2 belonging to D_2 , $\dots$ and n_4 belonging to D_4 ($n_1 + n_2 + n_3 + n_4 = n$), the ML estimates $\tilde{\mathbf{a}}_{0i}^*$ and $\tilde{\mathbf{a}}_i$ ($i = 1, 2, 3, 4$) are obtained by maximizing the following function (28,29):

$$\lambda = \prod_{i=1}^4 \prod_{j=1}^{n_i} \frac{\exp(\mathbf{a}_{0i}^* + \mathbf{a}_i^T\mathbf{s}_{ij})}{\sum_{g=1}^4 \exp(\mathbf{a}_{0g}^* + \mathbf{a}_g^T\mathbf{s}_{ij})} \quad (8)$$

with respect to the parameters. In other words, we maximize the probabilities in such a manner that all the samples are assigned to their prescribed categories. After the parameters $\tilde{\mathbf{a}}_{0i}^*$ and $\tilde{\mathbf{a}}_i$ being determined, the posterior probability of a sample in the prediction set (denoted by $\mathbf{s}^v$) belonging to a category D_k can be calculated as below:

$$\Pr(D_k|\mathbf{s}^v) = \frac{\exp(\tilde{\mathbf{a}}_{0k} + \tilde{\mathbf{a}}_k^T\mathbf{s}^v)}{\sum_{j=1}^4 \exp(\tilde{\mathbf{a}}_{0j} + \tilde{\mathbf{a}}_j^T\mathbf{s}^v)} \quad (9)$$

where $\tilde{\mathbf{a}}_{0k} = \tilde{\mathbf{a}}_{0k}^* + \log(n_1/n_k)$. Albert and Harris recommend this adjustment when the probability distribution functions $\Pr(\mathbf{s}|D_k)$ and $\Pr(\mathbf{s}|D_1)$ are not exactly known (28).

Computer software and hardware. All the mathematical calculations were completed by the self-developed programs written in MATLAB (MathWorks Inc., Natick, MA USA). The program was executed on an IBM-compatible personal computer with Intel Pentium 233 processor.

RESULTS

Fluorescence spectra

Figure 1a–c depicts the averaged fluorescence spectra of tissues in four categories at various excitation wavelengths. For simplicity, only those averaged spectra at 280, 320 and 360 nm excitation wavelengths are presented. The figures show that the patterns of the autofluorescence spectra were different at various excitation wavelengths. Figure 1a shows a prominent emission peak near 330 nm at 280 nm excitation. Figure 1b shows two emission peaks at 320 nm excitation: the first peak (near 385 nm) decreased, whereas the second peak (near 470 nm) increased in accordance with the cancer development stages (*i.e.* categories I–IV). This reversal of

Figure 1. Averaged spectra of samples in each category at (a) 280 nm, (b) 320 nm and (c) 360 nm excitation.

pattern might be due to collagen decrease and NADH increase in cancer tissues (8,30). Figure 1c shows two prominent emission peaks near 460 nm and 640 nm at 360 nm excitation for the samples in categories III and IV. The first peak (near 460 nm) decreased, whereas the second peak (near 640 nm) increased in accordance with the cancer development stages. The changes in these fluorescence patterns might be related to the decrease in collagen and accumulation of porphyrin compounds (7,16). The differences among the four averaged fluorescence spectra are obvious at 320 nm and 360 nm excitations (Fig. 1b,c) but are less apparent at 280 nm excitation (Fig. 1a).

Determination of the PLS factor number

Figure 2a shows the number of PLS factors chosen at different excitation wavelengths. For each excitation wavelength, the optimal PLS factor number was determined by using PLS analysis on the spectra of the total 96 samples based on cross validation. We found that no more than six

Figure 2. (a) The optimal PLS factor number for PLS analysis at different excitation wavelengths. (b) The number of correctly classified samples obtained by classifying a sample into a category if it has the highest probability of belonging to that category.

PLS factors were required for all excitation wavelengths used in this study, indicating that the dimensions of variables (the emission wavelengths in this study) of the data set could be significantly reduced by applying the PLS analysis. Notably, only three PLS factors were required at 320 nm excitation, revealing that the complexity of computation in the logistic regression could be minimized by employing the PLS factors of the spectra at 320 nm excitation.

Combination of PLS analysis and logistic regression

The classification algorithm, which combined PLS analysis and logistic regression, was developed using the autofluorescence spectra at all excitation wavelengths shown in Fig. 2. For each excitation wavelength, the samples were divided into 96 groups with each group consisting of only one sample. For each test, one sample was kept apart for prediction purposes, and the remaining samples were used for calibration; the validation procedure was repeated 96 times. The PLS scores obtained from the calibration set were employed to develop the logistic regression model, and these parameters were used to estimate the posterior probability for the sample in the prediction set of its belonging to each one of the four categories.

