DIFFERENCES BETWEEN THE DOUBLE AND SINGLE STRANDED DNA THROUGH THE RAYLEIGH LIGHT DISPERSION PHENOMENON♦

DIFFERENCES ENTRE L'ADN DOUBLE ET MONOCATENAIRE VUES A TRAVERS LE PHENOMENE RAYLEIGH DE DISPERSION DE LA LUMIERE

Mihaela Ilie[1]*, Daniela Coltuc[2]

[1]Université de Médicine et Pharmacie «Carol Davila», Faculté de Pharmacie, Traian Vuia No. 6, Bucarest, Roumanie
[2]Université "Politehnica" de Bucarest, Faculté d'Electronique, Télécommunications et de la Technologie d'Information, Roumanie

*Corresponding author: m16ilie@yahoo.com

Received: 22/05/2008
Accepted after revision: 16/06/2008

Abstract: The paper presents the results of a chemometric analysis with the aim of highlighting the differences between the double- and single stranded DNA by using Rayleigh Light Scattering (RLS) spectra.
The spectra were obtained by exciting the aqueous solutions of calf thymus DNA in the presence of a molecular probe (the Terbium chelate of the diethylenetriaminopentaacetic acid – Tb-DTPA). Each spectrum consisted

♦ ♦ Paper presented at the fifth edition of: "Colloque Franco-Roumain de Chimie Appliquée – COFrRoCA 2008", 25 – 29 June 2008, Bacău, Romania.

in 400 points representing the measured intensities in the range 200 – 400 nm, with a 0.5 nm step.
The paper presents two kinds of chemometric analysis: Principal Component Analysis (PCA) and Independent Component Analysis (ICA). PCA is well known for its ability in confining the information in a reduced number of uncorrelated coefficients, whilst ACI – lesser used in chemometrics – refines the analysis by imposing the condition of statistically independent components. Both PCA and ICA were used to reduce the dimension of each spectrum to two components. The two-dimensional representation of these components puts into evidence a clusterisation of the spectra following the excitation wavelength and, for the same excitation wavelength, a separation tendency between spectra belonging to a different kind of DNA (single-stranded and double-stranded).

Keywords: *pattern recognition, Principal Component Analysis, Independent Component Analysis, Rayleigh Light Scattering, DNA.*

INTRODUCTION

La chémométrie peut être décrite comme une application de la statistique en chimie, réalisée avec le but d'améliorer le processus de mesure et d'extraire l'information la plus utile et complète des données brutes, fournies par les mesures instrumentales physico-chimiques. En chémométrie, l'approche classique, qui fait référence à une mesure unique, est remplacée par l'analyse multi variée des données obtenues à la suite du processus de mesure [1, 2].
Une application fréquente de la chémométrie est la reconnaissance des formes. La reconnaissance des formes peut être non supervisée (ayant comme but la classification) ou supervisée (le but étant d'attribuer un échantillon méconnu à une classe, autrement dit, de reconnaître l'échantillon). Actuellement, une popularité croissante est enregistrée par l'analyse spectrale de l'infrarouge proche (NIR - Near InfraRed) par des techniques chémométriques de reconnaissance des formes, pour des applications dans les industries pharmaceutique et alimentaire [3 - 7].
Les travaux, présentés dans cet article, ont comme but de mettre en évidence le comportement différent de certaines solutions d'acide desoxiribonucléique (ADN) vis à vis du phénomène Rayleigh de dispersion de la lumière (RLS – Rayleigh Scattering Spectrum).

MATERIAUX ET METHODES

Description du formalisme mathématique utilisé

Pour l'analyse de spectres RLS, nous avons utilisé deux types de transformées: l'Analyse en Composantes Principales (ACP) et l'Analyse en Composantes

Indépendantes (ACI). Dans les deux cas, le rôle de la transformée a été de réduire la dimension des spectres, en gardant l'information essentielle pour les reconnaître.
L'ACP transforme une série de variables aléatoire corrélées dans une série non corrélée de même longueur. La décorrélation est accompagnée par la concentration d'information dans un petit nombre de variables, ce qui donne la possibilité de réduire la dimension de la série par élimination des variables pas significatives.
La transformation consiste à projeter la série $X^{(i)}$ sur les vecteurs propres Φ_k de sa matrice de covariance C:

$$C = \left[c_{i,j} = \mathrm{cov}\left(X^{(i)}X^{(j)}\right)\right] \text{ où } \mathrm{cov}\left(X^{(i)}, X^{(j)}\right) = \overline{\left(X^{(i)} - \overline{X}^{(i)}\right)\left(X^{(j)} - \overline{X}^{(j)}\right)} \ \forall i, j \qquad (1)$$

