UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE CIÊNCIAS EXATAS E DA TERRA
INSTITUTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

Programa de Pós-Graduação em Química

Identificação e rastreamento de câncer através da combinação de análise multivariada e técnicas bioespectroscópicas

Ana Carolina de Oliveira Neves Menezes

Tese de Doutorado

Natal/RN, agosto de 2017

ANA CAROLINA DE OLIVEIRA NEVES MENEZES

IDENTIFICAÇÃO E RASTREAMENTO DE CÂNCER ATRAVÉS DA COMBINAÇÃO DE ANÁLISE MULTIVARIADA E TÉCNICAS BIOESPECTROSCÓPICAS

Tese apresentada ao Programa de Pós-Graduação em Química da Universidade Federal do Rio Grande do Norte (PPGQ/UFRN), como parte dos requisitos necessários para obtenção do título de doutora em Química.

Orientador: Prof. Dr. Kássio Michell Gomes de Lima

NATAL – RN

2017

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI
Catalogação de Publicação na Fonte. UFRN - Biblioteca Central Zila Mamede

Menezes, Ana Carolina de Oliveira Neves.
Identificação e rastreamento de câncer através da combinação de análise multivariada e técnicas bioespectroscópicas / Ana Carolina de Oliveira Neves Menezes. - 2017.
112f.: il.

Universidade Federal do Rio Grande do Norte, Centro de Ciências Exatas e da terra, Programa de Pós-graduação em Química. Natal, RN, 2017.
Orientador: Kássio Michell Gomes de Lima.

1. Câncer - Tese. 2. Fluorescência molecular - Tese. 3. ATR-FTIR - Tese. 4. Espectrometria de massas - Tese. 5. Análise multivariada - Tese. I. Lima, Kássio Michell Gomes de. II. Título.

RN/UF/ CDU 54:616-006

Ana Carolina de Oliveira Neves Menezes

IDENTIFICAÇÃO E RASTREAMENTO DE CÂNCER ATRAVÉS DA COMBINAÇÃO DE ANÁLISE MULTIVARIADA E TÉCNICAS BIOESPECTROSCÓPICAS

Tese apresentada ao Programa de Pós-graduação em Química da Universidade Federal do Rio Grande do Norte, em cumprimento às exigências para obtenção do título de Doutora em Química.

Aprovada em: 09 de agosto de 2017.

Comissão Examinadora:

Dr. Kássio Michell Gomes de Lima – UFRN (orientador)

Dra. Lívia Nunes Cavalcanti – UFRN

Dra. Aurigena Antunes de Araújo – UFRN

Dr. Mario Cesar Ugulino de Araújo – UFPB

Dr. Ronei Jesus Poppi – UNICAMP

À minha família – meu marido, meus pais, meu irmão. Vocês são o amor que existe em mim.

A todos que de alguma forma lutam na guerra contra o câncer.

AGRADECIMENTOS

Neste momento em que escrevo esses agradecimentos, me passa um filme na cabeça. Foram quatro anos e, nesse período, muita coisa aconteceu. Algumas boas, muito boas, outras ruins, decepções, frustrações. Nessas horas, eu sempre tive com quem contar, um apoio, um incentivo. São essas pessoas que vêm à minha memória agora.

Meu orientador, prof. Kássio. Estamos trabalhando juntos há um bom tempo, hein?! Quero que saiba que sou muito grata pela oportunidade, por ter lhe conhecido tão despretensiosamente enquanto monitora de química analítica, e por você ter confiado em mim durante todos esses anos de orientação. A nossa colaboração aconteceu exatamente na hora certa para mim. E eu sempre tentei dar o meu melhor.

Ao meu marido, Fabrício. Você além de ser o "meu grande amor", sempre foi um grande, se não o maior, incentivador para mim! Muito obrigada por tudo que você fez e faz para me ajudar! Nunca vou esquecer nenhuma das vezes que você me ajudou a continuar na batalha, você sabe. Obrigada pela paciência e pelo amor. Eu te amo.

Aos meus pais, Gladson e Maria da Conceição. Obrigada por todo sacrifício que vocês fizeram para oferecer a mim e ao meu irmão a oportunidade de ter um futuro digno. Por terem nos educado com tanta honestidade e amor. Eu sempre vou amar vocês.

Aos meus amigos, de laboratório, da universidade, da vida – em especial, Priscila, Aline, Fernanda e Camilo – vocês foram essenciais principalmente quando "nada faz sentido"!

Aos técnicos de laboratório do Centro de Biociências e da Escola de Ciência e Tecnologia da UFRN, respectivamente, Flávio Maurílio e Marcos Araújo. Vocês fizeram tudo ser muito mais tranquilo! Quem dera que só tivesse gente como vocês pelo nosso caminho! Muito obrigada por todos os "galhos quebrados"!!

A todos os alunos, professores e voluntários envolvidos nas colaborações que foram realizadas durante o meu doutoramento.

Ao PPGQ-UFRN e à CAPES, pela bolsa concedida.

“Quanto maiores as dificuldades, maior a glória em superá-las. Os grandes navegadores devem sua reputação às tempestades e às tormentas.”

(Epicuro)

RESUMO

Esta tese relata a aplicação das espectroscopias no infravermelho médio, de fluorescência molecular e espectrometria de massas, combinadas a técnicas de análise multivariada, para classificação de células cancerosas em cultivo e de lesões pré-cancerosas através de plasma sanguíneo. Em um primeiro estudo, matrizes de excitação/emissão de fluorescência molecular foram obtidas para diferentes linhagens de células normais (3T3, ARPE, HEK) e cancerosas (HepG2, HeLa, HT-29, 786-0) e modelos de classificação foram construídos utilizando uma combinação dos algoritmos OPLS e UPLS-DA. Taxas de acerto de 100% e 75% foram obtidas para as classes Normal e Cancerosa, respectivamente. Ainda, foi avaliada a influência dos anticorpos anti-MMP-2 e anti-MMP-9 no desempenho dos modelos de classificação. Na presença dos anticorpos, as taxas de acerto nas classificações aumentaram consideravelmente atingindo 100% para ambas as classes, Normal e Cancerosa, através dos algoritmos OPLS/UPLS-DA. Em um segundo estudo, a espectroscopia ATR-FTIR foi utilizada para obtenção de espectros de plasmas sanguíneos de mulheres saudáveis (negativas para lesão intraepitelial ou malignidade, NILM) e portadoras de lesão intraepitelial cervical (SIL) de baixo (LSIL) ou alto grau (HSIL), causadas pelo vírus HPV. Modelos multivariados de classificação foram construídos, visando uma metodologia de rastreamento para o câncer cervical. Os algoritmos PCA-LDA/QDA, SPA-LDA/QDA e GA-LDA/QDA foram aplicados como ferramentas de classificação e seus desempenhos comparados. De maneira geral, os resultados obtidos através do algoritmo GA-QDA foram os mais satisfatórios, utilizando apenas variáveis espectrais selecionadas que puderam ser relacionadas a grupos funcionais pertencentes a diferentes biomoléculas. Os modelos GA-QDA classificaram corretamente NILM *vs.* SIL com sensibilidade e especificidade em torno de 90% e 83%, respectivamente. NILM *vs.* LSIL apresentaram sensibilidade e especificidade variando entre 67-94% e 82-94%, respectivamente. Para NILM *vs.* HSIL, os valores de sensibilidade e especificidade estiveram entre 76-97% e 73-100%, respectivamente. Em um terceiro estudo, a espectrometria de massas foi aplicada para obter os espectros de lipídios extraídos do plasma sanguíneo de mulheres da Classe NILM (n=42) e SIL (n=34). Modelos de classificação multivariados foram construídos utilizando os classificadores LDA, QDA e SVM. Os modelos baseados em SVM permitiram a discriminação das classes com sensibilidade e especificidade de 83.3% e 80.0% para NILM e SIL, respectivamente. Alguns possíveis lipídios foram associados a cada classe, tais como prostaglandinas, esfingolipídios e fosfolipídios, Tetranor-PGFM e um lipídio hidroxiperoxidado. Os resultados obtidos em todos os estudos evidenciam a potencialidade das técnicas espectroscópicas e multivariadas como possíveis metodologias de rastreamento e identificação de câncer, o que poderia contribuir fortemente para a redução da morbidade e mortalidade causadas pela doença.

Palavras-chave: Câncer. Fluorescência molecular. ATR-FTIR. Espectrometria de massas. Análise multivariada.

ABSTRACT

This thesis reports the application of both infrared and molecular fluorescence spectroscopy, as well as mass spectrometry, combined with multivariate analysis techniques for classification of cancerous cells in culture medium and precancerous lesions in blood plasma. In a first study, excitation/emission matrices of molecular fluorescence were obtained for normal (3T3, ARPE, HEK) and cancerous (HepG2, HeLa, HT-29, 786-0) cell lines and classification models were built by using a combination of the algorithms OPLS and UPLS-DA. Correct classification indexes of 100% and 75% were obtained for both classes, Normal and Cancer, respectively. In addition, it was evaluated the influence of the antibodies anti-MMP-2 and anti-MMP-9 in the performance of the classification models. In the presence of the antibodies, the correct classification indexes were considerably improved reaching 100% for both classes, Normal and Cancer, using the algorithms OPLS/UPLS-DA. In a second study, the ATR-FTIR spectroscopy was applied to obtain the spectra of blood plasma of both healthy women (negative for intraepithelial lesion or malignancy, NILM) and women with cervical intraepithelial lesion (SIL) of low grade (LSIL) or high grade (HSIL), caused by HPV virus. Multivariate classification models were built, aiming a screening methodology for cervical cancer. The algorithms PCA-LDA/QDA, SPA-LDA/QDA and GA-LDA/QDA were applied as classification tools and their performance was evaluated. In general, the results obtained by GA-QDA were the most satisfactory, by using only chosen spectral variables that could be related to chemical groups of different biomolecules. The models GA-QDA correctly classified NILM *vs*. SIL with sensitivity and specificity around 67-94% e 82-94%, respectively. For NILM *vs*. LSIL, sensitivity and specificity values were about 67-94% e 82-94%, respectively. For NILM vs. HSIL, the sensitivity and specificity values were 76-97% e 73-100%, respectively. In the third study, mass spectrometry was applied to obtain the spectra of lipids extracted from blood plasma of women of NILM (n=42) and SIL (n=34) classes. Multivariate classification models were built by using the classifiers LDA, QDA and SVM. SVM-based models allowed to discriminate the classes with sensitivity and specificity values of 83.3% and 80.0% for NILM and SIL, respectively. Some possible lipids were associated to each class, such as prostaglandins, phospholipids, sphingolipids, Tetranor-PGFM and a hydroperoxide lipid. The results achieved in all studies highlight the potentiality of the spectroscopic and multivariate techniques as possible methodologies for cancer screening, what could effectively contribute to reduce morbidity and mortality caused by cancer.

Keywords: Cancer. Molecular fluorescence. ATR-FTIR. Mass spectrometry. Multivariate analysis.

ÍNDICE DE ABREVIATURAS

ATR – reflectância total atenuada (do inglês, *attenuated total reflection*)

CIN – neoplasia intraepitelial cervical (do inglês, *cervical intraepithelial neoplasia*)

DA – análise discriminante (do inglês, *discriminant analysis*)

DNA – ácido desoxirribonucleico (do inglês, *desoxyribonucleic acid*)

EEM – matriz de excitação/emissão (do inglês, *excitation/emission matrix*)

EI – ionização por elétrons (do inglês, *electron ionization*)

ESI – ionização por eletrospray (do inglês, *electrospray ionization*)

FT – transformada de Fourier (do inglês, *Fourier transform*)

GA – algoritmo genético (do inglês, *genetic algorithm*)

HPV – vírus do papiloma humano (do inglês, *human papillomavirus*)

HSIL – lesão intraepitelial de alto grau (do inglês, *high grade squamous intraepithelial lesion*)

INCA – instituto nacional do câncer

IR – infravermelho (do inglês, *infrared*)

LDA – análise discriminante linear (do inglês, *linear discriminant analysis*)

LSIL – lesão intraepitelial de baixo grau (do inglês, *low grade squamous intraepithelial lesion*)

MALDI – ionização por dessorção a laser assistido por matriz (do inglês, matrix-assisted laser desorption ionization)

MIR – infravermelho médio (do inglês, *mid-infrared*)

OPLS – projeções ortogonais para estruturas latentes (do inglês, *orthogonal projections to latent structures*)

PARAFAC – análise de fatores paralelos (do inglês, *parallel factor analysis*)

PCA – análise por componentes principais (do inglês, *principal component analysis*)

PLS – mínimos quadrados parciais (do inglês, *partial least squares*)

PLS-DA – análise discriminante pelos mínimos quadrados parciais (do inglês, *partial least squares discriminant analysis*)

QDA – análise discriminante quadrática (do inglês, *quadratic discriminant analysis*)

SPA – algoritmo das projeções sucessivas (do inglês, *successive projections algorithm*)

SUS – Sistema Único de Saúde

SVM – máquina de vetores-suporte (do inglês, *support vector machine*)

TOF – analisador por tempo de voo (do inglês, time-of-flight)

UPLS-DA – análise discriminante pelos mínimos quadrados parciais em dados desdobrados (do inglês, *unfolded partial least squares discriminant analysis*)

WHO – organização mundial de saúde (do inglês, *World Health Organization*)

PREFÁCIO

Mesmo um laureado com Prêmio Nobel é chamado de Doutor. De fato, a conclusão do doutoramento se traduz na obtenção do título de maior relevância na formação acadêmica de pesquisadores em diversas áreas da ciência. Ao iniciar meu trabalho de doutorado em 2013, o Professor Doutor Kássio Michell Gomes de Lima, também meu orientador de Iniciação Científica e Mestrado, me propôs um projeto muito ambicioso e desafiador, que tinha como foco principal o rastreamento de câncer a partir do uso de técnicas instrumentais analíticas e quimiometria. O câncer é umas das doenças mais frequentes e mortais do século XXI em todo o mundo, cujo sucesso de cura está intrinsecamente associado à sua detecção precoce. Atualmente, os métodos de detecção de câncer envolvem procedimentos invasivos e de alto custo, e poder trabalhar em um projeto que visa detecção de câncer em seu estágio inicial ou mesmo pré-doença, a partir de métodos que possam vir a se tornar acessíveis para toda a população, se torna um desafio que vale a pena ser vivido.

Estes últimos quatro anos de minha vida foram de muitas emoções. Uma mescla de alegrias e decepções, tristezas e satisfações. Ah, se um trabalho publicado em um periódico representasse tudo o que se foi passado para chegar aos resultados finais. Quanta leitura. Quantos experimentos realizados; uma, duas, três e mais vezes. Trabalhar com amostras biológicas, de células, de sangue; uma mescla de estímulo e resguardo. Algoritmos e mais algoritmos, rotinas, modelos, modelos, modelos. Após tudo isso, colocar os resultados no papel, na forma de artigos, na forma de uma tese. Vivenciar o ambiente médico na Liga Contra o Câncer do Estado do Rio Grande do Norte certamente foi algo marcante. Pessoas idosas, pessoas jovens, definitivamente lutando contra essa doença terrível. E isso tudo é relacionado ao ambiente acadêmico. Mas todos têm uma vida fora da Universidade, com família, amigos, e diversas situações que mexem conosco diariamente, e simultaneamente a todas as nossas responsabilidades. Mas o Doutorado não tem a importância de maior título acadêmico à toa. E esse título de Doutora vem com todo o esforço devido.

Ao chegar próximo ao final de meu doutoramento, agradeço a todos os Doutores ou aspirantes a Doutores com os quais eu pude colaborar nestes últimos anos. E também a você, leitor, que despendeu de algum tempo para ler este documento simplesmente por achar que valia a pena. Por achar que era digno de me conceder o título de Doutora em Química. Eu espero que valha a pena. Para mim, valeu. E muito!

SUMÁRIO

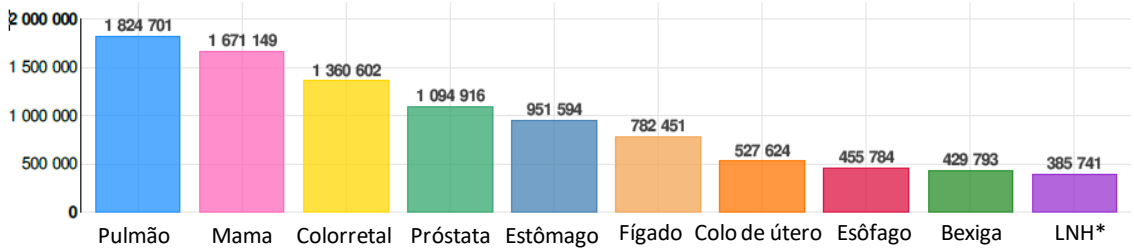
Capítulo 1 – Introdução Geral

1. Introdução

De acordo com estimativas da Organização Mundial de Saúde (WHO) foram diagnosticados 14,1 milhões de novos casos de câncer em 2012 e, ainda, 8,2 milhões de pessoas foram a óbito em virtude desta doença. Países menos desenvolvidos são responsáveis por grande parte destas estatísticas, representando um total de 57% e 65% dos novos casos e das mortes, respectivamente. Ademais, estima-se que nas próximas duas décadas, ocorra um aumento de 70% nos índices que contabilizam os novos casos de câncer no mundo [1,2].

No Brasil, o Instituto Nacional do Câncer (INCA) calculou aproximadamente 596 mil novos casos, entre homens e mulheres, somente no ano de 2016 [3]. Excluindo pele não melanoma, câncer de pulmão, próstata, colorretal, estômago e fígado são os tipos mais frequentes entre homens, enquanto em mulheres, câncer de mama, colorretal, pulmão, colo de útero e estômago apresentam as maiores incidências [2]. Na Figura 1 é apresentado um gráfico de barras da incidência dos dez tipos de câncer mais frequentes, entre homens e mulheres, em todo o mundo, no ano de 2012.

Figura 1: Distribuição proporcional dos dez tipos de câncer mais incidentes mundialmente, entre homens e mulheres, estimados para o ano de 2012, excluindo pele não melanoma.



*Linfoma não-Hodgkin

Fonte: Adaptado da referência [2].

Mais de um terço dos tipos de câncer podem ser prevenidos através de hábitos de vida que modifiquem ou evitem fatores de risco que podem contribuir para o desenvolvimento da doença, tais como tabagismo, obesidade, sedentarismo e infecções. Mais além, alguns dos tipos mais comuns de câncer (mama, colorretal e cervical) têm alto potencial de cura se detectados em estágios iniciais ou pré-cancerosos [1,4].

O câncer de colo de útero é o quarto tipo de câncer mais frequente em mulheres de todo o mundo, com uma estimativa de 528 mil novos casos em 2012. Ainda neste ano, cerca de 270 mil mulheres foram a óbito por câncer cervical, sendo que mais de 85% das mortes ocorreram em países menos desenvolvidos [1,2]. Praticamente todos os casos de câncer cervical podem ser atribuídos à infecção pelo vírus HPV, especialmente os subtipos 16 e 18, considerados de alto risco para desenvolvimento da doença por serem responsáveis por aproximadamente 70% dos casos. Embora seja a doença sexualmente transmissível mais comum, a infecção pelo vírus HPV na maioria das vezes regride espontaneamente dentro de poucos meses a dois anos após a aquisição, sem necessidade de nenhuma intervenção médica [5–9]. Contudo, em uma pequena proporção dos casos, certos tipos de HPV podem persistir e progredir causando o câncer de colo de útero. Essa infecção é essencialmente assintomática e evolui lentamente a partir de lesões pré-cancerosas, neoplasias intra-epiteliais cervicais (NIC), que podem levar de 5 a 20 anos, a depender do sistema imunológico da mulher, para se tornarem o câncer invasivo [10,11].

De acordo com a Organização Mundial de Saúde, as estratégias para a detecção precoce do câncer cervical envolvem o diagnóstico precoce – para pessoas que já apresentam sinais da doença – e os programas de rastreamento – para pessoas assintomáticas, saudáveis. No Brasil, o método oficial de rastreamento considerado pelo INCA, e inclusive oferecido pelo SUS, é o exame citopatológico (exame de Papanicolaou) que deve ser disponibilizado às mulheres na faixa etária de 25 a 64 anos e que já possuem vida sexualmente ativa [3,12]. A *Nomenclatura Brasileira para Laudos Citopatológicos Cervicais* (2012) classifica as lesões cervicais e suas equivalências de acordo com o Sistema de Bethesda (2001), que divide as lesões intraepiteliais escamosas em baixo (LSIL) e alto (HSIL) graus de capacidade de progressão para o câncer invasivo, sendo que as lesões LSIL englobam a categoria histopatológica NIC1 enquanto a HSIL corresponde a NIC2 /NIC3 [13]. Considerando o laudo obtido através do exame de rastreamento, diferentes condutas podem ser realizadas, e estas devem considerar diversos fatores no histórico da mulher, tais como:

idade, filhos, grau de infecção/lesão, histórico de doença cervical anterior e gravidez. A Tabela abaixo resume as diretrizes adotadas no Brasil em relação às recomendações de conduta inicial frente aos resultados alterados de exames citopatológicos realizados pelo SUS [12]:

Tabela 1: Resumo das recomendações para conduta inicial frente aos resultados alterados de exames citopatológicos nas unidades de atenção básica.

Diagnóstico citopatológico		Faixa etária	Conduta inicial
Células escamosas atípicas de significado indeterminado (ASCUS)	Possivelmente não neoplásicas (ASC-US)	< 25 anos	Repetir em 3 anos
		Entre 25 e 29 anos	Repetir a citologia em 12 meses
		≥ 30 anos	Repetir a citologia em 6 meses
	Não se podendo afastar lesão de alto grau (ASC-H)		Encaminhar para colposcopia
Células glandulares atípicas de significado indeterminado (AGC)	Possivelmente não neoplásicas ou não se podendo afastar lesão de alto grau		Encaminhar para colposcopia
Células atípicas de origem indefinida (AOI)	Possivelmente não neoplásicas ou não se podendo afastar lesão de alto grau		Encaminhar para colposcopia
Lesão de Baixo Grau (LSIL)		< 25 anos	Repetir em 3 anos
		≥ 25 anos	Repetir a citologia em 6 meses
Lesão de Alto Grau (HSIL)			Encaminhar para colposcopia
Lesão intraepitelial de alto grau não podendo excluir microinvasão			Encaminhar para colposcopia
Carcinoma escamoso invasor			Encaminhar para colposcopia
Adenocarcinoma *in situ* (AIS) ou invasor			Encaminhar para colposcopia

Fonte: Referência [12]

Dessa forma, através dos exames de rastreamento, as mulheres que apresentam alterações nos resultados podem ser acompanhadas e, quando necessário, submetidas a exames de diagnóstico, como a colposcopia e biópsia, de modo que o tratamento médico pode ser realizado de forma eficaz e providencial, evitando a evolução das lesões precursoras para o câncer invasivo. Quando as alterações que antecedem o câncer cervical são identificadas e tratadas adequadamente é possível prevenir a doença em praticamente 100% dos casos [3].

Nesse sentido, estratégias de rastreamento de câncer são fundamentais, pois possibilitam que novos casos sejam diagnosticados precocemente de modo que os pacientes tenham suas chances de cura consideravelmente aumentadas, uma vez que o

tratamento pode ser iniciado de maneira mais rápida e efetiva, menos custosa e complexa. Entretanto, para que os programas de rastreamento contribuam adequadamente é preciso que o maior número possível de pessoas seja alcançado e tenha acesso periódico aos exames, de modo que técnicas simples, pouco invasivas e de baixo custo, especialmente, podem favorecer sobremaneira a abrangência e sucesso dos rastreamentos [1,3,12].

O uso de técnicas espectroscópicas voltadas para a análise de material biológico, em um contexto biomédico, tem sido referido genericamente como bioespectroscopia [14]. As espectroscopias no infravermelho e de fluorescência molecular, assim como a espectrometria de massas, são exemplos de técnicas que vêm sendo cada vez mais utilizadas com abordagens clínicas, uma vez que permitem a obtenção de informação química sobre um grande e variado número de moléculas envolvidas direta ou indiretamente em processos biológicos. Estas, por sua vez, podem refletir a condição fisiopatológica de células, tecidos ou de um indivíduo em particular [15–25].

As ferramentas bioespectroscópicas possuem grande potencialidade para serem utilizadas na identificação de doenças, especialmente no que diz respeito a métodos de rastreamento, uma vez que apresentam características importantes tais como versatilidade de análises, alta sensibilidade analítica (ex. espectrometria de massas), rapidez e simplicidade na aquisição espectral, mínima ou inexistente necessidade de pré-tratamento de amostras e baixo custo de materiais e equipamentos (ex. infravermelho e fluorescência) [26–30]. Entretanto, tais técnicas frequentemente levam a dados espectrais de difícil interpretação e baixa especificidade, especialmente quando se analisa amostras biológicas reais, dada a complexidade intrínseca presente neste tipo de amostra e o fato de que uma grande fração das biomoléculas presentes estará contribuindo para o sinal como um todo. Desta forma, o uso de técnicas de análise multivariada configura-se como etapa imprescindível para o sucesso da bioespectroscopia.

Efetivamente, tais métodos computacionais permitem que dados espectroscópicos sejam interpretados com maior facilidade e que somente informações químicas relevantes sejam consideradas [31–34]. Com o propósito de estudo de doenças, por exemplo, diversos algoritmos podem ser utilizados tanto para pré-processamento de dados quanto para classificação de amostras em grupos que representem ausência ou presença de determinada patologia, estudo de progressão de doenças, identificação e caracterização de biomarcadores, entre outros. Alguns dos

algoritmos mais frequentemente utilizados para fins de classificação incluem o tradicional PLS-DA, e os métodos de redução de dimensionalidade (PCA) e seleção de variáveis (SPA e GA) associados a classificadores como LDA, QDA ou SVM [15,31,35–37].

Desta forma, o uso de técnicas espectroscópicas associadas à análise multivariada apresenta potencialidade considerável para ser utilizada no combate ao câncer, especialmente no que se refere aos métodos de rastreamento já existentes – por exemplo, o teste de Papanicolaou/exame preventivo – de maneira a atuar como alternativa de fácil implementação e baixo custo. Sendo assim, contribuindo para um maior alcance dos programas de rastreamento e, consequentemente, diminuição da morbidade e mortalidade causadas pelo câncer, especialmente em regiões rurais, pouco desenvolvidas e com escassez de recursos financeiros.

1.1 Técnicas instrumentais

1.1.1 Espectroscopia de fluorescência molecular

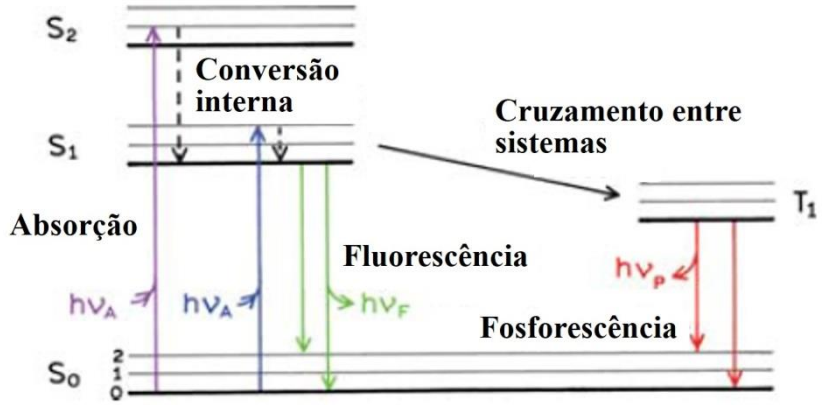
A espectroscopia de fluorescência molecular tem ganhado cada vez mais espaço dentro das áreas das ciências biológicas, nas mais variadas abordagens. Citometria de fluxo, sequenciamento de DNA, análises genéticas e variados exames médicos de diagnóstico são alguns exemplos de técnicas importantes em que a fluorescência molecular é extensivamente aplicada [27].

Muitas moléculas, especialmente aquelas que possuem estruturas aromáticas, rígidas, planares e que contêm duplas ligações conjugadas, possuem a habilidade de emitir radiação na forma de luz, um fenômeno conhecido como luminescência. Para que a luminescência ocorra é preciso inicialmente que as moléculas que se encontram em seu estado fundamental (de mais baixa energia) absorvam energia e atinjam estados eletronicamente excitados. Quando absorvida, a radiação na região do ultravioleta (acima de 200 nm) e visível do espectro eletromagnético leva a transições eletrônicas entre os elétrons não-ligantes (elétrons "*n*") e elétrons de ligações π do tipo $n \rightarrow \pi^*$ e $\pi \rightarrow \pi^*$. Em seguida, as moléculas excitadas podem retornar ao seu estado fundamental emitindo parte da energia (regra de Stokes) que foi absorvida na forma de luz.

A luminescência é dividida em fluorescência e fosforescência, dependendo da natureza do estado excitado que a molécula atinge. Se no estado excitado o elétron envolvido na transição mantém a sua orientação de spin no orbital excitado,

continuando emparelhado com o segundo elétron que está no orbital do estado fundamental, tem-se o estado excitado singlete, o que caracteriza a fluorescência. Já se o elétron no estado excitado inverte a sua orientação de spin, tem-se o estado excitado triplete que caracteriza a fosforescência[38,39]. Os processos que ocorrem entre a absorção e emissão de luz são usualmente ilustrados através do diagrama de Jablonski, conforme visto na Figura 2.

Figura 2: Diagrama de Jablonski ilustrando resumidamente o fenômeno de luminescência.



Fonte: Autor

Diversos processos moleculares podem ocorrer em estados excitados. As moléculas podem rapidamente perder energia através de colisões e cair para níveis vibracionais de menor energia. Além disso, a grande maioria das moléculas que ocupa um estado eletrônico excitado maior que o segundo (S_2), experimenta o processo de conversão interna e passa do menor nível vibracional do estado superior para o maior nível vibracional do estado inferior que possui a mesma energia. Uma vez que são possíveis transições do menor nível vibracional do primeiro estado excitado (S_1) para qualquer um dos níveis vibracionais do estado fundamental, os espectros de emissão têm o formato de bandas bastante alargadas e sobrepostas entre si [27,39].

Os espectrofluorímetros são os instrumentos utilizados para medidas de fluorescência. Grande parte é equipada com monocromadores, que permitem a escolha dos comprimentos de onda que serão usados na excitação e também da faixa de comprimentos de onda que será medida no espectro de emissão. Além disso, o detector consiste de tubos fotomultiplicadores que aumentam bastante a sensibilidade da técnica e podem detectar sinais de boa qualidade mesmo a partir de moléculas fracamente fluorescentes.

Uma das questões muito importantes que podem influenciar a qualidade do espectro de fluorescência é a concentração da amostra. É recomendado que as amostras analisadas estejam bastante diluídas (apresentem absorbância de até 0.05), especialmente para o estudo de compostos muito fluorescentes, de modo que possa existir uma relação linear entre a concentração e o sinal obtido. Além disso, fenômenos como o efeito do filtro interno (*inner filter effect*) e a supressão de fluorescência (*quenching*) podem comumente acontecer e levar a medidas incorretas, especialmente quando a diluição não é realizada adequadamente. Uma grande vantagem da espectroscopia de fluorescência molecular é que ela permite realizar medidas na ordem de parte por bilhão (ppb), devido a ser uma técnica de alta sensibilidade analítica, especialmente quando comparada às espectroscopias de absorção [27,39].

1.1.2 Espectroscopia no infravermelho médio

A espectroscopia no infravermelho médio (MIR) é uma técnica vibracional altamente conhecida e empregada tanto em nível de pesquisa científica quanto industrial. O conceito básico da técnica consiste em incidir sobre uma amostra um feixe de radiação, na região de 4000 a 400 cm^{-1} do espectro eletromagnético, de maneira que determinadas vibrações moleculares – vibrações permitidas que provoquem alteração no momento de dipolo da molécula – levarão a uma absorção da energia incidente em frequências específicas, gerando assim um espectro de infravermelho [38].

Esta técnica, altamente versátil, fornece um leque de informações químicas valiosas acerca de composição, estrutura e conformação de moléculas, o que permite que estas sejam caracterizadas através seus espectros. Cada molécula apresenta um espectro único e característico, em função de sua composição e estrutura química, que pode ser interpretado como sua impressão digital (*fingerprint*).

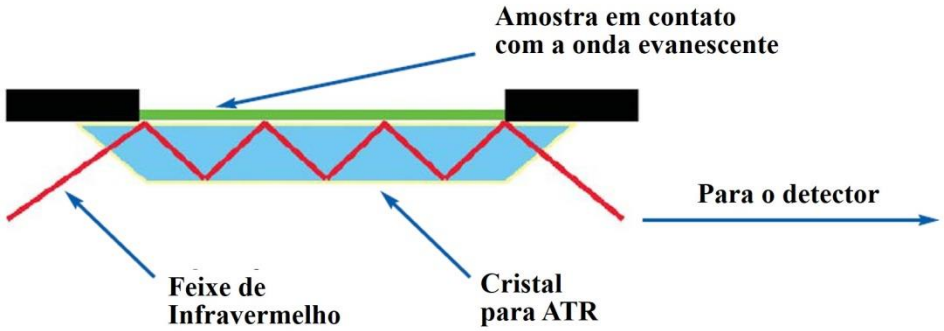
Além de análises qualitativas, a espectroscopia no infravermelho também permite que medidas quantitativas sejam realizadas, levando em consideração a fração da radiação incidente que foi absorvida pela amostra, através da bem estabelecida lei de Beer [26,38]. Atualmente, os espectrofotômetros de infravermelho são, em sua grande maioria, equipados com um sistema óptico (interferômetro de Michelson) que, associado a transformações matemáticas conhecidas como Transformada de Fourier (FT), permitem de forma rápida a obtenção de espectros com alta relação sinal/ruído e alta resolução.

A espectroscopia FTIR pode ser utilizada para analisar amostras sólidas, líquidas ou gasosas, entretanto, em muitos casos algum tipo de preparação de amostra é necessário para obtenção de espectros de boa qualidade. Tradicionalmente, as amostras são analisadas no modo de transmitância, porém fatores como reprodutibilidade espectral e formas de preparação de amostras podem levar a medidas incorretas, caso o analista não se atente a alguns cuidados.

A reflectância total atenuada (ATR) associada à espectroscopia no infravermelho por transformada de Fourier, ATR-FTIR, tem sido, nos últimos anos, a forma mais utilizada da espectroscopia no infravermelho médio [26]. Isso se deve ao fato de que para amostras sólidas e líquidas, incluindo pastas e materiais de difícil manuseio, essa técnica combate os aspectos mais desafiadores das análises por infravermelho, que são a preparação de amostras e a reprodutibilidade.

A técnica ATR permite a obtenção de espectros de forma estável, robusta, não-destrutiva e com mínima ou nenhuma preparação de amostras, uma vez que estas podem ser posicionadas diretamente como pós, líquidos, pastas, etc, sobre o cristal de ATR, e então a medida realizada. A reflectância total atenuada tem sido estudada desde muitos anos por grandes cientistas, como Isaac Newton, mas os maiores passos em seu desenvolvimento só aconteceram por volta do fim dos anos 1950 para o início da década 1960, quando muitos estudos foram publicados relatando aplicações da espectroscopia por ATR. A reflectância total é um caso especial de reflexão de uma onda eletromagnética em uma interface entre dois meios (Figura 3).

Figura 3: Processo de reflectância total atenuada.



Fonte: Autor

O princípio da espectroscopia ATR-FTIR consiste em incidir sob um ângulo específico um feixe de radiação infravermelha contra um meio de reflectância interna

opticamente denso (frequentemente um “cristal” de diamante) que reflete a radiação na interface com um meio mais denso opticamente, criando uma onda evanescente que se estende para além da superfície do cristal e pode interagir com as amostras colocadas na interface, permitindo, então, medidas de absorção.

Essa onda evanescente se sobressai apenas alguns microns (0.5 – 5 μ) para além do cristal, portanto, deve existir um bom contato entre a amostra e a superfície do mesmo. Nas regiões do infravermelho onde a amostra absorve energia, a onda evanescente será atenuada ou alterada, e isto será medido pelo detector, gerando um espectro de infravermelho [26]. Quando aplicada a amostras de natureza biológica, conforme foi realizado nesta tese, a espectroscopia ATR-FTIR também possui a vantagem de ser menos propensa à influência da água, o que é bastante importante considerando a grande proporção de água que existe nestas amostras, o que pode interferir fortemente no sinal observado.

1.1.3 Espectrometria de massas

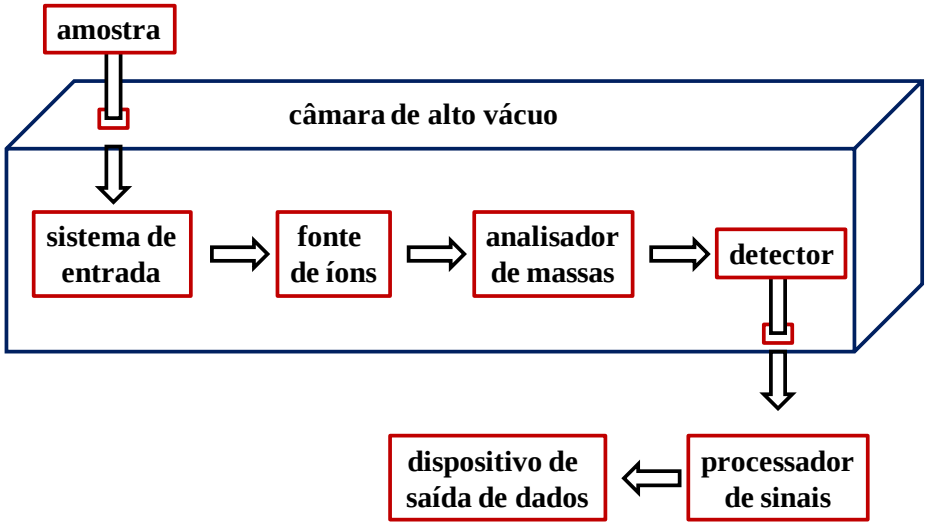
A espectrometria de massas moleculares consiste de uma técnica analítica extremamente valiosa para a elucidação estrutural de compostos moleculares, especialmente de natureza orgânica. A mesma é baseada na conversão de moléculas em íons em fase gasosa, os quais são subsequentemente separados no espectrômetro de massas de acordo com sua razão massa/carga (*m/z*).

Dentre as técnicas instrumentais descritas no presente documento, a espectrometria de massas é a que fornece informações mais precisas com relação aos aspectos estruturais do analito, especialmente com relação à massa molecular e padrão de fragmentação.

A espectrometria de massas tem seus princípios delineados há mais tempo que as demais técnicas espectroscópicas, quanto J. J. Thomson, em 1890, determinou a razão *m/z* do elétron, e após duas décadas, ao estudar gases atmosféricos, demonstrou a existência do néon-22 em uma amostra de néon-20, ou seja, estabelecendo assim a possibilidade dos elementos apresentarem isótopos. O primeiro espectrômetro de massas foi construído em 1918, por A. J. Dempster, e apesar das descobertas marcantes propiciadas pela técnica, a mesma só veio a se tornar popular há pouco mais de 50 anos, quando passaram a ser comercializados instrumentos a preços módicos e com alto grau de confiabilidade [40].

Os componentes básicos de um espectrômetro de massas estão representados na Figura 4 [38]. Inicialmente, a amostra é introduzida em uma câmara de alto vácuo, tendo seus componentes convertidos em íons gasosos (positivos ou negativos), que serão direcionados para o analisador de massas para serem dispersos em função das razões *m/z*. No detector, o feixe de íons é convertido em um sinal elétrico, que pode então ser processado, resultando em um espectro de massas.

Figura 4: Componentes básicos de um espectrômetro de massas.



Fonte: Adaptado da referência [38].

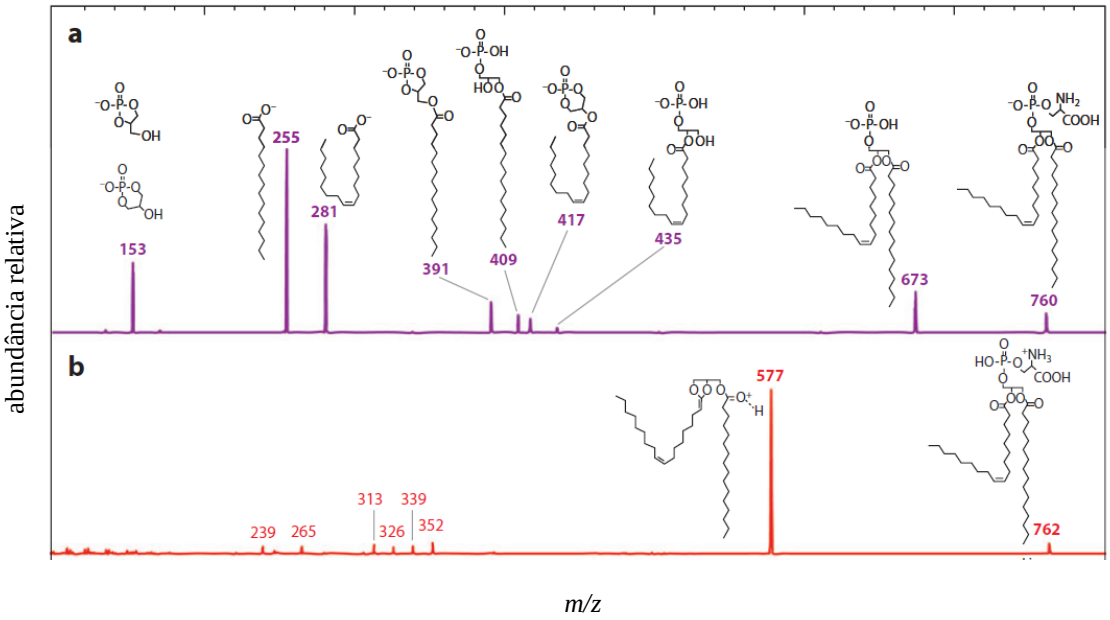
A fonte de ionização mais comumente empregada durante muito tempo na espectrometria de massas consiste na técnica baseada no impacto de elétrons (EI), todavia, a mesma se mostra muito energética e limitada a compostos voláteis e/ou de baixos pesos moleculares. Para superar estas limitações, muitos esforços foram despendidos no intuito de aprimoramento das técnicas de ionização. No final da década de 1980, uma das maiores revoluções dentro da espectrometria de massas ocorreu com o desenvolvimento das técnicas de ionização por eletrospray (ESI) e por dessorção (especialmente a MALDI), permitindo análises de materiais biológicos de alto peso molecular e baixa volatilidade, sendo possível, inclusive, manter intactas interações não covalentes existentes nestes sistemas químicos [41].

Adicionalmente, por se tratar de um método muito mais *soft* a ESI não acarreta em um alto grau de fragmentação, sendo assim muito interessante para análise de misturas complexas de moléculas. Com relação ao analisador de massas, são mais

frequentemente empregados o tipo quadrupolar e também o analisador por tempo de voo (TOF).

O resultado final de uma análise de espectrometria de massas se traduz em um gráfico que correlaciona a razão *m/z* com a estabilidade do íon associado à mesma. Como exemplo, a Figura 5 apresenta a análise de uma mistura de lipídeos por espectrometria de massas a partir da ionização do tipo ESI no modos negativo (a) e positivo (b), com as estruturas químicas dos principais fragmentos [42].

Figura 5: Espectro de massas de um mistura de fosfolipídeos obtidos através da ionização do tipo ESI.



