A RLS MODEL EXTRACTED BY ICA

Daniela Coltuc[1], Mihaela Ilie[2], Andreea Merlan[1]

[1]Faculty of Electronics, Telecommunications and Information Technology,
University "Politehnica" of Bucharest, Romania University Politehnica of Bucharest
Bd. Iuliu Maniu nr. 1-3, Bucharest, Romania
phone : (40) 21 40 24 83 73, email : coltucda@yahoo.com
andreea_mln, coltucda@yahoo.com
[2] Faculty of Pharmacy, "Carol Davila" University of Medicine and Pharmacy
Dionisie Lupu nr.37, Bucharest, Romania
phone: (40) 21 3111152, *m16ilie@yahoo.com*

Abstract: This paper presents some preliminary results regarding the ability of ICA in analyzing spectrometric data. A set of RLS spectra, measured for different concentration DNA solutions, in the presence of the probe TbDTPA, is decomposed like in a typical application of source separation and processed in order to get a model for RLS process in this case. The resulted curves fit well the model obtained by manually measuring and averaging the peaks of the RLS spectra.

Keywords: ICA, RLS spectra, DNA

1. INTRODUCTION

In the last two decades, chemometrics gained a well defined place among the methods used to assay materials, mainly in cosmetics, pharmaceutical and food industry. The chemometric methods replace the traditional procedures of measuring and interpreting the experimental data by an automatic analysis, better adapted to the increasing quantity and diversity of data provided by the modern instrumental methods.

A lot of definitions have been given to the word *chemometrics* (Workman, 2002). A common one describes chemometrics as an application of statistics to chemistry, aiming to extract more reliable information from raw instrumental data. Indeed, the statistical approach is another important advantage of chemometrics, since it is known that any measurement is intrinsically random. Last, but not least, is the ability of chemometric methods to analyze multivariate data, i.e. data issued from various physical phenomena or measured by different instruments.

In chemistry, spectra or chromatograms are sometimes inexpensive solutions for a first diagnosis, useful for further investigations. The spectra are high dimensional data, collecting hundreds or thousands

This paper was recommended for publication by Laurențiu Frangu

of spectral components. Consequently, their processing is very tedious if automatic procedures are not used. For this reason, at present, the spectrometry coupled with chemometrics, gets an increasing popularity, especially in pharmaceutical and food industry (Blăgoi *et al.*, 1998; Ilie *et al.*, 1998; Kolomiets and Siesler, 2005).

A frequently encountered chemometric aproach is *pattern recognition*. Usually, by means of pattern recognition techniques, new experimental data are classified in classes with well known properties. In this paper, the pattern recognition is used for deriving a heuristic model for the Rayleigh Light Scattering (RLS) process in the case of deoxyribonucleic acid (DNA) in aqueous solutions. Our approach is based on the Independent Component Analysis (ICA), a transform of the same type as Principal Component Analysis (PCA), largely used in chemometrics. The difference between ICA and PCA consists in the optimization criterion used for analysis. Thus, in deriving its basis, PCA minimizes the correlation, whilst ICA maximizes the statistical independence. Due to this property, ICA may be more useful than PCA in analysing data issued from complex phenomena, like those involved in RLS spectrometry.

We tested our method on RLS spectra, obtained for aqueous double stranded calf thymus DNA coupled with a small probe molecule - the terbium chelate of the diethylentriaminopentaacetic (Tb-DTPA). The purpose of our experiments was to test the abilities of ICA for chemometric applications and, in particular, to define a procedure for the automatic measurement of the DNA concentration. In the following sections, we present some preliminary results that demonstrate that ICA is able to extract significant information for this kind of application.

2. THE INDEPENDENT COMPONENT ANALYSIS

ICA is a recently developed method that proved to be quite useful in many applications. ICA is based on the "latent variables" statistical model, according to which the particular realization x_j of a random variable x are linear combinations of the latent variables $s_1,...,s_n$ (Hyvärinen *et al.*, 2001):

$$x_j = a_{j,1}s_1 + \ldots + a_{j,n}s_n, \quad j = 1 \ldots m \qquad (1)$$

where $a_{j,i}$ are real coefficients. By definition, the random variables s_i are statistically independent, hence the name of "independent components". They are also called "sources of x". The components s_i are latent because they cannot be observed directly. The mixing coefficients $a_{j,i}$ are also unknown. Consequently, based on the observed particular realizations x_j, one should estimate not only the s_i components, but also the coefficients $a_{j,i}$. Nonetheless, the estimation must be done under as general as possible assumptions.

