Journal of Cellular Biochemistry, Supplement 23:125-130 (1995)

Statistical Techniques for Diagnosing CIN Using Fluorescence Spectroscopy: SVD and CART

E. Neely Atkinson, PhD,[1] Michele Follen Mitchell, MD, MS,[2] Nirmala Ramanujam, MS,[3] and Rebecca Richards-Kortum, PhD[3]

[1] Department of Biomathematics, University of Texas M.D. Anderson Cancer Center, Houston, TX 77030
[2] Department of Gynecology, University of Texas M.D. Anderson Cancer Center, Houston, TX 77030
[3] The Biomedical Engineering Program, University of Texas, Austin, TX 78712

Abstract A quantitative measure of intraepithelial neoplasia which can be made *in vivo* without tissue removal would be clinically significant in chemoprevention studies. Our group is working to develop such a technique based on fluorescence spectroscopy. Using empirically based algorithms, we have demonstrated that fluorescence is discriminating normal cervix from low- and high-grade cervical dysplasias with similar performance to colposcopy in expert hands. These measurements can be made *in vivo*, in near real time, and results can be obtained without biopsy. This paper describes a new method using automated analysis of fluorescence emission spectra to classify cervical tissue into multiple diagnostic categories. First, data is reduced using the singular value decomposition (SVD), yielding a set of orthogonal basis vectors. Each patient's emission spectrum is then fit by linear least squares regression to the basis vectors, producing a set of coefficients for each patient. Based on these coefficient values, the classification and regression tree (CART) method predicts the patient's classification. These results suggest that laser-induced fluorescence can be used to automatically recognize and differentially diagnose cervical intraepithelial neoplasia (CIN) at colposcopy. This method of analysis is general in nature, and can analyze fluorescence spectra of suspected intraepithelial neoplasms from other organ sites. As a more complete understanding of the biochemical and morphologic basis of tissue spectroscopy is developed, it may also be possible to use fluorescence spectroscopy of the cervix as a surrogate endpoint biomarker in Phase I and II chemoprevention trials.

Key words: CIN, classification and regression trees, fluorescence spectroscopy, SIL

The development of invasive cervical neoplasia is believed to be preceded by a preinvasive stage, cervical intraepithelial neoplasia (CIN) [1]. Tertiary cancer prevention has focused on identifying and treating intraepithelial neoplasia in either the general population or in groups at high risk for developing carcinoma. Although screening, detection, and treatment programs have markedly reduced cervical carcinoma mortality, significant problems remain. Cervical carcinoma mortality rates are estimated to rise by 20% in the years 2000–2004 unless further improvements are made in current screening and diagnostic techniques [2]. Cytologic techniques used for initial CIN screening have a false-negative error rate of 20–30%. An abnormal Pap smear is followed by colposcopy, which has a limited predictive value even in experienced hands. Therefore, accurate diagnosis of CIN requires biopsy and histologic analysis. Improving the predictive value of CIN, particularly for less

Address correspondence to Rebecca Richards-Kortum, PhD, Biomedical Engineering Program, University of Texas at Austin, ENS 610, Austin, TX 78712.

experienced practitioners, could save patients from multiple biopsies and allow faster, more effective patient diagnosis and treatment, perhaps permitting diagnosis and treatment in a single visit. Optical diagnosis could be used to follow patients in chemoprevention trials without biopsy until the termination of the trial.

Toward this end, we are working to develop a quantitative measure of intraepithelial neoplasia which can be made *in vivo* without tissue removal. Our system is based on a technique known as fluorescence spectroscopy, in which tissue is illuminated with monochromatic light and the resulting fluorescence spectrum (the fluorescence intensity as a function of wavelength) is measured quantitatively. Because the cervix can be readily visualized, we have selected CIN as a model system to demonstrate the utility of diagnostic fluorescence spectroscopy.

A recent paper [3] reviewed our work in developing fluorescence spectroscopy for CIN diagnosis at colposcopy and discussed possibly using fluorescence spectroscopy as an SEB. In this paper, we present a method of developing classification algorithms which, based on its fluorescence spectrum, automatically identifies whether a cervical site is normal or contains a low- or high-grade dysplasia. Such algorithms are central to developing an optical system for real time automated diagnosis of cervical dysplasia at colposcopy.

