PARAFAC for Spectral Curve Resolution: A Case Study Using Total Luminescence in Human Dental Tartar

MARLON M. REIS, DÉBORA N. BILOTI, MÁRCIA M. C. FERREIRA,* FRANCISCO B. T. PESSINE, and GUSTAVO M. TEIXEIRA

Instituto de Química, Universidade Estadual de Campinas–UNICAMP, 13083-970, Campinas, SP, Brasil (M.M.R., D.N.B., M.M.C.F., F.B.T.P.); and Universidade Camilo Castelo Branco, São Paulo, SP, Brasil (G.M.T.)

Chromophore identification in biological samples often requires the physical separation of the compounds, which can be difficult. Although there are several advantages to hyphenated spectroscopic techniques for identification of substances, complex mixtures of chromophores presenting overlapped spectra cannot be identified directly through this method. This work presents an application of chemometrics to compound identification in biological samples by a spectroscopic hyphenated technique using a curve resolution method. The PARAllel FACtor analysis model (PARAFAC), which has no rotational indeterminacy, was used for curve resolution of excitation-emission spectra of human dental tartars. PARAFAC was applied under constraints (i.e., unimodality and non-negativity) and evaluated with a validation procedure. The resolved profiles are porphyrinic-like spectra presenting excitation band maxima at 407, 416, and 431 nm in the Soret band region (390–440 nm) of these substances.

Index Headings: **Porphyrinic excitation-emission spectra; PARAFAC; Curve resolution; Tartar.**

INTRODUCTION

Identification of chromophores in biological systems requires, in several cases, the physical separation of the substances, which is often difficult or impossible to achieve experimentally. Although the development of spectroscopic techniques such as hyphenated methods has generated several advantages for the identification of compounds, the data sets produced by those systems are, in general, complicated to deal with because of the amount of overlapping numerical information that is produced. In such cases, the direct identification of chromophores (i.e., from only spectroscopic techniques) depends on their spectral similarity; in other words, if the chromophores have overlapped spectra, direct identification is extremely difficult. On the other hand, chemometrics has presented several methods for dealing with such problems (i.e., overlapped spectra). This work presents an application of chemometrics to the identification of compounds in biological samples by a spectroscopic hyphenated technique and a curve resolution method.

Previous work has shown that feline and canine dental tartars, a well-known source of periodontal diseases, show red fluorescence when irradiated with ultraviolet light due to the presence of porphyrinic compounds.[1] In a later work, the analysis of total luminescence spectra of one human sample showed that the same porphyrins seem to be present in human tartars,[2] which was confirmed by the present work when three new human samples were studied.

Received 13 November 2000; accepted 13 February 2001.
* Author to whom correspondence should be sent.

The data set generated by hyphenated fluorescence spectroscopy for each tartar sample is a two-way data type, where an excitation wavelength range is scanned, producing an emission spectrum for each excitation wavelength. Therefore, an emission intensity surface is produced, where one dimension is the excitation wavelength and the second is the emission wavelength. The singular value analysis of tartar spectra matrices showed the presence of at least three chromophores that have the excitation and emission bands in the same spectral range. In order to perform the identification of the excitation and the emission spectra of each species, curve resolution was performed by using PARAllel FACtor analysis (PARAFAC).[3] This method, developed for psychometrics, has proved to be a useful tool for curve resolution and quantification of fluorophores in biological systems,[4–6] especially for cases where the spectra of more than one fluorophore are overlapped, which makes direct identification and quantification almost impossible.

The nonideal behavior of the experiment makes the data deviate from the theoretical model (i.e., low-wavelength component interference, scattering, and noisy data), bringing difficulties to the curve resolution. For the nonideal data, PARAFAC was therefore chosen for the curve resolution since it permits one to take advantage of prior information used in the form of constraints[6] such as non-negativity.

The final results show a good PARAFAC performance with a stable solution verified with a validation step. Finally, one resolved profile was due to the low-wavelength component interference, and three resolved profiles were attributed to emission and excitation profiles of porphyrinic species, since the excitation spectra appear in the Soret band region (390–440 nm), which is characteristic of the electronic transition of porphyrins.[7]

EXPERIMENTAL

Three human dental tartar samples were dissolved in hydrochloric acid 1:1 (v/v). The emission spectra were collected in the range from 460 to 750 nm, with 1 nm increments, on an SLM-AMINCO spectrofluorimeter (SPF-500C), with a Xe lamp (250 W) as the radiation source. These spectra were monitored in a range of excitation wavelengths (390–450 nm, with increments of 2 nm), producing a two-dimensional array for each sample, where each row is an emission spectrum and each column an excitation spectrum. The experiment was performed at room temperature (i.e., $\approx$ 25 °C).