En gardant le fait que les projections sur les vecteurs propres correspondant au plus élevées valeurs propres de C, on obtient les composantes principales de la série, autrement dit, un nombre réduit de variables décorrelées, qui concentre la plus part de l'information de la série initiale.
Dans notre application, l'ACP a été utilisée pour réduire la dimension de chaque spectre de 400 à 2 échantillons, en faisant la projection des spectres sur les deux vecteurs propres correspondant aux deux premières valeurs propres.
L'ACI est basée sur le modèle statistique des variables latentes, qui consiste à représenter les réalisations particulières x_j d'une variable aléatoire X, comme combinaisons linéaires de plusieurs variables aléatoires latentes s_i [8]:

$$x_j = \sum_i a_{i,j} s_i \qquad (2)$$

où $a_{i,j}$ sont des coefficients réels. Par définition, les variables s_i sont indépendantes, d'où le nom de composantes indépendantes. Comme les s_i ne sont pas observables directement, on les appelle aussi latentes. En partant d'un ensemble de réalisations particulières, l'ACI estime les variables s_i ainsi, que les coefficients $a_{i,j}$. Les estimations sont obtenues en optimisant une mesure de l'indépendance statistique entre les variables latentes. Dans notre expérimentation, nous avons utilisé FastICA [9] qui maximise la néguentropie des variables latentes [8]. A l'aide de FastICA, nous avons extrait de l'ensemble des spectres RLS, deux composantes indépendantes, qui ont servi comme base pour représenter les spectres dans un espace de dimension réduite à deux.
La réduction de la dimension des spectres à deux nous a permis de les représenter par des points dans un diagramme 2D. Afin de caractériser le regroupement des points dans des nuages, nous avons utilisé trois mesures: l'inertie totale d'un nuage et l'inertie intraclasse et l'inertie interclasse de l'ensemble des nuages.
Supposons que G est le centre de gravité d'un nuage de points $\Gamma = \{M_i, i = 1,...,n\}$. L'inertie totale du nuage, qui est une mesure de la dispersion des points M_i, est définie par:

$$\mathrm{I}(\Gamma) = \frac{1}{n}\left[d_2(M_1, G)^2 + ... + d_2(M_2, G)^2\right] \qquad (3)$$

où d_2 est la distance euclidienne. L'inertie intraclasse de l'ensemble des nuages $\Gamma_1 ... \Gamma_k$ est définie par la somme des inerties totales des nuages:

$$\mathrm{I}_{\mathrm{int}\,ra} = \mathrm{I}(\Gamma_1) + ... + \mathrm{I}(\Gamma_k) \qquad (4)$$

Le degré de séparation des nuages est mesuré par l'inertie interclasse :

$$I_{inter} = p_1 d_2(G_1, G) + ... + p_k d_2(G_k, G) \quad (5)$$

où G est le centre de gravité de l'ensemble des points, G_i sont les centres de gravité des nuages et p_i les poids des nuages.

Description de l'expérimentation

Les substances utilisées sont: de l'ADN double caténaire (dcADN) de thymus de veau, procuré chez la société Merck, de l'ADN monocaténaire (mcaDN) de thymus de veau, procuré chez la société Sigma et le complexe de terbium avec l'acide diéthilène triamino penta acétique (Tb-DTPA), obtenu par synthèse à l'Institut National de Recherche et Développement de Physique Nucléaire "Horia Hulubei" Bucarest. On a préparé des solutions de 4,5 μg/mL, 7,2 μg/mL, 9 μg/mL, 10,8 μg/mL et 13,5 μg/mL dans de l'eau distillée, pour tous les deux types de nucléoprotéine et l'on a ajouté, indépendamment de la concentration, une quantité de 0,1 mM Tb-DTPA.
Pour chaque solution, on a enregistré quatre spectres de diffusion Rayleigh de la lumière (RLS), correspondant à quatre longueurs d'onde d'excitation: 215 nm (spécifique au Terbium), 235 nm (quelconque), 255 nm (spécifique à l'ADN) et 285 nm (spécifique aux protéines qui accompagnent l'ADN). Les spectres ont été obtenus à l'aide d'un spectrofluorimètre Perkin Elmer modèle LS 50 B.
Chaque spectre consiste en 400 valeurs numériques représentant des intensités mesurées dans la bande 200 – 400 nm, avec un écart de 0,5 nm. L'expérimentation a été répétée neuf fois pour le dcADN et six fois pour le mcADN. On a obtenu, au total, 180 spectres de dcADN (9 expérimentations × 5 concentrations × 4 longueurs d'onde) et 117 spectres de mcADN (6 expérimentations × 5 concentrations × 4 longueurs d'onde, dont on a éliminé 3 spectres aberrants).
L'analyse chémomètrique a été effectué sous Matlab, avec des algorithmes décrits antérieurement [11 – 13]. En utilisant l'ACP et, alternativement, l'ICA, on a effectué une analyse de type reconnaissance non supervisée afin de classifier les spectres RLS selon la longueur d'onde d'excitation et, ensuite, pour 255 nm et 285 nm, selon le type de nucléoprotéine.