Fonte: Referência [42]

De fato, a espectrometria de massas se traduz em umas das técnicas mais revolucionárias associadas à identificação e caracterização estrutura de compostos orgânicos. Seu sucesso é justificado pelos cinco cientistas que foram agraciados com Prêmio Nobel por trabalhos diretamente associados à espectrometria de massas [40], além de tantos outros que fizeram uso da técnica para propósitos variados. Na área biológica, a espectrometria de massas propiciou um enorme avanço no entendimento de diversos sistemas de alta complexidade, incluindo aqueles que compõem as chamadas ciências ômicas [43], tais como proteômica, lipidômica, genômica, glicômica, transcriptômica, notavelmente quando associada à técnicas cromatográficas modernas, tais como as cromatografias gasosa e líquida de alta eficiência.

1.2 Quimiometria

1.2.1 Análise por componentes principais - PCA

A análise por componentes principais pode ser considerada como o método mais importante na área da quimiometria. A ideia principal por trás da PCA, isto é, de combinações lineares de variáveis originais representadas por componentes principais, foi introduzida por volta dos anos 1930 por Pearson e Hooteling, e diversas abordagens foram desenvolvidas desde então. Recentemente, alguns autores publicaram artigos de revisão detalhados em que os cálculos e fundamentos da técnica podem ser cuidadosamente apreciados [44,45].

De maneira geral, através da PCA é possível reduzir a dimensionalidade de um conjunto de dados. Assim, através de poucas componentes principais, amostras podem ser representadas por um conjunto de *scores* e *loadings* (ao invés de centenas ou milhares de variáveis originais) e visualizadas em um gráfico que retrata suas similaridades e/ou diferenças. Neste método, uma matriz espectral **X** é decomposta como:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{t} + \boldsymbol{E} \qquad (1)$$

onde **X** é a matriz espectral, **T** e **P** são as matrizes dos vetores *scores* e *loadings*, respectivamente, e **E** é a matriz dos resíduos [44-46]. Nesta tese, a PCA foi utilizada para fins de análise não supervisionada, de modo a se observar a presença ou ausência de grupos (*clusters*) entre as classes de amostras analisadas, baseando-se apenas nas informações químicas espectrais contempladas pelas componentes principais.

1.2.2 Seleção de variáveis – GA e SPA

Técnicas de seleção de variáveis permitem que sejam escolhidas uma ou mais partes do conjunto total de variáveis instrumentais medido, e estas variáveis selecionadas, por sua vez, é que serão utilizadas na construção dos modelos multivariados. Estas ferramentas são muito importantes na otimização dos resultados e podem contribuir significativamente para a obtenção de modelos mais simples, robustos e de fácil interpretação, uma vez que serão consideradas apenas as variáveis que representam informações úteis à análise, e excluídas aquelas redundantes e relacionadas a fontes de incertezas e ruídos [47,48].

O algoritmo genético (GA) está entre as técnicas de seleção de variáveis mais amplamente utilizada. Baseado no princípio da seleção natural de Darwin, o GA seleciona as variáveis por uma série de etapas iterativas, avaliando a aptidão de cada conjunto de variáveis candidato a uma possível solução. As etapas de seleção, cruzamento e mutação são os principais operadores que atuam aleatoriamente separando as variáveis por aptidão, e constituem um ciclo de evolução. Para um dado número máximo de ciclos, a melhor solução encontrada pelo GA será aquela que apresentar o melhor valor de aptidão [49-52].

O algoritmo das projeções sucessivas (SPA) foi inicialmente desenvolvido para ser aplicado em modelos de calibração multivariados, com o intuito de reduzir ao máximo a redundância e colinearidade entre as variáveis instrumentais, através de uma série de projeções vetoriais. Quando aplicado a problemas de classificação, o SPA escolhe as variáveis com menor colinearidade e que levam a menor valor de classificação incorreta para o modelo [53,54].

1.2.3 Projeções ortogonais para estruturas latentes - OPLS

Proposto por Trygg e Wold [55], o algoritmo das projeções ortogonais para estruturas latentes é um método de pré-processamento multivariado que tem por objetivo remover variação sistemática na matriz **X** que não é correlacionada (ortogonal) à matriz de respostas, **Y**, fornecendo modelos mais parcimoniosos e de fácil interpretação. OPLS é uma variação do tradicional PLS, adaptado para fins de pré-processamento. As primeiras etapas são iguais às do algoritmo NIPALS tradicional e, então, os *scores* e *loadings* das variações ortogonais são calculados, de modo que a matriz **X** pré-processada é obtida como segue:

$$\boldsymbol{X}_{OPLS} = \boldsymbol{X} - \boldsymbol{t}_{orto}\boldsymbol{p}_{orto}^{T} \quad (2)$$

em que $\mathbf{t}_{\text{orto}}$ e $\mathbf{p}^{\mathbf{T}}_{\mathbf{orto}}$ são os *scores* e *loadings* da matriz ortogonal, respectivamente [55-57]. Nesta tese, o OPLS foi aplicado aos espectros de fluorescência molecular (matriz **X**), de modo a aumentar a taxa de acerto dos modelos de classificação entre células cancerosas e normais.

1.2.4 Análise de fatores paralelos - PARAFAC

A análise de fatores paralelos consiste em um método multidimensional originado na psicometria, na década de 1960 [58,59]. Aplicado a dados químicos, PARAFAC é utilizado como um método de decomposição e pode ser considerado como uma generalização da PCA bilinear aplicada a dados multidimensionais, isto é, dados que resultam em uma matriz de resposta instrumental para cada amostra.

Matrizes de excitação-emissão de fluorescência molecular, EEM, são exemplos de dados multidimensionais para os quais o PARAFAC pode ser aplicado com sucesso. Cada amostra EEM produzirá uma matriz de tamanho *J* x *K*, onde *J* representa os comprimentos de onda de emissão e *K* os comprimentos de onda de excitação. Se *I* matrizes de amostras são agrupadas formando um paralelepípedo, então um conjunto de dados multidimensional, **X**, é obtido com dimensão [*I* x *J* x *K*]. Considerando que **X** siga o modelo PARAFAC trilinear, a sua decomposição é obtida minimizando a soma dos quadrados dos resíduos, e_{ijk}, conforme segue:

$$\boldsymbol{X}_{ijk} = \sum_{f=1}^{F} a_{if}\, b_{jf} c_{kf} + e_{ijk} \qquad (3)$$

em que F é o número de fatores, e_{ijk} é um termo de erro residual que tem as mesmas dimensões de **X**. Os vetores coluna a_{if}, b_{jf} e c_{kf} estão usualmente contidos em matrizes *scores* e *loadings* **A**, **B** e **C**, respectivamente [58-60].

Nesta tese, PARAFAC foi aplicado como método de decomposição para os dados EEM de fluorescência molecular e os scores obtidos foram usados como dados de entrada para modelos de classificação.

1.2.5 Análise discriminante pelos mínimos quadrados parciais - PLS-DA

Análise discriminante pelos mínimos quadrados parciais é um método quimiométrico relativamente novo e tem sido especialmente utilizado para fins de classificação, mas também para análise exploratória [61]. O PLS-DA é derivado da tradicional regressão PLS e consiste em construir um modelo de regressão entre a matriz de respostas instrumentais, **X**, contra um vetor de números (frequentemente de 0 e 1) que representam as duas classes que as amostras podem assumir, **y**.

De maneira geral, o algoritmo PLS-DA pode ser considerado como um método de classificação para conjuntos de dados que contenham apenas duas classes de

amostras (quando derivado do PLS1), através do estabelecimento de uma função discriminante linear. Através dessa função, as amostras são assinaladas a pertencerem ou não a uma classe específica em consequência de o valor calculado pelo modelo estar acima ou abaixo de um valor limite (*threshold*) previamente definido [61]. Nesta tese, o algoritmo PLS-DA foi aplicado para discriminação de células a partir dos *scores* de modelos PARAFAC previamente obtidos.

1.2.6 Análise discriminante pelos mínimos quadrados parciais em dados desdobrados - UPLS-DA

A análise discriminante pelos mínimos quadrados parciais em dados desdobrados consiste no mesmo princípio do PLS-DA, com a diferença de que o algoritmo agora é aplicado para dados de segunda ordem desdobrados. Nesta tese, o modelo UPLS-DA foi utilizado em matrizes desdobradas de excitação-emissão de fluorescência molecular, EEM.

Cada amostra EEM consiste em uma matriz de tamanho J x K, onde J representa os comprimentos de onda de emissão e K os comprimentos de onda de excitação. Para I matrizes de amostras agrupadas formando um paralelepípedo, um conjunto de dados multidimensional, **X**, é obtido com dimensão [I x J x K]. Previamente à construção dos modelos UPLS-DA, a matriz **X** é desdobrada (*unfolded*) de modo que sua dimensão passa a ser [I x JK], podendo então ser considerada estruturalmente como um conjunto de dados de primeira ordem [62].

1.2.7 Análise discriminante linear - LDA

A análise discriminante linear é um dos métodos de classificação mais utilizados na quimiometria. O objetivo geral de métodos de fronteira como o LDA é encontrar uma função discriminante que separe amostras em grupos [63,64]. Existem diferentes formas de se calcular a função discriminante através do LDA. Nesta tese, os algoritmos LDA que foram utilizados aplicaram a distância Mahalanobis [65] para obter a separação das classes.

Outra característica desse método é que ele emprega uma matriz total de covariância que é comum para as duas classes de amostras analisadas. Além disso, quando o conjunto de dados contém um número de variáveis muito maior do que o número de objetos (n), como pode ser o caso com espectros de infravermelho, a LDA

pode levar a altas taxas de erros de classificação, dessa forma, é indicado reduzir a dimensionalidade desse dos dados previamente à aplicação da LDA.

Algoritmos de seleção de variáveis tais como SPA e GA são ótimas alternativas para extrair somente variáveis representativas e tornar os dados menos ruidosos, menos colineares e com menor proporção de variáveis irrelevantes. A tradicional PCA também é uma boa opção visando o mesmo fim de redução da dimensionalidade dos dados e, nesse caso, os *scores* obtidos da PCA é que são utilizados como entrada para a LDA [63,64].

1.2.8 Análise discriminante quadrática - QDA

A análise discriminante quadrática é outro método de classificação multivariado, muito semelhante à LDA. A principal diferença entre os dois métodos, QDA e LDA, é que a QDA não assume que as duas classes possuam matrizes de covariância similares e, portanto, classifica as amostras utilizando uma função determinante que considera diferentes matrizes de covariância para cada classe. Assim, a função discriminante da QDA tem formato quadrático [64].

1.2.9 Máquinas de vetores-suporte – SVM

As máquinas de vetores-suporte são algoritmos criados inicialmente para problemas de reconhecimento de padrão, mas funcionalmente podem ser utilizados tanto para fins de classificação quanto para regressão multivariada [66]. SVM pode ser aplicado a problemas mais simples, de soluções lineares, da mesma forma que a problemas mais complexos, não lineares, como pode ser o caso de amostras reais, especialmente as de natureza biológica [31,67].

A ideia básica que fundamenta o SVM para classificação é o uso de hiperplanos que definem as fronteiras de decisão que separam os objetos (amostras) de diferentes classes. Isso significa dizer que com a ajuda de uma função Kernel os dados originais são projetados para um espaço multidimensional (ou seja, com dimensão superior àquela dos dados originais) de modo a criar regiões de fronteira no espaço que separam os grupos adequadamente. Dessa forma, um hiperplano otimizado é obtido e este, por sua vez, maximiza as margens de separação entre as duas classes de amostras conferindo generalização para o modelo e, consequentemente, aumentando sua habilidade de classificar amostras desconhecidas (conjunto de previsão) [66].

Diferentes funções Kernel podem ser utilizadas e cada uma delas fornece uma forma específica de separar as amostras no espaço multidimensional. O SVM é particularmente interessante para ser aplicado para modelagem de dados que apresentam comportamento não linear, seja por fatores intrínsecos ou experimentais [31].

2. Objetivos gerais

A presente tese de doutorado teve por objetivo geral avaliar a capacidade de técnicas espectroscópicas e de análise multivariada, utilizadas em conjunto, para classificação de câncer e de lesões pré-cancerosas em amostras biológicas reais (células em cultivo e plasma sanguíneo). Mais especificamente, aplicar:

- espectroscopia de fluorescência molecular e algoritmos de análise multivariada para classificação de diferentes linhagens celulares normais e cancerosas;
- espectroscopia ATR-FTIR e algoritmos de classificação multivariada a plasmas sanguíneos de mulheres saudáveis e portadoras de lesão cervical, como uma possível alternativa para rastreamento de câncer de colo de útero.
- espectrometria de massas por injeção direta e análise multivariada para discriminação entre plasmas sanguíneos de mulheres saudáveis e portadoras de lesão cervical uterina, através de uma abordagem lipidômica.

3. Organização da tese

A presente tese é apresentada em capítulos referentes aos trabalhos como primeira autora, que relatam estudos focados no tema central proposto para o projeto de pesquisa – identificação e rastreamento de câncer – essenciais para o desenvolvimento do doutorado. Ainda, nos apêndices são apresentados trabalhos realizados através de diferentes colaborações que ocorreram durante o período de doutoramento.

Capítulo 2 – "*The use of EEM fluorescence and OPLS/UPLS-DA algorithm to discriminate between normal and cancer cell lines: a feasibility study*" (publicado no periódico Analyst, DOI: 10.1039/c4an00296b) – relata uma aplicação de fluorescência molecular e análise discriminante via PLS para identificação de diferentes linhagens celulares normais e cancerosas, com comparação entre métodos multivariados e as principais vantagens e limitações do método proposto.

Capítulo 3 – "*ATR-FTIR and multivariate analysis as a screening tool for cervical cancer in women from northeast Brazil: a biospectroscopic approach*"

(publicado no periódico RSC Advances, DOI: 10.1039/c6ra21331f) – relata uma aplicação de espectroscopia no infravermelho e análise discriminante multivariada para rastreamento de lesões cervicais pré-cancerosas, com abordagem comparativa entre análise discriminante linear e quadrática e também seleção de variáveis.

Capítulo 4 – "*Mass spectrometry and multivariate analysis to classify cervical intraepithelial neoplasia from blood plasma: an untargeted lipidomic study*" (manuscrito em fase de submissão) – relata uma aplicação de espectrometria de massas por injeção direta e análise multivariada para classificação de lesões cervicais pré-cancerosas a partir dos lipídios extraídos do plasma sanguíneo, com enfoque na comparação entre diferentes métodos de análise discriminante multivariada.

Capítulo 5 – "*Conclusões e Perspectivas*" – apresenta um resumo dos principais resultados alcançados dentre os estudos, sua concordância com os objetivos propostos e a relevância para o tema central abordado, além de sugestões para trabalhos futuros.

Apêndice A – "*Area correlation constraint for the MCR-ALS quantification of cholesterol using fluorescence EEM data: a new approach*" (publicado no periódico Analytica Chimica Acta, DOI: 10.1016/j.aca.2016.08.011) – relata uma nova aplicação (restrição de correlação) do algoritmo MCR-ALS para quantificação de analitos em dados de segunda ordem.

Apêndice B – "*Biorganic concepts involved in the determination of glucose, cholesterol and triglycerides in plasma using the enzymatic colorimetric method*" (publicado no periódico Quimica Nova, DOI: 10.5935/0100-4042.20150040) – trabalho teórico-experimental voltado para o ensino de química, abordando os conceitos envolvidos nas determinações glicose, colesterol e triglicerídeos pelo método enzimático colorimétrico.

Apêndice C – "*Colorimetric determination of ascorbic acid based on its interfering effect in the enzymatic analysis of glucose: an approach using smartphone image analysis*" (publicado no periódico Journal of the Brazilian Chemical Society, DOI: 10.21577/0103.5053.20170086) – retrata a aplicação de espectroscopia UV-Vis e análise multivariada de imagens para quantificação do ácido *L*-ascórbico através de seu efeito interferente na análise de glicose pelo método enzimático colorimétrico.

Apêndice D – "*Determination of serum protein content using cell phone image analysis"* (publicado no periódico Analytical Methods, DOI: 10.1039/c6ay01783e) –

relata uma aplicação focada na análise de imagens para quantificação de proteínas em solução sintética e em amostras biológicas reais.

Apêndice E – "*Estimation of Brazilian charcoal properties using attenuated total reflectance-Fourier transform infrared (ATR-FTIR) spectrometry coupled with multivariate analysis*" (publicado no periódico Analytical Methods, DOI: 10.1039/c5ay01135c) – relata a quantificação do teor de carbono fixo, cinzas e materiais voláteis em carvão vegetal comercial, através de regressão multivariada e infravermelho médio.

4. Metodologia

Os estudos reportados nos Capítulos 2, 3 e 4 desta tese foram realizados pela colaboração entre o Instituto de Química (IQ/UFRN) e os Departamentos de Morfologia (DMOR/UFRN), Biofísica e Farmacologia (DBF/UFRN) e Análises Clínicas e Toxicológicas (DACT/UFRN) da Universidade Federal do Rio Grande do Norte, Natal, Brasil. Também se realizou colaboração com o Instituto de Química da Universidade Federal de Goiás (IQ/UFG), Goiânia, Brasil. As coletas de sangue das mulheres voluntárias foram realizadas na Maternidade Escola Januário Cicco (MEJC/UFRN), com aprovação do Comitê de Ética do Hospital Universitário Onofre Lopes (HUOL/UFRN), protocolo número #526/11.

4.1 Coleta e preparação das amostras

4.1.1 Células em cultivo

Células em cultivo em meio DMEM (Dulbecco´s modified Eagle's médium, Life Technologies, Inc., Grand Island, NY, USA) suplementado com soro fetal bovino 10% (v/v) (FBS Life Technologies, Inc., Grand Island, NY, USA) foram disponibilizadas pelo Departamento de Morfologia (DMOR/UFRN). Foram utilizadas as linhagens celulares de fibroblastos de ratos (3T3), retina humana (ARPE), renais humanas (HEK), câncer hepático humano (HepG2), câncer cervical humano (HeLa), câncer renal humano (786-0) e câncer colorretal humano (HT-29).

4.1.2 Plasma sanguíneo humano

As amostras de plasma sanguíneo humano foram obtidas através de mulheres voluntárias que habitam no estado do Rio Grande do Norte e frequentam a MEJC para atendimentos relacionados a patologias cervicais e exames de rastreamento, fornecidos

pelo SUS. A coleta de sangue em jejum foi realizada por venopunção no braço, anteriormente às consultas e exames médicos, em tubos contendo o anticoagulante EDTA. Após duas horas, o plasma foi coletado por gradiente de densidade e alíquotas foram transferidas para tubos criogênicos e armazenadas a -80ºC para análises posteriores.

4.2 Obtenção dos espectros

4.2.1 Fluorescência molecular

Espectros EEM das células em cultivo foram adquiridos através do espectrofluorímetro RF-5301 Shimadzu utilizando uma cubeta de quartzo com espessura de 0.5 mm. As larguras das fendas dos monocromadores de excitação e emissão foram mantidas em 1.5 e 3.0 nm, respectivamente. As superfícies espectrais EEM foram obtidas na região de 220 a 400 nm (incremento de 10 nm) para excitação e de 220 a 900 (incremento de 1 nm) para emissão, resultando em uma matriz de dimensão 19x680 para cada amostra.

4.2.2 ATR-FTIR

Espectros ATR-FTIR das amostras de plasma sanguíneo foram registrados a partir do espectrômetro FTIR Bruker Vertex 70 (Bruker Optics Ltd., Coventry, UK) com o acessório Helios ATR contendo um cristal de diamante, no ângulo de incidência de 45° do feixe de infravermelho. O instrumento foi configurado para realizar um total de 16 *scans* com resolução de 4 cm^{-1} para cada amostra e branco. Para cada amostra de plasma, 10 espectros foram coletados.

4.2.3 Espectrometria de massas

Os extratos lipídicos provenientes das amostras de plasma sanguíneo foram obtidos utilizando clorofórmio e metanol, conforme descrito pelo método Folch [68], e reconstituídos em isopropanol. As misturas foram analisadas através do espectrômetro de massas Q Quadrupole-Orbitrap Hybrid Exact (Thermo Scientific, Bremen, Germany) com ionização por eletrospray (ESI). As amostras foram injetadas diretamente na fonte de ionização e os espectros foram obtidos na faixa de m/z 200 a 1200, no modo positivo, com resolução de 140,000. Os espectros foram processados pelo software Xcalibur Analysis (version 2.0, Service Release 2, Thermo Electron Corporation).

4.3 Análise Computacional

Todo o tratamento computacional, desde a importação dos dados espectrais até a construção dos modelos multivariados, foi realizado através do software MATLAB versão 7.1 (MathWorks Inc., Natick, MA, USA) em conjunto com o pacote de algoritmos PLS-Toolbox versão 7.8 (Eigenvector Research Inc., Wenatchee, WA, USA). Diversos pré-processamentos foram aplicados em função do tipo de dado espectral utilizado, incluindo: correção de linha de base, normalização, correção de espalhamentos de luz (Rayleigh/Raman) e OPLS. Para construção dos modelos multivariados de classificação, os algoritmos (U)PLS-DA, PARAFAC, PCA-LDA/QDA, SPA-LDA/QDA, GA-LDA/QDA, PCA-SVM e GA-SVM foram utilizados.

Referências

[1] www.who.int.

[2] www.globocan.iarc.fr.

[3] www2.inca.gov.br.

[4] E.E. Calle, R. Kaaks, Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms, Nat. Rev. Cancer. 4 (2004) 579–591. doi:10.1038/nrc1408.

[5] C.M. de Oliveira, I.G. Bravo, N.C.S. e Souza, M.L.N.D. Genta, J.H.T.G. Fregnani, M. Tacla, J.P. Carvalho, A. Longatto-Filho, J.E. Levi, High-level of viral genomic diversity in cervical cancers: A Brazilian study on human papillomavirus type 16, Infect. Genet. Evol. 34 (2015) 44–51. doi:10.1016/j.meegid.2015.07.002.

[6] E.J. Nam, J.W. Kim, S.W. Kim, Y.T. Kim, J.H. Kim, B.S. Yoon, N.H. Cho, S. Kim, The expressions of the Rb pathway in cervical intraepithelial neoplasia; predictive and prognostic significance, Gynecol. Oncol. 104 (2007) 207–211. doi:10.1016/j.ygyno.2006.07.043.

[7] F. Cannella, A. Pierangeli, C. Scagnolari, G. Cacciotti, G. Tranquilli, P. Stentella, N. Recine, G. Antonelli, TLR9 is expressed in human papillomavirus-positive cervical cells and is overexpressed in persistent infections, Immunobiology. 220 (2015) 363–368. doi:10.1016/j.imbio.2014.10.012.

[8] S. Franceschi, S. Vaccarella, Beral's 1974 paper: A step towards universal prevention of cervical cancer, Cancer Epidemiol. 39 (2015) 1152–1156. doi:10.1016/j.canep.2015.10.019.

[9] B.F. Lees, B.K. Erickson, W.K. Huh, Cervical cancer screening: Evidence behind the guidelines, Am. J. Obstet. Gynecol. 214 (2016) 438–443. doi:10.1016/j.ajog.2015.10.147.

[10] C.P. Jenkins, DCrum, C.M. McLachlin, Cervical intraepithelial neoplasia., J. Cell. Biochem. 23 (1995) 71–79.

[11] J. Paavonen, Human papillomavirus infection and the development of cervical cancer and related genital neoplasias, Int. J. Infect. Dis. 11 (2007) S3–S9. doi:10.1016/S1201-9712(07)60015-0.

[12] T. Fascina, M.H.R. Oliveira, eds., Diretrizes brasileiras para o rastreamento do câncer do colo do útero, Second Edi, INCA, Rio de Janeiro, 2016.

[13] R. Nayar, D.C. Wilbur, The Pap test and Bethesda 2014, Acta Cytol. 59 (2015) 121–132. doi:10.1002/cncy.21521.

[14] N.C. Purandare, J. Trevisan, I.I. Patel, K. Gajjar, A.L. Mitchell, G. Theophilou, G. Valasoulis, M. Martin, G. von Bünau, M. Kyrgiou, E. Paraskevaidis, P. L Martin-Hirsch, W.J. Prendiville, F.L. Martin, Exploiting biospectroscopy as a novel screening tool for cervical cancer: towards a framework to validate its accuracy in a routine clinical setting, Bioanalysis. 5 (2013) 2697–2711. doi:10.4155/bio.13.233.

[15] G. Theophilou, M. Paraskevaidi, K.M. Lima, M. Kyrgiou, P.L. Martin-Hirsch, F.L. Martin, Extracting biomarkers of commitment to cancer development: potential role of vibrational spectroscopy in systems biology, Expert Rev. Mol. Diagn. 15 (2015) 693–713. doi:10.1586/14737159.2015.1028372.

[16] M. Monici, Cell and tissue autofluorescence research and diagnostic applications, Biotechnol. Annu. Rev. 11 (2005) 227–256. doi:10.1016/S1387-2656(05)11007-2.

[17] S. Andersson-Engels, C. af Klinteberg, K. Svanberg, C. Svanberg, In vivo fluorescence imaging for tissue diagnostics, Phys. Med. Biol. 42 (1997) 815–824. doi:10.1088/0031-9155/42/5/006.

[18] D.C.G. De Veld, M.J.H. Witjes, H.J.C.M. Sterenborg, J.L.N. Roodenburg, The status of in vivo autofluorescence spectroscopy and imaging for oral oncology, Oral Oncol. 41 (2005) 117–131. doi:10.1016/j.oraloncology.2004.07.007.

[19] M.A.M. Rodrigo, O. Zitka, S. Krizkova, A. Moulick, V. Adam, R. Kizek, MALDI-TOF MS as evolving cancer diagnostic tool: A review, J. Pharm. Biomed. Anal. 95 (2014) 245–255. doi:10.1016/j.jpba.2014.03.007.

[20] B. Flatley, P. Malone, R. Cramer, MALDI mass spectrometry in prostate cancer biomarker discovery, Biochim. Biophys. Acta. 1844 (2014) 940–949. doi:10.1016/j.bbapap.2013.06.015.

[21] K. Spalding, R. Board, T. Dawson, M.D. Jenkinson, M.J. Baker, A review of novel analytical diagnostics for liquid biopsies: spectroscopic and spectrometric serum profiling of primary and secondary brain tumors, Brain Behav. 6 (2016) 1–8. doi:10.1002/brb3.502.

[22] B. Chance, B. Thorell, Localization and kinetics of reduced pyridine nucleotide in living cells by microfluorometry, J. Biol. Chem. 234 (1959) 3044–3050.

[23] N. Ramanujam, Fluorescence Spectroscopy of Neoplastic and Non-Neoplastic Tissues, Neoplasia. 2 (2000) 89–117. doi:10.1038/sj.neo.7900077.

[24] K.M.G. Lima, K.B. Gajjar, P.L. Martin-Hirsch, F.L. Martin, Segregation of ovarian cancer stage exploiting spectral biomarkers derived from blood plasma or serum

analysis: ATR-FTIR spectroscopy coupled with variable selection methods, Biotechnol. Prog. 31 (2015) 832–839. doi:10.1002/btpr.2084.

[25] M.J. Walsh, M.J. German, M. Singh, H.M. Pollock, A. Hammiche, M. Kyrgiou, H.F. Stringfellow, E. Paraskevaidis, P.L. Martin-Hirsch, F.L. Martin, IR microspectroscopy: potential applications in cervical cancer screening, Cancer Lett. 246 (2007) 1–11. doi:10.1016/j.canlet.2006.03.019.

[26] G. Ramer, B. Lendl, Attenuated Total Reflection Fourier Transform Infrared Spectroscopy, 2013. doi:10.1002/9780470027318.a9287.

[27] J.R. Lakowics, Principles of Fluorescence Spectroscopy, (2006) 954.

[28] C.W. Huck, Y. Ozaki, V.A. Huck-Pezzei, Critical Review upon the Role and Potential of Fluorescence and Near-Infrared Imaging and Absorption Spectroscopy in Cancer Related Cells, Serum, Saliva, Urine and Tissue Analysis., Curr. Med. Chem. 23 (2016) 3052–3077. doi:10.2174/0929867323666160607.

[29] E.P. Diamandis, Mass Spectrometry as a Diagnostic and a Cancer Biomarker Discovery Tool: Opportunities and Potential Limitations, Mol. Cell. Proteomics. 3 (2004) 367–378. doi:10.1074/mcp.R400007-MCP200.

[30] A.L. Mitchell, K.B. Gajjar, G. Theophilou, F.L. Martin, P.L. Martin-Hirsch, Vibrational spectroscopy of biofluids for disease screening or diagnosis: Translation from the laboratory to a clinical setting, J. Biophotonics. 7 (2014) 153–165. doi:10.1002/jbio.201400018.

[31] L. Yi, N. Dong, Y. Yun, B. Deng, D. Ren, S. Liu, Y. Liang, Chemometric methods in data processing of mass spectrometry-based metabolomics: A review, Anal. Chim. Acta. 914 (2016) 17–34. doi:10.1016/j.aca.2016.02.001.

[32] R. Madsen, T. Lundstedt, J. Trygg, Chemometrics in metabolomics-A review in human disease diagnosis, Anal. Chim. Acta. 659 (2010) 23–33. doi:10.1016/j.aca.2009.11.042.

[33] R. Gautam, S. Vanga, F. Ariese, S. Umapathy, Review of multidimensional data processing approaches for Raman and infrared spectroscopy, EPJ Tech. Instrum. 2 (2015) 1–38. doi:10.1140/epjti/s40485-015-0018-6.

[34] A. Kottowska, Application of chemometric techniques in search of clinically applicable biomarkers of disease, Drug Dev. Res. 75 (2014) 283–290. doi:10.1002/ddr.21213.

[35] A.J. Lawaetz, R. Bro, M. Kamstrup-Nielsen, I.J. Christensen, L.N. J??rgensen, H.J. Nielsen, Fluorescence spectroscopy as a potential metabonomic tool for early

detection of colorectal cancer, Metabolomics. 8 (2012) 111–121. doi:10.1007/s11306-011-0310-7.

[36] L. Norgaard, G. Soletormos, N. Harrit, M. Albrechtsen, O. Olsen, D. Nielsen, K. Kampmann, R. Bro, Identification of genes involved in radiation-induced G 1 arrest y, J. Chemom. 21 (2007) 451–458. doi:10.1002/cem.

[37] N.C. Purandare, I.I. Patel, K.M.G. Lima, J. Trevisan, M. Ma'Ayeh, A. McHugh, G. Von Bünau, P.L. Martin Hirsch, W.J. Prendiville, F.L. Martin, Infrared spectroscopy with multivariate analysis segregates low-grade cervical cytology based on likelihood to regress, remain static or progress, Anal. Methods. 6 (2014) 4576–4584. doi:10.1039/C3AY42224K.

[38] D.A. Skoog, F.J. Holler, T.A. Nieman, Princípios de Análisis Instrumental, Quinta edi, Mc, Madrid, 2001. doi:10.1017/CBO9781107415324.004.

[39] M. Sauer, J. Hofkens, J. Enderlein, Handbook of Fluorescence Spectroscopy and Imaging: From Esemble to Single Molecules, Wiley, 2011.

[40] D.L. PAvia, G.M. Lampman, G.S. Kriz, J. Vyvyan, Introdução à Espectroscopia, Traduçlão, Cengage Learning, São Paulo, 2010.

[41] X. Han, A. Aslanian, J.R. Yates, Mass spectrometry for proteomics, Curr. Opin. Chem. Biol. 12 (2008) 483–490. doi:10.1016/j.cbpa.2008.07.024.

[42] S.J. Blanksby, T.W. Mitchell, Advances in Mass Spectrometry for Lipidomics, Annu. Rev. Anal. Chem. 3 (2010) 433–465. doi:10.1146/annurev.anchem.111808.073705.

[43] F. Girolamo, I. Lante, M. Muraca, L. Putignani, The Role of Mass Spectrometry in the "Omics" Era, Curr. Org. Chem. 17 (2013) 2891–2905. doi:10.2174/13852728178881311181627 25.

[44] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. doi:10.1039/c3ay41907j.

[45] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philos. Trans. R. Soc. A. 374 (2016) 20150202. doi:10.1098/rsta.2015.0202.

[46] M. Ringnér, What is principal component analysis?, Nat Biotechnol. 26 (2008) 303–304. doi:10.1038/nbt0308-303.

[47] C.A.D. Melo, P. Silva, A.A. Gomes, D.D.S. Fernandes, G. Vérasb, A.C.D. Medeiros, Classification of tablets containing dipyrone, caffeine and orphenadrine by

near infrared spectroscopy and chemometric tools, J. Braz. Chem. Soc. 24 (2013) 991-997. doi: 10.5935/0103-5053.20130127

[48] M.W. Mwadulo, A Review on Feature Selection Methods For Classification, A review on feature selection methods for classification tasks, Int. J. Computer Applicat. Technol. Res. 5 (2016) 395-402.

[49] R. Leardi, Genetic algorithms in chemometrics and chemistry: a review, J. Chemometrics 15 (2001) 559–569. doi: 10.1002/cem.651

[50] C. Huang, C. Wang, A GA-based feature selection and parameters optimization for support vector machines, Expert Systems with Applications 31 (2006) 231-240. doi:10.1016/j.eswa.2005.09.024

[51] C. Sukawattanavijit, J. Chen, H. Zhang, GA-SVM algorithm for improving land-cover classification using SAR and optical remote sensing data, IEEE Geoscience and Remote Sensing Letters 14 (2017) 284-288.doi: 10.1109/LGRS.2016.2628406

[52] C. B. Lucasus, G. Kateman, Gates Towards Volutionary Large-scale optimization: a software-oriented approach to genetic algorithms-I. general perspective, Computers Chem. 18, (1994) 127-136. doi: 10.1016/0097-8485(94)85006-2

[53] A.S. Soares, A.R. Galvão Filho, R.K.H. Galvão, M.C.U. Araújo, Improving the computational efficiency of the successive projections algorithm by using a sequential regression implementation: a case study involving NIR spectrometric analysis of wheat samples, J. Braz. Chem. Soc. 21 (2010) 760-763. doi: 10.1590/S0103-50532010000400024.

[54] S.F.C. Soares, R.K.H. Galvão, M.J.C. Pontes, M.C.U. Araújo, A new validation criterion for guiding the selection of variables by the successive projections algorithm in classification problems, J. Braz. Chem. Soc. 25 (2014) 176-181. doi: 10.5935/0103-5053.20130262.

[55] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), J. Chemom. 16 (2002) 119–128. doi:10.1002/cem.695.

[56] T. Verron, R. Sabatier, R. Joffre, Some theoretical properties of the O-PLS method, J. Chemom. 18 (2004) 62–68. doi:10.1002/cem.847.

[57] B. Wang, G. Liu, Q. Fei, Y. Ren, Orthogonal projection to latent structures combined with artificial neural networks in non-destructive analysis of Ampicilin powder, Spectrochim. Acta Part A. 71 (2009) 1695–1700. doi:10.1016/j.saa.2008.06.021.

[58] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38

(1997) 149–171. doi:10.1016/S0169-7439(97)00032-4.

[59] M.M. Sena, M.G. Trevisan, R.J. Poppi, PARAFAC: Uma ferramenta quimiométrica para tratamento de dados multidimensionais. Aplicações na determinaçãoo direta de fármacos em plasma humano por espectrofluorimetria, Quim. Nova. 28 (2005) 910–920. doi:10.1590/S0100-40422005000500032.

[60] S.K. Schmitz, P.P. Hasselbach, B. Ebisch, A. Klein, G. Pipa, R.A.W. Galuske, Application of Parallel Factor Analysis (PARAFAC) to electrophysiological data, Front. Neuroinform. 8 (2015) 1–10. doi:10.3389/fninf.2014.00084.

[61] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: Taking the magic away, J. Chemom. 28 (2014) 213–225. doi:10.1002/cem.2609.

[62] R. da Silva Fernandes, F.S.L. da Costa, P. Valderrama, P.H. Março, K.M.G. de Lima, Non-destructive detection of adulterated tablets of glibenclamide using NIR and solid-phase fluorescence spectroscopy and chemometric methods, J. Pharm. Biomed. Anal. 66 (2012) 85–90. doi:10.1016/j.jpba.2012.03.004.

[63] W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D.L. Massart, S. Heuerding, F. Erni, Comparison of regularized discriminant analysis, linear discriminant analysis and quadratic discriminant analysis, applied to NIR data, Anal. Chim. Acta. 329 (1996) 257–265. doi:10.1016/0003-2670(96)00142-0.

[64] S.J. Dixon, R.G. Brereton, Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on, Chemom. Intell. Lab. Syst. 95 (2009) 1–17. doi:10.1016/j.chemolab.2008.07.010.

[65] R.G. Brereton, G.R. Lloyd, Re-evaluating the role of the Mahalanobis distance measure, J. Chemom. 30 (2016) 134–143. doi:10.1002/cem.2779.

[66] H. Li, Y. Liang, Q. Xu, Support vector machines and its applications in chemistry, Chemom. Intell. Lab. Syst. 95 (2009) 188–198. doi:10.1016/j.chemolab.2008.10.007.

[67] J. Luts, F. Ojeda, R. Van de Plas Raf, B. De Moor, S. Van Huffel, J.A.K. Suykens, A tutorial on support vector machine-based methods for classification problems in chemometrics, Anal. Chim. Acta. 665 (2010) 129–145. doi:10.1016/j.aca.2010.03.030.

[68] R.E. Patterson, A.J. Ducrocq, D.J. McDougall, T.J. Garrett, R.A. Yost, Comparison of blood plasma sample preparation methods for combined LC-MS

lipidomics and metabolomics, J. Chromatogr. B. 1002 (2015) 260–266. doi:10.1016/j.jchromb.2015.08.018.

Capítulo 2

The use of EEM fluorescence data and OPLS/UPLS-DA algorithm to discriminate between normal and cancer cell lines: a feasibility study

Ana C. O. Neves

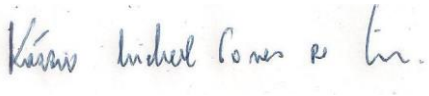
Raimundo F. A. Júnior Aurigena A. de Araújo

Ana L. C. S. L. Oliveira Kássio M. G. de Lima

Analyst, 2014, 139, 2423-2431.

Contribuição:

- Realizei a aquisição espectral;
- Realizei o processamento dos dados e construção dos modelos multivariados;
- Escrevi a primeira versão do manuscrito.



Ana C. O. Neves Prof. Kássio M. G. Lima

Analyst

PAPER

View Article Online
View Journal | View Issue

Cite this: *Analyst*, 2014, **139**, 2423

The use of EEM fluorescence data and OPLS/UPLS-DA algorithm to discriminate between normal and cancer cell lines: a feasibility study

Ana Carolina de Oliveira Neves,[a] Raimundo Fernandes de Araújo Júnior,[b] Ana Luiza Cabral de Sá Leitão Oliveira,[c] Aurigena Antunes de Araújo[c] and Kássio Michell Gomes de Lima*[a]

Excitation emission matrix (EEM) fluorescence spectroscopy combined with the OPLS method has been investigated as a promising tool to discriminate between normal and cancer cell lines in two datasets: (i) using several types of normal and cancer cells (including 3T3, ARPE, HEK, HepG2, HeLa, HT-29 and 786-0 cells); (ii) considering the expression of matrix metalloproteinase-2 and -9 (MMP-2 and MMP-9) in suspensions of HEK and 786-0 cell lines. Partial Least Squares-Discriminant Analysis (PLS-DA) using the score matrix from PARAFAC (Parallel Factor Analysis), UPLS-DA (Unfolded Partial Least Squares with Discriminant Analysis) and orthogonal projection to latent structures (OPLS) were used as the bases for the discrimination models. UPLS-DA presented relevant performance for cancer cells in both datasets, with 100% and 66.7% correct prediction for first and second cases, respectively, and poor discrimination relative to normal cells in the first dataset (25%). By using the OPLS, we achieved 75% correct prediction for normal cells and maintained 100% concordance for cancer objects. On applying OPLS to the second dataset, we obtained 100% correct prediction in both classes (normal and cancer) for calibration and prediction sets. These results suggest that EEM fluorescence spectroscopy combined with chemometrics could be used as a clinical tool for cancer cell detection based on intrinsic biomolecular signatures.

Received 10th February 2014
Accepted 13th March 2014

DOI: 10.1039/c4an00296b

www.rsc.org/analyst

Introduction

Cancer is one of the most frequent illnesses of the 21st century, leading to high death indexes mainly among the non-communicable diseases. This fact has generated progressive interest in many fields of science, and in recent years, metabolomics has expressly been dedicated to improve the comprehension of the "metabolic transformations" that occur in tumor cells.[1,2] The major difference between normal and neoplastic cells concerns the continuous growth and proliferation of tumor cells, even in the absence of growth factors. It has been suggested that the vast catalog of cancer cell genotypes is actually a manifestation of the 7 essential alterations in cell physiology that collectively dictate malignant growth: self-sufficiency in growth signals, avoidance of immuno-surveillance, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis.[3,4] Although all five major classes of proteases (serine, aspartic, cysteine, threonine, and metalloprotease) are involved in invasion and metastasis, matrix metalloproteinases (MMPs) in particular have been implicated in the degradation of various ECM components. Increased secretion of MMPs has been associated with tumour promotion, progression, invasion, angiogenesis, and metastasis.[5,6]

The current "gold standard" approach for cell detection is flow cytometry, which provides qualitative and quantitative information about each cell, allowing the identification of different subpopulations of normal or tumor cells.[7] However, flow cytometry exhibits a low cell-throughput rate. For example, high-speed sorters are only capable of a flow rate of less than a few thousand cells per second. This low throughput is a major disadvantage, given that many experiments require a very large number of cells.[8] On the other hand, spectroscopic methods have been applied as an efficient metabolomic tool in several works involving cancer detection.[2] Spectroscopic approaches do not require specific reagents and allow the possibility of nondestructive and noninvasive measurements. Nuclear magnetic resonance (NMR)[9] mass spectrometry (MS)[10] near-infrared (NIR)[11] attenuated total reflectance-Fourier transform infrared (ATR-FTIR)[12] and Raman spectroscopy (RS) methods[13]

[a]*UFRN-IQ-PPGQ, Biological Chemistry and Chemometrics, Natal, RN, Brazil. E-mail: kassiolima@gmail.com; Tel: +55 84 3342-2323*

[b]*Post Graduation Program in Functional and Structural Biology/Post Graduation Program Health Science/Department of Morphology, UFRN, Natal, RN, Brazil*

[c]*Post Graduation Program Public Health/Post Graduation Program in Pharmaceutical Science/Department of Biophysics and Pharmacology, UFRN, Natal, RN, Brazil*

are often applied to distinguish cancer cells, tissue, or blood. Successful use of these spectroscopic techniques requires the application of pre-processing and unsupervised or supervised chemometric algorithms. Among the generally used pre-processing methods, especially for purposes of classification, stands out the orthogonal projections to latent structures (OPLS).[14] Unsupervised methods that have been widely used in analyzing cell or tissue spectra include principal component analysis (PCA)[15] and hierarchical clustering.[16] Supervised methods include soft independent modeling class analogy (SIMCA),[17] partial least squares-discriminant analysis (PLS-DA)[18] and cluster analysis (CA).[9]

In the present report, we explored the possibility of introducing fluorescence spectroscopy of cell lines as an alternative tool for cancer research, considering that this technique has some advantages compared to flow cytometry, such as minimal sample preparation, which is limited to sample dilution, and reduced acquisition time, which is as low as a few minutes, depending on the spectral area and integration time. For this purpose, we applied Fluorescence Spectroscopy (FS) using a metabolomic approach, aiming to classify different cell lines based on their metabolic changes, leading to the possibility of using a simple analytical procedure as an alternative to sophisticated ones that are often used in metabolomics, such as NMR and MS, especially when a fast response is needed.