Figure 2b shows the summation of the number of correctly classified samples for each category at different excitation wavelengths. The samples were discriminated by means of

Figure 3. The posterior probabilities of belonging to four categories calculated by using logistic regression based on cross validation. The correctly classified samples are encircled in solid lines, and the incorrectly classified samples are encircled in dashed lines. (a) Probabilities of the samples belonging to category I. (b) Probabilities of the samples belonging to category II. (c) Probabilities of the samples belonging to category III. (d) Probabilities of the samples belonging to category IV.

Table 3. True and predicted samples in each category by means of the logistic regression at 320 nm excitation

True	Category I	Category II	Category III	Category IV
Category I	44	2	2	0
Category II	0	15	3	0
Category III	0	1	8	3
Category IV	0	1	2	15

the following principle: a sample is classified into a category if it has the highest probability of belonging to that category. We found that the excitation wavelength of 320 nm had the highest number of correctly classified samples (82 of the total 96 samples). Thus, in this study, 320 nm appears to be the optimal excitation wavelength.

Posterior probability and accuracy rates of sample classification at 320 nm excitation

Figure 3 shows the posterior probabilities of samples belonging to the four categories at 320 nm excitation. Figure 3a shows that most samples in category I (*i.e.* normal samples) have the highest probabilities of belonging to category I, and most samples in categories II, III and IV have lower probabilities of belonging to category I. Figure 3b, c and d displays the situation of samples belonging to categories II, III and IV, respectively.

Table 3 shows the number of correctly classified and misclassified samples for each category at 320 nm excitation. The accuracy rates of sample classification were 91.7% (44/48), 83.3% (15/18), 66.7% (8/12) and 83.3% (15/18) for categories I, II, III and IV, respectively.

DISCUSSION

This study developed a probability-based multivariate statistical algorithm for identifying development stages of oral cancer through analysis of the autofluorescence spectra of tissues taken from a DMBA-induced hamster buccal pouch carcinogenesis model. By combining PLS analysis and logistic regression, we achieved an average of 81.3% accuracy for identifying oral tissues at different stages of cancer development.

The classification accuracy rates for categories I, II and IV were high (>83%), but that for category III was relatively low (<70%). Because samples in category III included those with severe epithelial dysplasia, carcinoma *in situ* and early invasive SCC, the variations of their fluorescence spectra were inevitably greater than for samples belonging to other categories. The unsatisfactory results in category III might therefore be due to the collection of a variety of samples with a small sample population. The sample variety (from severe epithelial dysplasia to early invasive SCC) might cause the variations of the PLS scores of these samples. Even combined with the logistic regression, some of them are likely to be misclassified. Still, the good accuracy rates for discriminating samples in categories I and II indicate that this algorithm may be useful for diagnosing tissues in the very early stage of oral cancer development.

The autofluorescence spectra of the samples exhibited different patterns when excited at different wavelengths. This difference might be caused by the great diversity of the fluorophores existing in biological tissues: each fluorophore has its own emission pattern. For example, the fluorescence spectra of samples in the four categories (Fig. 1a) are all similar to that of tryptophan when excited at 280 nm (31). Because the fluorescence emission of most native proteins is dominated by tryptophan at 280 nm excitation (31), these results indicate that tryptophan may be one of the main fluorophores in oral tissues.

By contrast, the fluorescence intensities of the band 370–420 nm decrease from categories I to IV, whereas those of the band 470–540 nm increase from categories I to IV (Fig. 1b). Schomacker *et al.* and Ramanujam *et al.* have suggested that collagen, NADH and FAD contribute most to the fluorescence spectrum at 337 nm excitation (17,30), and the fluorescence peaks of collagen, NADH and FAD are at about 390, 460 and 540 nm, respectively (17). In this study, we found that the fluorescence spectra at 320 and 337 nm excitation were quite similar. These results indicate that collagen, NADH and FAD may be closely related to the stages of cancer development.

The fluorescence spectra of samples in categories III and IV exhibit a prominent peak at around 640 nm (Fig. 1c). Many investigators have also observed this emission peak from other cancerous tissues and concluded that it is mainly due to the presence of porphyrin compounds (32,33). Some researchers have used this peak to discriminate normal from cancerous tissues (16). The results of the present study, however, showed that this peak was inadequate for discriminating normal from hyperplastic tissues. Possibly, the porphyrin compounds that accumulate in the hyperplastic tissues are insufficient to differentiate the fluorescence spectra of these two types of tissues.