In many applications, x is random signal, depending on a parameter of time, space etc. In our case, this parameter is the wavelength. Without any loss of generality, we shall neglect, in the followings, the parameter dependency.

The equations in (1) may be written as (Hyvärinen *et al.*, 2001):

$$x = A \cdot s \qquad (2)$$

where A is a matrix containing the $a_{j,i}$ coefficients – called for this reason, mixing matrix – and x and s are two column vectors containing the particular realizations of x and the independent components s:

$$x = [x_1 \quad x_2 \quad \ldots \quad x_n]^T, \; s = [s_1 \quad s_2 \quad \ldots \quad s_n]^T \qquad (3)$$

Based on x, an ICA algorithm estimates both the matrix A and the vector s. The starting point is the assumption that s_i are statistically independent to each other. Once matrix A estimated, its pseudoinverse W is computed and the estimates of the independent components are obtained as follows:

$$y = W \cdot x \qquad (4)$$

The key of the estimation process is the components' nongaussianity (a consequence of the Central Limit Theorem says that the sum of two independent identical distristributed random variables is closer to a Gaussian than any distribution of the originating variables). Most of ICA algorithms estimate the independent components by optimizing a measure of nongaussianity (Hyvärinen *et al.*, 2001).

There are various methods to measure the nongaussianity of a random variable y. One of them is the negentropy, defined as (Hyvärinen *et al.*, 2001):

$$J(y) = H(y_{gauss}) - H(y) \qquad (5)$$

where y_{gauss} is a random gaussian variable with the same variance as y and H is the entropy. The algorithm FastICA, which has been used for our experiments, extracts the independent components by maximizing the negentropy (http://www.cis.hut.fi/projects/ica/fastica/). The definition in (5) has the drawback of a difficult computation. For this reason, simpler approximations of negentropy are a better choice. FastICA uses the following approximation, developed by Hyvarinen (Hyvärinen *et al.*, 2001):

$$J(y) \approx \left[E\{G(y)\} - E\{G(\upsilon)\}\right]^2 \qquad (6)$$

where υ is a gaussian random variable of zero mean and variance equal to unity and G is a nonquadratic function.

The search area of the FastICA is restrained by signal whitening, a preprocessing stage encountered in many algorithms. By whitening, the number of the computed parameters is reduced to half.

The extracted independent components s_i can be estimated until to a multiplicative constant and sign. Their order cannot be determined as well. FastICA is a projection pursuit method, allowing to evaluate also those situations when there are less components s_i then particular realizations x_j.

The data dimensionality reduction by ICA may be obtained in various ways. A first option would be to extract a smaller number of independent components, by taking the risk of loosing significant information. A second option would be to extract the maximum allowed number of independent components and to drop the insignificant ones, by defining a criterion. A third option is the synthetic description of sources by means, for instance, of negentropy. This way, a source, represented by hundreds of samples, may be reduced to single numerical value. In our approach, we have used the second method.

3. EXPERIMENTAL RESULTS

The RLS spectra used for tests were obtained from different concentration aqueous dsDNA solutions, in the presence of a constant concentration of Tb-DTPA. The dsDNA was purchased from Merck and the Tb-DTPA was obtained by synthesis, at "Horia Hulubei" National Institute for Nuclear Engineering. In fact, the dsDNA molecule is a nucleoprotein extracted and purified from calf thymus, which contains besides dsDNA, other biological molecules as hystones, polypeptides, separate polynucleotides, etc.

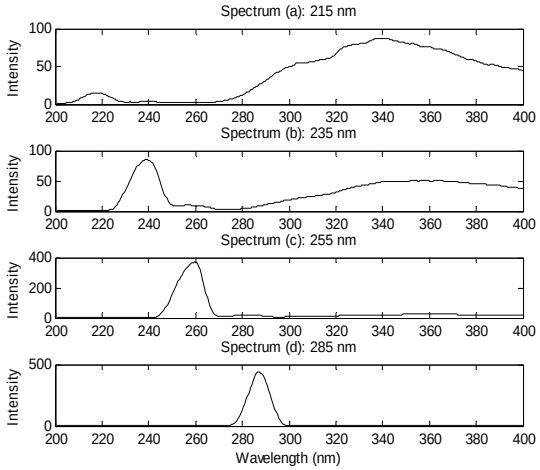

Fig.1. Experimental RLS spectra at 4 different wavelengths.