RESULTATS ET DISCUSSIONS

Dans une première étape, nous avons appliqué l'ACP à l'ensemble des spectres RLS (touts les deux types d'ADN, toutes les longueurs d'onde et toutes les concentrations ont été prises en compte). L'inspection des valeurs propres a montré que les deux premières, les plus élevées, représentent environ 80% du total des valeurs propres. En conséquence, nous avons restreint la représentation d'un spectre dans l'espace transformé, aux deux premières composantes principales, CP1 et CP2. Par la suite, les spectres RLS ont été représentés par des points de coordonnées CP1 et CP2 dans un diagramme 2D (Figure 1). On peut observer un regroupement selon les quatre longueurs d'onde d'excitation, un résultat attendu en tenant compte des différences observées au niveau de la forme des spectres lorsque la longueur d'onde d'excitation change [11].
A l'intérieur de chaque nuage, les spectres de dcADN et mcADN ne sont pas séparés d'une manière nette.

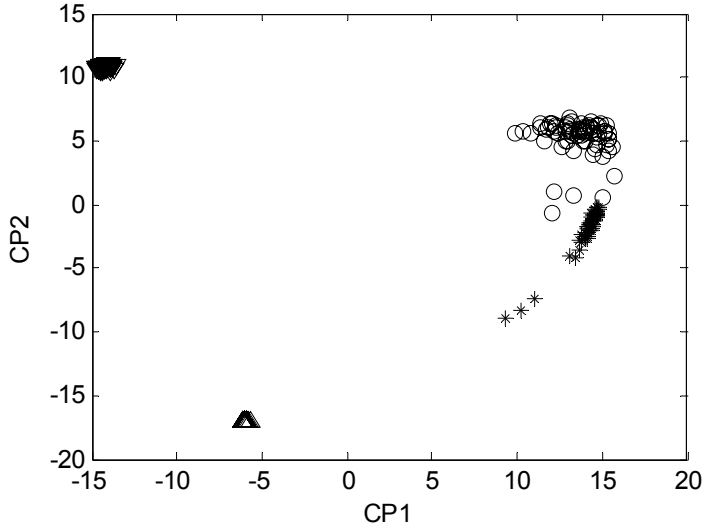
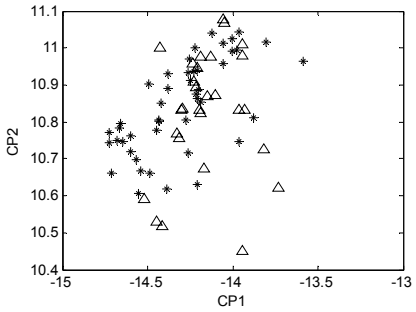
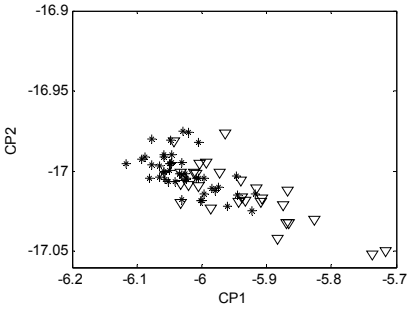


Figure 1. *Regroupement des spectres RLS selon la longueur d'onde (ACP)*
'' longueur d'onde d'excitation de 215 nm*
'○' longueur d'onde d'excitation de 235 nm
'Δ' longueur d'onde d'excitation de 255 nm
'∇' longueur d'onde d'excitation de 285 nm