Several groups have applied fluorescence spectroscopy to identify spectral features corresponding to metabolic changes and to discriminate cancer and non-cancer cell lines, as done by Britton Chance in a pioneer study.[19] The main focus of these approaches is to associate spectral differences as a consequence of pathological processes occurring in the cells, so that it may be possible to compare healthy and damaged organisms by analyzing their spectral profiles.[20] In this context, the application of multivariate analysis can be considered a powerful tool to improve classifications and the comprehension of biological processes involved, and we felt that its use is less described in the literature for cultured cells, except for studies using other kinds of biological samples, such as tissues and blood plasma. Lawaetz and co-workers[21] employed FS Excitation Emission Matrix (EEM) measurements on human blood plasma samples from a case control study on colorectal cancer. The objective of their study was to explore the possibilities for applying FS as a tool for detection of colorectal cancer. Nørgaard and co-workers[22] investigated whether FS of serum samples in combination with Extended Canonical Variates Analysis (ECVA) can be used to discriminate healthy females from breast cancer patients with solitary and multiple metastases. The FS results were compared with results based on the three tumor markers cancer antigen 15-3 (CA 15-3), carcinoembryonic antigen (CEA), and tissue polypeptide antigen (TPA).

However, our goal consists of investigating FS combined with multivariate analysis to discriminate cancer and normal objects of two datasets: (i) a dataset obtained from several types of cells (rat fibroblasts – 3T3; human retina – ARPE; human kidney cells – HEK; liver cancer cells – HepG2; cervical cancer cells – HeLa; colorectal cancer cells – HT-29; kidney cancer cells – 786-0), (ii) considering the expression of matrix metalloproteinase-2 and -9 (MMP-2 and MMP-9) in suspensions of human kidney normal cells – HEK and kidney cancer cells – 786-0. We demonstrated the use of an excitation emission matrix (EEM) to obtain spectra for normal and cancer cells in suspension. Unsupervised (PCA) and supervised (PLS-DA) pattern recognition methods were applied to extract spectral features and develop the classification models. Lastly, the focus of the present study was also to investigate how orthogonal projections to latent structures (OPLS)/PLS-DA could be applied to extract information from complex spectroscopy fingerprints. We conducted a comparison between the loading profile and pure standards for identified/confirmed possible biomarkers based on their fingerprints.

Experimental section

Pure standards

The reagents used in cell culture are listed as follows: Dulbecco's modified Eagle's medium (DMEM; Life Technologies, Inc., Grand Island, NY, USA) supplemented with 10% (v/v) heat-inactivated fetal bovine serum (FBS Life Technologies, Inc., Grand Island, NY, USA). Phosphate-buffered saline (PBS, pH 7.2 ± 0.1, LB Laborclin, PR Brazil), trypsin (MW: 23.3 kDa, SIGMA, SP Brazil), sodium azide 0.1% solution (Sigma, SP, Brazil), Trypan blue (Sigma, SP, Brazil), primary antibodies (rabbit polyclonal, MMP-2 and MMP-9, purchased from Santa Cruz Biotechnology, SP, Brazil), Alexa Fluor 488 goat anti-rabbit secondary antibody (purchased from Santa Cruz Biotechnology, SP, Brazil), and bovine serum albumin (BSA 1%; Life Technologies do Brazil LTDA, São Paulo, Brazil).

The following pure standard samples were used: acetyl-CoA (≥93%), adenosine diphosphate (ADP, ≥95%), adenosine monophosphate (AMP, ≥99%), adenosine triphosphate (ATP, ≥99%), glutathione oxidized (≥98%), glutathione reduced (≥98%), guanosine triphosphate (GTP, ≥98%) and nicotinamide adenine dinucleotide (NADH, ≥98%). All pure standards were purchased from Sigma-Aldrich. Pure standard samples were dissolved by shaking and stirring at room temperature under subdued lighting conditions using transparent solvents recommended by the supplier. The final concentration of pure standards was chosen to match 3.08×10^{-5} mol mL^{-1}. EEM data were collected immediately thereafter.

Cell preparation – dataset 1

Cell samples were selected from the Culture Collection of the Department of Morphology, Federal University of Rio Grande do Norte, Brazil. Normal cell lines included rat fibroblasts (3T3), human retina (ARPE), and human kidney cells (HEK). Human cancer cell lines included liver (HepG2), cervical (HeLa), colorectal (HT-29), and kidney (786-0) cells. All cell lines were maintained in Dulbecco's modified Eagle's medium (DMEM; Life Technologies, Inc., Grand Island, NY, USA) supplemented with 10% (v/v) heat-inactivated fetal bovine serum.

Cells were washed twice with PBS and harvested by trypsinization when they reached approximately 80% confluence. They were then immediately neutralized by the culture medium.

After this procedure, the cells were centrifuged and the supernatant was removed. The cell pellet was resuspended in 2 mL of DMEM and counted using a Neubauer chamber by means of the Trypan blue exclusion test to evaluate the total number of living cells. Cells were kept alive, preserving their cytoskeleton and organelles, in a culture medium for at least 60 minutes on ice or at a temperature of −4 °C.[23]

In this dataset, twenty-one cell samples were divided in two groups. Class 1 (C1) consisted of 12 cancer cell samples (6 calibration and 6 prediction) and class 2 (C2) consisted of 9 normal cell samples (5 calibration and 4 prediction).

Cell preparation – dataset 2

Cultured cells (HEK and 786-0) were plated on glass coverslips in 24-well plates (1×10^5 cells per well) in triplicate and grown for 3–7 days. After 24 hours, they were harvested by trypsinization, immediately neutralized by the culture medium and centrifuged, and the supernatant was then removed. Six samples from 786-0 cells and six samples from HEK cells were obtained. The cells were then incubated for at least 1 hour at 4 °C in a solution consisting of 200 μL of PBS, 10% FBS, 0.1% azide and 1 μL of primary antibodies (MMP-2 and MMP-9). Sections were washed gently, with 1–3 washes for 10 minutes each in 300–170 μL frozen PBS with 0.5% BSA. The primary antibody was detected with a Alexa Fluor 488 goat anti-rabbit secondary antibody diluted 1 : 3 in PBS containing BSA for 30 minutes at 4 °C. After this last step, cells were centrifuged, washed twice in frozen PBS and resuspended in 300 μL of PBS containing 10% FBS and 0.1% azide.

In this dataset, the twelve cell samples were divided in two groups. Class 1 (C1) consisted of 6 cancer cell samples (3 calibration and 3 prediction), and class 2 (C2) consisted of 6 normal cell samples (3 calibration and 3 prediction).

Fluorescence spectroscopy of cell culture samples and pure standards

This procedure was performed for obtaining the individual spectra of cells of both datasets. Spectra of cells and pure standards were acquired with a RF-5301 Shimadzu spectrofluorometer through a 0.5 mm quartz cuvette. The excitation and emission monochromator slit widths were fixed at 1.5 and 3 nm, respectively. The cells and pure standards were introduced into the cuvette with a disposable 1 mL syringe. After each measurement, the cell was cleaned by a sequence of 1 mol L^{-1} acetic acid, ultra-pure water, and acetone to dry the cell. The temperature was kept at 25 °C throughout the experiments. The emission values obtained for pure standards were acquired at an excitation of 290 nm and an emission range from 300 to 800 nm (1 nm steps). For cell culture samples, the spectral surfaces of emission/excitation were obtained in the excitation range from 220 to 400 nm (10 nm steps) and in the emission range from 220 to 900 nm (1 nm steps). This protocol resulted in a data matrix size of 19×680 for each cell.

Chemometrics

PCA. A dataset containing many variables can be simplified by data reduction, which makes the system more easily interpretable. PCA[15] is a well-known method to reduce the number of variables. In this method, a spectral matrix **X** is decomposed as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{t} + \mathbf{E}, \quad (1)$$

where **X** is the $I \times J$ data matrix; **T** is the $I \times A$ matrix of orthogonal score vectors $\mathbf{t}_a$, such that $\mathbf{T}^t\mathbf{T} = \mathrm{diag}(\lambda_a)$ and λ_a are eigenvalues of the matrix $\mathbf{X}^t\mathbf{X}$; P is the $J \times A$ matrix of loading vectors; E is the $I \times J$ residual matrix; I and J are the number of objects and variables, respectively; and A is the number of calculated components.

OPLS. Orthogonal projection to latent structures is a multivariate preprocessing method briefly described in this paper. This algorithm is an extension of traditional PLS regression that integrates an orthogonal signal correction filter to distinguish the variations in the data that are or are not useful for prediction.[14] The first steps are equivalent to the regular NIPALS method for a single **y** vector. After that, the scores and loadings of orthogonal variations are calculated so that the **X** filtered matrix, $\mathbf{X}_{\mathrm{OPLS}}$, is obtained:

$$\mathbf{X}_{\mathrm{OPLS}} = \mathbf{X} - \mathbf{t}_{\mathrm{orto}}\mathbf{p}^{\mathbf{T}}_{\mathrm{orto}} \quad (2)$$

where $\mathbf{t}_{\mathrm{orto}}$ and $\mathbf{p}^{\mathbf{T}}_{\mathrm{orto}}$ are the scores and loadings of orthogonal variation, respectively.

We applied OPLS to unfolded EEM fluorescence data of normal and cancerous cells previously to build UPLS-DA classification models. In this study, the objective of using OPLS consisted of distinguishing the **X** variation that is correlated to the response set **Y** (classes of samples) of the systematic variation that is orthogonal to **Y**. In mathematical terms, this is comparable to removing systematic variation in **X** that is orthogonal to **Y**.

PLS-DA. In this study, the PLS-DA method was used to differentiate between the 2 distinct cell types (normal *vs.* cancer cells). The PLS-DA method comes from the adaptation of PLS regression for pattern recognition. PLS-DA is performed by an exclusive binary coding procedure. During the calibration process, the PLS-DA method is trained to compute the "membership values," with one for each class. The sample is assigned to one class when the value is above a specific prediction threshold.[24] This method, adapted from PLS1 or PLS2 regressions, uses M spectral variables as predictors and q variables (0 or 1) as the variable response.[25]

The PLS-DA approach requires that the data of interest be split into 2 independent datasets: calibration and validation. Cross-validation is performed with random subsets; each test set is formed from a random selection of objects, such that a single object is included in only one test set. In this way, the prediction error is said to be representative of that of the new samples. The number of latent variables selected for each model is determined by using the smallest root mean square error of prediction (RMSEP).

UPLS-DA. The UPLS model consists of the first-order calibration PLS model when it is applied to second-order data that have been unfolded into vectors, $\mathbf{V}_x$. A usual PLS model is built with the $\mathbf{V}_x$ (JK, I) matrix. This model provides a set of loading P and weight loading W for both sizes (JK, A), where A is the number of latent variables.[26]

In the present study, the UPLS-DA model was built by unfolding the original data (*i.e.*, the matrix containing the fluorescence for each sample). The array of dependent variables contained the sample classes for the different cell samples (cancer and normal). A linear relationship was established in this model between the scores of **X** (*i.e.*, the vectorized matrix values for fluorescence of each sample) and the scores of **Y** (*i.e.*, the matrix containing the different class types).

Parallel factor analysis (PARAFAC)

Second-order data are obtained when a given sample produces a $J \times K$ data matrix or second-order array, where J and K denote the number of data points in the first and second dimensions, respectively. In EEM fluorescence measurements, J is the number of digitized emission wavelengths and K is the number of excitation wavelengths. If the I training matrices and the unknown sample matrix are stacked, then a 3-way data array **X** is obtained, with dimensions $[(I + 1) \times J \times K]$. Provided that **X** follows the trilinear PARAFAC model,[27] it can be represented by the sum of the Kronecker products of the 3 vectors for each responsive component, a_n, b_n, and c_n, which collect the relative concentrations or scores $[(I + 1) \times 1]$, the emission profiles $(J \times 1)$ and the excitation profiles $(K \times 1)$ for component n, respectively. Therefore, the specific expression can be written as follows:

$$\underline{X} = \sum_{n=1}^{N} a_n \otimes b_n \otimes c_n + \underline{E} \quad (3)$$

where $\otimes$ indicates the well-known Kronecker product, N is the total number of responsive components, and $\underline{E}$ is a residual error term of the same dimensions as $\underline{X}$. The column vectors a_n, b_n, and c_n are usually collected into the loading matrices A, B, and C, respectively.

By using a classification strategy, in the present study, the scores of matrices from the PARAFAC models were merged into a common score matrix containing all of the quantitative information extracted from the fluorescence measurements.

Software

Data were imported and the chemometric models (PCA, OPLS, PLS-DA, and UPLS-DA) were constructed in MATLAB version 7.1 software (Math-Works, Natick, USA) by using the PLS-toolbox (version 7.5.2, Eigenvector Research, Inc., Wenatchee, WA, USA). The PARAFAC calculations were performed with the N-way toolbox for Matlab, version 2.10.

Results and discussion

First dataset

Fig. 1 presents the excitation/emission fluorescence spectra (EEM) after removing Rayleigh and Raman scatterings of one sample of kidney cancer cells (Fig. 1A) and normal kidney cells (Fig. 1B) in DMEM.

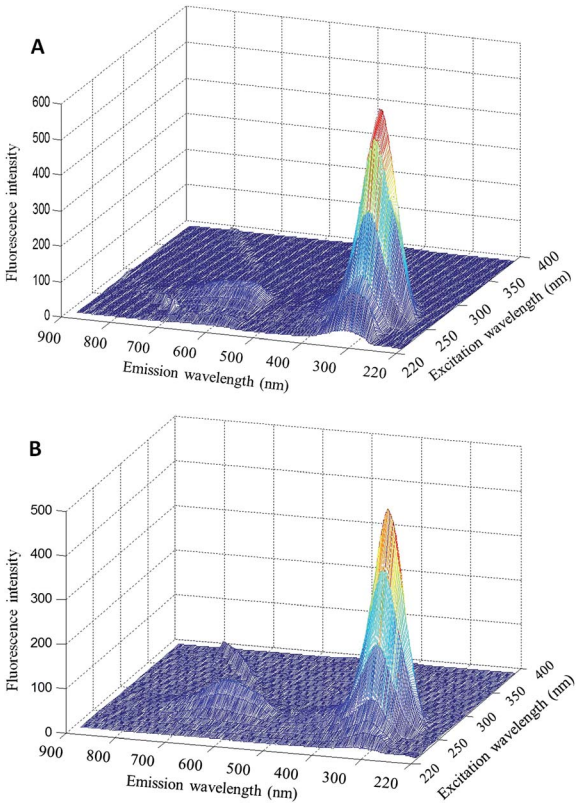


Fig. 1 Excitation–emission molecular fluorescence spectra obtained for both 786-0 cancer cells (A) and HEK normal cells (B) in DMEM. The Rayleigh and Raman scatterings have been removed from the spectra.

Most of the cell samples analyzed showed a well-defined and highly intense fluorescence area at an excitation range of 240–290 nm and emission range of 220–400 nm and a lower fluorescence in the 530–690 nm region. In the undiluted raw samples, the major peak verified in the fluorescence spectrum matched the region corresponding to endogenous fluorophores such as ADP, ATP, GTP, AMP, NADH, glutathione oxided, glutathione reduced, acetyl-CoA, structural proteins (collagen and elastin), amino acids (tryptophan, tyrosine, phenylalanine) and others.[28] In addition, there are a wide variety of organic compounds such as vitamins and lipids that may exhibit autofluorescence in the 530–690 nm region.[29]

A PCA analysis was performed on the EEM after removing the scatterings. Fig. 2 presents a direct comparison between normal and cancer cells. There was no separation or not enough separation between the two classes of samples (normal and cancer), with 97.31% of the explained variance seen in PC1 *versus* PC2. In fact, only 3 components were necessary to explain 99% of the variation.

Because no clear separation of cell types was found by the PCA analysis of the score matrix, we performed PLS-DA (score matrix from PARAFAC) and UPLS-DA analyses, using the unfolded matrix. The algorithm choice was based on the supervised character of the chemometric method. In advance to

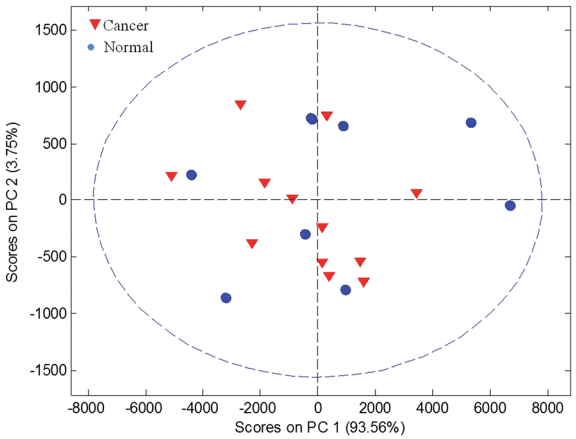


Fig. 2 Principal component analysis of all individual spectra from cancer (▼) and normal (●) cell samples.

multivariate model building, all twenty-one EEM spectra were preprocessed to remove the observed Rayleigh and Raman scatterings by using the eemscat algorithm (the excluded spectral regions were properly corrected by interpolation),[30] and then the classification models were built.

Among the twelve cancerous samples were the cell lines HepG2 (human liver hepatocellular carcinoma), HeLa (human cervical epithelial carcinoma), 786-0 (human renal cell adenocarcinoma) and HT-29 (human colon adenocarcinoma), with each contributing 3 samples. The normal cells were ARPE-19 (human retinal pigment epithelial), HEK-293 (human embryonic kidney 293) and 3T3 (rat fibroblast), and each cell line contributed 3 samples for a total of 9 objects. All twenty-one samples were divided into calibration and prediction sets, with 11 and 10 samples, respectively, by analyzing the PCA plot. In the calibration set, 6 samples represented cancer class (C) while 5 samples belonged to normal class (N). In the prediction set, cancer class accounts for 6 samples and normal class for other 4 samples.

Table 1 shows the main results obtained for models built to classify normal and cancer cell lines.

The UPLS-DA model used the unfolded matrix with a size of 21 × 12 939, while PARAFAC/PLS-DA only employed the five-factor score matrix from a previously constructed PARAFAC model. Considering the results of these two models, all the samples of cancer class were correctly classified in the calibration step. However, for both models, relevant difficulty was verified in classifying samples belonging to normal class, especially in the prediction phase. As an alternative aimed at decreasing the misclassifications related to normal class, the OPLS algorithm was applied to the dataset as a preprocessing method before building the UPLS-DA model. By using the orthogonalized unfolded matrix from OPLS, it was possible to observe a significant increase in the percentage of correct classification to normal class through the OPLS/UPLS-DA model when compared to the previous ones. In addition, this model was able to fully classify the cancer class samples and maintain the total concordance already reached with the previous models. It is important to mention that the only object of normal class that was misclassified in the prediction set was a rat fibroblast sample (that was represented by only 3 objects in the dataset), and thus, this result might not affect the model's ability to handle human cell samples. In classification of cases, OPLS identifies the variations in **X** that separates (correlated variation) and combines (orthogonal variation) the groups.

Metabolic transformations usually occur in cancer cells, such as uptake of glucose and glutamine and consequent rise in lactate production (Warburg effect); defective oxidative phosphorylation; inactivation of p53 protein; overexpression of transketolase-1 isoform (TKL-1); upregulation and activation of the enzymes fatty acid synthase (FASN), choline kinase (ChoK), ATP-citrate lyase (ACL), and acetyl-CoA carboxylase (ACC); activation and expression of matrix metalloproteinases (MMPs), nicotinamide adenine dinucleotide (NADH) and flavins, besides many other altered metabolic pathways that allow a tumor cell to duplicate its genome, proteins and lipids and assemble them to create daughter cells.[2,3] These transformations are the factors that allow OPLS to identify spectral contributions corresponding to each class of samples, thereby facilitating correct discrimination between normal and tumor cell lines. Since carcinogenesis and cancer biochemistry form a broad yet complex area, we have tried to identify some specific molecules that could be contributing to the overall EEM spectra differently after the OPLS/UPLS-DA model, sharpening the identification of biochemical differences.

In Fig. 3A, the first principal component loading profiles (p1-loading, black line) obtained by OPLS/UPLS-DA are plotted with eight pure standards (other colors), which correspond most

Table 1 Results obtained for classification models between several cancer (C) and normal (N) cell lines. The correct percentage of discrimination to each class is shown in parentheses

		Calibration		Prediction	
Model	Class	Cancer	Normal	Cancer	Normal
UPLS-DA (3)[a]	Cancer (C)	6 (100%)	0	6 (100%)	0
	Normal (N)	1	4 (80%)	3	1 (25%)
PARAFAC/PLS-DA (3)[a]	Cancer (C)	6 (100%)	0	6 (100%)	0
	Normal (N)	2	3 (60%)	4	0 (0%)
OPLS/UPLS-DA (3)[a]	Cancer (C)	6 (100%)	0	6 (100%)	0
	Normal (N)	0	5 (100%)	1	3 (75%)

[a] Number of latent variables.

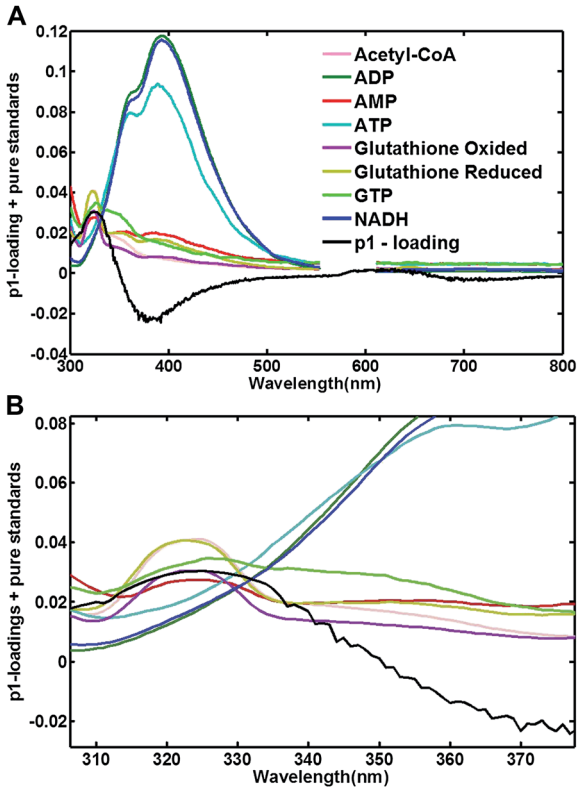


Fig. 3 (A) p1-loading and pure standard fluorescence emission plot, which indicate the metabolic differences between healthy and cancerous cell samples. (B) p1-loading and pure standard fluorescence emission plot at 305–380 nm. (C) p1-loading and pure standard fluorescence emission plot at 610–750 nm.

closely to endogenous fluorophores present in the cell constitution. The intensities of the spectra were normalized to enable comparison across all spectra, and the spectral ranges between 550 nm and 610 nm for pure standards were excluded due to the Rayleigh intensity. The p1-loading may be interpreted in relation to the more relevant spectral features of the components present in the analyzed cells. As can be seen in Fig. 3A and B, the p1-loading intensities presented higher weights at 320–330 nm (positive loadings) and 370–420 nm (negative loadings), respectively. In addition (Fig. 3C), the p1-loading presented slightly less intensity corresponding to 620–680 nm.

The p1-loading is in agreement with the fluorescence emission spectra of ADP (strong green line), ATP (cyan line) and NADH (blue line) pure standards at 410 nm but with an inverse relation. Our findings indicate a distinctive increase in the fluorescence intensity spectra for ADP, ATP and NADH in cancer cells compared to those in normal cells. These biomarkers are involved in energy metabolism (in general, shifting from aerobic respiration in normal cells to anaerobic respiration in cancer cells) and considered in the literature because of their diagnostic potential.[31] On the other hand, at 320 nm emission, the p1-loading is in agreement with the increased fluorescence emission spectra of acetyl-CoA (rose line), glutathione oxided (gold line), glutathione reduced (pink line), AMP (red line) and GTP (weak green line) pure standards. In addition, the differences observed at 320 nm emission, between normal and cancer cells could be attributed to tryptophan and micro-environmental changes in the amino acids.[32] This hypothesis indicates that cancer cells could consistently show higher fluorescence intensity than normal cells for these metabolites.[33] Further, the p1-loading presented a slight increase at 620–680 nm (Fig. 3C) where other metabolite standards such as acetyl-CoA, glutathione oxided and glutathione reduced also have shown increased fluorescence emission spectra. The differences observed at 620–680 nm emission could be also attributed to the intrinsic fluorescence of lipids and porphyrins present in cells.[20] The differences observed in the results obtained in both cells by application of OPLS/UPLS-DA were considered significant. However, a deeper interpretation of these small differences would require the study of a larger number of samples with better control of the variables that can influence these differences. Nevertheless, it is possible to conclude that at least in the analysis of pure standards (endogenous fluorophores) and in the analysis of the investigated normal and cancer cells, the OPLS/UPLS-DA method provided quantitative and qualitative information, including the loading spectra of the principal component, allowing their possible identification/confirmation.

Taking into account that renal cell carcinoma (RCC) is the most predominant malignancy of the kidneys and represents 2–3% of all cancers worldwide,[34,35] we decided to study the capacity of fluorescence spectroscopy and chemometric approaches to classify kidney cancer (786-0) and normal (HEK) cell lines based on the expression of matrix metalloproteinases MMP-2 and MMP-9. This is discussed in the "Second dataset" section, because it is well known that these zinc-containing endopeptidases are overexpressed in RCC and have a crucial role in tumor progression.

Second dataset

Fig. 4 shows the excitation/emission fluorescence spectra (EEM), after removing Rayleigh and Raman scattering, of one sample of kidney cancer cells (Fig. 4A) and normal kidney cells (Fig. 4B) with primary (rabbit polyclonal, MMP-2 and MMP-9) and secondary (Alexa Fluor 488 goat anti-rabbit) antibodies.

By analyzing the renal cancer cell EEM spectra with antibodies (primary and secondary), it is possible to observe a considerable increase in the fluorescence signal intensity in the whole spectrum but especially in the range 500–780 nm, compared to that in Fig. 1A, demonstrating that there were chemical interactions between the primary antibodies and the MMPs present into the cells. MMP-2 and MMP-9 reveal distinct structural domains. In tumor cells, they can be found in the membrane or cytoplasm according to their function and the degree of aggressiveness of these cells.[36]

As an initial exploratory analysis, PCA was performed using the twelve unfolded EEM spectra after removing the Rayleigh and Raman scatterings. Based on Fig. 5, it is possible to observe that 96.8% of the explained variance was covered by the two first principal components (PC1 and PC2).

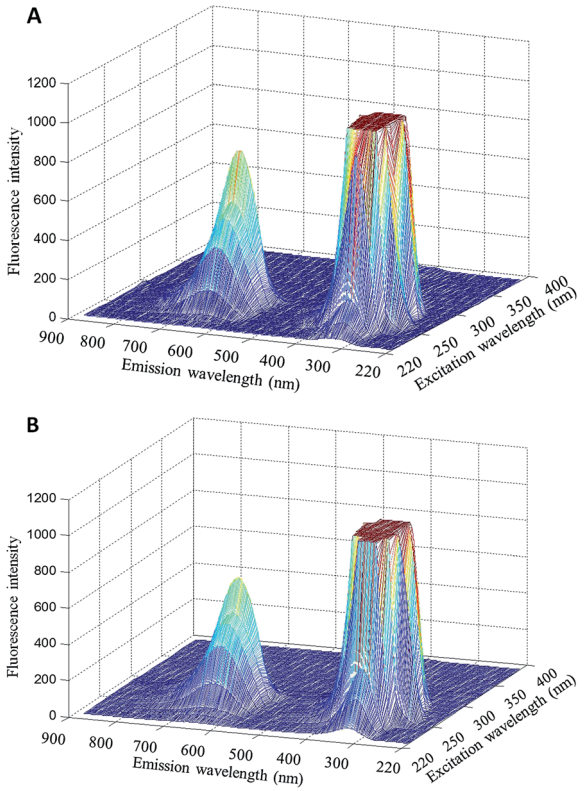


Fig. 4 Excitation–emission molecular fluorescence spectra obtained for both 786-0 cancer cells (A) and HEK normal cells (B) with primary antibodies (rabbit polyclonal, MMP-2 and MMP-9). The Rayleigh and Raman scatterings have been removed from the spectra.

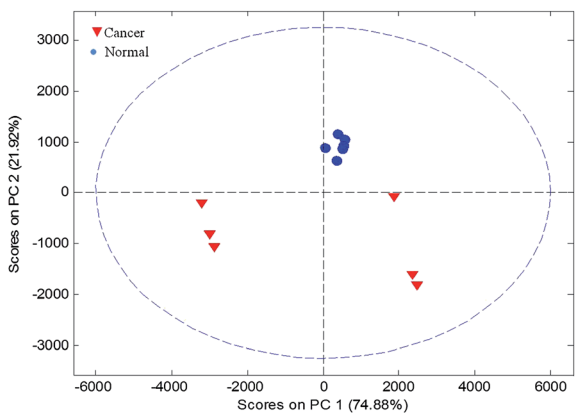


Fig. 5 Principal component analysis of 786-0 cancer cell (▼) and HEK normal cell (●) samples with primary antibodies.

From the PCA plot, it is remarkable that objects belonging to the normal class formed a well-defined cluster, while the samples of cancer class got divided into three groups wherein one object was expressively displaced toward a normal cluster. Several cells with malignant characteristics are found in a tumor. During several stages of cell division, numerous genes undergo mutations causing heterogeneous tumor cells with the capacity to proliferate, invade and metastasize. In fact, tumor cells from a human carcinoma alter their differentiation and function, for example normal human cells of the epithelial origin have the same morphology and function. These cells lose their cohesive-stationary characteristics and then exhibit a cohesive-migratory phenotype when their genes mutate. Cancer cells express important proteins, such as MMPs, for survival and invasion to the underlying tissues.

The dataset was divided into calibration and prediction sets, with 6 samples in each set, and a UPLS-DA classification model was built (Table 2).

As expected, the results of UPLS-DA are substantially in accordance with those visualized through the PCA, since the one object misclassified in prediction matches exactly the cancer sample displaced toward normal cells (PCA plot). Motivated by the improvements achieved through OPLS discussed in the First dataset, we applied this algorithm to preprocess the dataset and then build another UPLS-DA model (Table 3).

The obtained results demonstrate the OPLS's ability to find and distinguish the correlating chemical information provided by the interactions between the primary antibodies and the metalloproteinases MMP-2 and MMP-9 in X, hereby allowing 100% correct discrimination for both classes, cancer and normal, when calibrating or predicting the model. In addition, we have also tested PARAFAC/PLS-DA on this dataset in comparison to OPLS/UPLS-DA. The PARAFAC/PLS-DA model when used with two components as built for the OPLS/UPLS-DA model presents inferior performance for discriminating cancer and normal cells; however, the results are equivalent when five components are used in the former model.

In addition, to confirm our findings, we also have analyzed the first principal component loading profile (p1-loading, black line) obtained by the OPLS/UPLS-DA model adding pure standards, as shown in Fig. 6A. The p1-loading intensities presented higher weights at 550–780 nm (negative loadings) and a

Table 2 Results obtained for the classification model between 786-0 cancer cell (C) and HEK normal cell (N) samples. The correct percentage of discrimination to each class is shown in parentheses

		Calibration		Prediction	
Model	Class	Cancer	Normal	Cancer	Normal
UPLS-DA (2)[a]	Cancer (C)	3 (100%)	0	2 (66.7%)	1
	Normal (N)	0	3 (100%)	0	3 (100%)

[a] Number of latent variables.

Table 3 Results obtained for the classification model between 786-0 cancer cell (C) and HEK normal cell (N) samples after using the OPLS. The correct percentage of discrimination to each class is shown in parentheses

		Calibration		Prediction	
Model	Class	Cancer	Normal	Cancer	Normal
OPLS/UPLS-DA (1)[a]	Cancer (C)	3 (100%)	0	3 (100%)	0
	Normal (N)	0	3 (100%)	0	3 (100%)

[a] Number of latent variables.

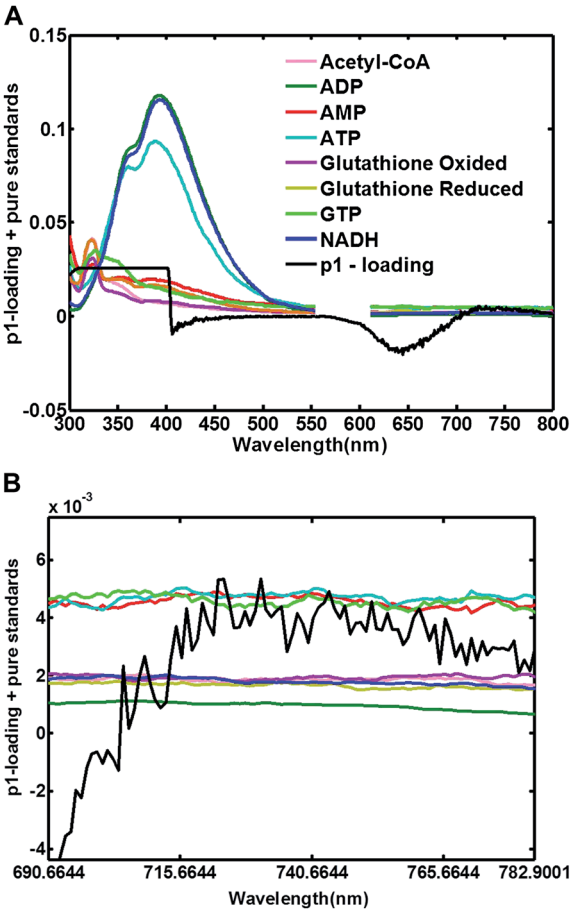


Fig. 6 (A) p1-loading and pure standard fluorescence emission plot, which indicate the metabolic differences between healthy and cancerous cell samples after addition of primary antibodies. (B) p1-loading and pure standard fluorescence emission plot at 690–780 nm.

saturated weight at 300–400 nm. The last interval range may be caused by high autofluorescence from many biomolecules. Nevertheless, the interval range between 690 nm and 780 nm (Fig. 6B) presents an interesting region for discrimination of cancer cells from normal cells. This finding may be evidence that there were chemical interactions between the primary antibodies and the MMPs present in the cells. Thus, these antibodies have specificities to the binding sites of metalloproteinases, and the p1-profile shows this effect. Regarding normal cells (HEK), it is well known that the actual concentrations of MMPs in healthy tissues are low or even undetectable;[37] nevertheless, there may be other interactions that cause an increase in fluorescence, as shown in Fig. 1B, besides contribution from the antibodies (that are quite fluorescent species). However, it is very unlikely that the interactions can be explained by merely using a simple technique like molecular fluorescence spectroscopy.

Conclusion

We introduced EEM fluorescence spectroscopy combined with multivariate analysis as a potential alternative method to discriminate between cancer and normal cell lines. We reported a fast, clean, and nondestructive methodology involving minimal sample preparation to categorize the samples. The results are very promising since the proposed methodology was able to classify many different types of cells (3T3, ARPE, HEK, HepG2, HeLa, HT-29 and 786-0) with satisfactory indexes of correct discrimination (100% for cancer and 75% for normal in the first dataset and 100% for both cancer and normal in the second dataset). The identification/confirmation of endogenous fluorophores using OPLS/UPLS-DA was also successful and close to the optimal fluorophores obtained by this model. This is especially relevant if taking into account the intrinsic difficulties inherent to OPLS/UPLS-DA to properly identify components (spectral confirmation) contributing to the measured spectra. In this work, we have presented a preliminary contribution to this complex biological system, and further work is needed to confirm the results obtained here. One point that might be improved in further studies is the low number of samples included in the models in such a way that more variability is incorporated and an alternative method could be widely applied. Also, we expect to expand this methodology to other kinds of biological samples (such as blood plasma and urine) that are frequently analyzed by sophisticated techniques such as NMR and MS, aiming to provide as much as possible a more convenient and less invasive analysis, especially when considering the possibility of applications in a screening approach where speed and simplicity are substantially important requirements.

Acknowledgements

A. A. O. Neves would like to acknowledge the financial support from PPGQ/UFRN/CAPES for a fellowship. K. M. G. Lima acknowledges the CNPq/CAPES project (Grant 070/2012) and FAPERN (Grant 005/2012) for financial support.

References

1 N. Vinayavekhin, E. A. Homan and A. Saghatelian, *ACS Chem. Biol.*, 2010, **5**, 91–103.
2 J. L. Griffin and J. P. Shockcor, *Nat. Rev. Cancer*, 2004, **4**, 551–561.
3 R. J. Deberardinis, N. Sayed, D. Ditsworth and C. B. Thompson, *Curr. Opin. Genet. Dev.*, 2008, **18**, 54–61.
4 G. Kroemer and J. Pouyssegur, *Cancer Cell*, 2008, **13**, 472–482.
5 T. D. Dung, C.-C. Feng, W.-W. Kuo, P. Pai, L.-C. Chung, S.-H. Chang, H.-H. Hsu, F.-J. Tsai, Y.-M. Lin and C.-Y. Huang, *Biosci., Biotechnol., Biochem.*, 2013, **77**, 1814–1821.
6 Z. Qian, X. Zhao, M. Jiang, W. Jia, C. Zhang, Y. Wang, B. Li and W. Yue, *BMC Cancer*, 2012, **12**, 442.
7 J. P. McCoy, *Hematol. Oncol. Clin. North Am.*, 2002, **16**, 229–243.
8 Z. Darzynkiewicz, H. Zhao, H. D. Halicka, P. Rybak, J. Dobrucki and D. Wlodkowic, *Crit. Rev. Clin. Lab. Sci.*, 2012, **49**, 199–217.

9 M. Cuperlović-Culf, N. Belacel, A. S. Culf, I. C. Chute, R. J. Ouellette, I. W. Burton, T. K. Karakach and J. A. Walter, *Magn. Reson. Chem.*, 2009, **47**, S96–S104.
10 A. L. Dill, L. S. Eberlin, C. Zheng, A. B. Costa, D. R. Ifa, L. Cheng, T. A. Masterson, M. O. Koch, O. Vitek and R. G. Cooks, *Anal. Bioanal. Chem.*, 2010, **398**, 2969–2978.
11 W. Yi, D. Cui, Z. Li, L. Wu, A. Shen and J. Hu, *Spectrochim. Acta, Part A*, 2013, **101**, 127–131.
12 M. Khanmohammadi, R. Nasiri, K. Ghasemi, S. Samani and A. Bagheri Garmarudi, *J. Cancer Res. Clin. Oncol.*, 2007, **133**, 1001–1010.
13 J. W. Chan, D. S. Taylor, T. Zwerdling, S. M. Lane, K. Ihara and T. Huser, *Biophys. J.*, 2006, **90**, 648–656.
14 J. Boccard and D. N. Rutledge, *Anal. Chim. Acta*, 2013, **769**, 30–39.
15 K. Odunsi, R. M. Wollman, C. B. Ambrosone, A. Hutson, S. E. McCann, J. Tammela, J. P. Geisler, G. Miller, T. Sellers, W. Cliby, F. Qian, B. Keitz, M. Intengan, S. Lele and J. L. Alderfer, *Int. J. Cancer*, 2005, **113**, 782–788.
16 L. Yu, L. Gao, K. Li, Y. Zhao and D. K. Y. Chiu, *Comput. Biol. Chem.*, 2011, **35**, 298–307.
17 Å. M. Wheelock and C. E. Wheelock, *Mol. BioSyst.*, 2013, **9**, 2589–2596.
18 D. A. MacIntyre, B. Jiménez, E. J. Lewintre, C. R. Martín, H. Schäfer, C. G. Ballesteros, J. R. Mayans, M. Spraul, J. García-Conde and A. Pineda-Lucena, *Leukemia*, 2010, **24**, 788–797.
19 B. Chance and B. Thorell, *J. Biol. Chem.*, 1959, **234**, 3044–3050.
20 N. Ramanujam, *Neoplasia*, 2000, **2**, 89–117.
21 A. J. Lawaetz, R. Bro, M. Kamstrup-Nielsen, I. J. Christensen, L. N. Jørgensen and H. J. Nielsen, *Metabolomics*, 2011, **8**, 111–121.
22 N. Harrit, M. Albrechtsen, O. Olsen, L. Nørgaard, D. Nielsen, K. Kampmann and R. Bro, *J. Chemom.*, 2007, **21**, 451–458.
23 P. Mazur, *Am. J. Physiol.*, 1984, **247**, C125–C142.
24 M. Barker and W. Rayens, *J. Chemom.*, 2003, **17**, 166–173.
25 R. Castillo, M. Otto, J. Freer and S. Valenzuela, *J. Chemom.*, 2008, **22**, 268–280.
26 R. S. Fernandes, F. S. L. da Costa, P. Valderrama, P. H. Março and K. M. G. de Lima, *J. Pharm. Biomed. Anal.*, 2012, **66**, 85–90.
27 R. Bro, *Chemom. Intell. Lab. Syst.*, 1997, **38**, 149–171.
28 M. W. Conklin, P. P. Provenzano, K. W. Eliceiri, R. Sullivan and P. J. Keely, *Cell Biochem. Biophys.*, 2009, **53**, 145–157.
29 A. P. Teixeira, C. A. M. Portugal, N. Carinhas, J. M. L. Dias, J. P. Crespo, P. M. Alves, M. J. T. Carrondo and R. Oliveira, *Biotechnol. Bioeng.*, 2009, **102**, 1098–1106.
30 R. Bro and M. Vidal, *Chemom. Intell. Lab. Syst.*, 2011, **106**, 86–92.
31 A. F. Santidrian, A. Matsuno-Yagi, M. Ritland, B. B. Seo, S. E. Leboeuf, L. J. Gay, T. Yagi and B. Felding-habermann, *J. Clin. Invest.*, 2013, **123**, 1068–1081.
32 F. Sun, W. Zong, R. Liu, J. Chai and Y. Liu, *Spectrochim. Acta, Part A*, 2010, **76**, 142–145.
33 S. Ganesan, P. G. Sacks, Y. Yang, A. Katz, M. Al-Rawi, H. E. Savage, S. P. Schantz and R. R. Alfano, *Cancer Biochem. Biophys.*, 1998, **16**, 365–373.
34 G. Kurban, B. L. Gallie, M. Leveridge, A. Evans, D. Rushlow, D. Matevski, R. Gupta, A. Finelli and M. A. S. Jewett, *Pathol., Res. Pract.*, 2012, **208**, 22–31.
35 J. M. Catania, G. Chen and A. R. Parrish, *Am. J. Physiol.*, 2007, **292**, F905–F911.
36 K. Kessenbrock, V. Plaks and Z. Werb, *Cell*, 2010, **141**, 52–67.
37 A. C. Pereira, S. U. Amadei, L. Eduardo, B. Rosa and B. Gelatinase, *Rev. Bras. Cancerol.*, 2006, **52**, 257–262.

Capítulo 3

ATR-FTIR and multivariate analysis as a screening tool for cervical cancer in women from northeast Brazil: a biospectroscopic approach

Ana C. O. Neves Cleine G. Miranda

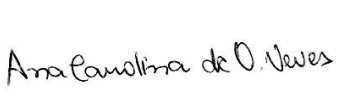
Priscila P. Silva Janaína C. O. Crispim

Camilo L. M. Morais Kássio M. G. de Lima

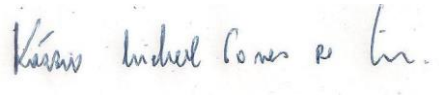
RSC Advances, 2016, 6, 99648--99655.

Contribuição:

- Realizei a aquisição espectral;
- Realizei o processamento dos dados e construção dos modelos multivariados;
- Escrevi a primeira versão do manuscrito.