This substance is a real "cocktail" of absorbing and fluorescent emitting molecules that are difficult to be separated by ordinary methods.

For the experiments, we used dsDNA solutions of five different concentrations: 4.5μg/mL, 7.2μg/mL, 9μg/mL, 10.8μg/mL and 13.5μg/mL. They were obtained in double distilled water from the same dsDNA stock solution of 100μg/mL. A constant concentration of 0.1mM Tb-DTPA was added to each solution. The Tb-DTPA was added to get a more accentuated increase of the signal with the increase in the dsDNA concentration (Ilie *et al.*, 2005b).

By using a Perkin Elmer LS 50B spectrofluorimeter, working at four different excitation wavelengths, seven replicates of the RLS spectrum of the dsDNA-Tb-DTPA solutions were acquired. Thus, a total of 35 spectra were collected, each spectrum consisting of 400 spectral components at a resolution of 0.5 nm. The excitation wavelengths were chosen following the terbium (215 nm), dsDNA (255 nm) and protein (285 nm) excitation wavelengths; the 235 nm was considered a "neutral" excitation wavelength (Ilie *et al.*, 2005a).

Fig. 1 presents the RLS spectra obtained for the 4.5μg/mL dsDNA solution at all the four excitation wavelengths. Spectrum (a), corresponding to the excitation wavelength of 215 nm, presents a maximum at 217 nm, consisting in the RLS signal of the substances in solution (dsDNA, Tb-DTPA and impurities), a second maximum at 237 nm, representing the water Raman band, and a large band - ranging approximately between 270nm and 400nm – that is a superposition of various emission bands. Spectrum (b) presents a first maximum at 237 nm (the RLS signal), a second one at 255 nm (the Raman

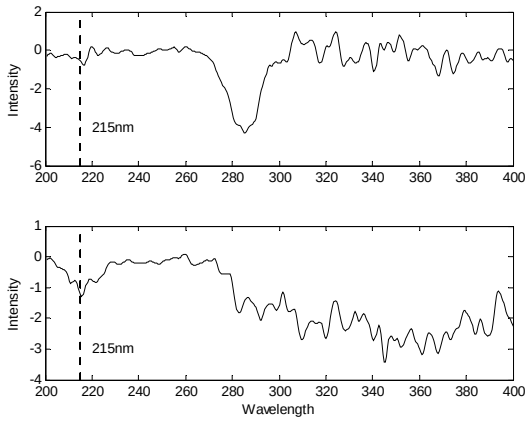

Fig.2. Two common independent components at 215nm excitation.

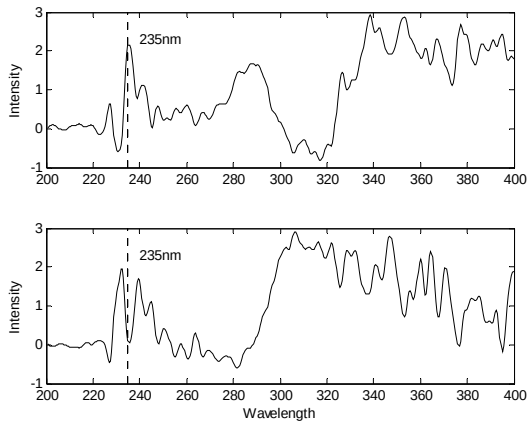

Fig. 3. Two common independent components at 235 nm excitation.

band of the water) and a rather large emission band belonging to dsDNA and impurities. Spectrum (c) has the RLS signal at about 258 nm and the Raman band at 284 nm, and, finally, spectrum (d) has the RLS band at 286 nm and the water Raman peak at 317 nm. With the increase of the dsDNA concentration, the molecular aggregate dsDNA-TB-DTPA becomes greater, resulting in a higher RLS signal (Ilie *et al.*, 2005a, 2005b); the Raman band does not depend upon the dsDNA concentration and the superposed emission bands are difficult to be separated and interpreted in terms of nucleoprotein concentration.