The calculations were done with the MATLAB (MathWorks) version for MS-Windows running on an

0003-7028/01/5507-0847$2.00/0

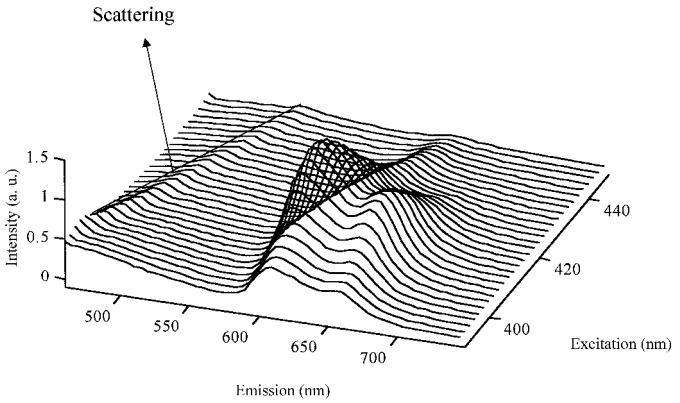


FIG. 1. Emission intensity surface for sample 1 (a.u., arbitrary unit).

IBM-compatible PC. The PARAFAC algorithm (see the Appendix) was obtained from a web site.[8]

Data. The performed experiment results in a data set whose emission intensity, μ, for each fluorophore at concentration, c_k, in a specific wavelength, λ_j^{em}, when excited at a wavelength, λ_i^{ex}, is described by a trilinear model:[4]

$$\mu_{ijk} = \epsilon_i \pi_j c_k \tag{1}$$

where ϵ_i is the extinction coefficient of the fluorophore at excitation wavelength λ_i^{ex}, π_j is the relative emission at detection wavelength λ_j^{em}, and c_k is the concentration of the fluorophore. If F fluorophores contribute to the intensity, the emission intensity (μ) can be written as

$$\mu_{ijk} = \sum_{1}^{F} \epsilon_{if} \pi_{jf} c_{kf} \tag{2}$$

Simple application of Eq. 2 requires small specimen absorbance, or diluted samples, and the excitation should not be transferable between chromofores.[4]

Figure 1 shows the emission spectra of sample 1. The lower wavelengths of the emission range present a band whose maximum intensity changes with the excitation wavelength due to Raman scattering.[9] The same kind of scattering is also observed in the other samples according to Fig. 2. Considering that the excitation band is the one used to identify the fluorophores, the emission range used in the curve resolution was kept between 580 and 749 nm to reduce the Raman scattering influence. Figure 2 shows the emission spectra of the three samples in the entire region (**a**) and in the region (**b**) used for the analysis.

It should be noted that the spectra of sample 3 in the spectral range used for the curve resolution show the pronounced influence of what is probably a low-wavelength component that is not as important in the other samples' spectra, making the problem rather difficult to solve. With this experimental behavior under consideration, curve resolution of the excitation and emission spectra was performed by employing a trilinear model.

Methods. ***PARAFAC Model.*** The decomposition model for a trilinear data used by PARAFAC is

$$x_{ijk} = \sum_{1}^{F} a_{if} b_{jf} c_{kf} \tag{3}$$

where x_{ijk} is the (i, j, k) original element in the trilinear data set, and a_{if}, b_{jf}, and c_{kf} are the loadings elements for