Ana C. O. Neves



Prof. Kássio M. G. Lima

RSC Advances

COMMUNICATION

ATR-FTIR and multivariate analysis as a screening tool for cervical cancer in women from northeast Brazil: a biospectroscopic approach

Ana C. O. Neves,[a] Priscila P. Silva,[a] Camilo L. M. Morais,[a] Cleine G. Miranda,[b] Janaina C. O. Crispim[b] and Kássio M. G. Lima*[a]

Cite this: *RSC Adv.*, 2016, **6**, 99648

Received 25th August 2016
Accepted 7th October 2016

DOI: 10.1039/c6ra21331f

www.rsc.org/advances

Cervical cancer is the fourth most frequent cancer in women worldwide and the third in Brazil. Screening methods can substantially reduce new cases of cervical cancer by identifying pre-cancerous lesions, making it possible to offer correct management and treatment. For this purpose, this work reports the use of attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy coupled with principal component analysis (PCA) and variable selection techniques, such as successive projections algorithm (SPA) and genetic algorithm (GA) associated to linear or quadratic discriminant analysis (LDA/QDA), to classify samples for negative for intraepithelial lesion or malignancy (NILM), $n = 43$, and squamous intraepithelial lesion (SIL), $n = 40$, directly from blood plasma. Furthermore, the possibility to categorize SIL subclasses according to low-grade squamous intraepithelial lesion (LSIL) and high-grade squamous intraepithelial lesion (HSIL) lesion degrees was evaluated. Application of variable selection algorithms, especially GA, considerably improved the classifications by choosing spectral variables that reflect the chemical differences between a healthy and pre-cancerous plasma sample. This method was able to correctly classify NILM *vs.* SIL with sensitivity and specificity for both classes varying around 77% using LDA. With QDA, the results were enhanced to sensitivity around 90% and specificity of 83%. NILM *vs.* LSIL presented sensitivity and specificity ranging between 67–94% and 82–94%, respectively. In addition, NILM *vs.* HSIL were found to have sensitivity and specificity from 76–97% to 73–100%, respectively, where QDA substantially provided better classifications. These findings highlight the potentiality of ATR-FTIR spectroscopy combined with multivariate analysis as a screening tool for pre-cancerous cervical lesions, which could contribute to reduce cervical cancer incidence.

Introduction

Cervical cancer is the fourth most frequent cancer in women, with a global estimate of 528 000 new diagnosed cases and 266 000 deaths in 2012. This disease is considerably more common in less developed regions, where it is responsible for 12% of all female cancers, accounting for almost 85% of all the cervical cancer statistics worldwide.[1] The Brazilian National Cancer Institute (INCA) expects 16 340 new cases of cervical cancer in 2016, corresponding to the third most incurring cancer type in Brazilian women.[2] Nowadays, it is well-known that the human papillomavirus (HPV) plays a very important role in the development of cervical cancer.

HPV is a small non-enveloped virus, a member of the family papilloma viruses, with a circular double-stranded DNA genome, which infects the epithelia of skin and mucosa. More than 180 HPV types have been identified, and can be separated into high-risk, intermediate-risk and low-risk, according to their potential to induce cancer in infected tissues. The high-risk HPVs most related to cervical carcinogenicity are HPV 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59 and 66, where HPV 16 and 18 are the most prevalent types and responsible for more than 70% of all cases of invasive cervical cancer.[3–7] Although HPV infection is the most frequent sexually transmitted disease worldwide, approximately 90% of all infected women are able to clear the virus within 2 years after infection by natural action of their immune system. If the immune system does not properly fight against the virus, the infection can develop to cause cervical intraepithelial neoplasia (CIN 1, CIN 2, CIN 3, according to the severity of the lesions), which is initially an asymptomatic condition that can either spontaneously regress to normal without any treatment, or can progress to invasive cervical cancer in 5–20 years.[8,9] These pre-cancerous lesions have different rates of progression to invasive cancer, where CIN 1 has a low rate, and CIN 3 has a high rate if left untreated.[10,11] In the Bethesda system of classification of cervical cytology, CIN 1 is classified as low-grade squamous intraepithelial lesion (LSIL), and CIN 2 and CIN 3 are grouped together as high-grade squamous intraepithelial lesions (HSIL).[12–16] In this context, it is essential to identify the occurrence/recurrence of these cervical lesions early to guarantee correct treatment and to avoid the risk of developing invasive cancer in the following years.

[a]*Institute of Chemistry, Biological Chemistry and Chemometrics, Federal University of Rio Grande do Norte, Natal 59072-970, RN, Brazil. E-mail: kassiolima@gmail.com; Tel: +55 84 3342 2323*

[b]*Healthy Sciences Center, Federal University of Rio Grande do Norte, Natal 59010-180, RN, Brazil*

Some screening methods commonly used today include tests for HPV, tests to detect cervical lesions by cytology (Pap smear) and unaided visual inspection with acetic acid (VIA), being the Pap smear most currently employed in developing countries.[17–19] Implementation of the Pap smear as a screening method worldwide has substantially decreased the morbidity and mortality from squamous carcinoma of the cervix. In the UK, the screening program using the Pap smear has considerably reduced the incidence of cervical cancer to become the eleventh most common female cancer in this region.[20,21] However, considering the human subjectivity present in this method due to the sampling and sample management being interpreted by the cytologist, the sensitivity (meaning the percentage of true positive cases detected) and/or specificity (meaning the percentage of true negative cases that are negative) of the Pap smear are 51% (30–87%) and 98% (86–100%), respectively. Sensitivity is particularly affected by the inter observer variability, and this lack of accuracy can lead to high false-negative rates that can induce failures in preventing cervical cancer, mainly in women that do not follow the correct periodicity of the screening programs.[7,20] Furthermore, some questions like poorly developed healthcare services, cultural and religious factors, limited resources and information can play a role in putting up barriers to implement Pap smearing as a screening method, especially in developing or rural regions where these issues are still a strong reality.[21]

Infrared spectroscopy is a vibrational technique that has the capacity to analyze biological systems, since complex molecules such as proteins, lipids, carbohydrates and nucleic acids exhibit distinct vibrational behaviors according to their molecular structure and conformation.[22] ATR-FTIR spectroscopy is a powerful alternative to be employed in resolving biological issues, considering its ability to reflect on the composition and variability of samples, and especially in the region of the "bio-fingerprint" (1800–900 cm^{-1}), where many important biomolecules have individual absorbing frequencies, thereby allowing scientists to search for biomarkers and metabolic profiles.[23] Remarkably, ATR-FTIR is a fast, non-destructive and clean method, making it possible to analyze a considerable number of samples in a day and to reuse them after spectra acquisition, thus avoiding the necessity of many reagents and sample handling steps, and promoting a reduction of waste generation and making the experiment more simple and cost-effective.[20,22] ATR-FTIR has been attracting great attention in cancer research as a powerful tool, leading to relevant publications over the last few years.[24] Theophilou and co-workers have used this technique to analyze ovarian tissues and to discriminate them between normal, borderline or malignant,[22] while Lima and co-workers have applied ATR-FTIR to classify blood plasma or serum samples according to their ovarian cancer stage.[23] Moreover, Purandare and co-workers have showed the capability of vibrational spectroscopy to segregate low-grade cervical cytology-based samples considering their potential to regress, remain static or progress.[25] Also in this field, Lima and co-workers have successfully classified cervical cytology specimens between high-risk or low-risk HPV infection.[26]

ATR-FTIR is definitely a remarkable tool for studying chemical species due to its ability to provide a high number of substantial information, however, when biological samples are taken into account, this technique itself may not provide enough specificity in the search for biomarkers since there are many biomolecules contributing to the whole signal, leading to a high amount of complex data. On the other hand, multivariate analysis has been proven to be effective in overcoming this drawback, allowing for the successful use of ATR-FTIR for biological purposes. This is especially evident in the possibility to extract essential information related to biomarkers, which reflects the particularity of each chemical system. In this context, SPA and GA have made it possible to select the most significant variables from complex spectral data, which can be associated to biomarkers. For classification, the employment of these algorithms is commonly associated to LDA, in the way that samples can be separated into groups based on their spectral similarities, and the classification model is used to predict unknown samples.[26]

The 21st century has been characterized by the search for alternative tools in several medical fields, and in this context the combination of inexpensive spectroscopic techniques and computational treatments emerges as a very promising strategy for screening cervical cancer. In this paper, we report our findings in the application of ATR-FTIR spectroscopy and multivariate analysis to differentiate NILM and SIL classes directly from blood plasma. Furthermore, we investigated the ability of this method to separate cervical squamous intraepithelial lesions into low-grade and high-grade lesions (LSIL and HSIL), respectively. Chemometric approaches were based on the use of PCA, SPA and GA algorithms associated to linear and quadratic discriminant analysis (LDA and QDA, respectively). To the best of our knowledge, this is the first work involving screening cervical pre-cancer stages in Brazilian women using ATR-FTIR and chemometrics. In addition, PCA-QDA, SPA-QDA and GA-QDA have never been reported in literature for this purpose. Considering the high incidence of cervical cancer all over the world and the relevance of its early detection, this fast, simple and inexpensive methodology may substantially contribute to cancer prevention, especially in developing countries.

Methods

Collection and preparation of specimens

This study involved women living in the state of Rio Grande do Norte/Brazil, attending the Maternidade Escola Januário Cicco (MEJC) of the Public Health System for cervical pathology screening consultations and reference services for colposcopy, from July 2014 to January 2016. All experiments were performed in compliance with the guideline "Biomedical research ethics review method involving people" (Brazil), and approved by the medical ethics committee at Hospital Universitário Onofre Lopes (HUOL), Brazil (protocol # 526/11), Federal University of Rio Grande do Norte. Informed consents were obtained from human participants of this study.

Collection of fasting blood samples (containing the anticoagulant EDTA) was performed per patient prior to cytology

smears or large loop excision surgery of the transformation zone (LLETZ), accounting for a total of 83 blood samples. In this study, atypical squamous cells (ASC) of undetermined significance (ASC-US and ASC-H) were excluded. Within two hours after blood collection by venipuncture, blood plasma was separated by density gradient, and aliquots were transferred into cryogenic tubes and stored at −80 °C until analysis. Before analysis, cytological samples (Pap smear) were obtained from women who were referred either in NILM or SIL groups. For women undergoing LLETZ surgery, histopathological analysis was performed on sections from paraffin blocks in 4 μm thickness and stained with hematoxylin/eosin. Cytology and histopathology are reported according to the Bethesda system:[27] 43 patients (NILM), 16 patients (LSIL) and 24 patients (HSIL).

ATR-FTIR spectroscopy

ATR-FTIR spectra [$n = 830$, 83 samples (NILM ($n = 43$), LSIL ($n = 16$), HSIL ($n = 24$))] were collected from a Bruker VERTEX 70 FTIR spectrometer (Bruker Optics Ltd., Coventry, UK) with Helios ATR attachment containing a diamond crystal internal reflective element using a 45 incidence angle of IR beam. The instrument was set up to perform a total of 16 scans with 4 cm^{-1} spectral resolution on both background and sample. Frozen plasmas were thawed at room temperature for 30–40 min, and 10 μL of each sample was transferred onto 10 different IR-reflective glass slides (Low-E; Kevley Technologies), resulting in a utilization of 100 μL of each plasma. Disposed plasmas were air-dried for approximately 30 min, until forming homogenous dried films[23,28] and immediately all samples were submitted to ATR-FTIR spectra acquisition. ATR crystals were washed with 70% v/v alcohol before each sample spectra acquisition and a background was collected.

Data analysis

MATLAB® R2010a software (Math Works Inc, Natick, MA, USA) was used for data import, pre-treatment and construction of multivariate classification models (PCA-LDA, SPA-LDA, GA-LDA, PCA-QDA, SPA-QDA and GA-QDA). ATR-FTIR spectra were pre-processed by cutting the region of interest between 1800 and 900 cm^{-1} (450 wavenumbers), baseline-corrected and normalized by the amide I peak (*i.e.*, ≈1650 cm^{-1}). Samples were divided into training (70%), validation (15%) and prediction (15%) sets for all classification models by applying the Kennard–Stone (KS) algorithm to the IR spectra.[29] Training samples were used in the model construction and optimization (variable selection by SPA and GA algorithms) while the prediction set was only applied to evaluate the classification model using LDA and QDA discrimination approaches. The optimal number of variables for SPA-LDA/QDA and GA-LDA/QDA was determined with an average risk G of LDA/QDA misclassification. Such a cost function is calculated in the validation set as:

$$G = \frac{1}{N_V}\sum_{n=1}^{N_V} g_n, \quad (1)$$

where g_n is defined as

$$g_n = \frac{r^2(x_n, m_{I(n)})}{\min_{I(m)\neq I(n)} r^2(x_n, m_{I(m)})} \quad (2)$$

where $I(n)$ is the index of the true class for the nth validation object x_n; $r^2(x_n,m_{I(n)})$ is the squared Mahalanobis distance between object x_n (of class index $I(n)$) and the sample mean $m_{I(n)}$ of this true class; and $r^2(x_n,m_{I(m)})$ is the squared Mahalanobis distance between object x_n and the center of the closest wrong class.[26]

To obtain a discriminant profile, the LDA classification score (L_{ij}) is calculated for a given class k by the following equation:

$$L_{ik} = (\mathbf{x}_i - \overline{\mathbf{x}}_k)^T \Sigma_{pooled}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}_k) - 2\log_e \pi_k \quad (3)$$

where $\mathbf{x}_i$ is an unknown measurement vector for sample i; $\bar{\mathbf{x}}_k$ is the mean measurement vector of class k; Σ_{pooled} is the pooled covariance matrix; and π_k is the prior probability of class k.[30]

On the other hand, the QDA classification score (Q_{ij}) is estimated using the variance–covariance for each class k and an additional natural logarithm term, as follows:

$$Q_{ik} = (\mathbf{x}_i - \overline{\mathbf{x}}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}_k) + \log_e|\Sigma_k| - 2\log_e \pi_k \quad (4)$$

where Σ_k is the variance–covariance matrix of class k; and $\log_e|\Sigma_k|$ is the natural determinant logarithm of variance–covariance matrix Σ_k.

Additionally, the main differences between these discrimination methods are that QDA forms a separated variance model for each class and does not assume classes having similar variance–covariance matrices; whereas LDA does not take into account different variance structures in each class, assuming that the analyzed classes have similar variance–covariance matrices.[31] The GA-LDA/QDA calculations were performed during 40 generations with 80 chromosomes each. One-point crossover and mutation probabilities were set to 60% and 10%, respectively. Moreover, the algorithm was repeated three times, starting from different random initial populations. The best solution (in terms of the fitness value) resulting from the three GA repetitions was employed.

The classification models were built for ATR-FTIR spectra pooled into three different cases:

(1) NILM (430 spectra) *vs.* SIL (LSIL and HSIL) (400 spectra);

(2) NILM (220 spectra) *vs.* low-grade lesions (LSIL) (160 spectra);

(3) NILM (220 spectra) *vs.* high-grade lesions (HSIL) (240 spectra);

Calculations of sensitivity (probability that a test result will be positive when disease is present) and specificity (probability that a test result will be negative when disease is not present) were performed for this study as important quality measures of model accuracy. Both parameters have a maximum value of 1 and a minimum of 0, and can be obtained by using the following equations:

$$\text{Sensitivity } (\%) = \frac{TP}{TP + FN} \times 100 \quad (5)$$

$$\text{Specificity } (\%) = \frac{TN}{TN + FP} \times 100 \quad (6)$$

where FN is defined as a false negative and FP as a false positive. TP and TN are defined as true positive and true negative, respectively.[26]

Results and discussion

Blood plasma ATR-FTIR mean spectra of NILM, LSIL and HSIL categories are shown in Fig. 1. A total of $n = 83$ specimens were collected generating 830 spectra to be analyzed, where the NILM class accounted for 43 samples, and LSIL and HSIL precancerous lesions represented 16 and 24 samples, respectively. In the region of interest between 900 and 1800 cm^{-1}, called the "bio-fingerprint region", some characteristic IR absorption bands can be observed in the spectra, such as the major peaks at $\approx$1650 cm^{-1} (amide I) and 1550 cm^{-1} (amide II) of aminoacids and proteins, as well as methylene groups of lipids at $\approx$1400–1470 cm^{-1}. Other important bands (although less intense) include the asymmetric and symmetric phosphate stretching vibrations at $\approx$1225 cm^{-1} and 1080 cm^{-1}, respectively, and also peaks at $\approx$1155 cm^{-1} corresponding to C–OH and C–O groups present in some aminoacids (such as serine, tyrosine, threonine) and carbohydrates, and the smooth band at $\approx$1030 cm^{-1} related to glycogen.[23]

It is possible to verify that the spectra present strong similarity related to absorption bands, in addition to being highly overlapped, in a way that it becomes difficult to categorize samples only considering the complex spectral information available. In this sense, application of multivariate algorithms is an essential strategy to extract the important spectral information, enabling the discrimination of samples between NILM or SIL classes based on their pathophysiological condition reflected in the spectral bands. Furthermore, variable selection algorithms such as SPA and GA are powerful tools to be used in the search for biomarkers in blood plasma, allowing that less complex models be obtained. In this study, all spectra were preprocessed by applying normalization (amide I band) and baseline correction, and the classification models (PCA-LDA/QDA, SPA-LDA/QDA and GA-LDA/QDA) were built using both the processed and the raw data in order to compare results. In general, sensitivity and specificity values of models were higher when classification was performed using the raw data, and the best results can be appreciated in the following discussions. Considering the importance of screening methods for reducing the new cases of cervical cancer, the main objective of this study was to apply chemometric tools to extract the biochemical information of samples representing women with or without cervical lesions, making it possible to separate samples in to the two classes of NILM and SIL. Additionally, more specific models were also investigated to categorize samples in attempt to show the potentiality of the proposed classification method, taking into account the existence of subgroups in the cervical lesion (SIL), LSIL and HSIL classes. This approach could be of great interest in clinical routine, since medical conduct is totally different in face of a patient with a low-grade lesion or high-grade lesion condition. Therefore, the whole NILM dataset (430 spectra) was divided approximately by half for this purpose, and a NILM dataset of only 220 spectra (22 samples) was used for models in order to have a similar data size compared to the LSIL and HSIL datasets. In all cases, a comparison between LDA and QDA models was performed by analyzing the sensitivity and specificity values obtained for both linear and quadratic models.

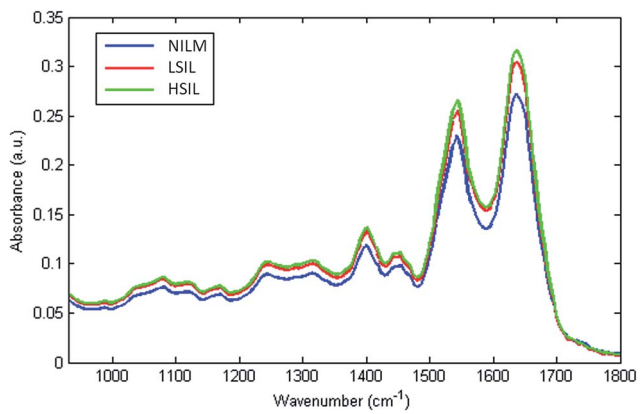


Fig. 1 ATR-FTIR mean spectra of NILM (blue), LSIL (red) and HSIL (green) samples, in the region of 900–1800 cm^{-1}.

NILM *vs.* SIL

PCA-LDA/QDA, SPA-LDA/QDA and GA-LDA/QDA were used to classify NILM *vs.* SIL. In this case, models using the non-processed dataset achieved better results for all algorithms, and especially for GA-LDA/QDA. In using 6 component scores (which accounts for >90% of the explained variance of the whole data), PCA-LDA produced a sensitivity and specificity of 37% for NILM class, and sensitivity and specificity values for SIL class were 80% and 75%, respectively. Application of QDA improved the sensitivity of NILM to 74%, however the specificity was lower than with LDA (26%), while application of QDA resulted in inferior classification rates for the SIL group (see Table 1). SPA-LDA and QDA were applied to the dataset in order to obtain a classification model using a considerably reduced number of variables, chosen by the minimum of the cost function G. Using only three selected wavenumbers (1404, 1508 and 1637 cm^{-1}), Fisher scores were obtained and the classification indexes (sensitivity and specificity, shown in Table 1) were very similar to those of PCA-LDA/QDA, both for NILM and SIL classes. Another variable selection strategy was the application of GA-LDA and GA-QDA to build classification models, considering only the variables most related to chemical information present in the dataset.

It is possible to observe from Table 1 that the GA-LDA model using the 68 selected wavenumbers from a whole 450 spectral variables improved classification rates for prediction samples when compared to PCA-LDA and SPA-LDA results. The GA-LDA model presented sensitivity of 77 and 78% for NILM and SIL, respectively, and also maintained very similar specificity results for both classes (75 and 78% for NILM and SIL classes, respectively). Using quadratic discriminant analysis associated to GA algorithm provided even better classification models, according to Table 1.

Table 1 Results (sensitivity and specificity) of prediction samples for the models PCA-LDA/QDA, SPA-LDA/QDA and GA-LDA/QDA: NILM *vs.* SIL; NILM *vs.* LSIL; NILM *vs.* HSIL

		LDA		QDA	
	Model	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
NILM *vs.* SIL	PCA	37/80	38/75	74/52	26/52
	SPA	40/78	40/78	61/42	37/58
	GA	77/78	75/78	89/83	90/82
NILM *vs.* LSIL	PCA	60/75	60/75	79/37	21/62
	SPA	63/71	60/71	76/47	24/58
	GA	76/83	82/87	94/67	94/83
NILM *vs.* HSIL	PCA	54/97	54/97	67/44	76/42
	SPA	45/94	45/94	48/28	51/72
	GA	76/94	73/86	88/97	91/100

The wavenumbers selected by GA are shown highlighted in Fig. 2A. GA-QDA model presented sensitivity and specificity values of 89 and 90%, respectively, for NILM class; and the model achieved sensitivity and specificity of 83 and 82% for SIL class, respectively, maintaining agreement between the classification indexes for both classes. Sample separation into the two categories is shown for GA-LDA and GA-QDA in Fig. 2B and C, respectively. Two clusters are adequately visualized, where samples are softly and more correctly grouped into their own classes with the GA-QDA model. GA-LDA/QDA have selected particularly interesting wavenumbers (Fig. 2A); namely, the variables at 1747 and 1724 cm^{-1}, associated to C═O stretching vibrations of lipids and aldehydes, respectively. The major peaks of 1639 cm^{-1} (amide I) of C═O stretching vibration of the amide group coupled to the N–H bond bending and the C–N bond stretching, as well as 1539 cm^{-1} (amide II) of C–N stretching and N–H deformation were observed. Finally, there are methyl and methylene groups of lipids and proteins at 1400 and 1454 cm^{-1}, respectively, asymmetric and symmetric stretching vibrations of phosphate at 1219 and 1080 cm^{-1}, respectively, and C–O groups of carbohydrates at 1155 cm^{-1} which also were observed.

NILM *vs.* LSIL (low-grade)

In this case, the classification between NILM and LSIL samples were investigated. As can be seen using the raw dataset in Table 1, PCA-LDA with 6 components presented sensitivity and specificity for NILM of 60%, while these values were superior (75%) for SIL. When QDA was applied, the sensitivity of NILM increased to 79%; however, a loss in specificity was observed (21%), and also the LSIL classification was impaired. Four variables were selected for SPA-LDA (1459, 1531, 1583 and 1641 cm^{-1}), and the sensibility and specificity values for both NILM and LSIL were very similar to those for PCA-LDA (around 60 and 70%, respectively). Better results were not found for QDA. Regarding Table 1, GA-LDA and QDA led to the best classification for this case in using 41 selected wavenumbers, as shown in Fig. 2D. Sensitivity and specificity for NILM were 76 and 82%, respectively, and these values were 83 and 87% for LSIL using the GA-LDA model. For the NILM class, GA-QDA was found presenting 94% of sensitivity and specificity, leading to better results. Separation of samples with GA-LDA/QDA models can be viewed in Fig. 2E and F, where both models have well defined clusters corresponding to NILM and LSIL samples.

In this case, GA-LDA/QDA selected some interesting variables (see Fig. 2D), namely the wavenumbers at 1724 and 1461 cm^{-1} associated to C═O stretching vibrations of aldehydes and methylene lipid groups; amide III from proteins at 1334 cm^{-1}; asymmetric and symmetric stretching vibrations of phosphate at 1221 and 1089 cm^{-1}; and out-of-plane C–H bending at 960 cm^{-1}. It is worth mentioning that some of these variables are coincident with those selected for NILM *vs.* SIL classification, as described above.

NILM *vs.* HSIL (high-grade)

Lastly, this event was performed in order to evaluate the classifications of NILM *vs.* HSIL. By observing Table 1, it can be seen that both PCA-LDA and SPA-LDA showed poor sensitivity and specificity for NILM, while these values were above 90% for HSIL prediction samples. However, calibration and validation sets were found to be poorly classified. PCA-LDA used 6 components and SPA-LDA selected 3 variables: 1483, 1535 and 1641 cm^{-1}. A gentle enhancement in results was achieved for the NILM class with the PCA-QDA model (see Table 1). As was observed in the previous cases, the best results for classification were obtained with the GA-LDA/QDA models. Using 45 selected wavenumbers (see Fig. 2G) for NILM, sensitivity and specificity were around 73%. For HSIL, these parameters have superior values of 94 and 86% (Table 1) with the GA-LDA model. In this particular case, application of QDA considerably improved results, leading to more accurate classifications for both classes. Regarding NILM, sensitivity and specificity values were 88 and 91%, respectively; while HSIL had 97% sensitivity and 100% specificity with the GA-QDA model. Fig. 2G and H show the classification of samples. A great separation may be observed, with two very-well defined clusters representing the analyzed classes.

In this case, GA-LDA/QDA have selected some interesting variables (see Fig. 2G), namely: variables at 1758 and 1729 cm^{-1} are associated to C═O stretching vibrations of lipids and aldehydes, respectively, major peaks at 1639 cm^{-1} (amide I) of C═O stretching vibration of the amide group coupled to the

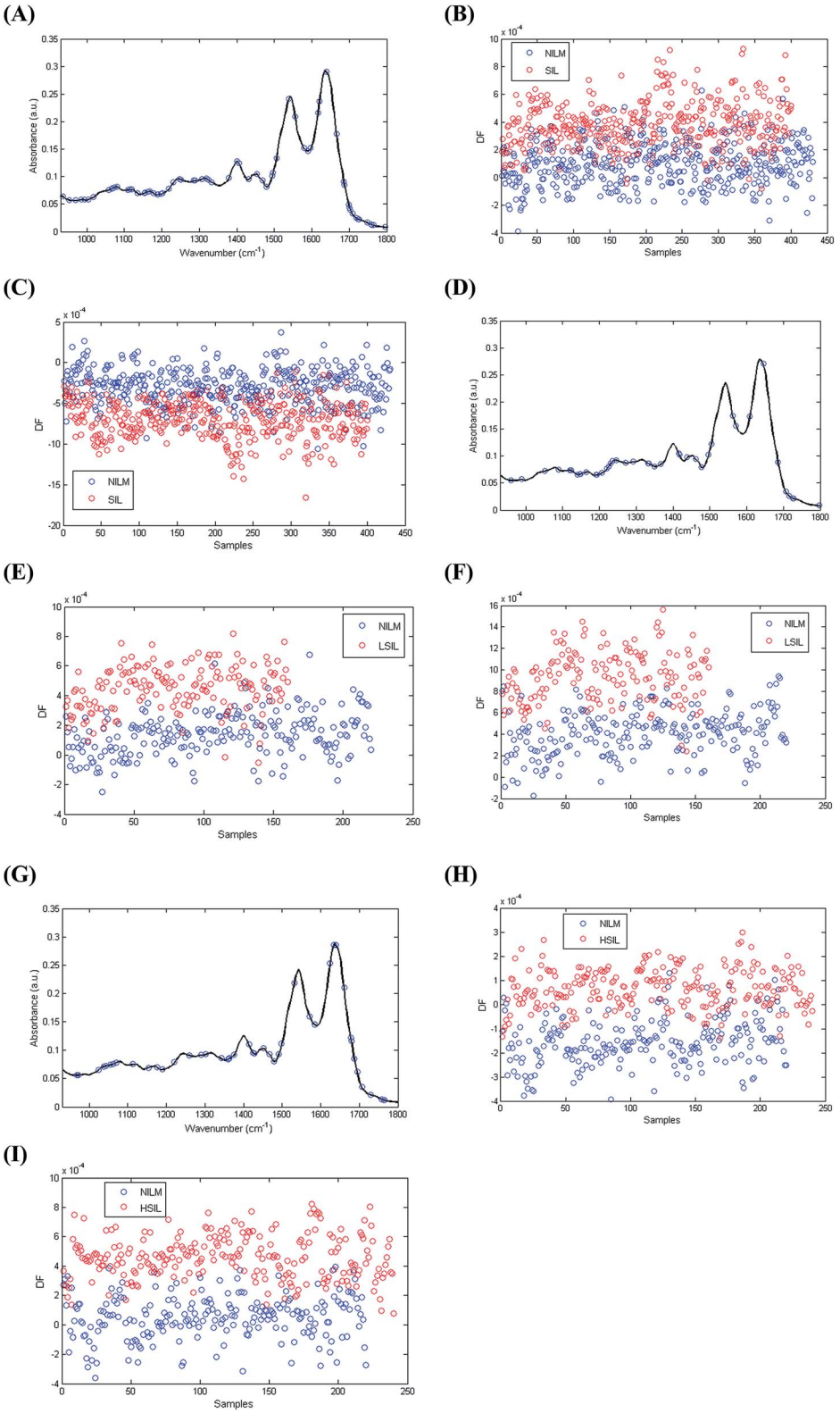


Fig. 2 Segregated categories based on the presence or absence of intraepithelial lesion of cervix, and selected wavenumbers by GA-LDA/QDA models. (A) 68 selected wavenumbers by GA-LDA/QDA models for NILM *vs.* SIL. (B) DF1 × specimens calculated by using the variables selected by GA-LDA from plasma samples segregated into NILM *vs.* SIL. (C) DF1 × specimens calculated by using the variables selected by GA-QDA from plasma samples segregated into NILM *vs.* SIL. (D) 41 selected wavenumbers by GA-LDA/QDA models for NILM *vs.* LSIL. (E) DF1 × specimens calculated by using the variables selected by GA-LDA from plasma samples segregated into NILM *vs.* LSIL. (F) DF1 × specimens calculated by using the variables selected by GA-QDA from plasma samples segregated into NILM *vs.* LSIL. (G) 45 selected wavenumbers by GA-LDA/QDA models for NILM *vs.* HSIL. (H) DF1 × specimens calculated by using the variables selected by GA-LDA from plasma samples segregated into NILM *vs.* HSIL. (I) DF1 × specimens calculated by using the variables selected by GA-QDA from plasma samples segregated into NILM *vs.* HSIL.

bending of the N–H bond and the stretching of the C–N bond, the right and side amide II at 1531 cm^{-1}, methylene lipid groups at 1467 cm^{-1}, amide III from proteins at 1342 cm^{-1}, out-of-plane C–H bending at 968 cm^{-1}, and the variables at 1043 and 1063 cm^{-1} representing glycogen band due to OH stretching coupled with bending and CO–O–C symmetric stretching of phospholipids and cholesterol esters, respectively.[23]

Conclusions

Results obtained in this study present the potentiality of ATR-FTIR spectroscopy associated with multivariate classification models (PCA-LDA/QDA, SPA-LDA/QDA and GA-LDA/QDA) as an alternative approach for cervical cancer screening. This method was able to correctly classify specimens between NILM and SIL (LSIL and HSIL) directly in blood plasma with sensitivity and specificity varying between 80 and 100% using a fast, clean, minimally invasive and cost-effective methodology. Application of variable selection algorithms (especially GA) considerably improved the classifications by choosing spectral variables that reflect the chemical differences between a healthy and pre-cancerous plasma sample. In addition, this study demonstrated utilization of quadratic discriminant analysis (QDA) and compared its results to those provided by LDA. These findings are very encouraging to be tested with much larger numbers of samples in order to robustly validate this method and better associate spectral variables to biomarkers of cervical intra-epithelial lesions. In this sense, this biospectroscopic approach could contribute to traditional screening methods for early detection of pre-cancerous lesions of cervix, and then reduce the number of new cases of this invasive disease worldwide.

Acknowledgements

A. A. O. Neves and Camilo L. M. Morais would like to acknowledge the financial support from PPGQ/UFRN/CAPES for the research grant. The authors would like to acknowledge the financial support from the Brazilian National Council for Scientific and Technological Development (CNPq). K. M. G. Lima acknowledges the CNPq Grant (305962/2014-0) for financial support. This work was funded by a grant from CNPq/Capes project (Grant 070/2012).

References

1 http://globocan.iarc.fr/, accessed on 18/08/2016.
2 http://www2.inca.gov.br/, accessed on 18/08/2016.
3 C. M. de Oliveira, I. G. Bravo, N. C. S. e. Souza, M. L. N. D. Genta, J. H. T. G. Fregnani, M. Tacla, J. P. Carvalho, A. Longatto-Filho and J. E. Levi, *Infect., Genet. Evol.*, 2015, **34**, 44–51.
4 E. J. Nam, J. W. Kim, S. W. Kim, Y. T. Kim, J. H. Kim, B. S. Yoon, N. H. Cho and S. Kim, *Gynecol. Oncol.*, 2007, **104**, 207–211.
5 F. Cannella, A. Pierangeli, C. Scagnolari, G. Cacciotti, G. Tranquilli, P. Stentella, N. Recine and G. Antonelli, *Immunobiology*, 2015, **220**, 363–368.
6 S. Franceschi and S. Vaccarella, *Cancer Epidemiol.*, 2015, **39**, 1152–1156.
7 B. F. Lees, B. K. Erickson and W. K. Huh, *Am. J. Obstet. Gynecol.*, 2016, **214**, 438–443.
8 C. P. Crum and C. M. McLachlin, *J. Cell. Biochem.*, 1995, **23**, 71–79.
9 J. Paavonen, *Int. J. Infect. Dis.*, 2007, **11**(2), S3–S9.
10 C. J. De Witte, A. J. M. Van De Sande, H. J. Van Beekhuizen, M. M. Koeneman, A. J. Kruse and C. G. Gerestein, *Gynecol. Oncol.*, 2015, **139**, 377–384.
11 T. M. Wilkinson, P. H. H. Sykes, B. Simcock and S. Petrich, *Am. J. Obstet. Gynecol.*, 2015, **212**, 769.e1–769.e7.
12 A. G. Waxman, D. Chelmow, T. M. Darragh, H. Lawson and A.-B. Moscicki, *Obstet. Gynecol.*, 2012, **120**, 1465–1471.
13 R. Nayar and D. C. Wilbur, *Cancer Cytopathol.*, 2015, **123**, 271–281.
14 S. Tabbara, A. B. D. Saleh, W. A. Andersen, S. R. Barber, P. T. Taylor and C. P. Crum, *Obstet. Gynecol.*, 1992, **79**, 338–346.
15 S. Mittal, I. Ghosh, D. Banerjee, P. Singh, J. Biswas, R. Nijhawan, R. Srinivasan, C. Ray and P. Basu, *Int. J. Gynecol. Obstet.*, 2014, **126**, 227–231.
16 N. Santesso, R. A. Mustafa, H. J. Schünemann, M. Arbyn, P. D. Blumenthal, J. Cain, M. Chirenje, L. Denny, H. De Vuyst, L. O. Eckert, S. E. Forhan, E. L. Franco, J. C. Gage, F. Garcia, R. Herrero, J. Jeronimo, E. R. Lu, S. Luciani, S. C. Quek, R. Sankaranarayanan, V. Tsu and N. Broutet, *Int. J. Gynecol. Obstet.*, 2015, **3**, 1–7.
17 B. S. Apgar and G. Brotzman, *Am. Fam. Physician*, 2004, **70**, 1905–1916.
18 R. H. Kaufman, E. Adam, J. Icenogle and W. C. Reeves, *Am. J. Obstet. Gynecol.*, 1997, **177**, 930–936.
19 R. A. Mustafa, N. Santesso, R. Khatib, A. A. Mustafa, W. Wiercioch, R. Kehar, S. Gandhi, Y. Chen, A. Cheung, J. Hopkins, B. Ma, N. Lloyd, D. Wu, N. Broutet and H. J. Schünemann, *Int. J. Gynecol. Obstet.*, 2016, **132**, 259–265.
20 M. J. Walsh, M. J. German, M. Singh, H. M. Pollock, A. Hammiche, M. Kyrgiou, H. F. Stringfellow, E. Paraskevaidis, P. L. Martin-Hirsch and F. L. Martin, *Cancer Lett.*, 2007, **246**, 1–11.
21 J. Martínez-Mesa, G. Werutsky, R. B. Campani, F. C. Wehrmeister and C. H. Barrios, *Prev. Med.*, 2013, **57**, 366–371.
22 G. Theophilou, M. Paraskevaidi, K. M. Lima, M. Kyrgiou, P. L. Martin-Hirsch and F. L. Martin, *Expert Rev. Mol. Diagn.*, 2015, **15**, 693–713.
23 K. M. G. Lima, K. B. Gajjar, P. L. Martin-Hirsch and F. L. Martin, *Biotechnol. Prog.*, 2015, **31**, 832–839.
24 N. C. Purandare, J. Trevisan, I. I. Patel, K. Gajjar, A. L. Mitchell, G. Theophilou, G. Valasoulis, M. Martin, G. von Bünau, M. Kyrgiou, E. Paraskevaidis, P. L. Martin-Hirsch, W. J. Prendiville and F. L. Martin, *Bioanalysis*, 2013, **5**, 2697–2711.
25 N. C. Purandare, I. I. Patel, K. M. G. Lima, J. Trevisan, M. Ma'Ayeh, A. McHugh, G. Von Bünau, P. L. Martin

Hirsch, W. J. Prendiville and F. L. Martin, *Anal. Methods*, 2014, **6**, 4576–4584.
26 K. M. G. Lima, K. Gajjar, G. Valasoulis, M. Nasioutziki, M. Kyrgiou, P. Karakitsos, E. Paraskevaidis, P. L. Martin and F. L. Martin, *Anal. Methods*, 2014, **6**, 9643–9652.
27 R. Nayar, *The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria and Explanatory Notes*, Springer, 3rd edn, 2015.
28 M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. a. Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G. D. Sockalingum, J. Sulé-Suso, R. J. Strong, M. J. Walsh, B. R. Wood, P. Gardner and F. L. Martin, *Nat. Protoc.*, 2014, **9**, 1771–1791.
29 R. Kennard and L. Stone, *Technometrics*, 1969, **11**, 137–148.
30 W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D. L. Massart, S. Heuerding and F. Erni, *Anal. Chim. Acta*, 1996, **329**, 257–265.
31 S. J. Dixon and R. G. Brereton, *Chemom. Intell. Lab. Syst.*, 2009, **95**, 1–17.

Capítulo 4

Classificação de lesões cervicais através de uma abordagem lipidômica pelo uso de espectrometria de massas e análise multivariada

Ana C. O. Neves Boniek G. Vaz

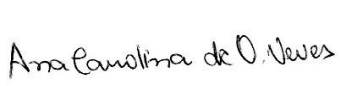
Thaís P. P. Mendes Kássio M. G. de Lima

Camilo L. M. Morais

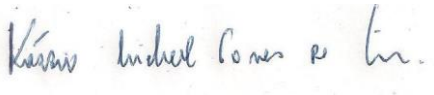
Manuscrito em fase de preparação.

Contribuição:

- Realizei o processamento dos dados e construção dos modelos multivariados;
- Escrevi a primeira versão do manuscrito.



Ana C. O. Neves



Prof. Kássio M. G. Lima

Classificação de lesões cervicais através de uma abordagem lipidômica pelo uso de espectrometria de massas e análise multivariada

1. Introdução

1.1 Câncer cervical: relevância do problema, origem e detecção

O câncer cervical é um dos maiores problemas de saúde da atualidade, sendo o tipo mais frequentemente em mulheres de todo o mundo, e o principal no Brasil [1]. Segundo a Organização Mundial de Saúde (OMS), em 2012 foram estimados 528.000 novos casos de câncer cervical, com 260.000 mortes, com grande contribuição dos países em desenvolvimento para estas estatísticas [2]. Dentro deste contexto, muitos esforços tem sido despendidos no combate deste tipo de câncer, com destaque para os exames Papanicolaou e também para a vacinação frente ao vírus Papilloma Humano (HPV; do inglês, *human Papillomavirus*). Como doença sexualmente transmitível, as infecções por HPV são as de maior incidência em todo o mundo, e também as maiores responsáveis pelo desenvolvimento de câncer cervical [3-5].

O HPV é um vírus relativamente pequeno e não envelopado, caracterizado por um genoma fita dupla circular de DNA, e com capacidade de infectar tecidos epiteliais da pele e mucosas. Atualmente, existem mais de 180 tipos de vírus HPV identificados, sendo o HPV 16 e o HPV 18 os mais comumente associados ao câncer cervical, responsáveis por 70% dos casos invasivos da doença [3-7]. Embora muito frequente, 90% das infecções por HPV podem ser naturalmente eliminadas pelo sistema imunológico da mulher em até dois anos. Todavia, da incapacidade de um combate efetivo de forma natural contra o vírus, a doença pode evoluir até quadro de neoplasias intraepiteliais, as quais são divididas em CIN 1, CIN 2 e CIN 3, de acordo com a severidade da infecção de um tecido. Estas condições são inicialmente assintomáticas, e possuem diferentes velocidades de evolução, sendo CIN 1 a mais lenta e a CIN 3 a mais agressiva [8,9]. Atualmente, existe uma convenção para classificação das lesões intraepiteliais de forma a contemplar tecidos e células. De acordo com o sistema Bethesda de classificação para a citologia, CIN 1 é classificada como lesão intraepitelial escamosa de baixo grau (LSIL) e CIN 2 juntamente com CIN 3, como lesão escamosa intraepitelial de alto grau (HSIL) [9-12].

O Papnicolaou é o exame oficial de rastreamento para o câncer cervical no Brasil, sendo inclusive oferecido pelo SUS. Especificamente, o exame consiste na obtenção de um esfregaço celular diretamente do colo do útero pelo uso de um espéculo. A amostra obtida é então analisada por um citologista, que avalia se existe indicativo de lesão cervical através da observação do formato e do tamanho das células. Certamente, este exame preventivo contribui efetivamente para o diagnóstico precoce do câncer e, inclusive, de lesões pré-cancerosas. Por outro lado, trata-se de um método altamente invasivo e dependente do examinador, no sentido de que pode haver dúvidas quanto ao lado e muita variação entre observadores. Sendo assim, sua sensibilidade (porcentagem de casos verdadeiros/positivos) e/ou especificidade (porcentagem de verdadeiros/negativos) giram em torno de 51% (30-807%) e 98% (86-100%), respectivamente [4,13]. De fato, esta subjetividade, associada ao fato da infecção evoluir de forma assintomática, podem atrasar o diagnóstico, ou mesmo induzir a um tratamento sem a devida necessidade.

Recentemente, a pesquisa visando o desenvolvimento de novas metodologias que possam complementar ou mesmo aprimorar os métodos de detecção de lesões cervicais tem se mostrado de muito interesse para as áreas clínicas e biomédicas. Técnicas mais baratas, sensíveis, rápidas e de amplo acesso estão emergindo com alternativas promissoras, especialmente quando na análise de fluidos biológicos, tais como plasma sanguíneo, urina, saliva, entre outros, que podem ser obtidos facilmente dos pacientes, e refletem condições fisiopatológicas dos mesmos.