DATA

Our spectroscopic system incorporates a pulsed nitrogen dye laser, an optical fiber probe and an optical multi-channel analyzer to record fluorescence spectra of the cervix *in vivo* [3]. Using this system, we collected fluorescence emission spectra from 276 tissue sites (each 2 mm in diameter) in a total of 93 patients. Emission spectra for excitation wavelengths of 337 nm, 380 nm, and 460 nm were measured for each sample. Emission intensities were measured from 360 nm to 650 nm in increments of 5 nm for the 337 nm excitation, from 400 nm to 680 nm for the 380 nm excitation, and from 480 nm to 680 nm for the 460 nm excitation. Each sample was graded as belonging to exactly one of the following categories: squamous normal, columnar normal, metaplasia, inflammation, low-grade squamous intraepithelial lesion (LGSIL), or high-grade SIL (HGSIL). Materials and methods used to obtain the spectra are described in detail elsewhere [3].

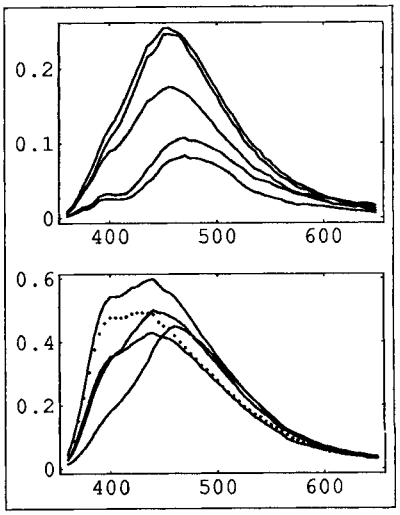

Fig. 1. Representative emission spectra from 337 nm excitation for two patients. In the top figure, the spectra are obtained from, in order of decreasing magnitude, LGSIL, squamous normal, squamous normal, HGSIL, and LGSIL samples. In the bottom figure, the four spectra represented by solid lines are from squamous normal samples, while the spectrum shown by the dotted line is from a columnar normal sample. These figures illustrate within-patient and between-patient variability of the data.

Figure 1 shows representative curves for two patients at the 337 nm excitation. In the top panel, the curves are, in descending order of magnitude, from LGSIL, squamous normal, squamous normal, HGSIL, and LGSIL samples. This decrease in emission magnitude with increasing pathologic severity is consistent throughout the data. It is clear from this figure that we cannot hope to distinguish squamous normal tissue from LGSIL, or to distinguish LGSIL from HGSIL at 337 nm; thus, it is necessary to examine the emission spectra at several excitation wavelengths. The lower panel shows curves from four squamous normal samples and one columnar normal sample; the columnar normal sample is indicated by the dotted line. This figure illustrates the level of variability which may be found even within a given patient.

A comparison of the y axis scales of the plots shows considerable patient-to-patient variation in the intensity of the curves. This interpatient vari-

ability is important for algorithm development; although magnitude decreases with increasing severity within each patient, different patients may well show overlap between magnitudes representing normal tissue in one patient and pathology in another. Thus, either each sample must be calibrated to the magnitude which represents normal for that patient, or characteristics other than overall magnitude of the emission spectra must be used in the analysis.

Previous studies have shown good success in classifying samples based on their emission spectra [3–6]; however, the techniques used in those studies assumed that approximately one-half of the samples obtained from any given patient will be squamous normal. This article describes two approaches which do not depend on such distribution. The first considers each sample individually, without reference to other samples obtained from the same patient. The second assumes that the technician performing the examination will obtain spectra from tissue which is clearly squamous normal; these spectra are then used to calibrate all other spectra from the same patient.