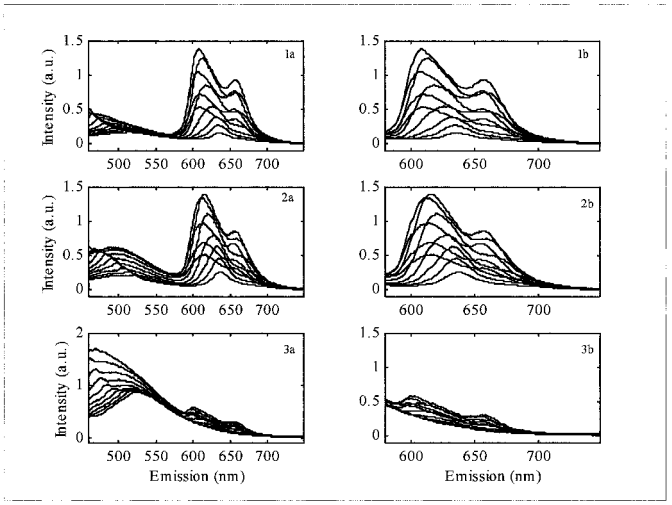


FIG. 2. Emission intensity spectra for different excitation wavelengths: 1a, 2a, and 3a correspond to entire spectral region where the spectra were recorded; 1b, 2b, and 3b show the spectral region used for the curve resolution of samples 1, 2, and 3, respectively (a.u., arbitrary unit).

the three dimensions of the data set. The loadings matrices **A**, **B**, and **C**, for which the F columns correspond to the elements a_{if}, b_{jf}, and c_{kf}, respectively, are found through an alternating least-squares (ALS) algorithm where the loss function (l)

$$l = \left\| x_{ijk} - \sum_{1}^{F} a_{if} b_{jf} c_{kf} \right\|^2 \tag{4}$$

is minimized.

PARAFAC and the Trilinear Fluorescence Data Set. The fluorescence data can be modeled by a three-way PARAFAC model, where the loadings **A** ($n \times F$), **B** ($m \times F$), and **C** ($r \times F$) correspond to an extinction coefficient of n excitation wavelengths for F fluorophores, relative emission at m detection wavelengths for F fluorophores, and concentrations of r samples of F fluorophores, respectively. For the tartar data set, the matrix **A** corresponds to the excitation profiles, **B** to the emission profiles, and **C** to the relative concentrations of the F fluorophores (a relative concentration is found since the scales of the extinction coefficient and relative emission are unknown). The full tartar data set is a (30 excitation wavelengths $\times$ 171 emission wavelengths $\times$ 3 samples) three-way array.

Constraints. The nonideal experimental behavior (e.g., low-wavelength component interference, noisy data) hinders the optimization step (i.e., local minima) in PARAFAC model fitting. In these cases, constraints

TABLE I. $P(F)$ values for PARAFAC models fitted with different numbers of factors (F).

F	$P(F)$[a]
1	2.0869
2	0.7962
3	0.1399
4	0.0222
5	0.0095
6	0.0045

[a] $P(F)$: see Eq. 5 in the text.

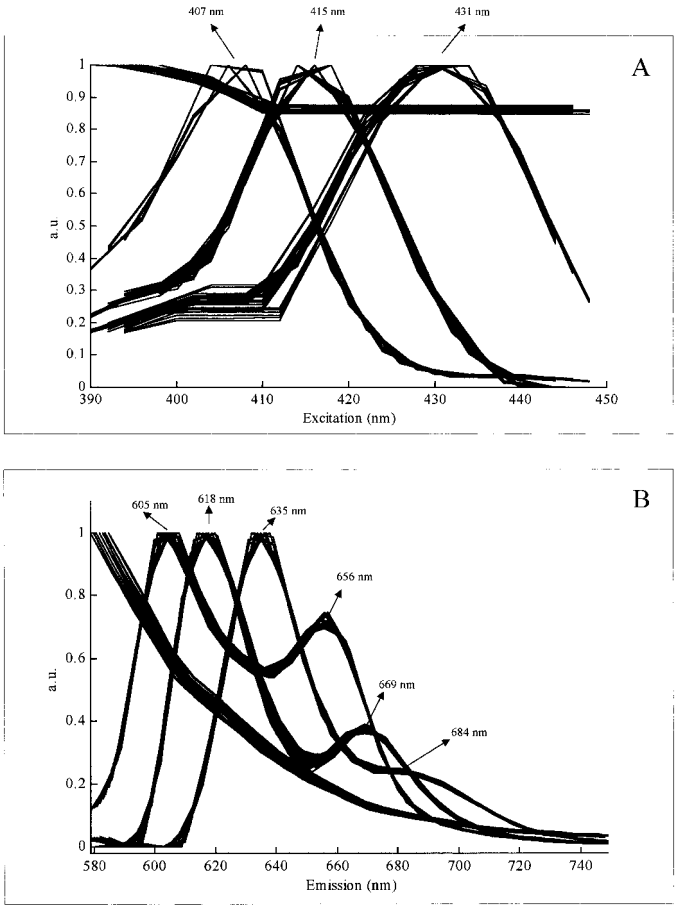


FIG. 3. Resolved spectral profiles by PARAFAC model for (**A**) excitation mode; (**B**) emission mode. The spectra corresponds to 35 validation arrays (a.u., arbitrary unit).