Le calcul de l'inertie intraclasse pour ces deux classes (Tableau 1) montre une dispersion plus élevée des points pour 215 nm (spécifique à Terbium) et 235 nm (neutre) que dans le cas de 255 nm (spécifique à l'ADN) et 285 nm (spécifique aux protéines).
L'inertie interclasse montre que la meilleure séparation est obtenue pour 285 nm (pour cette longueur d'onde, l'inertie interclasse est comparable avec les inerties totales des classes). En effet, l'agrandissement des nuages montre une tendance de séparation entre les spectres de dcADN et mcADN pour des longueurs d'onde d'excitation de 255 nm et 285 nm (Figure 2).





Nuage à 255 nm
'*' dcADN
'Δ' mcADN

Nuage à 285 nm
'*' dcADN
'∇' mcADN

Figure 2*. Agrandissement des nuages des spectres obtenus à 255 nm et 285 nm*

Tableau 1. *Inertie des nuages*

ACP				
	215 nm	235 nm	255 nm	285 nm
Inertie totale dcADN	4,852	4,906	0,087	0,002
Inertie totale mcADN	0,742	1,310	0,063	0,007
Inertie intraclasse	5,595	6,216	0,151	0,009
Inertie interclasse	0,068	0,172	0,005	0,002
ACI				
Inertie totale dcADN	28,129	70,903	1,733	0,031
Inertie totale mcADN	14,364	32,524	0,941	0,057
Inertie intraclasse	42,493	103,427	2,674	0,089
Inertie interlasse	2,260	2,603	0,101	0,006

Dans une deuxième étape, en utilisant fastICA, nous avons extrait de l'ensemble de spectres RLS deux composantes indépendantes qui ont servi de base dans un espace transformé, de dimension réduite à deux. En projetant chaque spectre sur ces deux composantes, nous avons obtenu deux valeurs numériques, qui sont les coordonnées du point représentant le spectre dans le diagramme 2D de la Figure 3. On peut observer, comme dans le cas de l'ACP, un regroupement des spectres selon la longueur d'onde d'excitation. Les nuages correspondants aux 255 nm et 285 nm sont mieux regroupés (les valeurs de l'inertie intraclasse du Tableau 1 le confirment).

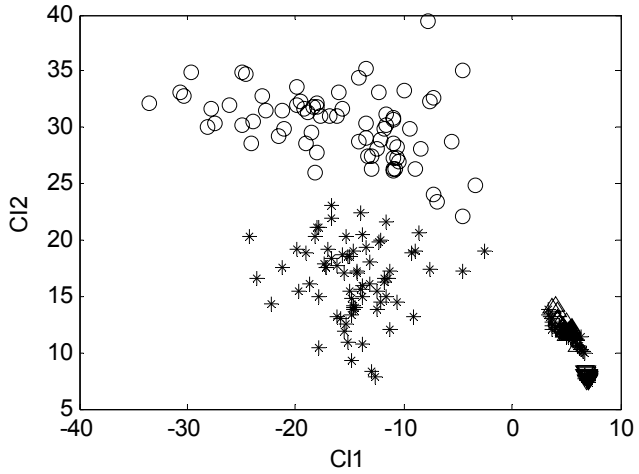
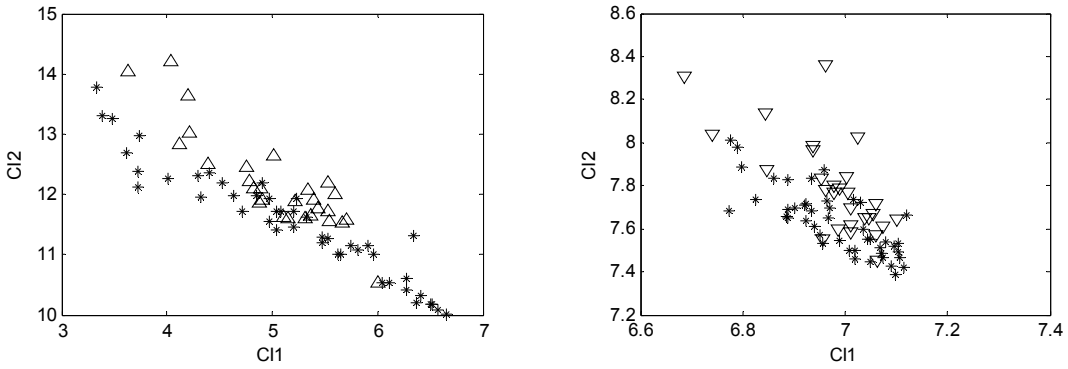