Figure 2 illustrates the data used for the first approach. The top panel shows the average spectrum (from the 337 nm excitation) for each tissue type. The topmost curve in the figure is the average for squamous normal and is well separated from curves for other tissue types. The next highest curve represents the LGSIL. The average curve for metaplastic samples is the bottommost curve. The average curves for columnar normal, inflammation, and HGSIL are not well separated at the excitation wavelength; indeed, the curves for columnar normal and HGSIL are virtually

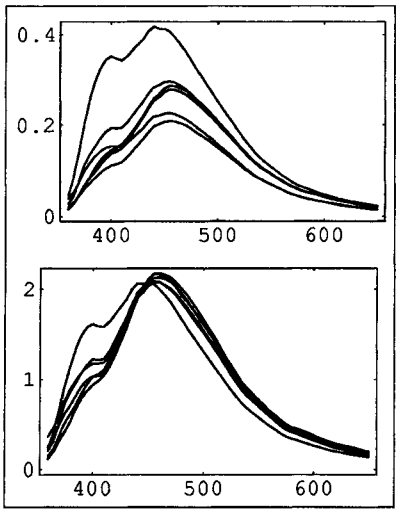

Fig. 2. Average emission spectra from excitation at 337 nm for the six tissue types. The upper panel is taken from the unpreprocessed data. The squamous normal curve is topmost and is well separated from the remaining types. The next highest curve represents LGSIL, while the bottommost curve represents metaplastic tissue. The remaining tissue types are not well separated at this excitation wavelength. In the bottom panel, each spectrum was normalized by the area of that spectrum before the averages were taken. The squamous normal curve remains well separated because peak emission occurs at a lower wavelength for squamous normal samples than for others.

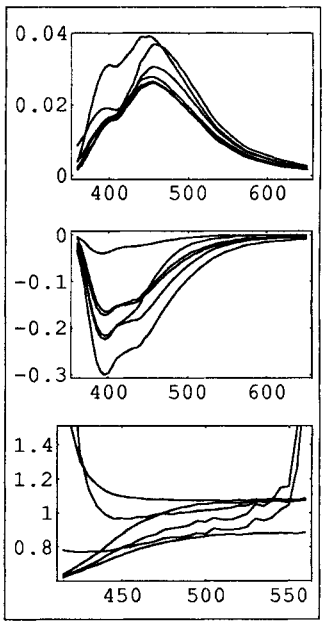

Fig. 3. Average emission spectra from excitation at 337 nm for the six tissue types with each curve calibrated by a sample known to be squamous normal. In the top panel, each curve was divided by the area of the reference sample. The topmost curve is squamous; next is HGSIL; the third is columnar normal; the remaining types are not well distinguished. In the middle panel, the reference curve is subtracted from each spectrum. The top curve is again squamous normal. The next two curves are metaplasia and inflammation. The next two are columnar normal and HGSIL. The lowest curve is LGSIL. In the bottom panel, each spectrum has been divided by the reference curve. Focusing on the 450 nm emission wavelength, the top curve is squamous normal. The next lower is HGSIL, followed by columnar normal, metaplasia and inflammation, with LGSIL at the bottom.

indistinguishable. In the lower panel, each spectrum was standardized by the area under the curve of that spectrum before the average was computed. This standardization removes between-patient variability in magnitude but also, unfortunately, removes information on peak intensity which may help discriminate between tissue types. The curve for squamous normal samples remains well separated because peak intensity occurs at a lower wavelength for these samples than for other tissue types, which indicates that the spectra are distinguished by their shapes as well as by their magnitudes.

Figure 3 illustrates the data used for the second approach, in which a spectrum from tissue known to be normal for a given patient is used to calibrate all other spectra for that patient. In this study, when more than one normal curve was available for a patient, the normal curve used for calibration was chosen at random. After one patient with no recorded normal curve was removed, 282 calibrated spectra were analyzed. The curves show the average calibrated spectra for each tissue type. In the top panel, each spectrum is divided by the area of the reference curve. In the middle panel, the reference curve is subtracted from each spectrum. In the bottom curve, each spectrum is divided by the reference curve; since this division is extremely unstable where the reference curve is close to 0, only the middle portion of each spectrum is used. These different adjustments appear to separate the curves in different ways. Only by actually using the calibrated data in a classification algorithm is it possible to determine which transformations are actually useful. The five data sets illustrated in Figures 2 and 3 provide the basic data used for classifying the samples.