1.2 Espectrometria de massas e lipidômica

Por definição, a lipidômica consiste no "entendimento compreensivo da influência de todos os lipídeos em um sistema biológico em relação à sinalização celular, arquitetura de membranas, modulação transcricional e translacional, interações célula-célula e célula-proteína, e resposta a modificações ambientais com o tempo" [14]. Lipídeos consistem de biomoléculas essenciais para o funcionamento de qualquer organismo vivo, uma vez que são constituintes fundamentais das membranas celulares, atuam como fonte de energia, são sinalizadores químicos, entre outros [14,15].

Nos últimos anos, o metabolismo desregulado de lipídeos tem sido associado à diversas doenças importantes, tais como o câncer [15,16], fazendo com que a abordagem lipidômica possa contribuir sobremaneira em diversos aspectos, tais como:

i) entendimento dos processos de iniciação da doença; ii) acompanhamento da progressão; iii) escolha do tratamento mais efetivo [17]. Mais além, a lipidômica tem sido diretamente aplicada na detecção precoce de doenças, especialmente pela identificação de metabólitos-chave e biomarcadores a partir da comparação de amostras normais e doentes.

A espectrometria de massas consiste em uma das técnicas de maior valor para a elucidação estrutural de moléculas orgânicas, sendo que o desenvolvimento métodos de ionização mais brandas tem possibilitado a análise de biomoléculas importantes, tais como proteínas, açúcares e lipídeos, dentre outras classes. Como consequência, o emprego da ESI-MS tem sido uma ferramenta muito valiosa para o estudo de quadros patológicos, incluindo o câncer [18-22].

Dentro da lipidômica, existem duas abordagens principais: *targeted* e *untargeted*, as quais se diferenciam, respectivamente, pelo estudo de classes de metabólitos específicos, ou de um sistema lipídico como um todo. Na abordagem *untargeted*, um perfil lipídico geral é obtido e utilizado diretamente para o estudo de alterações biológicas, ou mesmo para a identificação de lipídeos novos e inesperados [15,17]. Esta forma de análise é bastante útil quando empregada para a comparação de amostras saudáveis e doentes. Todavia, uma vez que os resultados são obtidos na forma de um espectro podendo conter informações de centenas ou mesmo milhares de lipídeos presentes em uma amostra, a interpretação dos resultados requer o uso de tratamentos computacionais e estatísticos sofisticados.

Os avanços da quimiometria tem impulsionado o desenvolvimento da lipidômica em muitas áreas específicas, uma vez que a mesma propicia a obtenção de informações relevantes provenientes de sistemas altamente complexos [23]. Neste contexto, a análise multivariada se torna imprescindível para a interpretação de espectros de massas de amostras biológicas provenientes de pacientes com diferentes condições clínicas. Algumas das etapas envolvidas na análise multivariada consistem de pré-processamento, seleção de variáveis, identificação de metabólitos e modelagem [23,24].

2. Rastreamento de câncer por abordagem lipidômica: um estudo de caso no estado do Rio Grande do Norte

Em um projeto pioneiro no Brasil, nosso grupo de pesquisa relatou o uso da técnica de infravermelho médio associada à análise multivariada na classificação de

lesões intraepiteliais cervicais em mulheres do estado do Rio Grande do Norte, como uma possível alternativa de rastreamento para o câncer de colo de útero (Capítulo 3 da presente tese) [1]. Como forma de ampliar a abordagem, o presente capítulo trata do empregado a espectrometria de massas juntamente com ferramentas quimiométricas para classificação de quadros normais, de LSIL e de HSIL a partir da análise de plasma sanguíneo. Diferentemente do trabalho, anterior, a presente proposta foca em uma abordagem lipidômica, a qual torna possível a identificação de possíveis metabólitos associados às diferentes condições.

2.1 Metodologia empregada

Inicialmente, amostras de sangue foram coletadas de 76 pacientes frequentando a Maternidade Escola Januário Cicco (Natal-RN) durante o período de julho de 2014 a março de 2016. Esta etapa foi realizada conforme aprovação do Comitê de Ética para pesquisa humana do Hospital Universitário Onofre Lopes (Protocolo # 526/11). Após a coleta, o sangue foi processado e o plasma obtido e armazenado a –80 ^{o}C para posterior análise. Cada voluntária foi encaminhada aos exames convencionais (Papanicolaou ou cirurgia de alta frequência) para obtenção de laudos de referência nos quais apontaram para 42 pacientes NILM (negativo para lesão intraepitelial ou malignidade) e 34 SIL (lesão intraepitelial escamosa), dos quais 13 são do grupo LSIL e 31 HSIL.

A extração dos lipídos foi realizada de acordo com o método Folch [25], utilizando uma mistura de metanol-clorofórmio 1:2 *v/v*. Após isso os solventes foram eliminados e os lipídeos resultados foram reconstituídos em isopropanol priori análise por espectrometria de massas.

As amostras foram analisadas em um equipamento Electrospray Ionization (ESI) Q Quadrupole-Orbitrap Hydribid Exact (Thermo Scientific). A mistura de lipídeos foi analisada por injeção direta no modo positivo e os espectros foram obtidos na faixa de 200-1200 *m/z*. Os espectros foram processados usando um *software* Xcalibur Analysis (versão 2.0, Thermo Electron Corporation). A base de dados Lipid Maps Lipidomic Gateway foi utilizada para identificação de metabólitos, considerando um erro de até 1.55 ppm.

A análise computacional foi realizada através do *software* MATLAB® R2010a (Math Works Inc., USA) utilizando a "caixa de ferramentas" PLS Toolbox versão 7.9.3 (Eingenvector Research Inc., USA) e rotinas feitas em laboratório. Foram realizados

pré-processamentos para montagem da matriz, normalização das intensidades, e seleção de variáveis. A matriz pré-processada foi utilizada para construção de modelos PCA-LDA, PCA-QDA, PCA-SVM, GA-LDA, GA-QDA e GA-SVM, afim de classificar as amostras entre NILM e SIL.

2.2 Principais resultados obtidos e discussões

Uma vez que os lipídeos possuem importantes funções associadas ao metabolismo, tanto em quadros fisiológicos normais quanto na presença de patologia, foi realizada uma análise lipidômica do plasma de 76 pacientes via ESI-MS por uma abordagem *untargeted*. Após os pré-processamentos, os espectros de massas médios das 42 amostras NILM e 34 SIL são apresentados nas Figuras 1A-B. Para ambas as classes, os espectros se mostram bastante similares, com maior concentração de sinais em regiões de menores valores de *m/z*. Como forma de diminuir o número de variáveis, foi aplicado o algoritmo ROI (região de interesse), resultando em uma matriz de dimensão 76x278, a qual foi posteriormente utilizada para construção dos modelos de classificação.

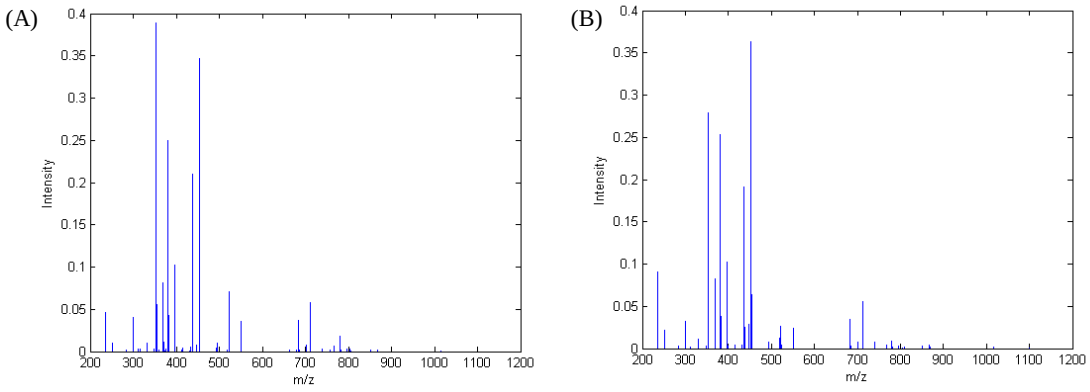


Figura 1. Espectros pré-processados das amostras NILM (A) e SIL (B).

A diferença entre os espectros médios de NILM e SIL, após pré-processamentos, é apresentada na Figura 2A. Como já esperado, apenas diferenças sutis foram observadas entre as duas classes. Os sinais positivos e negativos implicam que os valores de *m/z* específicos são mais intensos nas classes NILM e SIL, respectivamente. Foram identificados uma série de metabólitos associados a 5 valores de *m/z*, os quais estão foram relacionados às diferenças observadas após a subtração dos espectros

médios (Tabela 1). A Figura 2B apresenta as estruturas químicas de alguns dos possíveis metabólitos associados à diferenciação das classes NILM e SIL.

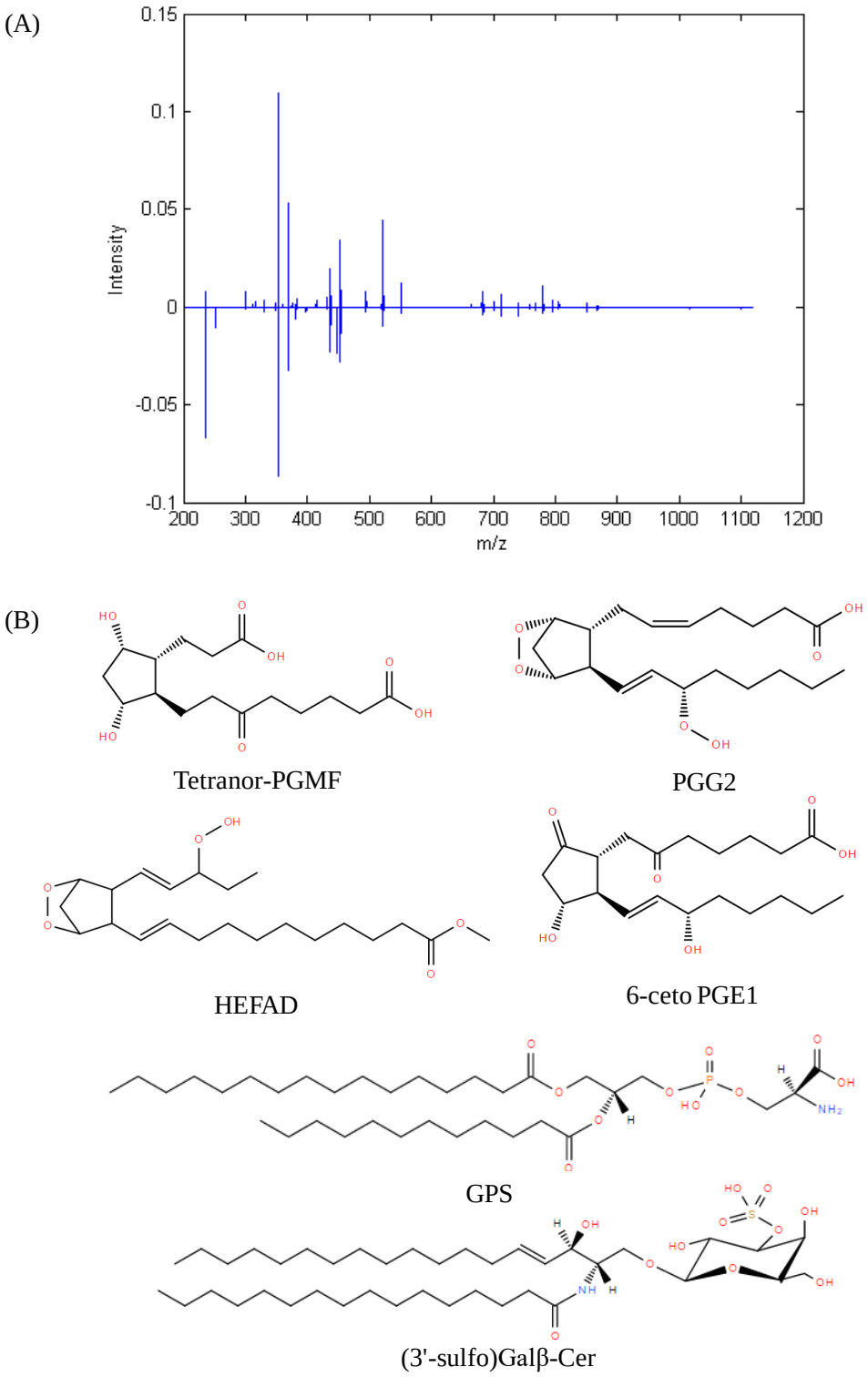


Figura 2. (A) Diferença entre os espectros médios das classes NILM e SIL; (B) estruturas químicas de alguns metabólitos selecionados.

Para a classe NILM, os lipídeos associados à fórmula $C_{20}H_{33}O_6$(*m/z* 369,227) consistem de metabólitos de prostaglandinas, tais como PGG2, 6-ceto PGE1, 20-hidroxi-PGE2, entre outros, além de tromboxanos B2 e B3, e derivados de ácido eicosanóico. Em geral, estas espécies são associadas a importantes fisiológicas, tais como sinalização e regulação de diversos processos homeostáticos e inflamatórios [26]. Também para a classe NILM, foram identificadas uma espécie glicerofosfolipídica derivada da serina ($C_{34}H_{67}O_{10}NP$, *m/z* 680,450) e um esfingolipídeo da subclasse dos sulfatídeos ($C_{40}H_{78}O_{11}NS$, *m/z* 780,526). Fosfolipídeos possuem papel crucial na estrutura das membranas celulares [27], enquanto os sulfatídeos são encontrados em diversas partes do corpo e compreendem espécies químicas multifuncionais [28]. Por outro lado, para a classe SIL, foi identificada a prostaglandina Tertanor-PGFM ($C_{16}H_{27}O_7$, *m/z* 331,117), um metabólito da $PGF_{2\alpha}$ [29]. A fórmula $C_{22}H_{37}O_6$ (*m/z* 397,258) foi supostamente associada a um derivado de ácido graxo hidroperoxidado e epoxidado, os quais podem ser formados a partir de processos de oxidação de lipídeos, incluindo aqueles associados ao desenvolvimento do câncer [30].

Tabela 1. Principais informações químicas associadas à diferenciação das classes NILM e SIL via abordagem lipidômica.

m/z	Erro[a]	Fórmula molecular	Possível lipídeo	Classe	Amostra
331.177	1.540	$C_{16}H_{27}O_7$	Tetranor-PGFM	FA[b]	SIL
369.227	0.853	$C_{20}H_{33}O_6$	PG	FA[b]	NILM
397.258	-0.643	$C_{22}H_{37}O_6$	HEFAD	FA[b]	SIL
680.450	0.490	$C_{34}H_{67}O_{10}NP$	GPS	GPL[c]	NILM
780.526	-1.249	$C_{40}H_{78}O_{11}NS$	(3'-sulfo)Galβ-Cer	SPL[d]	NILM

[a]Erro in ppm; [b]FA = ácidos graxos; [c]GPL = Glicerofosfolipídeos; [d]SPL = esfingolipídeos

Uma análise não supervisionada foi realizada através da PCA na busca por agrupamentos naturais entre as amostras em função da informação química relacionada à cada classe. As primeiras três componentes contabilizaram 64,8% da variança explicada e o gráfico dos *scores* para as primeiras duas componentes é apresentado pela Figura 3.

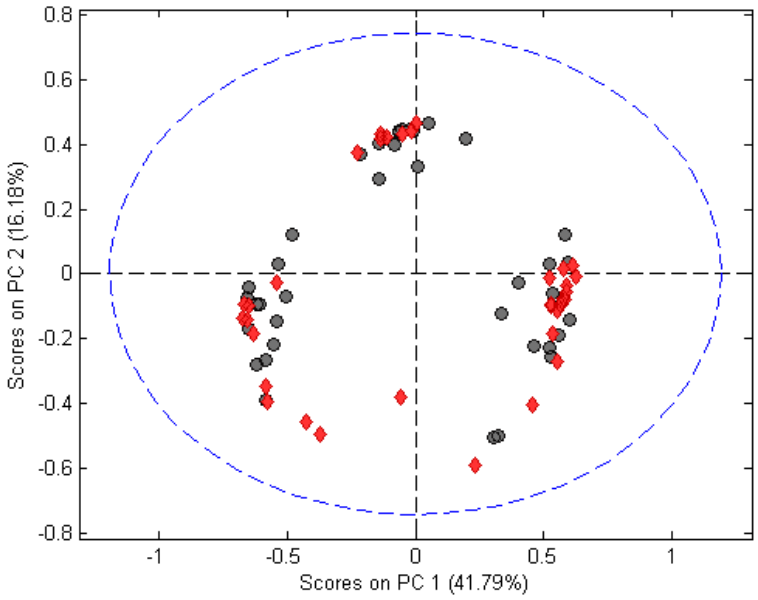


Figura 3. Gráfico dos *scores* da PCA para as amostras NILM (losangos vermelhos) e SIL (círculos cinzas).

Uma vez que a PCA não permitiu uma discriminação entre as classes NILM e SIL das amostras, foram aplicados os modelos de classificação supervisionados LDA, QDA e SVM. Foram utilizados tantos os *scores* provenientes da PCA quanto as variáveis selecionadas pelo GA. Os valores de sensibilidade e especificidade para os modelos LDA e QDA estão apresentados na Tabela 2. De modo geral, os modelos não apresentaram bons índices de classificação, com sensibilidade de 33-100% para NILM e 0-60% para SIL, e especificidade de 33-100% para NILM e 0-80% para SIL.

Tabela 2. Resultados de sensibilidade e especificidade para as amostras de previsão para classificação de NILM *vs* SIL utilizando modelos LDA e QDA.

	NILM		**SIL**	
	Sensibilidade[a]	Especificidade[a]	Sensibilidade[a]	Especificidade[a]
PCA-LDA	33.3	33.3	60.0	60.0
GA-LDA	50.0	83.3	60.0	80.0
PCA-QDA	100	100	0	0
GA-QDA	83.3	66.7	40.0	0

[a]Valores dados em porcentagem (%).

Como esperado, em função da grande variedade e complexidade com que os lipídeos podem existir em diferences organismos e condições (fisiológicas e patológicas), a abordagem *untargeted* pode se contemplar informação não diretamente associada à lesões cervicais. Além disso, a classe SIL também apresenta variação interna, uma vez que a mesma possui duas subclasses (LSIL e HSIL).

Dada a dificuldade observada nas classificações feitas por LDA e QDA, é de se considerar um método supervisionado não linear. Neste contexto, o SVM se mostra uma ferramenta poderosa para lidar com dados biológicos, uma vez que estes geralmente seguem respostas não lineares [23]. A Tabela 3 apresenta os resultados obtidos para classificar as amostras entre NILM e SIL utilizando modelos SVM.

Tabela 3. Resultados (em porcentagem) de sensibilidade e especificidade para as amostras de previsão para classificação de NILM *vs* SIL utilizando modelos SVM.

	NILM		**SIL**	
	Sensibilidade[a]	Especificidade[a]	Sensibilidade[a]	Especificidade[a]
PCA-SVM-L	33.3	33.3	60.0	60.0
PCA-SVM-Q	50.0	50.0	80.0	80.0
PCA-SVM-P	83.3	83.3	80.0	80.0
PCA-SVM-RBF	83.3	83.3	80.0	80.0
PCA-SVM-MLP	16.7	16.7	20.0	20.0
GA-SVM-L	50.0	66.7	80.0	40.0
GA-SVM-Q	16.6	83.3	80.0	20.0
GA-SVM-P	66.7	83.3	40.0	20.0
GA-SVM-RBF	66.7	50.0	40.0	40.0
GA-SVM-MLP	33.3	66.7	60.0	80.0

Foram testadas 5 funções Kernel diferentes visando o melhor ajuste para os dados: linear (L), quadrática (Q), polinomial de terceira ordem (O), função de base radial (RBF) e perceptron multicamadas (MLP). Da maneira similar aos modelos LDA e QDA, o SVM foi empregado utilizando ou os *escores* da PCA ou as variáveis selecionadas pelo GA. Conforme apresentado na Tabela 3, os valores de sensibilidade e especificidade para a classe NILM variaram de 16-83%, e para a classe SIL, de 20-80%. O melhor modelo foi o PCA-SVM-RBF, que alcançou valores de sensibilidade e especificidade de 83,3% e 80,0% para NILM e SIL, respectivamente.

A Figura 4 apresenta os *loadings* das três primeiras componentes principais utilizadas para o modelo PCA-SVM-RBF. As variáveis mais importantes (maiores coeficientes) para a diferenciação das classes estão na região de 200-450 *m/z*. De fato, esta observação está de acordo com a região onde encontram-se as principais diferenças entre as classes NILM e SIL.

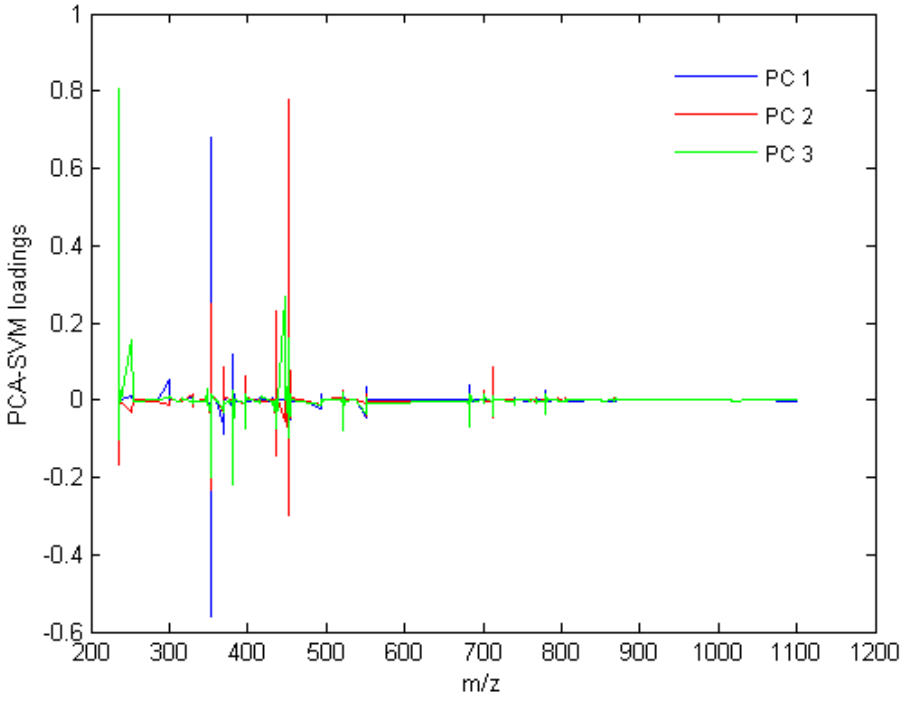


Figure 4. *Loadings* obtidos para o modelo PCA-SVM-RBF: PC1 (azul), PC2 (vermelho) e PC3 (verde).

3. Conclusões

Os resultados obtidos no presente estudo suportam a potencial aplicação da lipidômica, com base na combinação entre espectrometria de massas e análise

multivariada, como uma alternativa promissora para classificar plasma sanguíneo de mulheres entre NILM e SIL. O método se mostrou simples, rápido e com pré-tratamento de amostra bastante reduzido, especialmente por terem sido realizadas análises por injeção direta.

A presente abordagem lipidômica permitiu a identificação de possíveis metabólito associados à diferenciação das classes NILM e SIL, incluindo ácidos graxos, prostaglandinas e seus derivados/metabólitos, tromboxanos, glicerofosfolipídeos e esfingolipídeos.

O melhor modelo, PCA-SVM-RBF, conseguiu valores de sensibilidade e especificidade para NILM e SIL de 83,3% e 80,0%, respectivamente. Estes valores se mostram bastante satisfatórios, especialmente por se tratar de amostras biológicas de alta complexidade sem o uso de uma técnica de separação (tal como cromatografia líquida de alta eficiência).

4. Referências bibliográficas

[1] A. C. O. Neves, P. P. Silva, C. L. M. Morais, C. G. Miranda, J. C. O. Crispim and K. M. G. Lima, *RSC Adv.*, 2016, **6**, 99648–99655.

[2] www.globocan.iarc.fr, acessed in July, 05, 2017.

[3] S. Franceschi and S. Vaccarella, *Cancer Epidemiol.*, 2015, **39**, 1152–1156.

[4] B. F. Lees, B. K. Erickson and W. K. Huh, *Am. J. Obstet. Gynecol.*, 2016, **214**, 438–443.

[5] C. M. de Oliveira, I. G. Bravo, N. C. S. de Souza, M. L. N. D. Genta, J. H. T. G. Fregnani, M. Tacla, J. P. Carvalho, A. Longatto-Filho and J. E. Levi, *Infect. Genet. Evol.*, 2015, **34**, 44–51.

[6] E. J. Nam, J. W. Kim, S. W. Kim, Y. T. Kim, J. H. Kim, B. S. Yoon, N. H. Cho and S. Kim, *Gynecol. Oncol.*, 2007, **104**, 207–211.

[7] F. Cannella, A. Pierangeli, C. Scagnolari, G. Cacciotti, G. Tranquilli, P. Stentella, N. Recine and G. Antonelli, *Immunobiology*, 2015, **220**, 363–368.

[8] C. P. Crum and C. M. McLachlin, *J. Cell. Biochem.*, 1995, **23**, 71–79.

[9] J. Paavonen, *Int. J. Infect. Dis.*, 2007, **11**, S3–S9.

[10] C. J. De Witte, A. J. M. Van De Sande, H. J. Van Beekhuizen, M. M. Koeneman, A. J. Kruse and C. G. Gerestein, *Gynecol. Oncol.*, 2015, **139**, 377–384.

[11] A. G. Waxman, D. Chelmow, T. M. Darragh, H. Lawson and A.-B. Moscicki, *Obs. Gynecol.*, 2012, **120**, 1465–1471.

[12] R. Nayar and D. C. Wilbur, *Acta Cytol.*, 2015, **59**, 121–132.

[13] M. J. Walsh, M. J. German, M. Singh, H. M. Pollock, A. Hammiche, M. Kyrgiou, H. F. Stringfellow, E. Paraskevaidis, P. L. Martin-Hirsch and F. L. Martin, *Cancer Lett.*, 2007, **246**, 1–11.

[14] A. D. Watson, *J. Lipid Res.*, 2006, **47**, 2101–2111.

[15] S. M. Lam and G. Shui, *J. Genet. Genomics*, 2013, **40**, 375–390.

[16] M. R. Wenk, *Cell*, 2010, **143**, 888–895.

[17] N. Navas-Iglesias, A. Carrasco-Pancorbo and L. Cuadros-Rodríguez, *Trends Anal. Chem.*, 2009, **28**, 393–403.

[18] F. Perrotti, C. Rosa, I. Cicalini, P. Sacchetta, P. Del Boccio, D. Genovesi and D. Pieragostino, *Int. J. Mol. Sci.*, 2016, **17**, 1992.

[19] B. Flatley, P. Malone and R. Cramer, *Biochim. Biophys. Acta*, 2014, **1844**, 940–949.

[20] E. P. Diamandis, *Mol. Cell. Proteomics*, 2004, **3**, 367–378.

[21] M. A. M. Rodrigo, O. Zitka, S. Krizkova, A. Moulick, V. Adam and R. Kizek, *J. Pharm. Biomed. Anal.*, 2014, **95**, 245–255.

[22] U. Loizides-Mangold, *FEBS J.*, 2013, **280**, 2817–2829.

[23] L. Yi, N. Dong, Y. Yun, B. Deng, D. Ren, S. Liu and Y. Liang, *Anal. Chim. Acta*, 2016, **914**, 17–34.

[24] S. Datta and L. M. Depadilla, *Stat. Methodol.*, 2006, **3**, 79–92.

[25] R. E. Patterson, A. J. Ducrocq, D. J. McDougall, T. J. Garrett and R. A. Yost, *J. Chromatogr. B*, 2015, **1002**, 260–266.

[26] E. A. Dennis and P. C. Norris, *Nat. Rev. Immunol.*, 2015, **15**, 724.

[27] J. Li, X. Wang, T. Zhang, C. Wang, Z. Huang, X. Luo and Y. Deng, *Asian J. Pharm. Sci.*, 2015, **10**, 81–98.

[28] T. Takahashi and T. Suzuki, *J. Lipid Res.*, 2012, **53**, 1437–1450.

[29] W. Fenical, D. R. Kearns and P. Radlickal, *J. Am. Chem. Soc.*, 1969, **50**, 3398–3400.

[30] K. Fujimoto, W. E. Neff and E. N. Frankel, *Biochim. Biophys. Acta*, 1984, **795**, 100–107.

Capítulo 5 - Considerações finais e perspectivas

Câncer é uma das doenças que mais mata, na atualidade. Uma questão de fundamental importância na luta contra o câncer consiste no diagnóstico precoce. Programas de rastreamento podem contribuir fortemente para o diagnóstico precoce do câncer, entretanto, muitas vezes estas técnicas exigem infraestrutura adequada, além de pessoal especializado, o que pode dificultar a implementação e limitar o alcance de pessoas. Além disso, algumas são propensas à subjetividade do analista, podendo levar a resultados falsos negativos. Nesta tese, foi proposto o uso de técnicas espectroscópicas e multivariadas para a identificação e rastreamento de câncer.

A fluorescência molecular foi aplicada em uma metodologia muito simples, rápida e de baixo custo para o estudo de células em cultivo (Capítulo 2). Foi possível a classificação de diferentes linhagens celulares normais e cancerosas, através dos algoritmos OPLS/UPLS-DA, com taxas de acerto satisfatórias, acima de 75%. Ainda, foi avaliada a influência dos anticorpos anti-MMP-2 e anti-MMP-9 na eficiência dos modelos de classificação, de modo que foi possível classificar corretamente todas as amostras analisadas. Os resultados alcançados através desta metodologia mostram a possibilidade da fluorescência molecular ser utilizada no estudo de células, como uma alternativa à citometria de fluxo, por exemplo, para fins de análises de rastreamento, mais rápidas e menos custosas.

A espectroscopia ATR-FTIR foi aplicada juntamente às técnicas multivariadas PCA-LDA/QDA, SPA-LDA/QDA e GA-LDA/QDA para a classificação de plasmas sanguíneos de mulheres saudáveis e portadoras de lesão cervical (Capítulo 3). Os resultados mostraram que a metodologia proposta é capaz de classificar corretamente as amostras com valores de sensibilidade e especificidade comparáveis às metodologias de rastreamento tradicionais (variando de 67% a 100%). A metodologia proposta é realizada de maneira rápida, não exigindo a participação de pessoal altamente especializado e de infraestrutura e/ou materiais de alto custo. Também, elimina algumas etapas que podem ser influenciadas pela subjetividade do analista, como existentes na metodologia tradicional do teste Papanicolaou. Ainda, foi possível discriminar as amostras não somente pela ausência ou presença de lesão, mas também pelo grau da lesão cervical, resultado que possui bastante influência na clínica, uma vez que o tratamento realizado

para pacientes com alto e baixo grau de lesão cervical intraepitelial é essencialmente diferente.

A espectrometria de massas foi aplicada juntamente à análise multivariada em uma abordagem lipidômica, por injeção direta, dos lipídios extraídos do plasma sanguíneo de mulheres saudáveis e portadoras de lesão cervical (Capítulo 4). Diversos métodos de classificação multivariados foram comparados (LDA, QDA e SVM) e os resultados obtidos através do PCA-SVM permitiram classificar as amostras de NILM e SIL com sensibilidade e especificidade de 83.3% e 80.0%, respectivamente. Alguns possíveis lipídios foram atribuídos às diferenças espectrais existentes entre as duas classes, tais como prostaglandinas, fosfolipídios e esfingolipídios relacionados à Classe NILM, enquanto para SIL relacionou-se o lipídio Tetranor-PGFM e um lipídio hidroperoxidado.

Os estudos propostos mostraram através de seus resultados a possibilidade de introdução da bioespectroscopia como ferramenta para o estudo do câncer, no Brasil. Entretanto, muitos esforços ainda são necessários principalmente no sentido de ampliar o número de amostras analisadas e, então, prosseguir para os métodos de validação. Ainda, as abordagens relatadas nesta tese podem ser extrapoladas para outros tipos de amostras biológicas, como tecido, urina, saliva, e para outros tipos de câncer. Com este propósito, foi estabelecida uma parceria com a Liga Contra o Câncer do Rio Grande do Norte, em 2015, para realização de um projeto de rastreamento para o câncer de mama, através da espectroscopia ATR-FTIR e análise multivariada. A metodologia utilizada para aquisição e tratamento dos dados espectrais será a mesma, conforme mostrado no segundo estudo relatado nesta tese (Capítulo 3) e, pretende-se obter uma amostragem de 500 mulheres voluntárias, incluindo saudáveis e portadoras de câncer de mama recém- diagnosticado.

Apêndice A

Area correlation constraint for the MCR-ALS quantification of cholesterol using EEM fluorescence data: A new approach

Ana C. O. Neves

Romá Tauler

Kássio M. G. de Lima

Analytica Chimica Acta, 2016, 937, 21-28.

Contribuição:

- Realizei a preparação dos experimentos e a aquisição espectral;
- Participei do processamento dos dados e da construção dos modelos multivariados;
- Escrevi a primeira versão do manuscrito.

Ana C. O. Neves

Prof. Kássio M. G. Lima

Analytica Chimica Acta 937 (2016) 21–28

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Area correlation constraint for the MCR–ALS quantification of cholesterol using EEM fluorescence data: A new approach

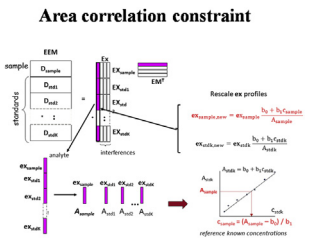
Ana Carolina de Oliveira Neves [a], Romá Tauler [b], Kássio Michell Gomes de Lima [a, *]

[a] Institute of Chemistry, Post Graduation Program in Chemistry, Biological Chemistry and Chemometrics Research Group, CEP 59072-970, Natal, RN, Brazil
[b] IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain

HIGHLIGHTS

- A new additional constraint for the MCR–ALS is proposed.
- Area correlation constraint has been applied to EEM data.
- Two data sets were investigated.
- This constraint approach may be extrapolated to the general case of second-order multivariate calibration data.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 29 May 2016
Received in revised form
9 August 2016
Accepted 12 August 2016
Available online 13 August 2016

Keywords:
Cholesterol
Excitation-emission matrix
Area correlation constraint
MCR–ALS

ABSTRACT

This work demonstrates the use of a new additional constraint for the Multivariate Curve Resolution–Alternating Least Squares (MCR–ALS) algorithm called "area correlation constraint", introduced to build calibration models for Excitation Emission Matrix (EEM) data. We propose the application of area correlation constraint MCR–ALS for the quantification of cholesterol using a simulated data set and an experimental data system (cholesterol in a ternary mixture). This new constraint includes pseudo-univariate local regressions using the area of resolved profiles against reference values during the alternating least squares optimization, to provide directly accurate quantifications of a specific analyte in concentration units. In the two datasets investigated in this work, the new constraint retrieved correctly the analyte and interference spectral profiles and performed accurate estimations of cholesterol concentrations in test samples. This the first study using the proposed area constraint using EEM measurements. This new constraint approach emerges as a new possibility to be tested in general cases of second-order multivariate calibration data in the presence of unknown interferents or in more involved higher order calibration cases.

1. Introduction

Multivariate Curve Resolution–Alternating Least Squares (MCR–ALS), nowadays, is widely used in the chemical sciences for the analysis of very diverse kinds of data such as chromatography [1], kinetic modeling [2], calibration [3], spectroscopic image [4], multilinear fluorescence spectra [5], environmental [6], multiset data [7] and others [8]. In 2014–2015, 149 Scopus-indexed journal papers were published with keywords "MCR–ALS" or "Multivariate Curve Resolution–Alternating Least Squares". MCR–ALS was used for developing or comparing algorithms ($n = 28$) in some research fields (pharmaceutical, biological science, environmental, nanoparticle, and food).

As it has been shown in these studies, the inclusion of knowledge about the system in the form of natural constraints (non-negativity in the concentration and spectral profiles, unimodality,

* Corresponding author.
E-mail address: kassiolima@gmail.com (K.M.G. de Lima).

http://dx.doi.org/10.1016/j.aca.2016.08.011
0003-2670/

closure, selectivity, trilinearity, local rank, …) is enough to take the resolution to optimal solutions. A variant of MCR–ALS with a correlation constraint was proposed by Antunes et al. [9] to obtain quantitative information of analytes in real concentration units in the presence of interferences for the calibration of first order data. The idea behind of MCR–ALS using correlation constraint is to build an internal calibration model relating the real concentration values of calibration samples to the ones in arbitrary units provided by the constrained least-squares calculated concentration profiles during the MCR optimization procedures. This constraint was used for first order data and correction of matrix effect as can be found in some studies [3,10–13].

In the present report, we explored the possibility of introducing a new additional constraint for the MCR-ALS algorithm called "area correlation constraint" which is presented aiming to improve the properties of this method for the quantification of a specific analyte in the presence of unknown interferences. This constraint includes pseudo-univariate local regressions using the area of resolved profiles against reference values during the alternating least squares optimization, to provide directly accurate quantifications of a specific analyte in concentration units. For this new correlation constraint, we propose the application of MCR-ALS for the quantification of cholesterol by means of Excitation Emission Matrix (EEM) data. Cholesterol is a steroidal alcohol, highly controversial biomolecule present in humans and crucial to synthesis of several hormones and bile salts, besides being part of cell membranes [14]. Chemical information can be extracted from both excitation and emission spectra which allows improving the selectivity of the technique [15].

Two data sets were investigated. The first data set was obtained by simulation of the EEM fluorescent spectral response of mixtures of cholesterol, triglycerides and glucose at different concentrations and using their pure EEM spectra previously measured in the laboratory. This will be the simulated samples data set. The second case was similar to the previous one but preparing the mixtures in the laboratory and measuring their EEM spectra. This will be the synthetic samples data set. The first data set will be used specially to test the area correlation constraint in MCR–ALS proposed in this work, while the second dataset will be used to test the proposed method in a system composed by real synthetic samples. To the best of our knowledge, there is no report in literature focusing in quantification of cholesterol by EEM fluorescence.

2. Experimental section

2.1. Pure standards and mixtures

Standards of glucose (100 mg dL^{-1}), triglycerides (200 mg dL^{-1}) and cholesterol (20 mg dL^{-1}) were obtained by enzymatic assay kits supplied by Labtest, Brazil. The samples were prepared by adding varying amounts of the three pure standards solutions and the final volume of each mixture was completed to 2 mL using deionized water.

2.2. Experimental data sets

2.2.1. Dataset 1: simulated samples

EEM fluorescence response of 16 mixtures of 3 constituents glucose, cholesterol and triglycerides were simulated using their previously known pure emission spectra (Fig. 1a, b and 1c) at eight

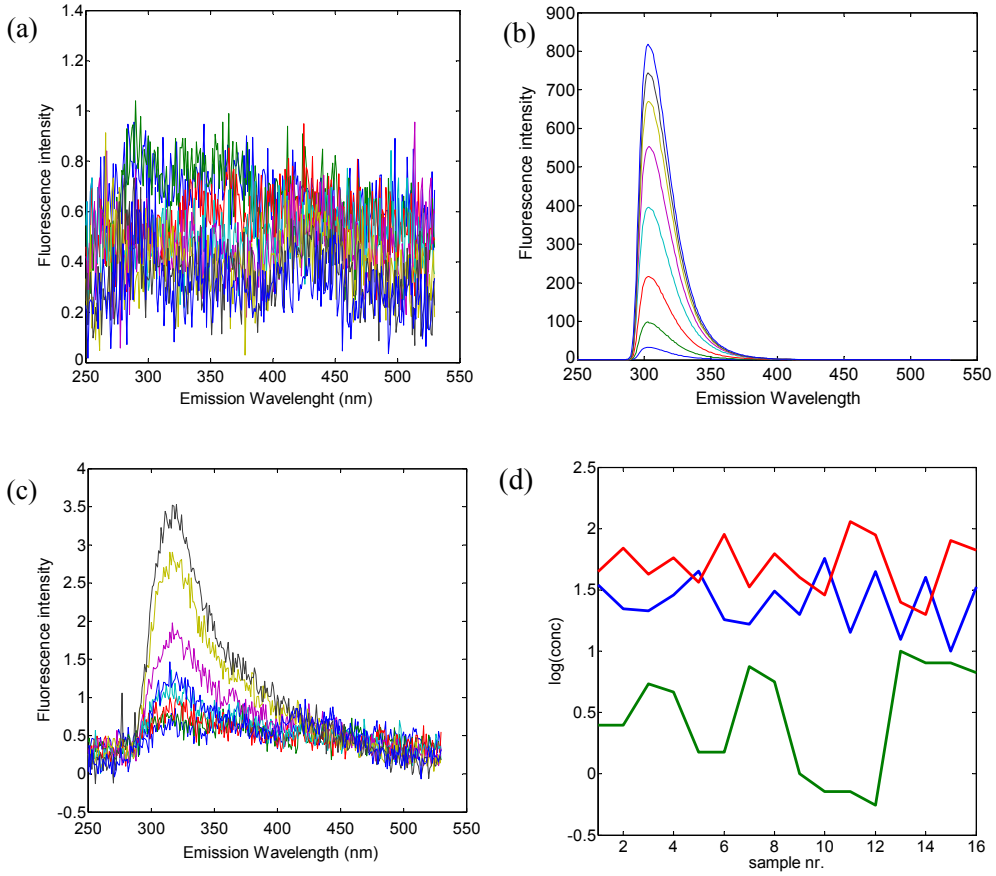


Fig. 1. Pure emission spectra simulated of 16 mixtures of 3 constituents glucose (a), cholesterol (b) and triglycerides (c) at eight different excitation wavelengths. (d) Respective concentrations (in logarithmic scale) in the mixtures changing randomly at different levels.

different excitation wavelengths (220–290 nm, with 10 nm step) and considering their respective concentrations in the mixtures changing randomly at different levels as shown in Fig. 1d (in logarithmic scale, to show them in the same plot, since triglycerides was at higher concentrations, see below). Pure EEM fluorescence spectra of the three constituents were obtained from experimental measurements of them at 5.0 mg dL^{-1}, 10.0 mg dL^{-1} and 1.0 mg dL^{-1} for glucose, triglycerides and cholesterol, respectively. The experimental study of this ternary system was carried out considering the presence of one analyte and two interferents. Fig. 1a is showing only noise, since glucose is not a fluorescent compound. In fact, glucose was included in attempt to simulate a more complex situation since glucose can be excited by UV radiation. On the other hand, as can be seen in Fig. 1b, triglycerides is fluorescent, which could lead to interference in the quantification of cholesterol from both excitation and emission signals. In this case, the goal of the study was the investigation of the effectiveness of the application of the new area correlation constraint implemented in the MCR-ALS method by checking the recovery of the concentrations of cholesterol in these mixtures. This simulated data set resulted in a data matrix size of 8 × 280 for each sample.

2.2.2. Dataset 2: laboratory synthetic samples

Determination of cholesterol levels in laboratory synthetic samples was performed taken into account the presence of two biochemical compounds commonly quantified in clinical analysis, glucose and triglycerides, in attempt to induce a possible matrix effect. The cubic D-optimal mixture design was developed to set the composition of the calibration and validation samples, totalizing fifteen mixtures. The calibration samples were chosen to cover all the concentration range. On the other side, validation samples were chosen randomly inside this range to avoid extrapolation. The experimental design used three factors (glucose, cholesterol and triglycerides) and six levels for each factor. The concentrations were chosen to avoid signal detection saturation problems related to fluorescence analysis (especially for the emission of cholesterol). As can be seen in Table 1, the diluted concentration ranges of cholesterol, glucose and triglycerides in the samples were 0.01–0.04 mg dL^{-1}, 0.05–0.1 mg dL^{-1} and 0.1–0.2 mg dL^{-1}, respectively.