The two most widely used methods for spectral analysis are PLS and PCA. Because PLS uses the information of spectra and the constituents simultaneously when extracting factors, this method typically requires fewer factors than does PCA, and thus faster and more parsimonious results can be obtained. However, some researchers have shown that the prediction results are not optimal when the relationship between spectra and the constituent deviates from linearity (34). The combination of some other nonlinear statistical algorithms (*e.g.* artificial neural network) and PCA or PLS yields significantly better prediction results than does use of PCA or PLS alone (34,35). In this study, we combined PLS and logistic regression, which in nature is a nonlinear regression model, to obtain better prediction results than use of PLS alone.

One additional advantage of the algorithm developed in this study is that it provides the posterior probability of samples belonging to the diagnostic categories. Thus, clinicians could recheck those samples having lower posterior probability of belonging to one category by using a traditional biopsy method. This probability-based algorithm can quantitatively provide diagnostic information and might therefore be clinically useful for classifying tissues at different stages of cancer development.

REFERENCES

1. Crean, D. H., C. Liebow, R. B. Penetrante and T. S. Mang (1993) Evaluation of porfimer sodium fluorescence for measuring tissue transformation. *Cancer* **72**, 3068–3077.
2. Montan, S., K. Svanberg, and S. Svanberg (1985) Multicolor imaging and contrast enhancement in cancer-tumor location using laser-induced fluorescence in hemotoporphyrin derivative-bearing tissue. *Opt. Lett.* **10**, 56.
3. Kessel, D. (1984) Hematoporphyrin and HpD: photophysics, photochemistry and phototherapy. *Photochem. Photobiol.* **39**, 851–859.
4. Alfano, R. R., B. B. Das, J. Cleary, R. Prudente and E. Celmer (1991) Light sheds light on cancer-distinguishing malignant tumors from benign tissues and tumors. *Bull. N.Y. Acad. Med.* **67**, 143–150.
5. Sliberberg, M., B., H. E. Savage, G. C. Tang, P. G. Sacks, R. R. Alfano and S. P. Schantz (1994) Detecting retinoic acid-induced biochemical alteration in squamous cell carcinoma using intrinsic fluorescence spectroscopy. *Laryngoscope* **104**, 278–282.
6. Hung, J., S. Lam, J. C. LeRiche and B. Palcic (1991) Autofluorescence of normal and malignant bronchial tissue. *Lasers Surg. Med.* **11**, 99–105.
7. Romer, T. J., M. Fitzmaurice, R. M. Cothren, R. Richards-Kortum, R. Petras, M. V. Sivak and J. R. Kramer (1995) Laser-induced fluorescence microscopy of normal colon and dysplasia in colonic adenomas: implication for spectroscopic diagnosis. *Am. J. Gastroenterol.* **90**, 81–87.
8. Ramanujam, N., M. F. Mitchell, A. Mahadevan, S. Thomsen, E. Silva and R. Richards-Kortum (1994) Fluorescence spectroscopy: a diagnostic tool for cervical intraepithelial neoplasia (CIN). *Gynecol. Oncol.* **52**, 31–38.
9. Vo-Dinh, T., M. Panjehpour, B. F. Overholt and P. Buckley III (1997) Laser-induced differential fluorescence for cancer diagnosis without biopsy. *Appl. Spectrosc.* **51**, 58–63.
10. Kolli, V. R., H. E. Savage, T. J. Yao and S. P. Schantz (1995) Native cellular fluorescence of neoplastic upper aerodigestive mucosa. *Arch. Otolaryngol. Head & Neck Surg.* **121**, 1287–1292.
11. Sally, J. J. (1954) Experimental carcinogenesis in the cheek pouch of the Syrian hamster. *J. Dent. Res.* **33**, 153–162.
12. Morris, A. L. (1961) Factors influencing experimental carcinogenesis in the hamster cheek pouch. *J. Dent. Res.* **40**, 3–15.
13. Shin, D. M., I. B. Gimenez, J. S. Lee, K. Nishioka, M. J. Wargovich, S. Thacher, R. Lotan, T. J. Slaga and W. K. Hong (1990) Expression of epithelial growth factor receptor, polyamine levels, ornithine decarboxylase activity, micronuclei, and transglutaminase I in a 7,12-dimethylbenz[a]anthracene-induced hamster buccal pouch carcinogenesis model. *Cancer Res.* **50**, 2505–2510.
14. Chen, C. H. (1987) An epidemiological study of oral squamous cell carcinoma in south Taiwan. *J. Formosan Dent. Assoc.* **10**, 268–274.
15. Kwan, H. W. (1976) A statistical study on oral carcinomas in Taiwan with emphasis on the relationship with betel nut chewing: a preliminary report. *J. Formosan Med. Assoc.* **75**, 497–506.
16. Balasubramanian, S., V. Elangovan and S. Govindasamy (1995) Fluorescence spectroscopy identification of 7,12-dimethylbenz[a]anthracene-induced hamster buccal pouch carcinogenesis. *Carcinogenesis* **10**, 2461–2465.
17. Schomacker, K. T., J. K. Frisoli, C. C. Compton, T. J. Flotte, J. M. Richter, N. S. Nishioka and T. F. Deutsch (1992) Ultraviolet laser-induced fluorescence of colonic tissue: basic biology and diagnostic potential. *Lasers Surg. Med.* **12**, 63–78.
18. Carlsson, A. E. and K. L. R. Janne (1995) Near-infrared spectroscopy as an alternative to biological testing for quality control of hyaluronan: comparison of data preprocessing methods for classification. *Appl. Spectrosc.* **49**, 1037–1040.
19. Haaland, D. M., H. D. T. Jones and E. V. Thomas (1997) Multivariate classification of the infrared spectra of cell and tissue samples. *Appl. Spectrosc.* **51**, 340–345.