DATA REDUCTION

For each sample, intensities are recorded for 59 emission wavelengths for 337 nm excitation, 57 wavelengths for 380 nm, and 41 wavelengths for 460 nm; a total of 157 intensities was recorded for each sample. To avoid the instability associated with overfitting a model, and to derive a model simple enough to be implemented in a clinical setting, these data should be reduced to a few summary measures for each patient. We accomplish this reduction by approximating each spectrum as the weighted sum of a small number of basic curves; that spectrum can then be represented by the coefficients used in the sum. The basic curves for each wavelength are derived from the (possibly preprocessed) emission data at that wavelength, using the singular value decomposition (SVD) [7]. If the emission spectra are used as the columns of a matrix, the SVD of that matrix will provide a set of orthogonal basis vectors which span the column space of the data matrix, ordered by how well those basis vectors can be used to fit, in the least squares sense, the original data. By using the first few basis vectors, we retain most of the information about the spectra but with many fewer data points.

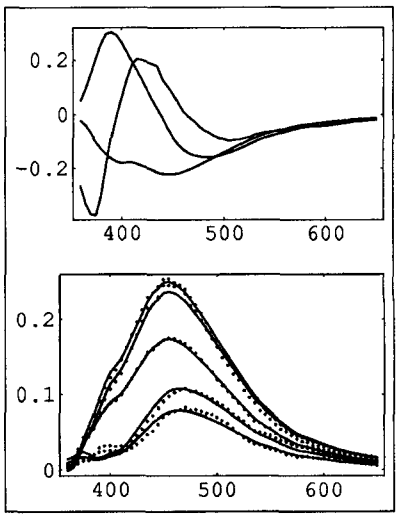

Fig. 4. Representing the spectra using orthonormal basis vectors. The top panel shows the three basis vectors from the unpreprocessed data at 337 nm with the largest singular values. The bottom panel shows the spectra for a representative case (in dotted lines) along with the least squares fits produced using the three basis vectors (in solid lines.)

Figure 4 illustrates this process. The upper panel shows the first three basis curves for the unpreprocessed data at 337 nm. The bottom panel shows the original spectra and the least squares fit using the three basis curves for a representative patient. Since the basis curves are orthonormal, the least squares fits are trivial to compute. The emission spectra seem to be fit reasonably well using three basis curves for each excitation wavelength. Thus, we represent each sample by three coefficients for each excitation wavelength, for a total of nine data points per sample.

TABLE I. Results of Applying the CART Method to Calibrated and Uncalibrated Data Sets

Data Set	Five Categories % correctly classified	Five κ	Three Categories % correctly classified	Three κ
Uncalibrated	72.3	0.59	83.0	0.62
Uncalibrated, areas normalized	74.2	0.62	84.6	0.63
Divided by area of reference normal	69.5	0.61	79.5	0.63
Reference normal subtracted	65.6	0.55	82.3	0.67
Divided by reference normal	69.1	0.60	80.1	0.64

*The κ statistic give a measure of agreement adjusted for the agreement which would be expected due to chance alone.

TABLE II. Results of Applying the CART Method to Uncalibrated Data With Normalized Areas

Actual Histology	Predicted Histology						
	Squamous normal	Columnar normal	Metaplasia	Inflammation	LGSIL[a]	HGSIL[b]	Total
Squamous normal	168	0	2	2	11	5	186
Columnar normal	7	14	0	0	4	4	29
Metaplasia	6	0	10	0	3	1	20
Inflammation	6	4	3	7	0	7	27
LGSIL	3	1	2	2	35	2	45
HGSIL	8	4	4	1	7	45	69
Total	**198**	**23**	**21**	**10**	**60**	**64**	**276**

[a] LGSIL = low-grade squamous intraepithelial lesion; [b] HGSIL = high-grade squamous intraepithelial lesion

CLASSIFICATION

The samples were classified into predicted histologies using a classification and regression tree (CART) algorithm from the S-Plus statistical package [8] which produces a series of binary branch points; at each branch, deciding which branch to take depends on the value of a single covariate. The same covariate may be used at more than one branch. At the end of a series of such branches, a terminal node is reached and a classification assigned to all cases which meet the criteria defining that node. CART is an extremely flexible modeling tool which makes few parametric assumptions and automatically accommodates variable interactions.