2.3. Instrumentation

EEM fluorescence spectra of these synthetic samples were acquired with a RF-5301 Shimadzu spectrofluorometer, through a 1 cm quartz cuvette. The excitation and emission monochromator slit widths were fixed at 1.5 and 3 nm, respectively. The temperature was kept at 25 °C throughout the experiments. The EEM spectra were recorded at excitation wavelengths from 220 to 320 nm at regular steps of 10 nm; the emission wavelengths were from 220 to 700 nm at regular steps of 1 nm. This protocol resulted in a data matrix size of 11 × 480 for each sample. All EEM spectra were preprocessed to remove the observed Rayleigh and Raman scatterings by using the algorithm proposed by Zepp et al., as can be appreciated in Fig. 2. The algorithm removes emission peaks ±10–15 nm at each excitation wavelength from the EEM matrix and fills the missing regions using Delaunay interpolation of the surrounding data points [16].

Table 1
Experimental design of the synthetic samples: calibration and validation sets.

	Glucose (mg dL^{-1})	Triglycerides (mg dL^{-1})	Cholesterol (mg dL^{-1})
Calibration			
1	0.083	0.133	0.020
2	0.066	0.133	0.030
3	0.091	0.117	0.015
4	0.083	0.117	0.035
5	0.050	0.100	0.010
6	0.050	0.200	0.010
7	0.050	0.100	0.040
8	0.100	0.200	0.040
Validation			
9	0.066	0.167	0.020
10	0.083	0.167	0.030
11	0.083	0.183	0.015
12	0.091	0.183	0.035
13	0.100	0.100	0.010
14	0.100	0.200	0.010
15	0.100	0.200	0.040

2.4. Theory

2.4.1. Multivariate curve resolution-alternating least squares (MCR-ALS)

Multivariate curve resolution (MCR) is a bilinear decomposition method, which assumes the additivity of the individual responses of each component present in the investigated system. MCR is proposed to resolve the mixture an**a**lysis problem in multicomponent systems identifying the main sources of variation, constituents and profiles contributing to the raw measurements [17]. MCR is performed by decomposing the experimental data matrix **D,** into the product of the pure component response profiles, grouped in a column matrix, **C,** commonly associated with the concentration profiles, and in a row matrix, $\mathbf{S}^T$, usually associated with the spectra of the constituents, leading therefore to a simple bilinear model, according to Equation (1a):

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \quad (1a)$$

Dimensions of **D** matrix are (I,J), where I are its number of rows, for instance number of spectra measured on samples and J number of wavelengths or spectroscopic channels. Dimensions of **C** and $\mathbf{S}^{\mathbf{T}}$ matrices are respectively (I,N) and (N,J), where N are the number of components contributing to the measured signal. Matrix **E,** of the same dimensions (I,J) as **D**, has the residuals not explained by the model $\mathbf{CS}^{\mathbf{T}}$, which should be small and ideally close to the experimental noise. The model parameters, **C** and $\mathbf{S}^{\mathbf{T}}$, are estimated using an iterative optimization by an Alternating Least Squares (ALS) algorithm that fits **C** and $\mathbf{S}^T$ matrices to the experimental data **D**, for a proposed number of components, N, using initial estimates of either **C** or $\mathbf{S}^T$. These initial estimates can be obtained in different manners, for instance using the purest emission or excitation spectra of the raw (measured) data. See references [4,18] and [29] for more details. This selection was not critical for the data examples studied in this work since rotation ambiguities were very low (see below in the results section) and the experimental design used in the preparation of samples. During the ALS optimization, a variety of constraints can be applied, such as non-negativity, unimodality, closure, trilinearity and/or selectivity, aiming to facilitate the convergence of algorithms, resolve rotational ambiguities and ensure chemically meaningful solutions. See Refs. [18,19] for more details about application of constraints in MCR-ALS method.

In the case of the analysis of a single sample by EEM fluorimetry, the data matrix **D** has the emission spectra at every excitation wavelength in its rows, and the excitation spectra at every emission wavelength in the columns, therefore Equation (1a) is adapted to the EEM notation in the analogous Equation (1b):

$$\mathbf{D} = \mathbf{E}_{\mathbf{X}}\mathbf{E}_{\mathbf{M}}^{\mathbf{T}} + \mathrm{E} \quad (1b)$$

where now, $\mathbf{E}_{\mathbf{X}}$ is the column matrix of the pure excitation spectra

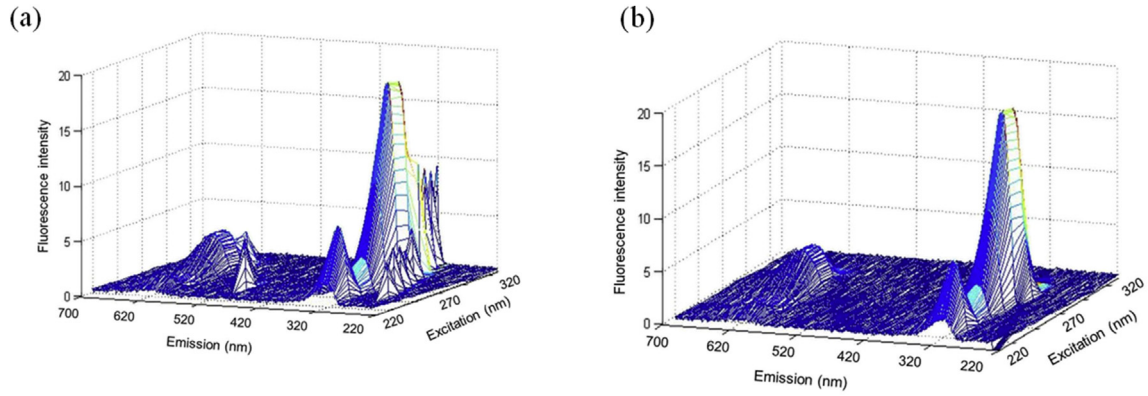


Fig. 2. EEM fluorescence spectra obtained for sample 1 (from Table 1): before (a) and after (b) removing Rayleigh and Raman scatterings.

and $\mathbf{E_M^T}$ is the row matrix of pure emission spectra of the fluorescent components in the sample.

The MCR model of Equations (1a) and (1b) can be extended to the analysis of more complex data arrangements, where multiple data sets or data matrices coming from the analysis of multiple samples or from multiple experiments (k = 1, …,K) are simultaneously analyzed. In these situations, the data can be structured as a column-wise augmented data matrix, $\mathbf{D}_{aug}$, of dimensions (K x I,J) where model of Equation (1a) is adapted to the new model of Equation (2a):

$$\mathbf{D_{aug}} = \begin{pmatrix} \mathbf{D_1} \\ \mathbf{D_2} \\ \dots \\ \mathbf{D_K} \end{pmatrix} = \begin{pmatrix} \mathbf{C_1} \\ \mathbf{C_2} \\ \dots \\ \mathbf{C_K} \end{pmatrix} \mathbf{S^T} + \mathbf{E_{aug}} = \mathbf{C_{aug}} \mathbf{S^T} + \mathbf{E_{aug}} \qquad (2a)$$

where $\mathbf{C_{aug}}$ is the augmented concentration matrix giving the concentration profiles of every component in every individual data matrix $\mathbf{D_k}$ (k = 1,..,K) and $\mathbf{S^T}$ is the matrix of spectra of the resolved components. Observe that in the model of Equation (2a), the concentration profiles of every component in the different samples simultaneously analyzed are allowed to be different, whereas the spectra of the resolved components are forced to be the same. The flexibility of Equation (2a) allows the correct application of the bilinear model to many situations in chemistry where the concentration profiles of the same component in different situations may change in position and in shape, like in chromatography or in many chemical reaction systems. In the case of multiple EEM data matrices simultaneously analyzed, the extended model of Equation (2a) is adapted to the new notation:

$$\mathbf{D_{aug}} = \begin{pmatrix} \mathbf{D_1} \\ \mathbf{D_2} \\ \dots \\ \mathbf{D_K} \end{pmatrix} = \begin{pmatrix} \mathbf{E_{X1}} \\ \mathbf{E_{X2}} \\ \dots \\ \mathbf{E_{XK}} \end{pmatrix} \mathbf{E_M^T} + \mathbf{E_{aug}} = \mathbf{E_{X\ aug}} \mathbf{E_M^T} + \mathbf{E_{aug}} \qquad (2b)$$

where now $\mathbf{E_{Xaug}}$ is the augmented matrix of the pure excitation spectra, and $\mathbf{E_M^T}$ is the matrix of the emission spectra of the resolved components, and $\mathbf{E_{aug}}$ is the matrix of the augmented residuals. In the case of EEM data, since the two modes (directions, ways) of the data matrix are spectral modes, excitation and emission fluorescent spectra, the bilinear model of previous Equations can be further extended to the so-called trilinear model, where now the excitation matrices of Equation (2b) only differ in the relative intensities/scales of the excitation profiles. These scale changes between **Ex** matrices contain therefore the relative quantitative information of every fluorescent component in the different samples. This can be written by

$$\mathbf{D_k} = \mathbf{E_X} \cdot \mathbf{C_k} \mathbf{E_M^T} + \mathbf{E_k} \qquad (3)$$

where $\mathbf{C_k}$ is now a diagonal matrix of dimensions (N,N), with the relative concentrations of the different resolved components. When experimental data follow this model, MCR-ALS can be easily adapted to it, and an appropriate trilinearity constraint can be applied to enforce the common excitation profiles of the same component in the different sample data matrices, $\mathbf{E_{Xk}}$, to be the same. More details about the trilinear and other multilinear extensions of MCR-ALS are given in detail elsewhere [20–23].

In this work, non-negativity constraints were applied to both emission and excitation spectral profiles by using a fast non-negativity least squares algorithm, during the ALS optimization of the augmented data matrices. When the relative differences in standard deviations of the residuals between experimental and ALS calculated data values are less than a previously selected value, commonly used as 0.1%, in two consecutive cycles, the ALS convergence is achieved. In this phase, when MCR-ALS results are obtained and the components are identified, for quantification purposes, the MCR-ALS scores obtained per analyte and sample as integrated area under the related resolved excitation or emission spectrum, can be regressed against nominal concentration values to build a calibration curve and, then, to predict the analyte concentration of unknown samples [3].

2.4.2. New area correlation constraint for the MCR-ALS analysis of EEM second order data

In this work anew constraint has been introduced to build calibration models for second order data. In previous works [11,24], it has been shown that first order calibration models (calibration models adequate for instruments providing a data vector in the measure of a single sample, like for instance a spectrum) can be built using MCR-ALS with an internal correlation constraint, which at every iteration of the ALS optimization establishes a linear local model between the concentration values of the analyte in a set of calibration samples and the currently estimated concentration (scores) values estimated by ALS. Unknown concentration values of this analyte in other samples are then also estimated by this local model. By the application of this correlation constraint, the obtained results were found to be analogous to those obtained with other well established first order calibration methods like PLSR [3] and the ubiquitously present rotation ambiguities can be eliminated [25] for the concentrations of the calibrated component (analyte) in all samples.

On the other hand, MCR-ALS analysis of second order data (obtained by instruments which provide a data matrix in the

analysis of a single sample, like in EEM spectroscopy or in hyphenated chromatography LC-DAD, GC-MS, LC-MS, …), relative quantitative estimations of one constituent in the different samples (different data matrices) simultaneously analyzed by MCR-ALS, can be easily derived from the relative areas or heights of the concentration (score) profiles of this resolved component in the different samples (data matrices) (see Fig. 3 below). A calibration model can be built also in this case if the concentration of this constituent (analyte) is known in some of the samples (calibration samples), and the derived model used to predict then the concentration of this analyte in the other samples where the concentration of this constituent is unknown. This method has already been used successfully in previous works [11,24] and can give the so-called second order advantage (calibration in the presence of unknown interferences), similarly to other chemometric methods like GRAM [26] or PARAFAC [27]. This approach can be adapted to the different analytical strategies like the standard addition and the internal standard methods [28].

In this work, a step further in this approach is proposed which consists in the implementation of the second order calibration model during every iteration of the ALS optimization as a constraint, and not as post-processing strategy as it was performed until now. As it is shown in Fig. 3, the areas (or heights) of the excitation profiles of the analyte are regressed to the known analyte concentrations at every iteration with the linear model: $A_{stdk} = b_0 + b_1 c_{stdk}$ where A_{stdk} is the area of the emission profile of the sample standard k, b_0 and b_1 are respectively the offset and slope of the regression line, and c_{stdk} is the known concentration of the analyte in standard sample k, and the model parameters b_0 and b_1 are obtained for the k = 1, … K standard samples. $\mathbf{ex_{stdk}}$ analyte profiles in the standard samples are then rescaled according to the model regression parameters, as well as $\mathbf{ex_{sample}}$ analyte profiles in the unknown samples, not used to build the calibration model. In Fig. 3, the inverse linear equations are given to rescale these profiles. In this way, the excitation profiles are properly scaled according to the analyte concentration in the standards and a calibration model is built which can be afterwards used for the estimation of the analyte concentration in the unknown samples. This step is performed at every ALS step and therefore at convergence to optimal data fit, the estimation of the $\mathbf{E_X}$ and $\mathbf{E_M}$ profiles will be also optimal, as well as the analyte concentration estimates. In the preliminary implementation of this constraint in this work, the use of this constraint is only shown as a way to improve quantitative estimations of the selected component for calibration (analyte) and further studies are pursued to show more clearly the effect of this constraint in the reduction of ambiguities and uncertainties in the final results.

The area correlation constraint presented in this work is a new variant of the correlation constraint for MCR-ALS. Previous implementation of the correlation constraint was only for first order data (one sample one response vector). This means that the correlation was established between individual concentration values of the considered analyte in some of the samples and the corresponding known values (e.g. from standards). In this case, the correlation constraint is implemented for second order data (one sample gives a data matrix) and it is applied on the area of the whole (concentration) profile of the analyte, not on the individual values. In the examples shown in this work, the concentration information is from the area of the emission profile of the analyte, and the constraint is applied to this area, not to the individual emission values. This represents a completely different approach and it is only valid for second order data, not for first order data as it was in previous versions.

2.5. Data analysis software

Data were imported and the chemometric models were constructed in MATLABversion7.1 software (Math-Works, Natick, USA). MCR–ALS toolbox works also under MATLAB computation and visualization environment and it was freely obtained at www.mcrals.info. MCR-ALS toolbox can be downloaded from the MCR-ALS web page http://mcrals.wordpress.com/download/mcr-als-toolbox/. MCR-ALS including the area correlation constraint is under development and can be only obtained as a command line

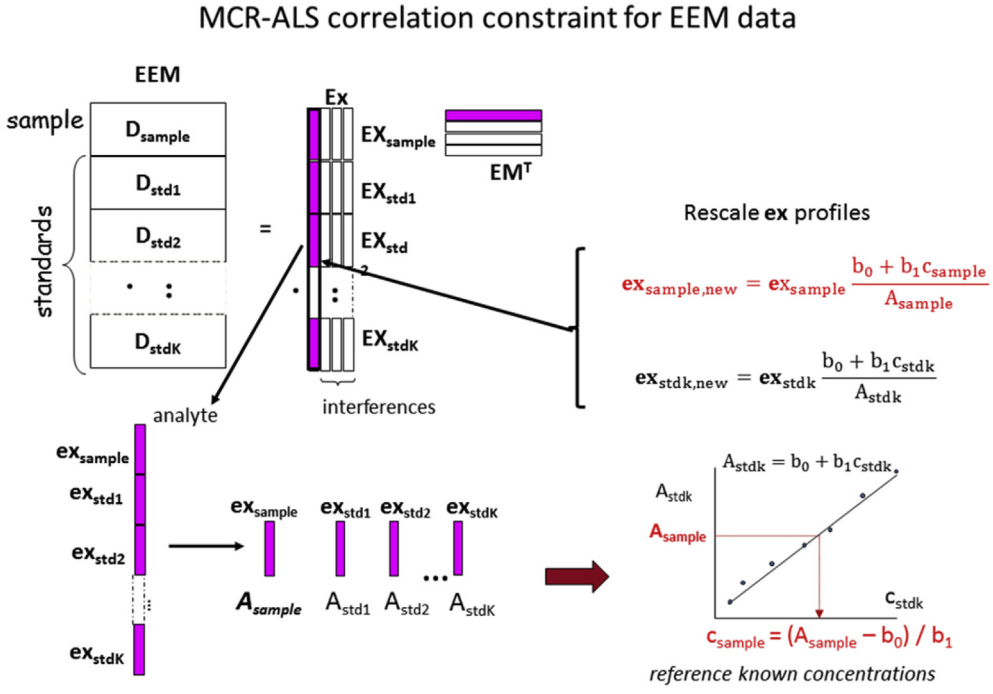


Fig. 3. Area correlation constraint applied on EEM dataset. Detailed description see Experimental Section.

m.file MATLAB version under request to one of the authors (RT).

3. Results and discussion

3.1. Determination of cholesterol in simulated mixture samples

In this first data example, MCR–ALS with the area correlation constraint was tested for the analysis of EEM fluorescent spectra from the simulated mixtures of cholesterol, triglycerides and glucose. Pure emission spectra of these three constituents were shown in Fig. 4 from 250 to 550 nm. In Fig. 5 results obtained by the application of the new introduced area correlation constraint are given for the case of the simulated mixture samples. Calibration concentrations for cholesterol were given for eight of the 16 samples (considered alternatively from the whole data set). Applied constraints were the non-negativity constraints for excitation and emission spectra, the area correlation constraint for the cholesterol concentration values for the calibration samples and the equal relative maximum intensity in the emission spectra of the three components. Lack of fit was 2.58%, R^2 explained variance was 99.93%, and convergence was achieved after a number (46) of ALS iterations.

As seen in Fig. 5a, resolution of emission spectra for cholesterol with band maximum at 304 nm was excellent, coincident with the experimental one shown in Fig. 4b. Although triglycerides had a much lower signal (Fig. 4c) and they were highly overlapping with cholesterol band, they were resolved with a band maximum at 319 nm (Fig. 5a) which coincides also with the experimental one. Third component was the background signal with no shape and only accounting for the background glucose signal (Fig. 4a). In Fig. 5b, MCR-ALS predicted concentrations of cholesterol are compared with those used for the data simulation. Agreement between MCR-ALS calculated and simulated concentrations (Fig. 5b) was excellent, with very low (below 0.1%) relative errors in the prediction of cholesterol concentrations in validation samples. Regression line of predicted versus nominal values was also excellent (with unit slope, zero offset and regression coefficient very close to one). Results obtained for this simulated data example confirms therefore the correct behavior of the newly proposed area correlation constraint, which allowed recovering the correct quantitative concentration of cholesterol in the presence of the other unknown interferences by means of the MCR-ALS method.

These results are added to previous results obtained by MCR–ALS where quantitative information was correctly recovered for first order calibration data sets [11,24]. In this paper this type of quantitative constraint is extended to second order data, where one analytical sample gives a data matrix, like it is the case for excitation-emission fluorescent analysis of individual samples. This result was achieved without the need of application of the trilinearity constraint [21] and it may be therefore extended to the common cases where second order data do not follow the trilinear model, like in chromatographic analysis of samples using multichannel spectral detection (i.e. in LC–DAD, GC–MS or LC–MS, among other). In case of not applying the correlation constraint and considering only bilinear modeling of the data, results were also good, with practically the same data fitting as when correlation constraint was applied (proving that it was fulfilled), but concentrations were given in arbitrary units and need a post-processing step of calibration of the obtained results. In case of the presence of rotation ambiguities [29,30] in results, this post-processing step would not guarantee an optimal recovery of the quantitative information. The incorporation of the correlation constraint during the ALS optimization will decrease therefore possible rotation ambiguities, since force the profiles of the calibrated analyte to have the correct areas. In our case, this could be checked comparing the recovery of cholesterol excitation-emission profiles and of

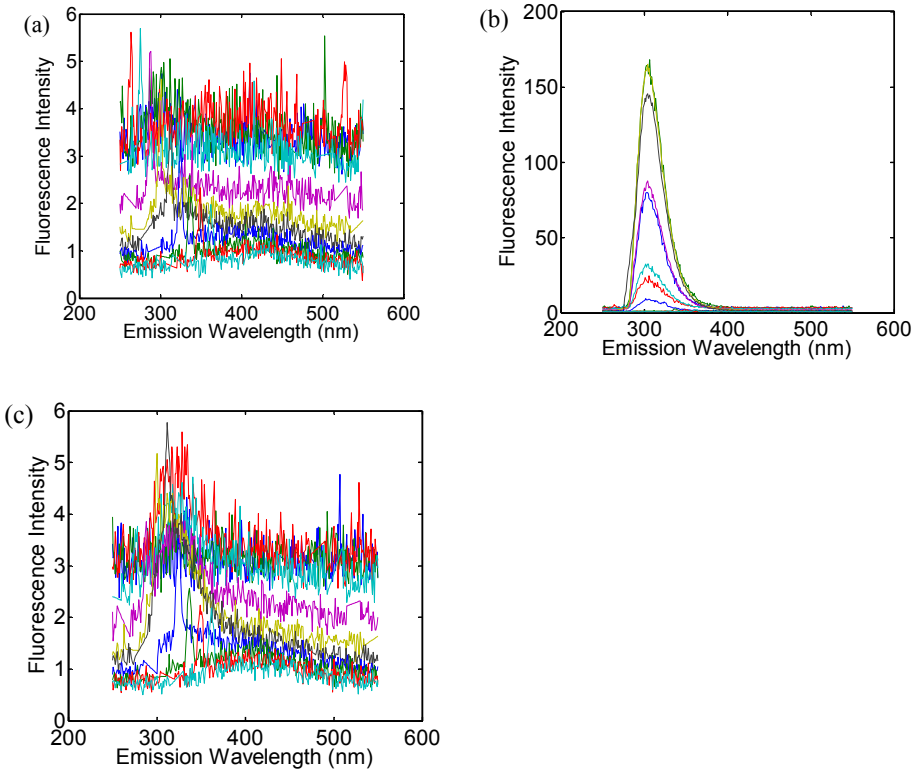


Fig. 4. Pure emission spectra from 250 to 550 nm of: (a) glucose; (b) cholesterol; (c) triglycerides constituents. Excitation wavelengths from 220 to 290 nm (10 nm step).

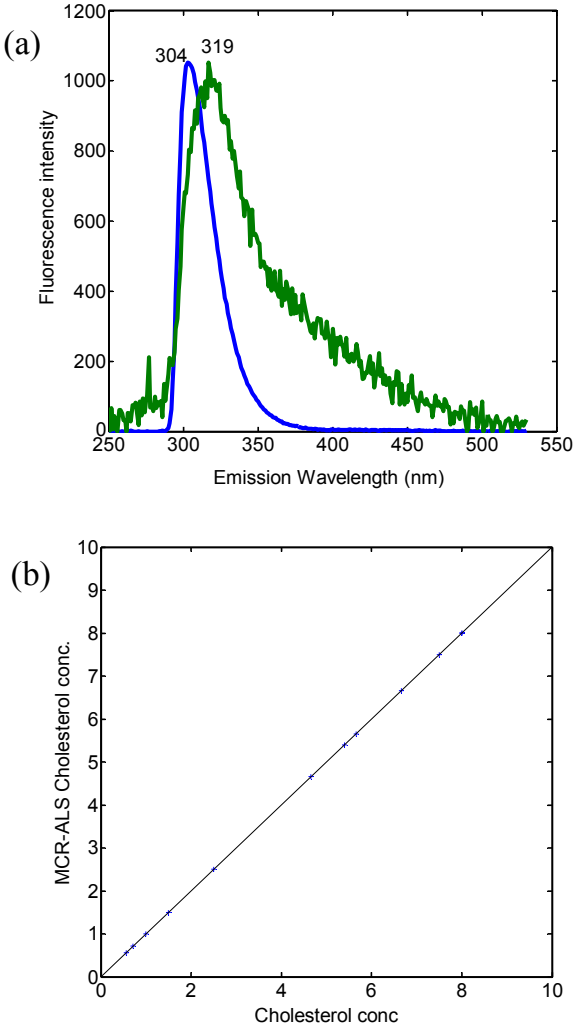


Fig. 5. MCR–ALS results of simulated data using the area correlation constraint: (a) Retrieved MCR–ALS cholesterol (blue line) and triglycerides (green line) profiles when analysing simulated samples; (b) Cholesterol concentration values predicted by MCR–ALS vs. concentration reference values for simulated data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

concentrations in both approaches. It was confirmed that results were more accurate when correlation constraint was applied with better recoveries of concentration (excitation profile areas). Further investigation of the effects on rotation ambiguity of the newly introduced area correlation constraint for second order data is being performed at present, as well as its extension to chromatographic data. In this work, the main research goal was testing its possible implementation and use in the estimation of cholesterol in synthetic samples, results which are given below in the next section.

3.2. Determination of cholesterol in laboratory synthetic samples

In this case, excitation emission fluorescent spectra of 15 synthetic mixture samples were measured in the laboratory. These 15 mixture samples had the three constituents, cholesterol, triglycerides and glucose at concentrations given for the 8 calibration samples and 7 validation samples (in Table 1). Quantification of cholesterol was performed using the areas below the excitation-emission profiles resolved by MCR-ALS and using reference values

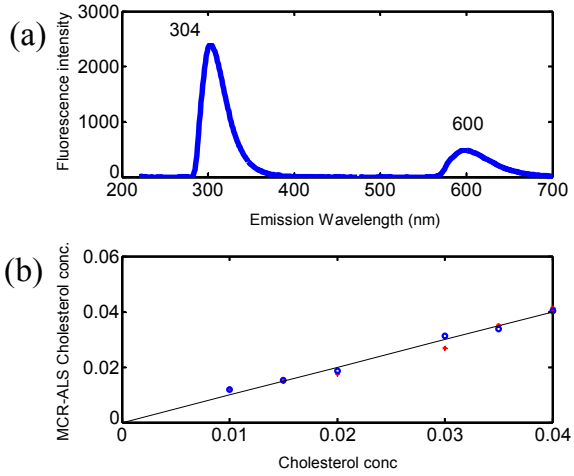


Fig. 6. (a) Retrieved MCR–ALS cholesterol profile when analysing synthetic samples. (b) Cholesterol concentration values predicted by MCR–ALS vs. concentration reference values in calibration samples (○) and validation samples (+).

of Table 1 for the 8 calibration samples (see method section). In this case the experimental measurement of the emission spectra was extended to the region 220 nm–700 nm (in the previous study of simulated samples, the emission spectra covered the region between 250 and 550 nm).

Fig. 6a, the cholesterol emission spectrum recovered by MCR-ALS had total agreement with the true one (the two peak maxima at 304 and 600 nm coincide exactly with cholesterol experimental one). In Fig. 6b, MCR-ALS cholesterol predicted values for calibration (red crosses) and prediction (blue circles) samples are plotted against the known values used for their preparation. The agreement was very good, with prediction errors below 6% for cholesterol concentration in the validation samples, a regression line of predicted versus actual values with a slope of 0.98, an offset of 0.0007, and a correlation coefficient of 0.988. Therefore, as a conclusion, the application of the area correlation constraint was working well also in this case and it allowed the correct recovery of cholesterol concentrations in the presence of triglycerides and glucose for these real synthetic samples prepared in the laboratory from their mixtures (Table 1).

4. Conclusions

In this study, the main conclusion achieved in the analysis of the two data examples is that the area correlation constraint can be used as a new variant of MCR-ALS for EEM measurements. This new extension of the MCR-ALS method allowed the recovery of the spectral information of cholesterol in the analyzed samples. This new constraint approach emerges as a new possibility to be tested in general cases of second-order multivariate calibration data (when one data matrix per sample is determined) in the presence of unknown interferents or in more difficult cases.

Acknowledgements

A. A. O. Neves would like to acknowledge the financial support from PPGQ/UFRN/CAPES for a fellowship. The authors would like to acknowledge the financial support from Brazilian National Council for Scientific and Technological Development (CNPq). K.M.G. Lima acknowledges the CNPq Grant (305962/2014-0) for financial support. This work was funded by grants from CNPq/Capes project (Grant 070/2012) and by the CHEMAGEB project (FP/2007–2013)/ERC Grant Agreement n.320737).

References

[1] K.M.G. Lima, C. Bedia, R. Tauler, A non-target chemometric strategy applied to UPLC-MS sphingolipid analysis of a cell line exposed to chlorpyrifos pesticide: a feasibility study, Microchem. J. 117 (2014) 255–261.

[2] A. De Juan, M. Maeder, M. Martínez, R. Tauler, Application of a novel resolution approach combining soft- and hard-modelling features to investigate temperature-dependent kinetic processes, Anal. Chim. Acta 442 (2001) 337–350.

[3] T. Azzouz, R. Tauler, Application of multivariate curve resolution alternating least squares (MCR-ALS) to the quantitative analysis of pharmaceutical and agricultural samples, Talanta 74 (2008) 1201–1210.

[4] J. Felten, H. Hall, J. Jaumot, R. Tauler, A. de Juan, A. Gorzsás, Vibrational spectroscopic image analysis of biological material using multivariate curve resolution–alternating least squares (MCR-ALS), Nat. Protoc. 10 (2015) 217–240.

[5] M.C.G. Antunes, J.C.G. Esteves Da Silva, Multivariate curve resolution analysis excitation-emission matrices of fluorescence of humic substances, Anal. Chim. Acta 546 (2005) 52–59.

[6] E. Peré-Trepat, R. Tauler, Analysis of environmental samples by application of multivariate curve resolution on fused high-performance liquid chromatography-diode array detection mass spectrometry data, J. Chromatogr. A 1131 (2006) 85–96.

[7] M. De Luca, G. Ragno, G. Ioele, R. Tauler, Multivariate curve resolution of incomplete fused multiset data from chromatographic and spectrophotometric analyses for drug photostability studies, Anal. Chim. Acta 837 (2014) 31–37.

[8] A.C. Olivieri, G.M. Escandar, A.M.D. La Peña, Second-order and higher-order multivariate calibration methods applied to non-multilinear data using different algorithms, Trac. - Trends Anal. Chem. 30 (2011) 607–617.

[9] M.C. Antunes, J.E.J. Simão, a C. Duarteb, R. Tauler, Multivariate curve resolution of overlapping voltammetric peaks: quantitative analysis of binary and quaternary metal mixtures, Analyst 127 (2002) 809–817.

[10] J. Jaumot, B. Igne, C.A. Anderson, J.K. Drennen, A. De Juan, Blending process modeling and control by multivariate curve resolution, Talanta 117 (2013) 492–504.

[11] H.C. Goicoechea, A.C. Olivieri, R. Tauler, Application of the correlation constrained multivariate curve resolution alternating least-squares method for analyte quantitation in the presence of unexpected interferences using first-order instrumental data, Analyst 135 (2010) 636–642.

[12] S.E. Richards, E. Becker, R. Tauler, A.D. Walmsley, A novel approach to the quantification of industrial mixtures from the Vinyl Acetate Monomer (VAM) process using Near Infrared spectroscopic data and a Quantitative Self Modeling Curve Resolution (SMCR) methodology, Chemom. Intell. Lab. Syst. 94 (2008) 9–18.

[13] L.B. Lyndgaard, F. Van den Berg, A. De Juan, Quantification of paracetamol through tablet blister packages by Raman spectroscopy and multivariate curve resolution-alternating least squares, Chemom. Intell. Lab. Syst. 125 (2013) 58–66.

[14] T. Gonçalves, M.B.P.P. Oliveira, A. Sanches-silva, H.S. Costa, Cholesterol determination in foods : comparison between high performance and ultra-high performance liquid chromatography, Food Chem. 193 (2014) 18–25.

[15] A.C.D.O. Neves, R. Fernandes de Araújo, A. Luiza Cabral de Sá Leitão Oliveira, A. Antunes de Araújo, K.M.G. de Lima, The use of EEM fluorescence data and OPLS/UPLS-DA algorithm to discriminate between normal and cancer cell lines: a feasibility study, Analyst 139 (2014) 2423–2431.

[16] R.G. Zepp, W.M. Sheldon, M.A. Moran, Dissolved organic fluorophores in southeastern US coastal waters: correction method for eliminating Rayleigh and Raman scattering peaks in excitation-emission matrices, Mar. Chem. 89 (2004) 15–36.

[17] M. Garrido, F.X. Rius, M.S. Larrechi, Multivariate curve resolution–alternating least squares (MCR-ALS) applied to spectroscopic data from monitoring chemical reactions processes, Anal. Bioanal. Chem. 390 (2008) 2059–2066.

[18] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, Chemom. Intell. Lab. Syst. 76 (2005) 101–110.

[19] M.H. Van Benthem, M.R. Keenan, Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems, J. Chemom. 18 (2004) 441–450.

[20] A. de Juan, S.C. Rutan, R. Tauler, D.L. Massart, Comparison between the direct trilinear decomposition and the multivariate curve resolution-alternating least squares methods for the resolution of three-way data sets, Chemom. Intell. Lab. Syst. 40 (1998) 19–32.

[21] R. Tauler, I. Marqués, E. Casassas, Multivariate curve resolution applied to three-way trilinear data: study of a spectrofluorimetric acid–base titration of salicylic acid at three excitation wavelengths, J. Chemom. 12 (1998) 55–75.

[22] A. Malik, R. Tauler, Extension and application of multivariate curve resolution-alternating least squares to four-way quadrilinear data-obtained in the investigation of pollution patterns on Yamuna River, India—A case study, Anal. Chim. Acta 794 (2013) 20–28.

[23] A. Malik, R. Tauler, Performance and validation of MCR-ALS with quadrilinear constraint in the analysis of noisy datasets, Chemom. Intell. Lab. Syst. 135 (2014) 223–234.

[24] R.R. de Oliveira, K.M.G. de Lima, R. Tauler, A. de Juan, Application of correlation constrained multivariate curve resolution alternating least-squares methods for determination of compounds of interest in biodiesel blends using NIR and UV–visible spectroscopic data, Talanta 125 (2014) 233–241.

[25] G. Ahmadi, R. Tauler, H. Abdollahi, Multivariate calibration of first-order data with the correlation constrained MCR-ALS method, Chemom. Intell. Lab. Syst. 142 (2015) 143–150.

[26] N. (Klaas), M. Faber, The price paid for the second-order advantage when using the generalized rank annihilation method (GRAM), J. Chemom. 15 (2001) 743–748.

[27] P.C. Damiani, A.J. Nepote, M. Bearzotti, A.C. Olivieri, A test field for the second-order advantage in bilinear least-squares and parallel factor analyses: fluorescence determination of ciprofloxacin in human urine, Anal. Chem. 76 (2004) 2798–2806.

[28] E. Peré-Trepat, S. Lacorte, R. Tauler, Alternative calibration approaches for LC-MS quantitative determination of coeluted compounds in complex environmental mixtures using multivariate curve resolution, Anal. Chim. Acta 595 (2007) 228–237.

[29] R. Tauler, Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution, J. Chemom. 15 (2001) 627–646.

[30] J. Jaumot, R. Tauler, MCR-BANDS: a user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution, Chemom. Intell. Lab. Syst. 103 (2010) 96–107.

Apêndice B

Biorganic concepts involved in the determination of glucose, cholesterol and triglycerides in plasma using the enzymatic colorimetric method

Fabrício G. Menezes — Sheeza D. Lourenço

Ana C. O. Neves — Lilian C. Silva

Djalan F. Lima — Kássio M. G. Lima

Química Nova, 2015, 38(4), 588-594.

Contribuição:

- Participei da realização dos experimentos.
- Participei da escrita do manuscrito.

Ana C. O. Neves — Prof. Kássio M. G. Lima

BIOORGANIC CONCEPTS INVOLVED IN THE DETERMINATION OF GLUCOSE, CHOLESTEROL AND TRIGLYCERIDES IN PLASMA USING THE ENZYMATIC COLORIMETRIC METHOD

Fabrício G. Menezes, Ana C. O. Neves, Djalan F. de Lima, Sheeza D. Lourenço, Lilian C. da Silva and Kássio M. G. de Lima*

Instituto de Química, Universidade Federal do Rio Grande do Norte, 59072-970 Natal – RN, Brasil

Recebido em 13/08/2014; aceito em 12/01/2015; publicado na web em 23/03/2015

Bioorganic and biological chemistry have been found to be highly motivating to undergraduate students and in this context, biochemical blood parameter analysis emerges as highly attractive content. In this proposal, several aspects related to analyses of glucose, cholesterol and triglycerides using the enzymatic colorimetric method were involved, and the findings have at least two relevant implications: i) introducing students to connections between organic chemistry and biology based on enzymatic processes, including reactivity and mechanistic aspects; ii) performing a micro scale bioassay analysis. The proposal requires two theoretical classes (2 h per class) and one practical class (4 h).

Keywords: graduate education; bioorganic/biological chemistry; blood analysis; enzymatic colorimetric method.

INTRODUCTION

Bioorganic has been pointed to as a driving force of stimulation for students of different areas to learn chemistry.[1-3] In this context, the subject of "blood" which is minimally approached in undergraduate courses, promisingly emerges to be implemented as programmatic content. Blood is an aqueous solution, basically consisting of a liquid part (plasma) and figured elements (cells and fragments). Several organic and inorganic compounds in varied structural complexity are present in the blood, such as proteins, sugars, lipids and salts, among others, which must be at adequate levels for good health. Some of the most common biological parameters found in blood are glucose, cholesterol and triglycerides, and their periodic control is highly relevant to diagnostics, prevention and treatment of several pathologies.[4,5] The gold standard for analysis of these three parameters is called the enzymatic colorimetric method (ECM), and the principle of the method as the name suggests is based on reactions that lead to chromophore species, which may be quantified by molecular spectroscopy.[5]

From our point of view, organic and biological chemistry applied to routine biochemical analysis could fill a gap in chemistry teaching in a fairly attractive way, and in this work we describe a simple and multidisciplinary didactic proposal concerned with the theoretical and practical aspects of biochemical blood parameters analysis (BBPA). The strategy is easily executable in a conventional teaching laboratory and consists of two theoretical classes (2h per class) and one practical class (4h). Firstly it is proposed an intensive discussion about many classical and relevant theoretical aspects of organic chemistry based on enzymatic reactions behind the analysis of glucose, cholesterol and triglycerides in blood through ECM. In the second part, students may experience the analysis of these three biochemical parameters in real plasma samples by collecting experimental data, using analytical techniques, performing statistical treatment and working on interpreting the results. The background required for the present activity consists in dominating the basic concepts of: i) atomic and molecular structure; ii) functional groups, their reactions and mechanisms; iii) catalysis; and iv) UV-vis spectroscopy. To evaluate if the objectives were achieved, the students were asked to answer a questionnaire of the experiment performed in the laboratory.

*e-mail: kassiolima@gmail.com

THEORY

General aspects of biochemical blood parameters analysis

In the first step of this proposal, a discussion of the reactions involved in the analysis of glucose, cholesterol and triglycerides in the blood plasma by the traditional ECM is suggested. This can be performed in two two-hour classes.

Analysis of glucose

Figure 1 presents the reactions involved in the quantification of glucose through ECM.

Through this schematic illustration, several interesting aspects of chemical stability and reactivity may be discussed. Glucose itself is an open-chain monosaccharide coexisting with α- and β-glucopyranose, and this equilibrium is called mutarotation. More specifically, in an aqueous solution glucose undergoes chemoselective intramolecular nucleophilic addition to afford two stable 6-membered cyclic hemiacetals, in which the only difference arises from the stereochemistry of the anomeric carbon. Since hemiacetal formation proceeds under thermodynamic control, it is expected that axial-OH substituent is itself due to the bonding interaction between the lone pair on oxygen (axial) and C-O antibonding molecular orbital (anomeric effect) as well as favorable dipole interactions.[6] However, the proportion of α and β glucopyranose forming in water is about 2:1, which indicates that both stereoeletronic effects are not decisive. In fact, there is a solvent effect favoring equatorial hydroxyl group which is supported by theoretical studies in gas and aqueous phases.[7]

Oxidation of several organic compounds is usually found in organic text books, but not so deeply in mechanistic terms. This fact becomes even more pronounced when biological processes and enzymatic action are taken into account. In the formation of gluconolactone from glucopyranose, there is an enzyme-induced chemoselectivity factor presented, as only one carbon (the anomeric one) is oxidized facing the other six. Mechanistically, this reaction proceeds as illustrated in Figure 2.[8-10] Initially, glucose and glucose oxidase rapidly form an enzyme-substrate complex, and the oxidation process takes place through a hydride transfer from the anomeric carbon of the carbohydrate to the FAD co-factor

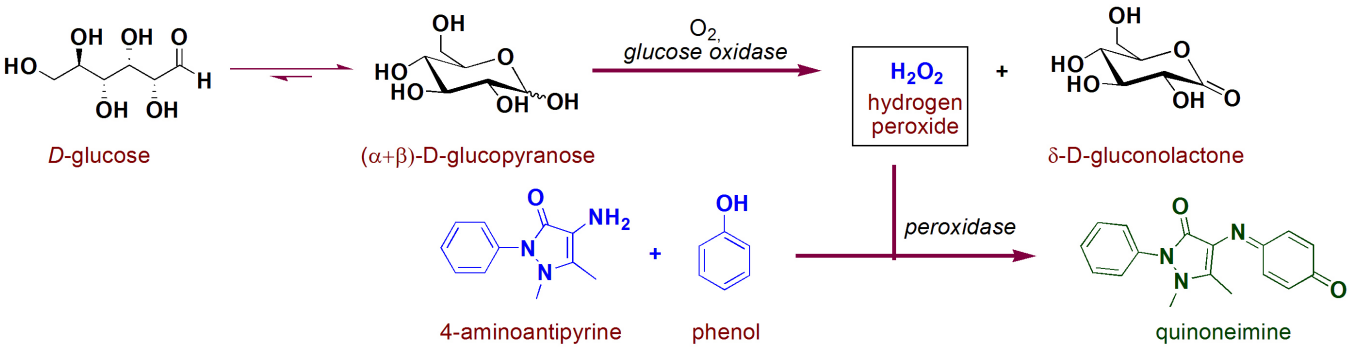


Figure 1. *Reactions involved in the quantification of glucose by the ECM*

(presented in the glucose oxidase enzyme close to the active site) to generate $FADH^-$. Sequentially, this anionic form reacts with molecular oxygen to form a radical pair (RP), which is converted into flavinhydroperoxide (FHPO) and then releases hydrogen peroxide in the biological media. Structural aspects presented in the system allow catalytic and selectivity of the process, such as hydrogen bond, non-covalent polar and hydrophobic interactions, which are illustrated in the work of Witt and colleagues.[11]

Figure 3 presents a literature-based overall view of peroxidase-catalyzed co-oxidation of phenol and 4-aminoantypirine to antipyrilquinoneimine dye.[12-15] In this chemical transformation, an equivalent amount of phenol and 4-aminoantipyrine reacts in the presence of four equivalents of hydrogen peroxide and peroxidase, resulting in dye formation and four molecules of water (Figure 3A). The term peroxidase is related to a class of enzymes with potential to induce oxidative processes by using hydrogen peroxide, and being a heme-moiety (prophirin-iron complex - Por-Fe(III)) in the active site. In the glucose quantification by ECM, two molecules of phenol are transformed in the respective phenoxy radicals, first by action of ϖ-cation radical, which was generated from hydrogen peroxide and Por-Fe(III), and then in the regeneration of the native enzyme (Figure 3B). Subsequent reaction of the antipyril radical, generated by action of the phenoxy radical, reacts with another (resonance stabilized) phenoxy radical to afford a pyrazolone derivative, which is converted into quinoneimine dye by the action of hydrogen peroxide (Figure 3C).The generated antipyrilquinoneimine dye has maximum absorbance at approximately 500 nm.