based on prior physical information (e.g., non-negativity of spectra) are used in the optimization of the function l (see Eq. 4), providing a stable solution for the trilinear decomposition. Previous tests with unconstrained models resulted in excitation and emission profiles with some small negative values. These negative values do not affect the chromophore identification since they do not change the profiles' shape and peak position, but both non-negativity and unimodality constraints were used to ensure that the final results have a physical meaning. In this work, the PARAFAC algorithm was initialized with random values and the non-negativity constraint applied to the three modes (excitation and emission wavelengths and concentration). The resulting profiles of this model were used as starts for the final model, where the unimodality constraint was applied in the excitation wavelengths (i.e., it is assumed that only one band is present in the excitation range for each fluorophore) and the non-negativity constraint for the other modes. It is important to note that the unimodality constraint was chosen on the basis of the results obtained with the non-negativity constrained model, since it showed only one band in the excitation range.

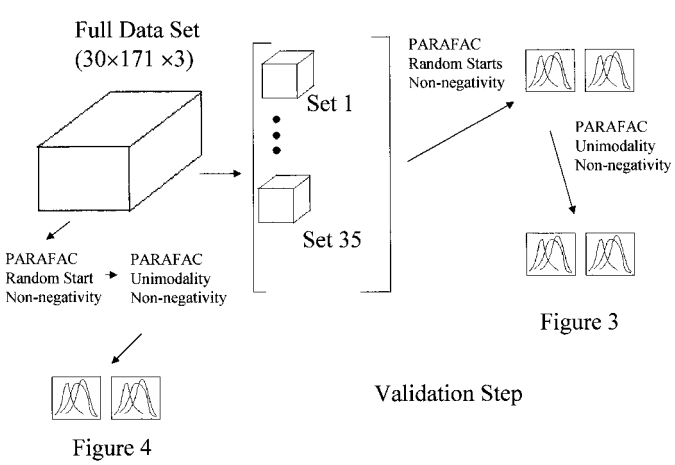
Validation. Validation is a fundamental step in PARAFAC modeling to identify local minima in the optimization of the loss function. The PARAFAC model was validated by using a resampling procedure. In order to do that, each original matrix having the dimension 30 × 171 (i.e, 30 rows and 171 columns) was divided into 35 matrices having one of the following dimensions: (9 × 24), (9 × 25), (10 × 24), or (10 × 25), which depends on the validation set. The first matrix was generated from the full matrix by taking the rows 1, 4, 7, . . . 28 (leaving out the rows 2, 3, 5, 6, . . . 29, 30) and the columns 1, 8, 15, . . . 169 (leaving out the columns 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, . . . 170, 171), resulting in a matrix with dimension (10 × 25). The second matrix was generated from the full matrix by taking the rows 1, 4, 7, . . . 28 (leaving out the rows 2, 3, 5, 6, . . . 29, 30) and the columns 2, 9, 16, . . . 170 (leaving out the columns 1, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, . . . 171). This analysis is carried out in such way that the last set corresponds to the rows 5, 8, 11, . . . 29 (leaving out the rows 1, 2, 3, 4, 6, . . . 30) and taking the columns 7, 14, 21, 28, 35, . . . 154, 161, 168 (leaving out the columns 1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, . . . 169, 170, 171) having dimension (9 × 24). In this way, for each set of rows, seven sets of columns were built. Thus, the original matrix was divided into five sets of rows, which was divided into 35 matrices. This procedure generates 35 three-way arrays that are used to fit 35 PARAFAC models. The results of these 35 PARAFAC models are used to evaluate the goodness of the fit—in other words, to verify whether the resolved profiles represent the same kind of information for the 35 subsets.