In order to put into evidence the DNA contribution, we have analyzed the RLS spectra as in a typical application of source separation. Since we had not a

theoretical model for the dsDNA-Tb-DTPA spectral behaviour, we were constrained to perform a blind extraction of spectra sources. In a subsequent stage, we tried to identify among them one or more sources that are significant for DNA concentration.

The processing was done in two steps: for a fixed excitation wavelength, the spectra of each concentration were analyzed separately and, afterwards, by using the correlation coefficient, the sources appearing at all concentrations were identified. As DNA is a common constituent for our solutions, one expects to find its contribution at all concentrations and, consequently, among the common sources. We have considered two sources as being similar, in the case of a coefficient of correlation greater than 0.7 (the extracted independent components are only estimates of the spectra sources).

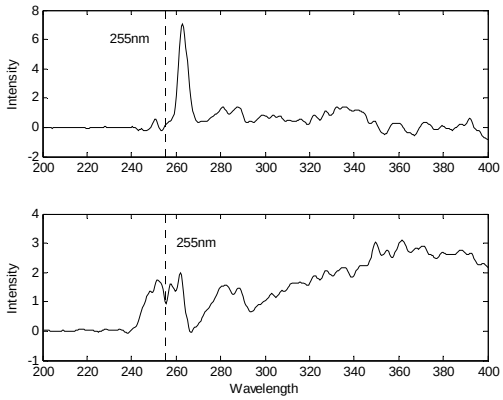

Fig.4. Two common independent components at 255nm excitation.

For each wavelength and for each concentration, seven replicates of the RLS spectrum were available. They have been collected in a matrix X , each row of X containing a replicate. A maximum number of independent components were extracted by using the FastICA algorithm, i.e. seven for each concentration. We asked to FastICA to use a diagonal matrix as initial A .

At 215 nm excitation, two sources were present at all concentrations. They are shown in Fig. 2. Their amplitude and polarity are not significant since ICA cannot determine the energy and the sign of the extracted components. Only their shape is important. The source on the first position consists in a peak around 285 nm, whilst the other one seems to be a residue composed of signals that could not be separated.

Figure 3 shows the common sources found by analyzing the spectra at 235 nm excitation. The first source exhibits two peaks, one at the excitation wavelenght and the other one around 280 nm.

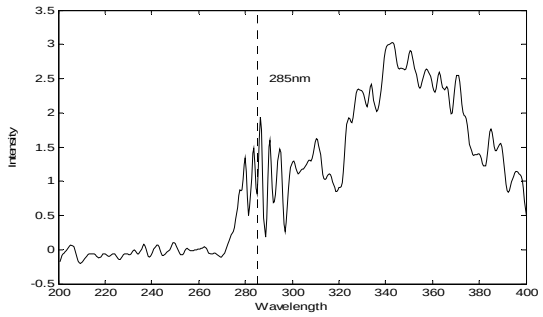

Fig. 5. A common component at 285 nm excitation.

Again, the second source seems to be a residue.Two common sources were found also for the spectra recorded at 255 nm excitation (Fig. 4). The first source has an isolated peak, slightly displaced by respect to the excitation wavelength. The second source exhibits wide peaks around 255 and 280 nm and, probably, mixed contributions in the upper band.

At 285 nm excitation, a single source has been present at all concentrations. It is shown in Fig. 5.

Each set of seven components, extracted by analyzing the RLS spectra at fixed excitation wavelength and DNA concentration, constitutes a basis. It is, of course, an adapted basis. For a fixed excitation wavelength, one may use one of the five bases (one for each concentration), in order to represent the RLS spectra obtained at the same excitation wavelength. The spectra projections on the coordinate representing a common source might be relevant for the DNA contribution to the RLS spectrum.

In order to get a model for the DNA contribution at each excitation wavelength, we have projected the corresponding RLS spectra on a common source and we have computed the mean of the projections obtained at the same concentration. Although the concentration is theoretically the same, random variations appear always from one experiment to another. By getting the mean, the estimation error due to these variations is reduced. The obtained five values are represented, versus the DNA concentration, in a plot and, by linearly interpolating the values, a polygonal approximation of a curve curve was drawn.