Analysis of cholesterol

Figure 4 presents a schematic illustration of the reactions involved in the total cholesterol quantification by the ECM.

Most of this lipid is found flowing in the blood as the cholesterol ester, which is enzymatically converted into free cholesterol in the first step of the protocol. Hydrolyses of esters are found in many biological processes and from an academic point of view, they are one of the most contributing reactions to mechanistic studies involving nucleophilic substitution at the carbonyl carbon, with special emphasis on Hammet relationship studies.[16-18] In the chemical route presented in Figure 4, cholesterol hydrolysis is performed by action of cholesterol esterase, an enzyme which is also capable of inversely catalyzing esterification reactions. Enzymatic hydrolyses of esters are chemical processes which require high synergy of the active site components. To exemplify this cooperation, Figure 5 presents a schematic illustration of the mechanism involved in the transformation of a general ester into carboxylic acid and alcohol by carboxylesterase action through a process governed by a catalytic triad formed by Glu, His and Ser aminoacids.[19] Initially, a hydroxyl group of Ser side chain attacks the carbonyl carbon of the substrate, through a Glu-His cooperative general base catalysis type (stage I), leading to a tetrahedral intermediate which is decomposed by cationic

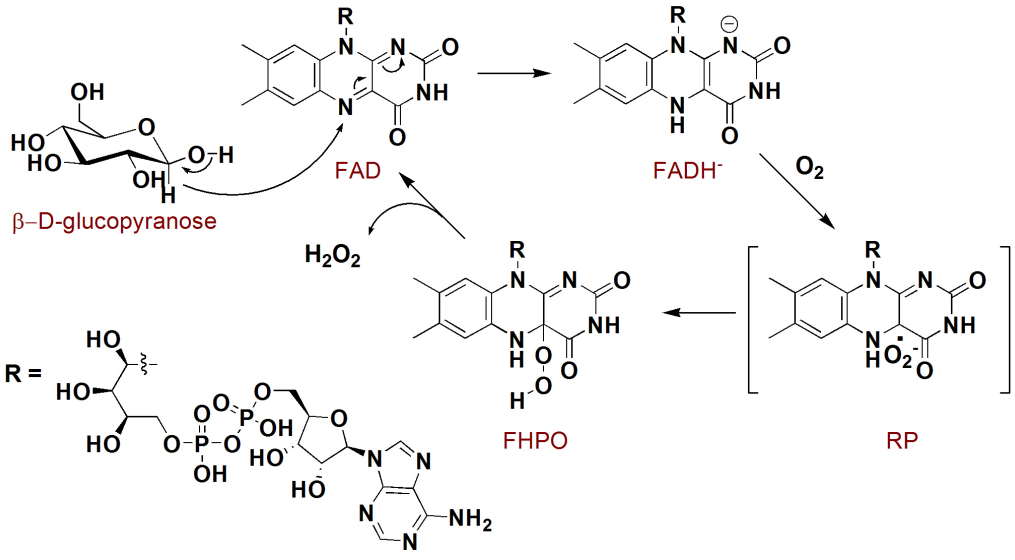


Figure 2. *Mechanism of glucose oxidation by action of glucose oxidase in biological media*

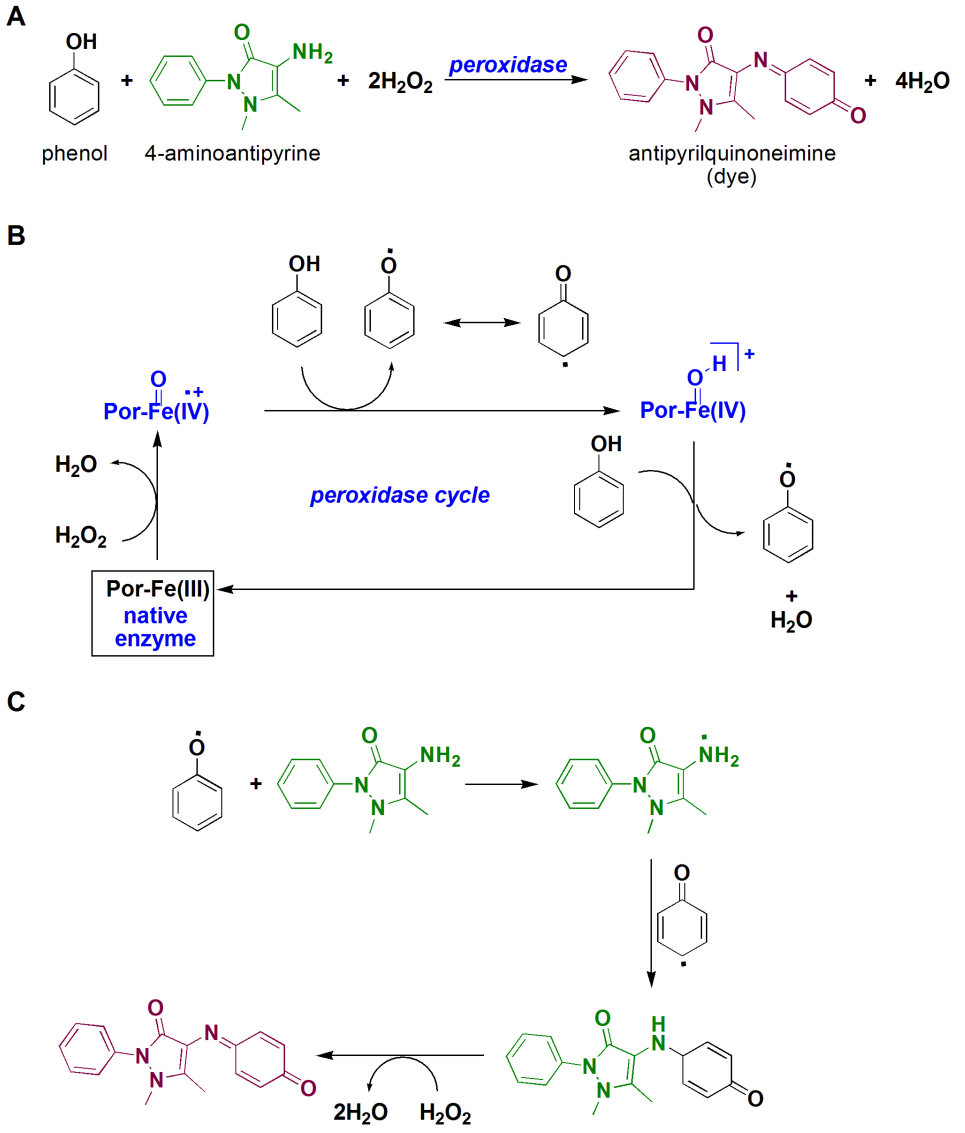


Figure 3. *A) Overall quinoneimine dye formation; B) peroxidase cycle mechanism; C) mechanism associated to quinoneimine formation*

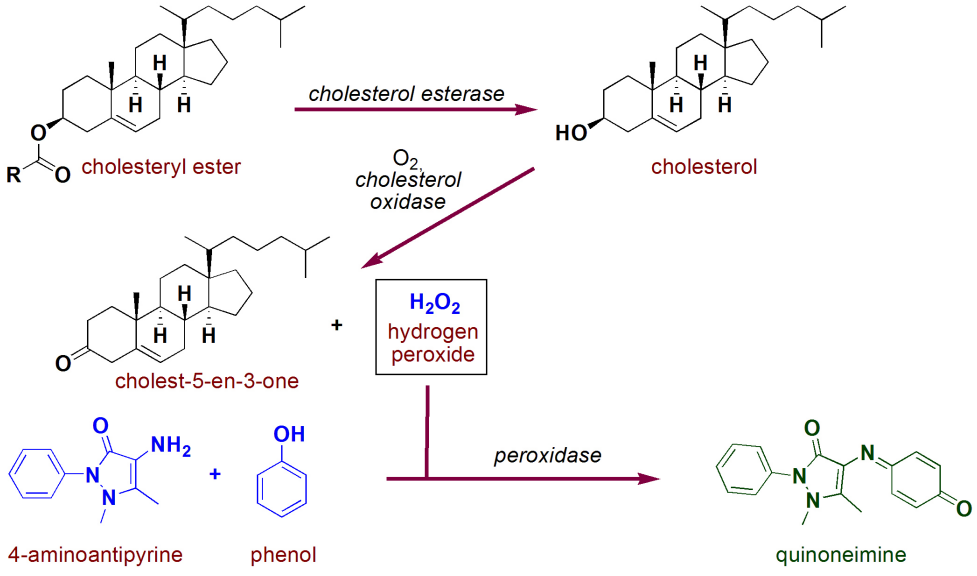


Figure 4. *Reactions involved in the quantification of total cholesterol by the ECM*

His assistance leaving group departure (stage II). Then, similarly to stages I and II, hydrolysis of O-acylated serine by water proceeds as demonstrated in stages III and IV, respectively.

After being released in its free form, cholesterol is then oxidized to cholest-5-en-3-one by action of cholesterol oxidase. This process is very similar to those presented in

Figure 2b for glucose (except for the nature/specificity of the substrate/enzyme), *i.e.* first a hydride is transferred to FAD to be able to afford $FADH_2$, which is then oxidized by molecular oxygen.[20] In this reaction, some interesting mechanistic insights are discussed in literature, such as key acid/base catalysis role and stabilization interaction of aminic hydrogen of the side chain aspargine amino acid residue (Asn-485) and FAD π-system (pyrimidine ring), which greatly contributes to this redox process.[21,22] Since hydrogen peroxide is also produced in this latter reaction, it will be acting directly in the formation of spectroscopically measurable quinoneimine chromophore (as presented in Figure 3).

Analysis of triglycerides

An illustration of the reactions involved in the protocol for quantification of triglycerides based on the ECM is presented in Figure 6. Initially, triglycerides are hydrolyzed by lipoprotein lipase, an enzyme containing a site for interaction with the lipids as well as a catalytic domain formed by Ser^{132}, Asp^{156} and His^{241}.[23,24]

In the second stage, glycerol from triglycerides hydrolyses is converted into glycerol-3-phosphate by reaction with ATP mediated by glycerol kinase and magnesium cation. Among the different complexes which could be formed in this reaction, Figure 7 shows a mechanistic illustration based on X-ray crystallography work reported by Mao and collaborators.[25] After complexation, nucleophilic attack of the 1-OH from glycerol to magnesium-activated ATP is suggested to proceed assisted by Asp^{245} through general base catalysis. In the transition state, Asp^{10} residue also complexes to Mg^{2+} ion and helps in the orientation and stabilization of the γ-phosphoryl group transfer, resulting in the reaction products. In fact, several metalloenzymes are present in biological processes, including those associated to cleavage of P-O bonds in organophosphorous compounds, playing the role of the specific metallic cation in each system.[26]

The second last step in triglycerides analysis by ECM consists of an oxidative process involving FAD, molecular oxygen and glycerol 3-phosphate oxidase. The oxidation of C2 from glycerol 3-phosphate was proposed by Yeh and colleagues to proceed through a catalyzed general base (by Arg^{317} action) hydride transfer from C2-H to FAD,[27] as briefly presented in Figure 8. In this study, the active site was modeled based on experimental and theoretical data, and a highly complex system was found with at least fourteen amino acid residues interacting in the system.[27] In this oxidation process, hydrogen peroxide is generated and then used in quinoneimine chromophore formation step, exactly as mentioned in the glucose and cholesterol analysis and presented in Figure 3.

Some considerations

In fact, there are several relevant chemical aspects involved in the protocols of quantification of glucose, cholesterol and triglycerides by the ECM, especially by the presence and effective action of enzymes. For example, mechanisms involving simple ester hydrolyses are often presented in text books proceeding with specific basic or acid catalysis, however as demonstrated herein for biochemical reactions, they are usually found proceeding with general catalysis in highly ordered processes mediated by enzymes. Nowadays, these reactions have also been adapted and applied as green processes for biodiesel production.[28]

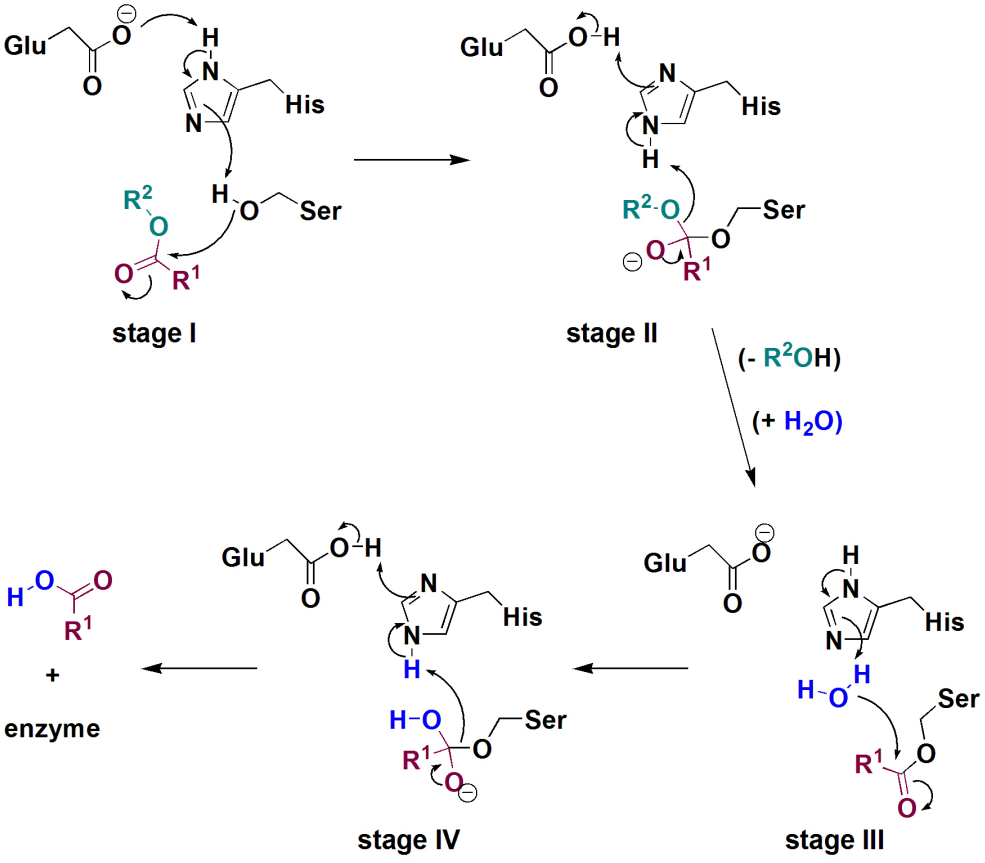


Figure 5. *Example of an enzymatic ester hydrolyses mechanism*

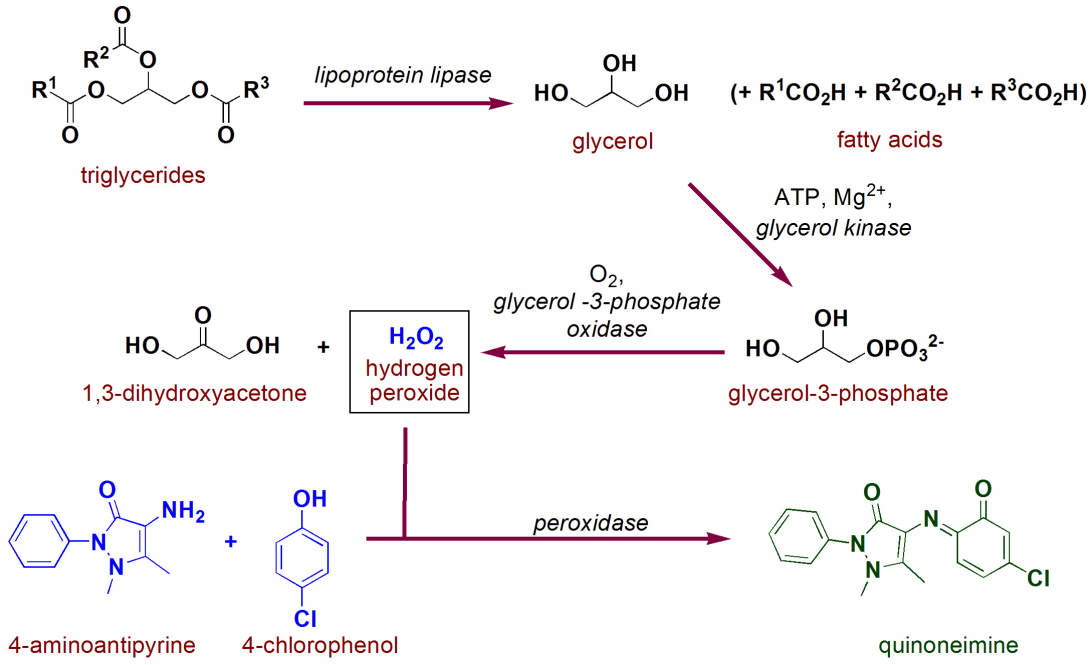


Figure 6. *Reactions involved in the quantification of triglycerides by the ECM*

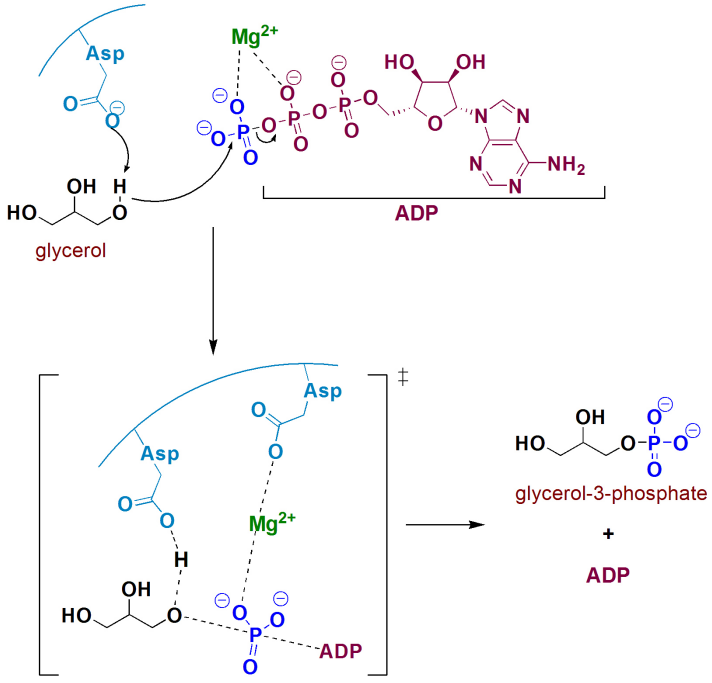


Figure 7. *Phosphorylation of glycerol by ATP, mediated by glycerol kinase and Mg^{2+}*

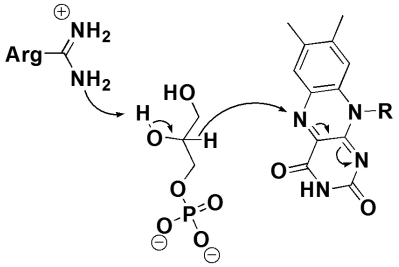


Figure 8. *Hydride transfer by glycerol 3-phosphate to FAD in the active side of glycerol 3-phosphate oxidase*

Phosphoryl transfer, such as those presented in the triglycerides protocol are quite minimally approached in conventional organic chemistry material, even though being very relevant to several areas. In fact, these reactions are remarkably reported in preparative and mechanistic studies, including those related to molecular biology and new development of therapeutic agents, pesticides and chemical weapons degradation.[29,30]

Oxidative processes and their correlated reduction ones are carried out in biological media with an impressive chemical balance, including substrates, enzymes, metallic cations and reactive species such as molecular oxygen, hydrogen peroxide and hydride transfer.

Some competitive aspects related to enzymatic reactions may be presented, and at this point enzyme promiscuity, related to capacity of enzymes catalyzing distinct chemical transformation (*i.e.* reactions involving different substrates) in different rates can also be discussed.[31] A classic example arises from ascorbic acid, which can undergo fast oxidation in the reaction media due to its higher affinity for peroxidase when compared to phenolic compounds, impairing quinoneimine chromophore formation.[32] In fact, ascorbic acid when present in high levels is treated as an interferent in biochemical analysis.[33] Dopamine is another example of chemical species that produces negative interference in hydrogen peroxide/peroxidase-based biochemical analysis (especially those based on 4-aminophenazone), due to the possibility of catechol moiety oxidation to *o*-quinone, and further reactions, including alternative dyes with distinct absorptivity (Figure 9).[34]

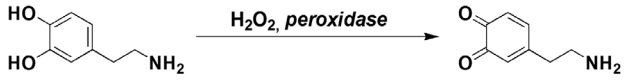


Figure 9. *Conversion of dopamine into dopamine o-quinone by action of hydrogen peroxide and peroxidase*

On the other hand, catalytic promiscuity is not always an adverse event, such as in the case of lipoprotein lipase and its hydrolylitc capacity toward triglycerides derived from distinct fatty acids along with di- and mono-acylglycerol.[35] Another example of non-prejudicial catalytic promiscuity is the action of cholesterol oxidase. This enzyme also acts in the isomerization of cholest-5-en-3-one (C-5-en-3-one) to cholest-4-en-3-one (C-4-en-3-one), as presented in Figure 10.[21,22] Mechanistically, after cholesterol is oxidized, Glu/His-assisted enolization of C-5-en-3-one leads to C-dienol which is then converted to C-4-en-3-one.

Since spectroscopically quantitative detectable quinoneimine dyes are generated in the ECM, the possibility of exploring molecular spectroscopy in class emerges, including its applications in biochemistry.[36] In fact, the color of solutions containing quinoneimine is due to the presence a ϖ-conjugated system, and this feature is made possible by the reaction of phenol and 4-aminoantipyrine which was recently pioneered and explored in the chemosensor area.[37] For the quinoneimines involved in the herein presented analysis, energy is related to an absorption in the visible green region (λ_{max} = 500-505 nm).[38]

Laboratory experiment: A suggested model for implementation

One more goal for this proposal consists in the possibility of performing a practical activity involving biochemical blood analysis. More specifically, this experiment refers to analysis of glucose, cholesterol and triglycerides in rat blood plasma through ECM and can be performed in a conventional teaching laboratory. All real samples were provided by the Animal Department of Biophysics and Pharmacology of the Federal University of Rio Grande do Norte, however with the impossibility of using animal models, synthetic or isolated samples or even human blood can be applied in this study.[39-45] Experimental details are presented in the supporting information.

The activity was applied to a total of 9 undergraduate students of bioorganic chemistry (fifth semester) from the 2013–2014 course, working in teams of three, and guided by the instructor. Each group received two samples of rat blood plasma and then performed the quantification of glucose, total cholesterol and triglycerides by ECM. Table 1 shows the results achieved by the students for this experiment. To demonstrate the influence of interfering species, a spatula tip of ascorbic acid was added to one of the samples, which measured zero mg dL^{-1} of glucose after enzymatic colorimetric method.

After the practical activity, students were gathered to discuss the results, under supervision of the instructor. Firstly, working with real rat blood plasma samples was treated as a very motivating factor, since

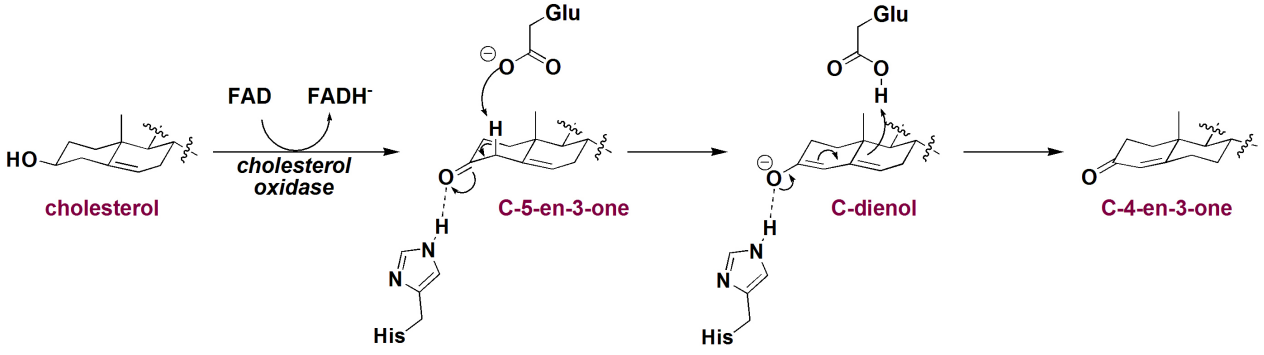


Figure 10. *Reported mechanism for isomerization by action of cholesterol oxidase*

Table 1. Results obtained by the students in the glucose, cholesterol and triglycerides analyses by enzymatic colorimetric method

Sample[a]	Glucose			Total Cholesterol			Triglycerides		
	MV[b]	RV[c]	RE[d]	MV[b]	RV[c]	RE[d]	MV[b]	RV[c]	RE[d]
G1-S1	106	107	-0.93	46	46	0.00	96	92	4.53
G1-S2	130	156	-16.67	24	36	-33.33	40	53	-24.53
G2-S3	122	111	9.91	42	44	-4.54	75	79	-5.06
G2-S4	120	125	-4.00	55	63	-12.70	61	59	3.28
G3-S5	109	105	3.67	48	51	-5.88	60	84	-28.57
G3-S6	118	93	26.88	44	40	10.00	75	81	-7.41

[a]G-A means Group-Sample; [b]measured value in mg dL^{-1}; [c]reference value in mg dL^{-1}. [d]relative error in %.

it is in fact unusually applied in undergraduate courses. The students proved to be quite enthusiastic about the experiment, especially due to the new view of organic chemistry and its reactions, and also due to the biological factor being far from the basic concepts usually presented in text books. Also, the possibility of applying the content in routine activities when associated to enzymatic processes and UV-vis spectroscopy became motivating. Further, they were able to formulate some hypotheses to explain those results that significantly differed from the values taken as reference. The first factor to be suggested was the time of reaction, followed by temperature of water bath. However, assuming that colorimetric reagents were in enough excess, these parameters could only explain lower measured values in relation to the reference ones. Possible operation errors were also taken into account, which enables an interesting discussion concerned about the reliability of clinical analysis. Another relevant aspect was related to when the blood was collected/stored and the relative stability of the analyzed biochemical parameters (degradation). In the discussion, problems due to contamination and validity of the reagents, as well as undesirable reactions were minimized.

Finally, the students were required to answer a questionnaire (see supporting information) in an attempt to be able to evaluate the whole activity on many aspects, such as novelty, new view of organic chemistry relevance and its application in biological chemistry, multidisciplinarity, stimulus to learning, among others. The results were accepted as quite motivating for continuity of the activity.

CONCLUSIONS

This manuscript describes a theoretical/experimental didactic proposal to be applied in biological chemistry disciplines for students of different courses. Classical organic chemistry reactions and the correlated enzymatic processes involved and BBPA of glucose, cholesterol and triglycerides were discussed in two theoretical classes and then during a practical activity students were allowed to directly quantify these biochemical parameters through gold standard ECM. The present proposal is coherent to the need for the development of educational materials exploring numerous chemical concepts associated to student´s everyday themes in undergraduate courses.

SUPPLEMENTARY MATERIAL

Student handout, experimental section and post-lab questionnaire are available in http://quimicanova.sbq.org.br, PDF format, with free access.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Brazilian entities CAPES, FAPERN and UFRN for financial and structural support, as well as all students who participated in the activity. Also, thanks to the editor and referees for substantial comments and suggestions for the improvement of the manuscript.

REFERENCES

1. Whitesides, G. M.; Deutch, J.; *Nature* **2011**, *469*, 21.
2. Reingold, I. D.; *J. Chem. Educ.* **2001**, *78*, 869.
3. Kirk, S. R.; Silverstein, T. P.; Willemsen, J. J.; *J. Chem. Educ.* **2006**, *83*, 1171.
4. Selby, J. V; Krumholz, H. M.; Kuntz, R. E.; Collins, F. S.; *Sci. Transl. Med.* **2013**, *5*, 1.
5. Neves, A. C. O.; de Araújo, A. A.; Silva, B. L.; Valderrama, P.; Março, P. H.; de Lima, K. M. G.; *J. Pharm. Biomed. Anal.* **2012**, *66*, 252.
6. Gonthier, J. F.; Steinmann, S. N.; Wodrich, M. D.; Corminboeuf, C.; *Chem. Soc. Rev.* **2012**, *41*, 4671.
7. Ha, S.; Gao, J.; Tidor, B.; Brady, J. W.; Karplus, M.; *J. Am. Chem. Soc.* **1991**, *113*, 1553.
8. Mattevi, A.; *Trends Biochem. Sci.* **2006**, *31*, 276.
9. Roth, J. P.; Klinman, J. P.; *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 62.
10. Wilson, R.; Turner, A. P. F.; *Biosensors and Bioeletronics* **1992**, *7*, 165.
11. Witt, S.; Wohlfahrt, G.; Schomburg, D.; Hecht, H.; Kalisz, H. M.; *Biochem. J.* **2000**, *347*, 553.
12. Berglund, G. I.; Carlsson, G. H.; Smith, A. T.; Szöke, H.; Henriksen, A.; Hadju, J.; *Nature* **2002**, *417*, 463.
13. Uliana, M. P.; Vieira, Y. W.; Donatoni, M. C.; Corrêa, A. G.; Brockson, U.; Brockson, T. J.; *J. Braz. Chem. Soc.* **2008**, *19*, 1484.
14. Nicell, J. A.; Wright, H.; *Enzyme Microb. Technol.* **1997**, *21*, 302.
15. Vojinović V.; Azevedo, A. M.; Martins, V. C. B.; Cabral, J. M. S.; Gibson, T. D.; Fonseca, L. P.; *J. Mol. Catal. B: Enzym.* **2004**, *28*, 129.
16. Smith, M. B.; March, J.; *Advanced Organic Chemistry: Reactions, Mechanisms and Structure*; 6th ed.; John Wiley and Sons Inc.: New Jersey, 2007.
17. Carey, F. A.; Sundberg, R. J.; *Sundberg, Advanced Organic Chemistry. Part A: Structure and Mechanisms*; 5th ed.; Springer: New York, 2007; p. 1199.
18. Orth, E. S.; Medeiros, M.; Souza, B. S.; Caon, N. B.; Kirby, A. J.; Nome, F.; *J. Phys. Org. Chem.* **2012**, *25*, 939.
19. Montella, I. R.; Schama, R.; Valle, D.; *Mem. Inst. Oswaldo Cruz* **2012**, *107*, 437.
20. Kumari, L.; Kanwar, S. S.; *Adv. Microbiol.* **2012**, *2*, 49.
21. Yin, Y.; Sampson, N. S.; Vrielink, A.; Lario, P. I.; *Biochemistry* **2001**, *40*, 13779.
22. Sampson, N. S.; Vrielink, A.; *Acc. Chem. Res.* **2003**, *36*, 713.
23. Emmerich, J.; Beg, O. U.; Peterson, J.; Previato, L.; Brunzell, J. D.; Brewer, H. B.; Santamarina-Fojo, S.; *J. Biol. Chem.* **1992**, *267*, 4161.
24. van Ilbeurgh, H.; Roussel, A.; Lalouel, J.; Cambillau, C.; *J. Biol. Chem.* **1994**, *269*, 4626.
25. Mao, C.; Ozer, Z.; Zhou, M.; Uckun, F. M.; *Biochem. Biophys. Res. Commun.* **1999**, *259*, 640.
26. Domingos, J. B.; Longhinotti, E.; Machado, V. G.; Nome, F.; *Quim. Nova* **2003**, *26*, 745.
27. Yeh, J. I.; Chinte, U.; Du, S.; *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 3280.
28. Straathof, A. J. J.; *Chem. Rev.* **2014**, *114*, 1871.
29. Orth, E. S.; Medeiros, M.; Bortolotto, T.; Terenzi, H.; Kirby, A. J.; Nome, F.; *J. Org. Chem.* **2011**, *76*, 10345.
30. Medeiros, M.; Orth, E. S.; Manfredi, A. M.; Pavez, P.; Micke, G. A.; Kirby, A. J.; Nome, F.; *J. Org. Chem.* **2012**, *77*, 10907.
31. Hult, K.; Berglund, P.; *Trends Biotechnol.* **2007**, *25*, 231.
32. Martinello, F.; da Silva, E. L.; *Clin. Chim. Acta* **2006**, *373*, 108.
33. Vasilarou, A. G.; Georgiou, C. A.; *J. Chem. Educ.* **2000**, *77*, 1327.
34. Karon, B. S.; Daly, T. M.; Scott, M. G.; *Clin. Chem.* **1998**, *44*, 155.
35. Kapoor, M.; Gupta, M. N.; *Process Biochem.* **2012**, *47*, 555.
36. Penzer, G. R.; *J. Chem. Educ.* **1968**, *45*, 692.
37. Kim, H. Y.; Lee, H. J.; Chenag, S.; *Talanta*, **2014**, *in press*.
38. Penzer, G. R.; *J. Chem. Educ.* **1968**, *45*, 692.
39. Erikson, J. M.; Biggs, H. G.; *J. Chem. Educ.* **1973**, *50*, 631.
40. Daines, T. L.; Morse, K. W.; *J. Chem. Educ.* **1976**, *53*, 126.
41. Taylor, R. P.; Broccoli, A. V.; Grisham, C. M.; *J. Chem. Educ.* **1978**, *55*, 63.
42. Barreto, M. C.; *J. Chem. Educ.* **2005**, *82*, 103.
43. Perles, C. E.; Volpe, P. L. O.; *J. Chem. Educ.* **2008**, *85*, 686.
44. Taber, D. F.; Li, R.; Anson, C. M.; *J. Chem. Educ.* **2011**, *88*, 1580.
45. Valdez, H. C.; Amado, R. S.; de Souza, F. C.; D'Elia, E.; *Quim. Nova* **2012**, *35*, 601.

Apêndice C

Colorimetric determination of ascorbic acid based on its interfering effect in the enzymatic analysis of glucose: an approach using smartphone image analysis

Mayra S. Coutinho

Fabrício G. Menezes

Ana C. O. Neves

Kássio M. G. Lima

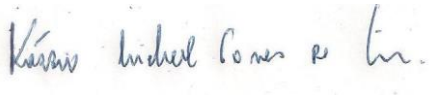
Camilo L. M. Morais

Journal of the Brazilian Chemical Society, 2017, 00(0), 1-6.

Contribuição:

- Participei do planejamento dos experimentos
- Participei da escrita do manuscrito.

Ana C. O. Neves



Prof. Kássio M. G. Lima

http://dx.doi.org/10.21577/0103-5053.20170086

J. Braz. Chem. Soc., Vol. 00, No. 00, 1-6, 2017.
Printed in Brazil - ©2017 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

Short Report

Colorimetric Determination of Ascorbic Acid Based on Its Interfering Effect in the Enzymatic Analysis of Glucose: An Approach Using Smartphone Image Analysis

Mayra S. Coutinho, Camilo L. M. Morais, Ana C. O. Neves, Fabrício G. Menezes and Kássio M. G. Lima*

Química Biológica e Quimiometria, Instituto de Química, Universidade Federal do Rio Grande do Norte, 59072-970 Natal-RN, Brazil

The use of digital image analysis as an analytical tool is a reality nowadays, and the use of smartphones stands out due to its high accessibility and practicality. Ascorbic acid (AA) is a natural and essential vitamin available as a supplement as a result of its use in preventing and treating several pathologies. This paper reports a simple, fast and low cost method using smartphone image analysis for quantification of AA based on its interfering effect in the enzymatic colorimetric detection of glucose. Commercial vitamin C tablets were used as prediction set for AA quantification, showing very satisfactory results (relative errors < 4%), where no statistical difference at a 95% confidence level was observed between the AA content estimated by the imaging method and the labeled reference values. As advantages, this method does not use expensive reagents neither laborious procedures to carry out the analysis.

Keywords: ascorbic acid, glucose analysis interfering, enzymatic colorimetric method, image analysis

Introduction

The development of new analytical methodologies based on digital imaging has been attracting considerable attention in the last decade due to its low cost, simplicity, non-destructive aspect and speed of analysis. Common digital image capturing devices such as cameras, webcams, scanners, and smartphones are used to register images from colorimetric reactions, in which the intensity of the developed color is directly proportional to the analyte concentration. To demonstrate these features, several papers involving colorimetric detection have been published such as for the quantification of synthetic dyes;[1] monitoring of organic reactions;[1,2] evaluation of food quality;[3] characterization of drug authenticity;[4] and determination of biochemical parameters.[5-9] Among several devices applied in image-based analytical methodologies, the use of smartphones has emerged as a notable tool due to the massive amount of users all over the world, easy image acquisition and high accessibility for data transmission.[4-8]

Ascorbic acid (AA) is a natural water-soluble vitamin widespread in nature that plays a crucial role in biological processes and metabolism,[10] AA is a common constituent of the human diet and commercialized as a supplement (vitamin C) in high levels, since its ingestion is strongly suggested for prevention and treatment of several pathologies.

Several approaches have been reported in an attempt to determine AA levels in different matrices, such as the high-performance liquid chromatography (HPLC) and ultra performance liquid chromatography (UPLC) methods, electrochemical measurements, as well as molecular spectroscopy such as UV-Vis and fluorescence.[11-18] These methods can be very attractive in terms of limit of detection (LOD) and quantification (LOQ), which can include values around 0.3 and 1.0 μg dL^{-1}, respectively, making analysis of samples containing very low concentrations levels possible.[11-15] However, they present some considerable drawbacks, such as the dependence on sophisticated instruments; time-consuming analysis; need for expensive reagents in high volumes; and have laborious procedures. In addition to these reported methods for quantification of AA, there are some commercially available assays, usually based in enzymatic reagents presented in very specific medium, involving colorimetric/fluorometric spectroscopies, and although they allow quantification in low levels, they are very expensive, which may limit their use for some routine applications.

*e-mail: kassiolima@gmail.com

The advantages of the recent methods based in image analysis consist in overcoming at least some of these drawbacks. In the literature, there are only a few works which have reported the quantification of AA in solution by image analysis.[17-20] In fact, these methods are very interesting since they involve modern aspects of chemistry, such as nanoparticle formations and click reactions. On the other hand, some drawbacks remain such as the need for expensive reagents, laborious procedures and being time-consuming. Also, the limits of quantification are not as low as conventional analysis; however, this does not exclude the relevance of the method since some fruits and commercial supplements are composed of high levels of AA.

Therefore, in this paper we present a new method based on digital images acquired with a smartphone for determining AA in aqueous solution. For this purpose, we take advantage of the effect that AA causes in the analysis of glucose through the gold standard spectroscopic-based enzymatic colorimetric method (ECM).[21] More specifically, AA acts on suppressing the formation of quinoneimine dye (essential for UV-Vis analysis), which induces lower color intensity of a glucose solution as the concentration of AA increases. The results suggest that it is possible to apply the method herein described in the preliminary analysis of commercial vitamin C. Although the drawback originated by the presence of AA in samples subjected to glucose analysis has been known for decades, to the best of our knowledge, herein we report the first time that this interference effect was measured and associated to quantitative purposes.

Experimental

Reagents and instrumentation

L-Ascorbic acid and D-glucose were both acquired from Sigma-Aldrich and used without further purification. Glucose mono reagent (color reagent) kit was acquired from Bioclin (Bioclin Quibasa, Brazil) and stored as instructed by the manufacturer. This color reagent is composed of a buffer (pH 7.0); phenol (10 mmol L^{-1}); 4-aminoantipyrine (0.3 mmol L^{-1}); sodium azide (7.7 mmol L^{-1}); glucose oxidase (> 10.000 UI); and peroxidase (> 700 UI). All experiments were carried out in deionized water. Commercial vitamin C tablets were purchased in drugstores. A Hemoquímica model HM0064D water-bath was used in the experiments. UV-Vis analyses were carried out in an Evolution 60S model Thermo Scientific spectrophotometer using 1 cm length quartz cuvettes. All images were acquired by using a Sony Xperia C2304 smartphone with camera resolution of 8 megapixels.

Samples

The calibration set was built from five calibration samples in triplicate containing 10.0 mg L^{-1} of glucose and different amounts of AA (6.0, 7.0, 8.0, 9.0 and 10.0 mg L^{-1}). All solutions were prepared in deionized water, which is very important since AA undergoes oxidation in the presence of metal, such as copper. On the other hand, in the absence of any catalyst (such as copper or hydrogen peroxide), the presence of dissolved oxygen is usually insignificant in the oxidation of AA.[22] All solutions were prepared just before use. Analysis of vitamin C in commercial tablets (brands A-C) was performed by dissolving their content in water until concentration of 8.0 mg L^{-1}.

UV-Vis spectroscopy

UV-Vis analysis was performed as a reference method for the quantification of AA through image analysis. UV-Vis spectra were acquired for the calibration samples by adding 20 μL of the solutions containing D-glucose (10.0 mg L^{-1}) and L-ascorbic acid (varying from 6.0 to 10.0 mg L^{-1}) to 2 mL of the color reagent after heating at 37 °C for 10 minutes. The blank was the color reagent for all cases. Changes in the absorbance were analyzed at 508 nm of the UV-Vis spectra.

Image acquisition

The images from the same samples measured by UV-Vis spectroscopy were also captured by using a smartphone. The same cuvettes analyzed by the spectrometer were placed onto a white surface with a white background and the pictures were taken with the smartphone held at a statistic position of 19.5 cm from the bucket, avoiding blurriness. In addition, blank images (images of cuvettes filled with deionized water) were also recorded in order to avoid ambient lighting interference and further used to obtain the RGB (red-green-blue) absorbance. The spectra and images were acquired immediately after the samples were prepared. This experimental setup is illustrated in Figure 1a.

After acquiring the sample images in .JPEG format with 96 × 96 dpi resolution, they were loaded into a personal computer to extract the region of interest (ROI) composed of a square of 10 × 10 pixels (Figure 1b). This procedure was done using GIMP 2.0 software.[23]

Digital image analysis

The ROI of the images (Figure 1b) were processed

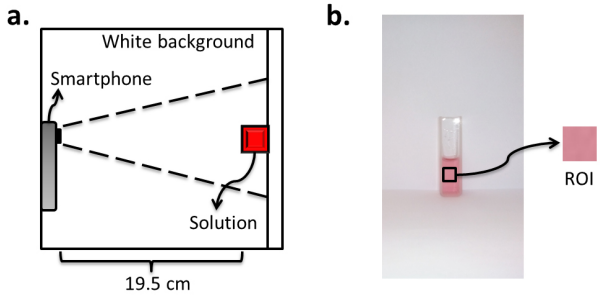


Figure 1. (a) Illustration of experimental setup for image acquisition; (b) example of region of interest (ROI) cropped from a sample image of AA.

using MATLAB R2014a environment (Math Works, USA) through the calculation of the RGB-resolved absorbance as the analytical signal univariately related to the concentration of AA, according to equation 1:[24]

$$A_k = -\log \frac{I_k}{I_{0,k}} \quad (1)$$

where, A_k is the RGB-resolved absorbance for a given color channel k (red, green or blue); I_k is the intensity for a given color channel k; and $I_{0,k}$ is the intensity of blank for a given color channel k. The intensity value is calculated as the mean for its related color channel.