Number of Fluorophores. The number of fluorophores was chosen by comparing the $P(F)$ values, as described by Eq. 5, fitted for six models with the number of factors (i.e., F in Eq. 5) varying from one up to six. By the end, six $P(F)$ values were calculated and the variation analyzed. To confirm the number of fluorophores, we tested three models: (1) three fluorophores, (2) four fluorophores, and (3) five fluorophores, for each of the 35 models described in the validation section.

$$P(F) = \sum_{w=1}^{35} \left\| {}^{(w)}x_{ijk} - \sum_{1}^{F} {}^{(w)}a_{if}\,{}^{(w)}b_{jf}\,{}^{(w)}c_{kf} \right\|^2 \qquad (5)$$

where (w) indicates the data set from the 35 different three-way arrays described in the validation section. The outline of the analysis is summarized in Scheme I.

SCHEME 1



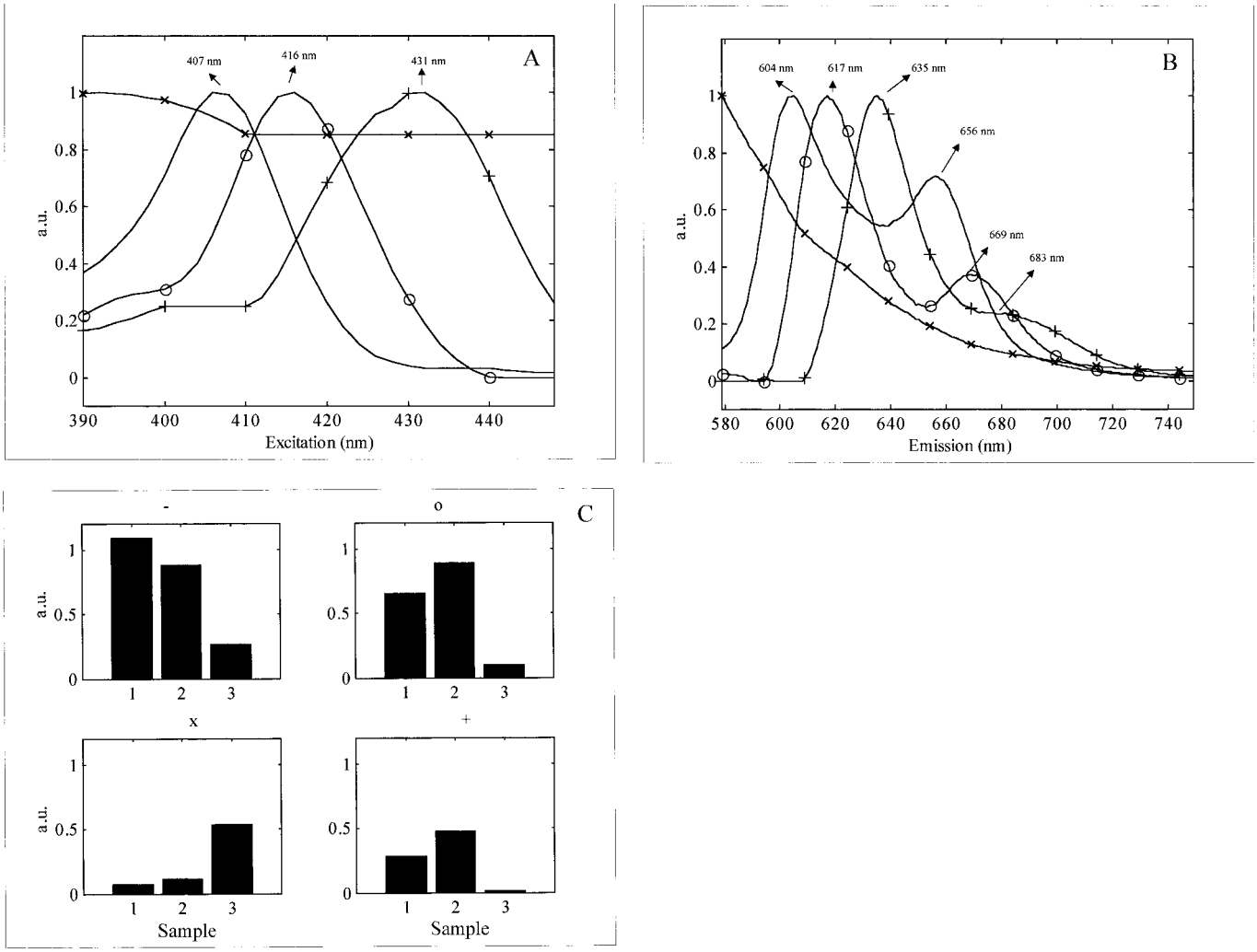


FIG. 4. Resolved spectral profiles by PARAFAC model applied to the full array. (**A**) Excitation; (**B**) emission; (**C**) concentration modes. The symbols used to distinguish profiles in **A** and **B** are the same used for **C**, so that the plot on top right of **C** shows the relative concentration of the porphyrinic species which has the spectral profile marked with the circle symbol in **A** and **B** (a.u., arbitrary unit).

RESULTS

The number of fluorophores was identified, first, by the evaluation of $P(F)$ values (see Eq. 5) of models fitted by using from one to six factors ($F = 1, 2, \ldots 6$), as shown in Table I. The variation of these values indicates that the number of fluorophores should be three, four, or five, since for more than five factors this variation became very small. In others words, using six or more factors does not decrease the $P(F)$ values significantly. The analysis of $P(F)$ values cannot show the best number of factors for the PARAFAC model but helps to reduce the number of choices. Thus, three PARAFAC models were fitted with three, four, and five factors for each of the 35 different three-way arrays described in the validation section. The best result was obtained for the four-fluorophores model, where the resolved profiles for the 35 arrays (see Scheme I) are in agreement (Fig. 3).