RESULTS

The data were classified using coefficients from each of the five data sets described above, two uncalibrated and three calibrated by a reference curve. Outcomes were classified in two different ways. In the first scheme, all six diagnostic categories were considered as separate; in the second, columnar normal, squamous normal, metaplasia, and inflammation were treated as

one category, while LGSIL and HGSIL were each treated as separate categories, for a total of three possible diagnoses. The results from the analyses are summarized in Table I. The κ statistic gives a measure of agreement between observed and predicted values adjusted for the agreement which would be expected due to chance alone [9]. The lowest misclassification rate was given by the data set which was not calibrated to a reference curve but in which the area of each curve was normalized; a summary of the classifications made using this data set are given in *Table II*.

DISCUSSION

The results given in Table II are not sufficiently accurate for use in a clinical setting. Nonetheless, we can draw some interesting conclusions and indicate directions for further development. First, the spectra can all be fit quite well using a small number of basis vectors; thus, the curves can be described in a low dimensional space. Second, calibrating the spectra to known normals did not improve the performance of the classification. This may indicate that the shape and the location of the curves, as well as the magnitude, carry a great deal of information. By closely considering which categories are well distinguished and which are not, it may be possible to extract specific features of the curves to discriminate certain diagnoses. Thus different transformations, different excitation wavelengths, or different portions of the emission spectra could be used to perform different discriminations.

CONCLUSIONS

SVD seems to be a practical and powerful technique for reducing the dimensionality of excitation-emission spectra. CART provides a simple and flexible method for classifying spectral samples based on the results of the SVD. A number of transformations may be applied to the data to adjust for between- and within-patient variability; different transformations have different effects at different emission wavelengths. Additional work is needed to determine which transformations and which wavelengths are most effective for identifying specific sample types. If appropriate transformations can be developed, fluorescence spectroscopy will prove useful as a surrogate endpoint biomarker in chemoprevention trials.

ACKNOWLEDGMENTS

This work was supported in part by Patient Technologies, Inc.

REFERENCES

1. Boone CW, Kelloff GJ, Steele VE: The natural history of intraepithelial neoplasia: Relevance to the search for intermediate endpoint biomarkers. J Cell Biochem 16G (Suppl):23–26, 1992.
2. Beral V, Booth M: Prediction of cervical cancer incidence and mortality in England and Wales (letter) Lancet i:495, 1986.
3. Richards-Kortum R, Mitchell MF, Ramanujam N, Mahadevan A, Thomsen T: *In vivo* fluorescence spectroscopy: Potential for non-invasive, automated diagnosis of cervical intraepithelial neoplasia and use as a surrogate endpoint biomarker. J Cell Biochem 19(Suppl):111–119, 1994.
4. Ramanujam N, Mahadevan A, Mitchell MF, Thomsen S, Silva E, Richards-Kortum R: Fluorescence spectroscopy of the cervix. Clin Consul Obstet Gynecol 6: 62–69, 1994.
5. Ramanujam N, Mitchell MF, Mahadevan A, Warren S, Thomsen S, Silva E, Richards-Kortum R: *In vivo* diagnosis of cervical intraepithelial neoplasia using 337 nm-excited laser-induced fluorescence. Proc Natl Acad Sci 91:10193–10197,1994.
6. Ramanujam N, Mitchell MF, Mahadevan A, Thomsen S, Silva E, Richards-Kortum R: Fluorescence spectroscopy: A diagnostic tool for cervical intraepithelial neoplasia (CIN). Gynecol Oncol 52:31–38, 1994.
7. Stewart, GW: "Introduction to Matrix Computations." In the series "Computer Science and Applied Mathematics." New York: Academic Press, 1973, pp 317–326.
8. Clark LA, Pregibon D: Tree-based models. In Chambers JM, Hastie TJ (eds): "Statistical Models in S." Pacific Grove: Wadsworth & Brooks/Cole Advanced Books and Software, 1983, pp 377–419.
9. Altman DG (ed): Some common problems in medical research. In "Practical Statistics for Medical Research." London: Chapman and Hall, 1991, pp 403–409.