Figure 3. *Regroupement des spectres RLS selon la longueur d'onde (ACI)*
'' longueur d'onde d'excitation de 215 nm*
'○' longueur d'onde d'excitation de 235 nm
'Δ' longueur d'onde d'excitation de 255 nm
'∇' longueur d'onde d'excitation de 285 nm



Nuage à 255 nm
'*' dcADN
'Δ' mcADN

Nuage à 285 nm
'*' dcADN
'∇' mcADN

Figure 4. *Agrandissement des nuages des spectres obtenus à 255 nm et 285 nm*

L'inertie interclasse, comparée à celle intraclasse, montre une meilleure séparation à 255 nm et 285 nm (comme pour l'ACP), mais des performances plus faibles que dans le cas de l'ACP. La tendance de séparation est quand même gardée, comme on peut l'observer dans la Figure 4, ou les nuages ont été agrandis.

CONCLUSIONS

Nous avons effectué une analyse chémométrique des spectres RLS des solutions aqueuses de mcADN et dcADN, avec des concentrations variables entre 4 et 15 μg/mL, en présence de Tb-DTPA (concentration constante de 0,1mM). Dans le cas de l'ACP, ainsi que dans celui de l'ACI, l'analyse de type reconnaissance de formes non supervisée a mis en évidence le regroupement des spectres selon la longueur d'onde d'excitation. Pour 255 nm et 285 nm, même si les nuages ne sont pas nettement délimités, on observe une tendance de séparation entre le mcADN et dcADN. Le calcul des inerties intra et interclasse montre que, dans le cas de l'ACP, la séparation est meilleure. Cette expérimentation démontre qu'une démarche basée sur la reconnaissance des formes peut mener à des résultats intéressants dans l'analyse automatique des spectres RLS.

BIBLIOGRAPHIE

1. Brereton, R.G.: *Chemometrics: applications of mathematics and statistics to laboratory systems,* E. Horwood, New York, **1990**;
2. Workman, J. Jr.: *Chem. Intell. Lab. Syst.,* **2002**, **60**, 13–23;
3. Mânzatu, I., Ioniță-Mânzatu, V., Ioniță-Mânzatu, M., Vasilescu, M., Nuşescu, G., Puică, M., Ilie, M.: *Roumanian Biotechnological Letters,* **1997**, **2** (3), 193-200;
4. Mânzatu, I., Ioniță-Mânzatu, V., Dumitraşcu, M., Ilie, M., Nuşescu, G., Vasilescu, M., Puică, M.: *Roumanian Biotechnological Letters,* **1997**, **2** (3), 217-225;

5. Ilie, M., Ioniță-Mânzatu, M., Vasilescu, M., Puică, M., Blăgoi, G.: *J. NIR Spectroscopy*, **1998**, **6** (1-4), A175-A179;
6. Blăgoi, G., Bleotu, A., Puică, M., Vasilescu, M., Ilie, M.: *J. NIR Spectroscopy,* **1998**, **6** (1-4), A285-A290;
7. Kolomiets, O., Siesler, H.W.: *J. NIR Spectroscopy*, **2005**, **12** (5), 271-278;
8. Hyvärinen, A., Karhunen, J., Oja, E.: *Independent Component Analysis,* John Wiley & Sons, **2001**;
9. The FastICA package for Matlab, http://www.cis.hut.fi/projects/ica/fastica/
10. Université de Nice, Département de Mathématiques, Cours de mathématiques appliqués à la biologie, http://www-math.unice.fr/~diener/SV1-04/COURS11.pdf
11. Ilie, M., Fugaru, V., Baconi, D., Bălălău, D., Boscencu, R.: *Revista de Chimie (Bucureşti),* **2005**, **56** (4), 355-358;
12. Ilie, M., Colțuc, D., Bălălău, D., Boscencu, R., Baconi, D.: *Revista de Chimie (Bucureşti)*, **2005**, **56** (12), 1226 – 1230;
13. Coltuc, D., Merlan, A., Ilie, M.: *Analele Universitatii Dunarea de Jos din Galați*, **2007**, **Fascicle III**, ISSN 1221-454X, 53-58.