Results for the three-fluorophores model presented profiles with a wide band, suggesting that more of them could be resolved. The five-fluorophores model resulted in different emission profiles for the same excitation profile when different arrays, found in the validation step, were used. In this case, two or more profiles are considered "equal" if their shapes are similar and their maxima positions appear at the "same" wavelength (i.e., the position of two maxima must differ by at least 4 nm to be considered different, since the resolution in the excitation mode is 2 nm). Although the measurements are different among the 35 arrays of the validation step, the maxima position of the profiles must appear at the "same" wavelengths and the profiles must have similar shape. Thus the four-factors model is the best one.

The final model was fitted by using the full array (30 $\times$ 171 $\times$ 3) with four factors. The results are shown in Fig. 4. One of the four profiles shown in Fig. 4A and Fig. 4B (identified with the symbol "$\times$") does not look like a porphyrinic profile and is regarded as an interference from the band in the emission range 460–579 nm (see Fig. 2A). The loadings presented in Fig. 4C represent the relative amount of fluorophores in the samples, since the true concentration is not available, thus the fluorophores' molar absorptivities are unknown, and the correct scale for these concentrations therefore cannot be found. These concentrations give the information about the relative composition among the three samples.

Figure 5 presents the emission spectrum of haemato-

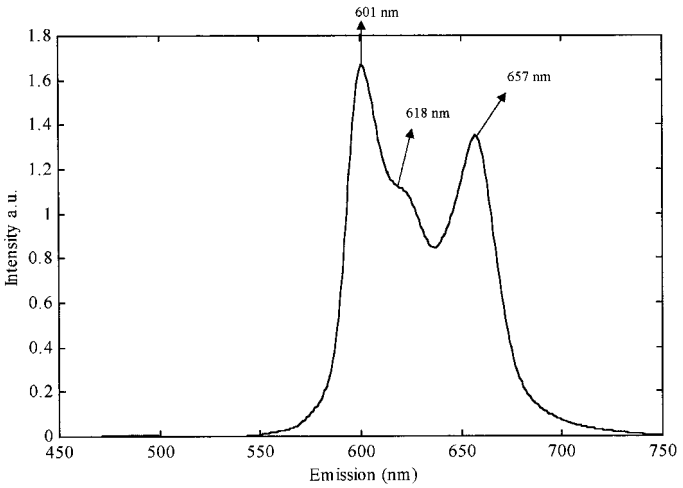


FIG. 5. Emission spectrum of haematoporphyrin excited at 417 nm (a.u., arbitrary unit).

porphyrin (hematoporphyrin dihydrochloride-sigma) excited at 417 nm. By comparing this spectrum with those found by PARAFAC it is possible to confirm that the resolved profiles do look like porphyrinic spectra and that the band in the emission range 460–579 nm (*compare* Figs. 2A and 5) is due to the presence of unknown interferent(s) present in the tartar samples.