From the seven possibilities (Fig. 2, 3 and 4), two provided coherent results: the first common source at 255 nm and the single common source at 285 nm. In the first case, the approximated curve obtained by linear interpolation is shown in Fig. 6. Its shape fits the model obtained by manually measuring and averaging the peaks of the RLS spectra at 255 nm (a procedure that is, at present, in use) (Ilie, 2007).

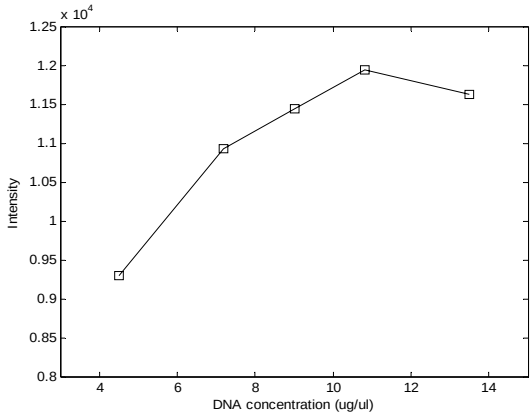

Fig. 6. Approximated curve at 255 nm.

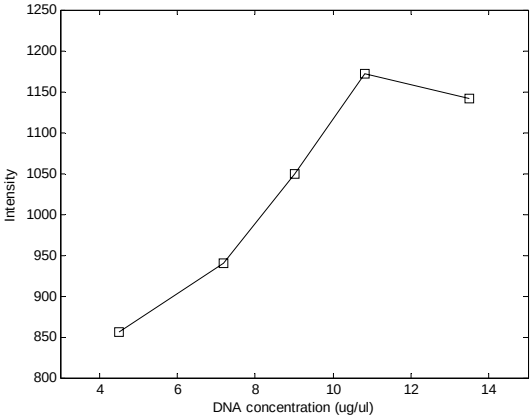

Fig. 7. Approximated curve at 285 nm.

The second approximated curve (Fig. 7), obtained at 285 nm excitation, has the same type of variation, but a different convexity. This suggests either another DNA type of response, or the response of another component, most probably, the protein.

4. CONCLUSIONS

The conclusion of our experiments is that ICA is able to extract, from RLS spectra, significant information for the DNA concentration in aqueous solution. Indeed, the spectra projections on the common components of ICA basis, has put in evidence a dependency that is similar to those obtained by manual measurements. Based on these results, we intend to develop, in the future, automatic procedures for processing data from RLS spectrometry.

REFERENCES

Blăgoi, G., Bleotu A., Puică M., Vasilescu M.and Ilie M. (1998). NIR Investigation of some Lipid Extracts in Order to Ascertain Their Quality. *J. NIR Spectroscopy*, 6 (1-4), pp. A285-A290.

Hyvärinen A., Karhunen J., Oja E. (2001). *Independent Component Analysis*. John Wiley & Sons

Ilie, M., Ioniță-Mânzatu M., Vasilescu M., Puică M.and Blăgoi G. (1998). Comparison of Different Modalities of Outlier Treatment for Qualitative NIR Spectra. *J. NIR Spectroscopy*, 6(1-4), pp. A175-A179.

Ilie, M., Fugaru V., Baconi D., Bălălău D. and Boscencu R. (2005a). Fluorescence Resonance Energy Transfer and Light Scattering Study of Irreversible DNA Conformational Changes in the Presence of a Terbium Chelate. *Revista de Chimie*, 56 (4), pp. 355-358.

Ilie, M., Colțuc D., Bălălău D., Boscecncu R. and Baconi D. (2005b). Analysis of fluorescence – scattering spectra for certain nucleoprotein – terbium chelate complexes: a chemometric approach. *Revista de Chimie*, 56 (12), pp. 1226 – 1230.

Ilie, M. (2007). Studiul chemometric al spectrelor de împrăştiere ale ADN-ului din nucleoproteine – Ed. Tehnoplast Company SRL, Bucureşti

Kolomiets, O., H. Siesler, W. (2005). The influence of spectral resolution on the quantitative near infrared spectroscopic determination of an active ingredient in a solid drug formulation. *J. NIR Spectroscopy*, 12 (5), pp. 271-278.

Workman, J. Jr. (2002). The state of multivariate thinking for scientists in industry. *1980–2000, Chem. Intell. Lab. Syst.*, 60, pp.13–23.

The FastICA package for Matlab, http://www.cis.hut.fi/projects/ica/fastica/