20. Martens, H. and T. Naes (1989) *Multivariate Calibration*. Wiley Inc., London.
21. Sjöström, M. and S. Wold (1980) A pattern recognition method based on principal component models. In *Pattern Recognition in Practice* (Edited by E. S. Gelsema and L. N. Kanal), pp. 351–359. North-Holland Publishing Company, Amsterdam.
22. Sjöström, M., S. Wold and B. Söderström (1986) PLS discriminant plots. In *Pattern Recognition in Practice II* (Edited by E. S. Gelsema and L. N. Kanal), pp. 461–470. Elsevier Science Publishers, B.V., Amsterdam.
23. Ramanujam, N., M. F. Mitchell, A. Mahadevan, S. Thomsen, A. Malpica, T. Wright, N. Atkinson and R. Richards-Kortum (1996) Development of a multivariate statistical algorithm to analyze human cervical tissue fluorescence spectra acquired in vivo. *Lasers Surg. Med.* **19**, 46–62.
24. Ramanujam, N., M. F. Mitchell, A. Mahadevan-Jansen, S. L. Thomsen, G. Staerkel, A. Malpica, T. Wright, N. Atkinson and R. Richards-Kortum (1996) Cervical precancer detection using a multivariate statistical algorithm based on laser-induced fluorescence spectra at multiple excitation wavelengths. *Photochem. Photobiol.* **64**, 720–735.
25. C. Y. Wang, C. T. Chen, C. P. Chiang, S. T. Young, S. N. Chow and H. K. Chiang (1998) Partial least-squares discriminant analysis on autofluorescence spectra of oral carcinogenesis. *Appl. Spectrosc.* **52** (9), 1190–1195.
26. Willard, H. H., L. L. Merritt, J. A. Dean and F. A. Settle (1981) Fluorescence and phosphorescence spectrophotometry. In *Instrumental Methods of Analysis*, 6th ed. Van Nostrand, New York.
27. Wold, S. (1978) Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* **20**, 397–405.
28. Albert, A. and E. K. Harris (1987) *Multivariate Interpretation of Clinical Laboratory Data*, Chapter 5. Marcel Dekker, New York.
29. Anderson, J. A. (1972) Separate sample logistic discrimination. *Biometrika* **59**, 19–35.
30. Ramanujam, N., M. F. Mitchell, A. Mahadevan, S. Warren, S. Thomsen, E. Silva and R. Richards-Kortum (1994) In vivo diagnosis of cervical intraepithelial neoplasia using 337-nm-excitation laser-induced fluorescence. *Proc. Natl. Acad. Sci. USA* **91**, 10193–10197.
31. Lakowicz, J. R. (1984) Protein fluorescence. In *Principles of Fluorescence Spectroscopy*. Plenum Press, New York.
32. Yuanlong, Y., Y. Yanming, L. Fuming, L. Yufen and M. Poozhong (1987) Characteristic autofluorescence for cancer diagnosis and its origin. *Lasers Surg. Med.* **7**, 528–532.
33. Marchesini, R., M. Brambilla, E. Pignoli, G. Bottiroli, A. C. Croce, M. D. Fante, P. Spinelli and S. D. Palma (1992) Light-induced fluorescence spectroscopy of adenomas, adenocarcinomas and non-neoplastic mucosa in human colon. *J. Photochem. Photobiol.* **14**, 219–230.
34. Bhandare, P., Y. Mendeison, R. A. Peura, G. Janatsch, J. D. Kruse-Jarres, R. Marbach and H. M. Heise (1993) Multivariate determination of glucose in whole blood using partial least-squares and artificial neural networks based on mid-infrared spectroscopy. *Appl. Spectrosc.* **47**, 1214–1221.
35. Borggaard, C. and H. H. Thodberg (1992) Optimal minimal neural interpretation of spectra. *Anal. Chem.* **64**, 545–551.