Table II presents the maxima position for the excitation spectra obtained in this and in the previous works, suggesting that the same porphyrinic species are present in the human, feline, and canine tartar samples.

CONCLUSION

The identification of the compounds of tartar samples should involve physical separation techniques, since their fluorescence spectra are overlapped. Alternatively, PARAFAC allows such identification by fluorescence spectroscopy. The complete identification of the porphyrinic species requires a further step, since the biological samples can be formed by mixtures of very similar porphyrins which are difficult to resolve by the spectral resolution method used in this experiment. These similar porphyrins should have the same porphyrin framework, differing only slightly in the peripheral side groups (e.g., a peripheral side group of acetic acid vs. propionic acid), as pointed out by Ferreira et al.[1] In this way, the primary goal of this work is to present three excitation and emission profiles that resulted from curve resolution by the validated PARAFAC model of a nonideal data set.

TABLE II. Soret bands for the porphyrinic species resolved and previous results from the literature.

References	Excitation wavelengths (nm)		
Ferreira et al.[1]	410	417	436
Reis and Ferreira[2]	410	417	436
Present work	407	416	431

ACKNOWLEDGMENT

The authors acknowledge the financial support from FAPESP for carrying out this work.

1. M. M. C. Ferreira, M. L. Brandes, I. M. C. Ferreira, K. S. Booksh, W. C. Dolorwy, M. Gauterman, and B. R. Kowalski, Appl. Spectrosc. **49,** 1317 (1995).
2. M. M. Reis and M. M. C. Ferreira, Química Nova **22,** 11 (1999).
3. R. A. Harshman and M. E. Lundy, Comp. Stat. Data Anal. **18,** 39 (1994).
4. R. T. Ross and S. Leurgans, Methods Enzymol. **246,** 679 (1995).
5. R. Bro and H. Heimdal, Chemom. Intell. Lab. Syst. **34,** 85 (1996).
6. R. Bro, Chemom. Intell. Lab. Syst. **38,** 149 (1997).
7. J. E. Falk, *Porphyrins and Metalloporphyrins* (Elsevier, Amsterdam, 1964), Vol. 2.
8. http://www.models.kvl.dk/users/rasmus/ (Version 1.03, October 1998).
9. J.C. Andre, M. Bouchy, and M. L. Viriot, Anal. Chim. Acta **105,** 297 (1979).

APPENDIX: PARAFAC ALS ALGORITHM

Initialize $\mathbf{B}$ and $\mathbf{C}$

Step 1.

$$\mathbf{Z} = (\mathbf{C} \,|\otimes|\, \mathbf{B})$$
$$\mathbf{A} = {}_{(n,mr)}\mathbf{X}\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}$$

Step 2.

$$\mathbf{Z} = (\mathbf{C} \,|\otimes|\, \mathbf{A})$$
$$\mathbf{B} = {}_{(m,nr)}\mathbf{X}\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}$$

Step 3.

$$\mathbf{Z} = (\mathbf{B} \,|\otimes|\, \mathbf{A})$$
$$\mathbf{C} = {}_{(r,nm)}\mathbf{X}\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}$$

If the relative change in the l value (see Eq. 4 in the text) between two iterations is sufficiently small, then stop; otherwise go to step 1, where ${}_{(n,mr)}\mathbf{X}$ denotes the three-way data array unfolded in an $(n \times mr)$ matrix:

$$(\mathbf{B} \,|\otimes|\, \mathbf{A}) = [\mathrm{vec}(\mathbf{a}_1\mathbf{b}_1^{\mathrm{T}}) \quad \mathrm{vec}(\mathbf{a}_2\mathbf{b}_2^{\mathrm{T}}) \cdots \mathrm{vec}(\mathbf{a}_F\mathbf{b}_F^{\mathrm{T}})];$$

$$\mathrm{vec}(\mathbf{a}_1\mathbf{b}_1^{\mathrm{T}}) = \begin{pmatrix} a_{11}b_{11} \\ a_{21}b_{11} \\ \vdots \\ a_{n1}b_{11} \\ a_{11}b_{21} \\ \vdots \\ a_{n1}b_{m1} \end{pmatrix}$$