Statistical evaluation

In this study, the performance of the calibration model was evaluated by comparing the AA predicted values by the imaging method with the AA measured values. This was made by calculating the statistic RMSEC (root mean square error of calibration), relative standard-deviation (RSD), bias (average difference between predicted and measured values), relative error, accuracy, R (correlation coefficient), R^2 (coefficient of determination), limit of detection (LOD) and limit of quantification (LOQ). Additionally, the Student's *t*-test was performed to find any significant difference in the results of RGB prediction and AA measurements. The LOD and LOQ were calculated as follows:

$$LOD = \frac{3.3 S_{y/x}}{A} \sqrt{1 + h_0 + \frac{1}{I}} \quad (2)$$

$$LOQ = \frac{10 S_{y/x}}{A} \sqrt{1 + h_0 + \frac{1}{I}} \quad (3)$$

where, $S_{y/x}$ is the residual standard deviation; A is the slope of the univariate calibration curve; h_0 is the leverage for the blank sample; and I is the number of calibration samples. The leverage for the blank sample h_0 can be estimated as:[25]

$$h_0 = \frac{\bar{c}_{cal}^2}{\sum_{i=1}^{I} (c_i - \bar{c}_{cal})^2} \quad (4)$$

where, $\bar{c}_{cal}$ is the mean calibration concentration and c_i is each of the calibration concentration values.

Results and Discussion

In the ECM method for quantification of glucose, the referred carbohydrate (as (α+β)-D-glucopyranose) is oxidized to δ-gluconolactone by molecular oxygen in the presence of enzyme glucose oxidase. Sequentially, hydrogen peroxide, which is also generated in this first stage, induces oxidative condensation of 4-aminoantipyrine and phenol by the action of enzyme peroxidase to afford a quinoneimine dye (Figure 2a).[21] Since both glucose oxidation and synthetic dye formation in the ECM are based on oxidative processes, it is possible to assume that AA can compete with the substrates present in the reaction media and be previously oxidized to the dye formation to afford dehydroascorbic acid (DHA, Figure 2a, inset). Colors of the solutions containing available enzymatic reagent with a specified amount of glucose (10.0 mg L^{-1}) will become less intense as the concentration of AA increases in the solution, and this effect may be followed by both UV-Vis spectroscopy and image analysis (Figure 2b).

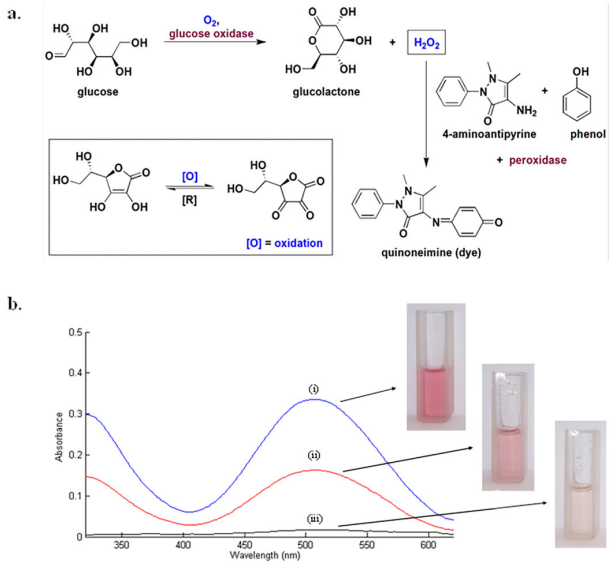


Figure 2. (a) Reactions involved in the formation of quinoneimine from initial oxidation of glucose and subsequent condensation of 4-aminoantipyrine and phenol. Inset: oxidation of AA into dehydroascorbic acid in the reaction media; (b) changes in UV-Vis spectra (320-650 nm) and color intensity of the solutions of colorimetric reagent containing 20 µL of glucose solution (10.0 mg L^{-1}) in 2 mL of color reagent upon addition of AA: (*i*) 0 mg L^{-1}; (*ii*) 5.0 mg L^{-1}; (*iii*) 10.0 mg L^{-1}.

Calibration set using standard ascorbic acid

Since the gold standard ECM for determination of glucose is based in UV-Vis spectroscopy measurements, we started our studies by evaluating the changes in the UV-Vis spectra of the enzymatic color reagent after adding different amounts of AA (6.0-10.0 mg L^{-1}) in solutions containing 10.0 mg L^{-1} of glucose. A calibration curve ($y = -0.0310x + 0.336$) was found with an R^2 equal to 0.998 (Figure 3a) for UV-Vis data, presenting an RMSEC of 0.06 mg L^{-1} and a relative error for calibration of 0.76%. After UV-Vis measurements, the images captured with the smartphone were analyzed and the same trend was observed using this method, where the G-B (green-blue) resolved absorbance values were used to build the calibration curve ($y = -0.00526x + 0.0560$) with an R^2 equal to 0.972 (Figure 3b). The G-B resolved absorbance represents the region between 546.1 and 435.8 nm, covering the region of the absorption band for AA chromogen (λ_{max} = 508 nm). The RMSEC found using the image method was equal to 0.24 mg L^{-1} and a relative error for calibration of 2.91% was found.

Although the UV-Vis method presented better sensitivity and smaller error, the relative error found in the calibration set using the image method was significantly small (< 3%), indicating good accuracy using the synthetic samples. The linearity of the curves were also evaluated according to the statistical *F* test.[25,26] For both techniques, the calculated *F* values (0.050 and 0.164 for UV-Vis and image method, respectively) were smaller than the critical *F* value (6.39, $\alpha = 0.05$) for a confidence level of 95%, proving the linearity of the calibration curves. In addition, the low values of bias found for both methods indicate absence of systematic error. The figures of merit for the methods using both UV-Vis and image-based approaches are shown in Table 1.

Table 1. Figures of merit for the calibration set based on UV-Vis spectroscopy and image method

Figure of merit	UV-Vis method	Image method
RMSEC / (mg L^{-1})	0.06	0.24
Bias / (mg L^{-1})	0.002	–0.006
Relative error / %	0.76	2.91
R	0.999	0.986
R^2	0.998	0.972
Sensitivity	0.0310	0.00526

RMSEC: root mean square error of calibration; R: correlation coefficient; R^2: coefficient of determination.

Image analysis applied to commercial vitamin C tablets

After confirming a linear trend found as a consequence of AA in the formation of quinoneimine dye, we then aimed to validate this methodology. Firstly, an external validation test was performed using three synthetic samples with 7.5 mg L^{-1} of ascorbic acid. Validation response of 7.7 ± 0.1 mg L^{-1} (relative error of 2.67%) and 7.5 ± 0.3 mg L^{-1} (relative error of 0.04%) was found for UV-Vis and image method, respectively. These results confirmed the ability of the proposed method using images to predict test samples with low error. Thereafter, commercial effervescent vitamin C samples (brands A-C) were analyzed in order to evaluate the proposed method in real environment. The nominal and predicted AA concentrations for these commercial samples are shown in Table 2.

As shown in Table 2, the nominal and predicted concentrations of AA in commercial vitamin C samples were very similar using the imaging method, with relative errors smaller than 4%. These relative errors were a little higher than using UV-Vis spectroscopy, which could be caused due to ambient lighting interference during image acquisition although most of this effect was corrected

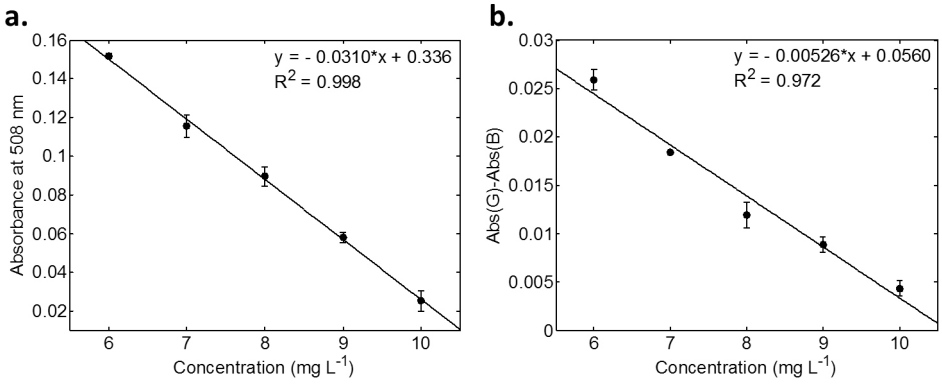


Figure 3. Effect of AA (6.0-10.0 mg L^{-1}) in the formation of quinoneimine chromophore resulting from the reaction of 20 µL of glucose solution (10.0 mg L^{-1}) in 2 mL of color reagent: (a) calibration curve obtained from UV-Vis data; (b) calibration curve using image data, where Abs(G) is the resolved absorbance for G channel and Abs(B) is the resolved absorbance for B channel. Vertical bars represent the standard deviation obtained in a triplicate of each point.

Table 2. Nominal and predicted concentrations for UV-Vis and imaging method. Average relative error in percentage and *t*-value calculated by Student's *t*-test are shown

Method	Brand		
	A	B	C
UV-Vis			
Nominal concentration / (mg L^{-1})	8.0	8.0	8.0
Predicted concentration / (mg L^{-1})	8.0 ± 0.3	8.0 ± 0.2	7.9 ± 0.2
Relative error / %	0.01	0.04	1.25
t-Value	0.01	0.03	–0.87
Imaging			
Nominal concentration / (mg L^{-1})	8.0	8.0	8.0
Predicted concentration / (mg L^{-1})	8.2 ± 0.5	8.1 ± 0.6	8.3 ± 0.6
Relative error / %	2.50	1.25	3.75
t-Value	0.69	0.29	0.87

during absorbance calculation. A Student's *t*-test was applied to the samples' triplicate and the calculated *t*-value is also shown in Table 2. All calculated *t*-values for imaging method were smaller than the critical *t*-value (4.3, *p*-value = 0.975), therefore showing no statistical difference at a 95% confidence level between the predicted concentrations of AA calculated by the imaging method and the nominal concentrations for the three brands analyzed.

The precision of imaging method was evaluated by relative standard deviation (RSD). For brands A, B and C the RSD were equal to 6.1, 7.4 and 7.2%, respectively. Using UV-Vis spectroscopy, the RSD were smaller due to its better sensitivity: 3.7% (brand A), 2.5% (brand B), and 2.5% (brand C). Although the RSD is higher in imaging method, the values obtained (< 10%) seems to be adequate to this type of analysis, since the speed and portability of this technique competes with its sensitivity. The accuracy of imaging method was evaluated by its recovery. Recovery ranging from 96.2-108.7% was found for brand A; 93.7-108.7% for brand B; and 96.2-111.2% for brand C. Using UV-Vis spectroscopy, the recoveries range were shorter: 96.2-103.7% (brand A); 97.5-102.5% (brand B); 96.2-101.2% (brand C). These recovery values are close to 100%, proving the capability of both curves to predicted concentrations of AA with low error.

The limit of detection (LOD) and quantification (LOQ) were estimated for the image method using the modern IUPAC recommendation,[25] and they were equal to 0.055 and 0.166 mg L^{-1}, respectively. Using UV-Vis spectroscopy, the LOD and LOQ values were equal to 0.045 and 0.138 mg L^{-1}, respectively. Although the LOD and LOQ values calculated for the image method are slightly elevated, they do not affect the analysis of commercial vitamin C samples, since its content in solution is much higher. The AA content calculated by the image method in commercial vitamin C samples is summarized in Table 3.

In literature, there are only a few works reporting the quantification of AA from digital image analysis, however, these methodologies are mainly based on the use of noble metal-based nanoparticles and click chemistry reactions, which require specific and expensive reagents, in addition to relatively laborious and time-consuming procedures.[17-20] In this context, we can attest that the results herein are very interesting due to three main facts: (*i*) we have worked with both standard and real commercial samples of AA, with the latter being a more complex chemical composition; (*ii*) we made use of commercially available and very low cost reagent ($15.00USD/250 mL); (*iii*) our analyses were performed using a very portable image acquisition device (smartphone); and (*iv*) the method requires little handling and only about fifteen minutes to be completed.

Table 3. Reference and predicted mass content of AA using UV-Vis spectroscopy and image method for samples of brand A, B, and C

Brand	Reference labeled content of ascorbic acid	Predicted content by UV-Vis	Relative error by UV-Vis / %	Predicted content by image	Relative error by image / %
A	1.0 g	1.0 ± 0.04 g	0.01	1.0 ± 0.1 g	2.50
B	500 mg	500 ± 12 mg	0.04	506 ± 37 mg	1.20
C	1.0 g	1.0 ± 0.03 g	1.25	1.0 ± 0.1 g	3.75

Conclusions

We have reported herein a new methodology for determining AA in solution based on the analysis of digital images obtained with a smartphone. The principle of this proposed strategy is based on a lower quantity of chromophore quinoneimine (resulting from the ECM approach for determining glucose) being formed as the AA concentration in media is increased. One of the main goals of this work was to explore a well established adverse effect for the analysis of glucose and successfully apply to the quantitative analysis of AA in aqueous media, especially from samples of vitamin C supplement in tablets. An image analysis approach was investigated in order to predict concentrations of AA in aqueous solution and the results found were very satisfactory for both calibration (standard AA) and prediction (commercial vitamin C supplements) sets. Therefore, this method can be highlighted due to its short analysis time, low cost, high availability of enzymatic colorimetric color reagent, and the simplicity and portability of the instrumentation being utilized (a smartphone). Furthermore, this new approach may contribute to the development and consolidation of a new field for chemical analysis, which includes quantitative evaluation of a broad spectrum of chemical analytes based on the use of mobile phones and image analysis.

Acknowledgments

The authors would like to acknowledge the structural and financial support from UFRN, CAPES and CNPq. M. S. C. would like to thank UFRN for the scholarship. C. L. M. M. and A. C. O. N. thank the Post-Graduate Program in Chemistry (PPGQ) of UFRN and CAPES. K. M. G. L. thanks the CNPq (305962/2014-0).

References

1. da Silva, L. C.; de Lima, D. F.; Silva, J. A.; de Morais, C. L. M.; Albuquerque, B. L.; Bortoluzzi, A. J.; Domingos, J. B.; Araújo, R. M.; Menezes, F. G.; de Lima, K. M. G.; *J. Braz. Chem. Soc.* **2016**, *26*, 1067.
2. Hemmateenejad, B.; Akhond, M.; Mohammadpour, Z.; Mobaraki, N.; *Anal. Methods* **2012**, *4*, 933.
3. Russ, J. C.; *J. Food Sci.* **2015**, *80*, 1974.
4. Yu, H.; Le, H. M.; Kaale, E.; Long, K. D.; Layloff, T.; Lumetta, S. S.; Cunningham, B. T.; *J. Pharm. Biomed. Anal.* **2016**, *125*, 85.
5. Martinez, A. W.; Phillips, S. T.; Carrilho, E.; Thomas, S. W.; Sindi, H.; Whitesides, G. M.; *Anal. Chem.* **2008**, *80*, 3699.
6. Zhu, H.; Sencan, I.; Wong, J.; Dimitrov, S.; Tseng, D.; Nagashima, K.; Ozcan, A.; *Lab Chip* **2013**, *13*, 1282.
7. Morais, C. L. M.; Neves, A. C. O.; Menezes, F. G.; Lima, K. M. G.; *Anal. Methods* **2016**, *8*, 6458.
8. Morais, C. L. M.; Lima, K. M. G.; *Anal. Methods* **2015**, *7*, 6904.
9. Morais, C. L. M.; Lima, K. M. G.; *Talanta* **2014**, *126*, 145.
10. Du, J.; Cullen, J. J.; Buettner, G. R.; *Biochim. Biophys. Acta* **2012**, *1826*, 443.
11. Klimczak, I.; Gliszczyńska-Świgło, A.; *Food Chem.* **2015**, *175*, 100.
12. Kemmegne-Mbouguen, J. C.; Angnes, L.; *Sens. Actuators, B* **2015**, *212*, 464.
13. Santos, D. A.; Lima, K. P.; Março, P. H.; Valderrama, P.; *J. Braz. Chem. Soc.* **2016**, *27*, 1912.
14. VanderJagt, D. J.; Garry, P. J.; Hunt, W. C.; *Clin. Chem.* **1986**, *32*, 1004.
15. Vermeir, S.; Hertog, M. L. A. T. M.; Schenk, A.; Beullens, K.; Nicolaï, B. M.; Lammertyn, J.; *Anal. Chim. Acta* **2008**, *618*, 94.
16. Wu, X.; Diao, Y.; Sun, C.; Yang, J.; Wang, Y.; Sun, S.; *Talanta* **2003**, *59*, 95.
17. Zhang, Y.; Li, B.; Xu, C.; *Analyst* **2010**, *135*, 1579.
18. Hemmateenejad, B.; Shakerizadeh-Shirazi, F.; Heidari, S.; Shahrivar-kevishahi, A.; *Anal. Methods* **2015**, *7*, 6318.
19. Gomes, M. S.; Trevizan, L. C.; Nóbrega, J. A.; Kamogawa, M. Y.; *Quim. Nova* **2008**, *31*, 1577.
20. Ferreira, D. C. M.; Giordano, G. F.; Soares, C. C. D. S. P.; de Oliveira, J. F. A.; Mendes, R. K.; Piazzetta, M. H.; Gobbi, A. L.; Cardoso, M. B.; *Talanta* **2015**, *141*, 188.
21. Menezes, F. G.; Neves, A. C. O.; de Lima, D. F.; Lourenço, S. D.; da Silva, L. C.; de Lima, K. M. G.; *Quim. Nova* **2015**, *38*, 588.
22. Jansson, P. J.; Jung, H. Y. E. R.; Lindqvist, C.; Nordstro, T.; *Free Radical Res.* **2004**, *38*, 855.
23. https://www.gimp.org/, accessed in May 2017.
24. Christodouleas, D. C.; Nemiroski, A.; Kumar, A. A.; Whitesides, G. M.; *Anal. Chem.* **2015**, *87*, 9170.
25. Olivieri, A. C.; *Anal. Chim. Acta* **2015**, *868*, 10.
26. Danzer, K.; Currie, L. A.; *Pure Appl. Chem.* **1998**, *70*, 993.

Submitted: February 13, 2017
Published online: May 17, 2017

Apêndice D

Determination of serum protein content using cell phone image analysis

Camilo L. M. Morais

Fabrício G. Menezes

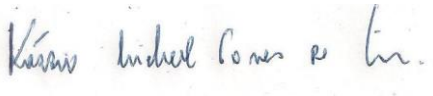
Ana C. O. Neves

Kássio M. G. Lima

Analytical Methods, 2016, 8, 6458-6462.

Contribuição:

- Participei da realização dos experimentos;
- Participei da escrita do manuscrito.



Ana C. O. Neves

Prof. Kássio M. G. Lima

Analytical Methods

TECHNICAL NOTE

View Article Online
View Journal | View Issue

Cite this: *Anal. Methods*, 2016, **8**, 6458

Determination of serum protein content using cell phone image analysis

Camilo L. M. Morais, Ana C. O. Neves, Fabrício G. Menezes and Kássio M. G. Lima*

This paper presents a simple, fast and inexpensive way to measure serum protein content (albumin and total proteins) by integration of color images acquired with a cell phone camera and multiple linear regression (MLR). MLR models were built using the RGB-HSV image systems from the color changes induced by bromocresol green and Biuret methods on solutions containing albumin (0.24 to 5.03 g dL^{-1}) and total proteins (0.25 to 8.20 g dL^{-1}), respectively, acquired from a cell phone camera. These methods were adapted to a microscale analysis by using an ELISA 96-microwell plate of 250 μL as a reaction container, and subsequently were statistically validated. The RMSEP values obtained for albumin and total protein models were respectively equal to 0.28 and 0.12 g dL^{-1}, and their precision was respectively equal to 6.16 and 4.85%. No statistical difference was observed at a confidence level of 95% between the protein concentration calculated by the proposed method and those found by the reference method, proving the reliability of the imaging method as an alternative approach for these assays.

Received 23rd June 2016
Accepted 19th July 2016

DOI: 10.1039/c6ay01783e

www.rsc.org/methods

Introduction

Proteins are macromolecules (essentially polypeptides) found in all living organisms with key roles in their structures and metabolism.[1] Quantification of serum proteins performed in clinical analysis is based on two main parameters: albumin and total proteins and is crucial to prevention, diagnosis and monitoring of several pathologies.[2] For example, the concentration of total protein serum can be an indicator of chronic infection and cryoglobulinemia;[3] while albumin levels are associated with the nutritional state, inflammatory state and mortality especially in the elderly.[4,5]

In laboratories of clinical analysis, quantification of total proteins and albumin is based on colorimetric reactions and UV-Vis spectroscopy. The reference method for the determination of the total protein content is the Biuret reaction,[6] in which Cu(II) ions coordinate to the nitrogen of the peptide bonds of proteins in strongly basic media, leading to a violet-colored product with an absorption maximum at 540 nm. Although this method is actually applied in clinical analysis, it has some advantages such as the relative specificity of the Biuret reagent for proteins and the reproducibility of absorbance values.[7] On the other hand, for albumin determination, the bromocresol green (BCG) indicator[8] is usually applied since it leads to colored anionic species in the presence of the referred protein possible to be measured by UV-Vis spectroscopy through a band that emerges at 630 nm.

Nowadays, image processing has attracted great attention as an analytical tool for colorimetric analysis due its versatility comprising, simple, fast and low-cost procedures. In fact, there are several relevant examples reported in the literature, such as for the determination of nitrite in clinical, food and environmental samples,[9] determination of sodium diclofenac, sodium dipyrone and calcium gluconate in injection drugs,[10] screening analysis of teas,[11] quantification of ethanol in drinks,[12] and quantification of sulfite in beverages.[13] Furthermore, the use of an image analysis technique for biochemical parameters can be a possible alternative for use in clinical assays, as demonstrated recently for creatinine[14] and glucose[15] determination. Digital images are basically based on three color systems: RGB (R – red, G – green and B – blue), HSV (H – hue, S – saturation and V – value) and grayscale.[14] These color systems provide information about the color itself, defined as color intensity, and about the spatial distribution of the object to be analyzed. Usually, analytical calibration models are built using intensity-related values associated with univariate curves;[16] however, since the linearity of single color channels is affected by the media in some cases, the use of multivariate calibration procedures, such as multiple linear regression (MLR), becomes necessary.[17] MLR is one of the simplest multivariate regression techniques, consisting of a series of linear equations systems resolved by the minimum square method for obtainment of regression coefficients that correlate instrumental measurements with some parameters of interest, such as concentration.

In this paper, we propose an analytical method integrating image analysis techniques based on MLR with colorimetric reactions to determine total proteins and albumin in serum. This alternative can provide an alternative approach for the

Biological Chemistry and Chemometrics, Institute of Chemistry, Federal University of Rio Grande do Norte, Natal 59072-970, RN, Brazil. E-mail: kassiolima@gmail.com

determination of the serum protein content presenting several advantages, such as the low cost of analytical instrumentation, reduction of the reactant volume spent during analysis and increase of the analytical frequency.

Materials and methods

Samples

Albumin and total protein determination was performed using synthetic and real serum samples. The samples utilized in the calibration stage were acquired from Labtest® (Labtest Diagnóstica SA, Brazil) with reference codes ref. 19 for albumin and ref. 99 for total proteins. The standard albumin utilized is traceable to the Certified Reference Material (CRM) 470 from the Institute for Reference Materials and Measurements/International Federation of Clinical Chemistry (IRMM/IFCC), and the standard total protein is traceable to the Standard Reference Material (SRM) 927 from the National Institute of Standards and Technology (NIST). The sample selection into calibration ($n = 9$) and prediction ($n = 5$) sets was performed using the Kennard–Stone algorithm.[18] The calibration set was composed of 5 synthetic and 4 real samples and the prediction set was composed of 5 real samples. The concentration ranges studied were from 0.24 to 5.03 g dL^{-1} for albumin, and 0.25 to 8.20 g dL^{-1} for total proteins.

Reference methods based on UV-Vis spectroscopy

The colorimetric reaction for the determination of albumin was based on the BCG reaction. For this reaction, 1.0 mL of color reagent (citrate buffer 60 mmol L^{-1} pH 3.8, bromocresol green 300 μmol L^{-1}, and Brij® 35 ≥ 6.0 mmol L^{-1}) reacted with 10 μL of the target-sample for 10 minutes at 37 °C. After the reaction, the reference measurements for albumin were carried out using a UV-Vis spectrophotometer Evolution 60S (Thermo Scientific, USA) at 630 nm. For this, the concentration of albumin was estimated according to the Labtest® kit (eqn (1)):

$$c_{alb} = \frac{A_{630}}{A^0_{alb}} \times 3.8 \quad (1)$$

where c_{alb} is the concentration of albumin in g dL^{-1}; A_{630} is the sample absorbance at 630 nm; A^0_{alb} is the absorbance at 630 nm of the albumin standard solution with a concentration of 3.8 g dL^{-1}.

For total proteins, the colorimetric method was based on the Biuret reaction, in which 1.0 mL of the color reagent (sodium hydroxide 600 mmol L^{-1} and copper sulfate 12 mmol L^{-1}) reacted with 20 μL of target-sample for 10 minutes at 37 °C. The reference measurements for total proteins were carried out using the UV-Vis spectrophotometer at 545 nm according to the Labtest® kit (eqn (2)):

$$c_{prot} = \frac{A_{545}}{A^0_{prot}} \times 4.0 \quad (2)$$

where c_{prot} is the concentration of total proteins in g dL^{-1}; A_{545} is the sample absorbance at 545 nm; A^0_{prot} is the absorbance at 545 nm of the total protein standard solution with a concentration of 4.0 g dL^{-1}.

Image acquisition

After the spectrophotometric measurement, 250 μL of the reactant content for albumin and total proteins were transferred to an ELISA microplate (Fisher Scientific, USA) being used as a reaction container for image acquisition. The images of the ELISA microplate were acquired by using a cell phone Samsung Galaxy Y 2.0 megapixel camera. The pictures were taken from a distance of 15 cm from the camera to the top of the microplate above a white surface by using 2× zoom, with the device held in static position to avoid blur. This procedure was performed under ambient lighting and immediately after the reaction content was transferred to the microplate. The images were saved in .JPG format with 96 × 96 dpi resolution. The regions of interest (ROIs) used to perform the computational analysis were acquired from the center of each microwell on the ELISA microplate as a square region of 10 × 10 pixels. In order to avoid ambient light interference in the microplate images, blank images were acquired for each sample, those being composed of a microwell filled with deionized water.

Computational analysis

The cell phone images were processed using MATLAB® software version 7.12 (MathWorks, USA) with PLS Toolbox 7.0.3 (Eigenvector Research, Inc. USA). The intensities for red-green-blue-hue-saturation-value (RGB-HSV) channels were computed as the mean for each of these matrices and then normalized by dividing their values by the blank intensities. The intensity values for each sample were concatenated into a row vector, $\mathbf{x}\{1 \times 6\}$, representing the color information for each sample. These row vectors were arranged into an $\mathbf{X}\{m \times 6\}$ matrix composed of m samples, which were divided into calibration ($\mathbf{X}_{cal}$) and prediction ($\mathbf{X}_{pred}$). MLR calibration was performed by estimating the regression coefficients as follows:

$$\mathbf{y}_{cal} = \mathbf{X}_{cal}\mathbf{b} \quad (3)$$

$$\hat{\mathbf{b}} = (\mathbf{X}^T_{cal}\mathbf{X}_{cal})^{-1}\mathbf{X}^T_{cal}\mathbf{y}_{cal} \quad (4)$$

where $\mathbf{y}_{cal}$ is the concentration vector for the calibration set and $\mathbf{b}$ is the vector containing the regression coefficients. The estimated concentration for the prediction set ($\hat{\mathbf{y}}_{pred}$) was calculated according to eqn (5).

$$\hat{\mathbf{y}}_{pred} = \mathbf{X}_{pred}\mathbf{b} \quad (5)$$

Multivariate analytical validation

The imaging method was evaluated and validated according to the following figures of merit: root-mean-square error of cross-validation (RMSECV) and prediction (RMSEP), linearity, trueness, precision, accuracy, sensitivity and prediction uncertainty. In addition to correlation coefficients, the linearity was

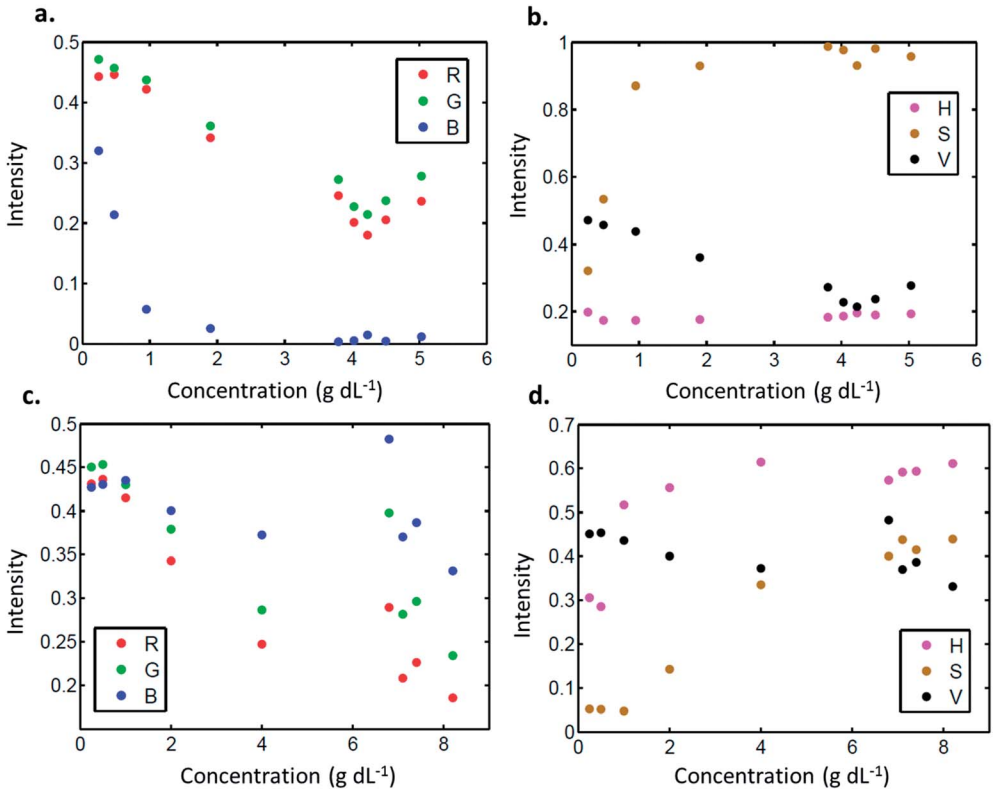


Fig. 1 RGB (a) and HSV (b) intensity responses for albumin at each concentration level of the calibration set; RGB (c) and HSV (d) intensity responses for total proteins at each concentration level of the calibration set.

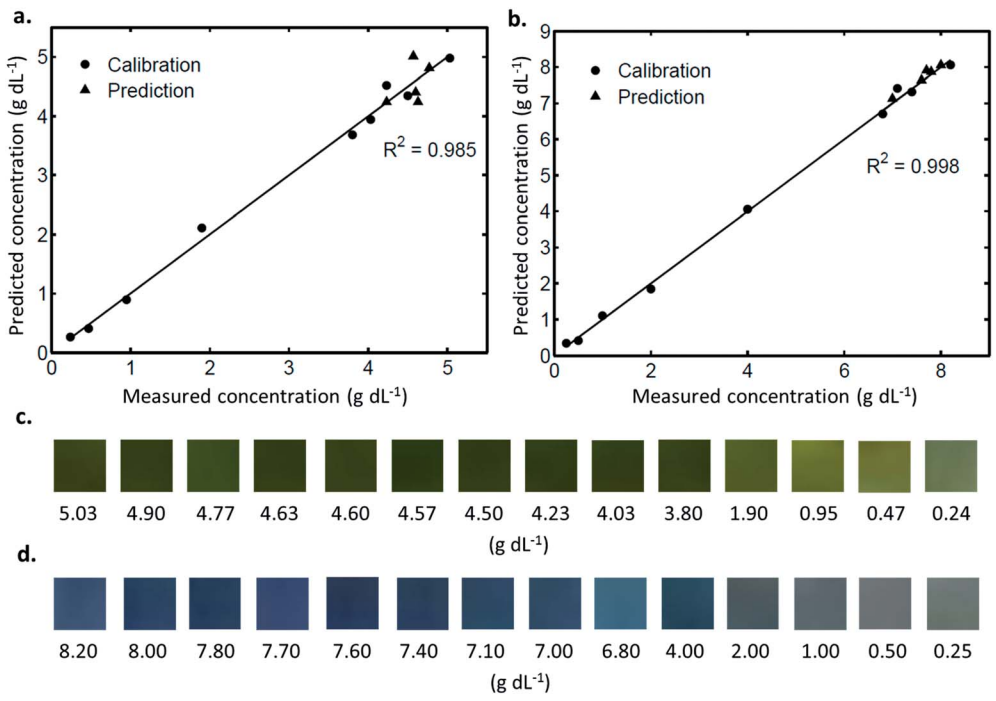


Fig. 2 Measured *versus* predicted concentrations calculated using MLR for RGB-HSV imaging models for albumin (a) and total proteins (b) and sample images at different concentration levels for albumin (c) and total proteins (d).

estimated following the experimental F test, the best criteria to evaluate this figure of merit.[19,20] This statistical parameter is calculated according to eqn (6):

$$F_{\text{exp}} = \left(\frac{S_{y/x}}{S_y}\right)^2 \quad (6)$$

where F_{exp} is the experimental F ratio; $S_{y/x}$ is the residual standard deviation; S_y is the pure error, a measure of the instrumental noise. The $S_{y/x}$ and S_y can be estimated from the calibration set as follows:

$$S_{y/x} = \sqrt{\frac{\sum_{i=1}^{I}\left(y_i - \hat{y}_i\right)^2}{I-2}} \quad (7)$$

$$S_y = \sqrt{\frac{\sum_{l=1}^{L}\sum_{q=1}^{Q}\left(y_{lq} - \bar{y}_l\right)^2}{I-Q}} \quad (8)$$

in which y_i and $\hat{y}_i$ are the experimental and estimated response for sample i; y_{lq} is the calibration response for replica q at level l; $\bar{y}_l$ is the mean response at level l; I is the total number of calibration samples; L is the total number of levels; Q is the total number of replicates at each level.

The trueness, precision and accuracy were estimated following criteria such as relative error (RE), relative standard-deviation (RSD) and β-expectation tolerance intervals.[14] Also, the imaging method was evaluated using a set of global desirable indexes: dosing range index (I_{DR}), trueness index (I_{T}), precision index (I_{P}) and accuracy index (I_{A}).[21] These indexes provide quantitative information varying from 0 to 1 regarding how the model responds to the validation quality parameters.

Results and discussion

Univariate analysis

Initially, the RGB-HSV responses for albumin and total protein images were evaluated univariately to investigate any linear relationship between single color channels and concentration values as demonstrated. However, as can be seen in Fig. 1, no linear relationship is present on single color channels, especially in high concentrations, where the color tends to saturate or assumes a disorder pattern. This fact induces the use of a more robust calibration model based on multivariate regression.

Multivariate analysis: MLR

MLR was used to build a multivariate calibration model for the quantification of albumin and total proteins. The calibration samples were used to obtain the regression coefficients and the prediction samples to test the model. Cross-validation leave-one-out was used during the calibration step. Fig. 2 shows the measured *versus* predicted concentration for calibration and prediction sets obtained using MLR.

It is clear, from Fig. 2, that there is a linear tendency in the whole concentration range for both parameters being analyzed, which shows that the estimated concentrations for both calibration and prediction sets were very similar to the reference values. Both models were linear, presenting an F value near to zero and way below the critical F. In addition, excellent correlation and determination coefficients were observed. The relative error found for both models was quite small (<7%), especially for total protein determination (1.54%). These error values were close to those of the reference method, which has presented average relative errors of 3.73% for albumin and 0.88% for total protein. Additionally, the precision and accuracy were very satisfactory as well, where the RSD for six replicas was smaller than 7% and the recovery near to 100% for both parameters. Table 1 provides the figures of merit used to validate both models.

The limits of detection (LOD) and quantification (LOQ) found for the MLR model of both biochemical parameters were very satisfactory. These values of the LOD and LOQ found in the imaging method do not interfere in the clinical analysis, since

Table 1 Results found for MLR models

Model parameters	Albumin	Total protein
Working range (g dL^{-1})	0.24–5.03	0.25–8.20
Regression coefficients	[2.06, 1599, −8.15, 0.30, −0.32, −16.06]	[−4.34, 7.22, −16.96, 0.10, −0.72, 13.92]
RMSECV (g dL^{-1})	0.14	0.15
Bias at calibration (g dL^{-1})	0.00	0.00
Prediction parameters		
RMSEP (g dL^{-1})	0.28	0.12
Bias at prediction (g dL^{-1})	0.01	−0.10
Linearity		
Slope	0.994	0.998
Intercept	0.016	0.008
R	0.996	0.998
R^2	0.993	0.997
F experimental	0.034	0.018
Trueness		
Relative error (%)	6.09	1.54
I_{T}[a]	0.837	0.989
Precision		
RSD (%)	6.16	4.85
I_{P}[a]	0.947	1.000
Accuracy		
Recovery (%)	91.67–109.76	100.56–102.70
β-TI	[0.30; 11.88]	[−4.03; 7.11]
I_{A}	0.924	0.995
Others FOM		
LOD[b] (g dL^{-1})	0.09	0.08
LOQ[b] (g dL^{-1})	0.27	0.27
I_{DR}	0.994	0.997

[a] Trueness index (I_{T}) and precision index (I_{P}) for an acceptable error of 15%. [b] Statistical significance was determined at the 0.05 level.

the critical values of albumin and total proteins in serum are respectively equal to 3.5–5.5 g dL^{-1} and 6.0–8.0 g dL^{-1}, for an adult person.

The desirable indexes (I_T, I_P, I_{DR} and I_A) found using MLR for both parameters were also very good, being close to 1. The trueness index (I_T) close to 1 indicates that the method is almost not biased and the precision index (I_P) indicates the precision of the method. Also, the dosing range index (I_{DR}) equal to one indicates that the method is valid in the whole concentration range studied, and the accuracy index (I_A) indicates the total error expected in the dosing range, showing the overall quality of the method.[21] These indexes were especially good for total protein determination, being all over 0.98. They are calculated based on β-TI (using three series of triplicates, $p = 3$ and $n = 3$), which is a quality parameter that indicates all predicted values obtained for each biochemical parameter which are expected to present relative errors within the accepted limits, showing the reliability and feasibility of use of image analysis for these assays. In addition, a paired *t*-test was performed and no statistical difference was observed at a confidence level of 95% between the results found using the imaging method and the spectrophotometric reference values.

Conclusion

A multivariate calibration method based on RGB and HSV color channels obtained from cell phone images was developed and validated for serum protein content determination (albumin and total proteins). The results presented in this paper show that image analysis with chemometric techniques can be a potential alternative, or even a substitute, for spectrophotometric measurements used traditionally in these kinds of assays. The advantages of this method rely on the reduction in the amount of reagent and waste in the analysis, since a 250 μL two-dimensional microplate array was used as a reaction container, lower cost due to the simplicity of the utilized equipment, and increased analytical frequency since the imaging method records a 96-microwell based image once, which can be analyzed in a few seconds. Furthermore, this method has potential to be used during on-site analysis in remote or non-developed regions, as a quick and non-expensive clinical assay for proteins.

Acknowledgements

The authors would like to acknowledge the structural and financial support from UFRN, CAPES and CNPq. Camilo L. M. Morais and Ana C. O. Neves thanks the Post-Graduate Program in Chemistry (PPGQ) of UFRN and CAPES; Kássio M. G. Lima thanks the CNPq (305962/2014-0).

References

1 R. F. Murray, *Harpe's Illustred Biochemistry*, Lange Medical Books/McGraw-Hilll, New York, 2006.
2 J. L. García-García, D. Pérez-Guaita, J. Ventura-Gayete, S. Garrigues and M. de la Guardia, *Anal. Methods*, 2014, **6**, 3982.
3 G. Motyckova and M. Murali, *Am. J. Hematol.*, 2011, **86**, 500–502.
4 G. A. Kaysen, K. L. Johansen, S.-C. Cheng, C. Jin and G. M. Chertow, *J. Renal Nutr.*, 2008, **18**, 323–331.
5 M. C. Corti, J. M. Guralnik, M. E. Salive and J. D. Sorkin, *JAMA, J. Am. Med. Assoc.*, 1994, **272**, 1036.
6 R. F. Itzhaki and D. M. Gill, *Anal. Biochem.*, 1964, **9**, 401–410.
7 B. T. Doumas, D. D. Bayse, R. J. Carter, T. Peters and R. Schaffer, *Clin. Chem.*, 1981, **27**, 1642–1650.
8 S. Ito and D. Yamamoto, *Clin. Chim. Acta*, 2010, **411**, 294–295.
9 T. M. G. Cardoso, P. T. Garcia and W. K. T. Coltro, *Anal. Methods*, 2015, **7**, 7311–7317.
10 W. da Silva Lyra, F. A. Castriani Sanches, F. Antônio da Silva Cunha, P. H. Gonçalves Dias Diniz, S. G. Lemos, E. Cirino da Silva and M. C. Ugulino de Araujo, *Anal. Methods*, 2011, **3**, 1975.
11 P. H. G. D. Diniz, H. V. Dantas, K. D. T. Melo, M. F. Barbosa, D. P. Harding, E. C. L. Nascimento, M. F. Pistonesi, B. S. F. Band and M. C. U. Araújo, *Anal. Methods*, 2012, **4**, 2648.
12 L. Pires dos Santos Benedetti, V. Bezerra dos Santos, T. A. Silva, E. Benedetti-Filho, V. L. Martins and O. Fatibello-Filho, *Anal. Methods*, 2015, **7**, 4138–4144.
13 L. Pires dos Santos Benedetti, V. Bezerra dos Santos, T. A. Silva, E. Benedetti-Filho, V. L. Martins and O. Fatibello-Filho, *Anal. Methods*, 2015, **7**, 7568–7573.
14 C. de Lelis Medeiros de Morais and K. M. Gomes de Lima, *Anal. Methods*, 2015, **7**, 6904–6910.
15 M. Xia, L. Wang, Z. Yang and H. Chen, *Anal. Methods*, 2015, **7**, 6654–6663.
16 D. C. Christodouleas, A. Nemiroski, A. A. Kumar and G. M. Whitesides, *Anal. Chem.*, 2015, **87**, 9170–9178.
17 T. Næs, T. Isaksson, T. Fearn and T. Davies, *A User-friendly Guide to Multivariate Calibration and Classification*, NIR publications, Charlton, Chichester, UK, 2002.
18 R. Kennard and L. Stone, *Technometrics*, 1969, **11**, 137–148.
19 A. C. Olivieri, *Anal. Chim. Acta*, 2015, **868**, 10–22.
20 K. Danzer and L. A. Currie, *Pure Appl. Chem.*, 1998, **70**, 993–1014.
21 E. Rozet, V. Wascotte, N. Lecouturier, V. Préat, W. Dewé, B. Boulanger and P. Hubert, *Anal. Chim. Acta*, 2007, **591**, 239–247.

Apêndice E

Estimation of Brazilian charcoal properties using attenuated total reflectance-Fourier transform infrared (ATR-FTIR) spectrometry coupled with multivariate analysis

Rafaela M. R. Bezerra

Alexandre S. Pimenta

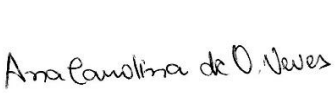
Ana C. O. Neves

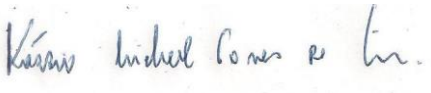
Kássio M. G. de Lima

Analytical Methods, 2015, 7, 5695-5701.

Contribuição:

- Coordenei o processamento dos dados e a construção dos modelos multivariados;
- Participei da escrita da primeira versão do manuscrito.





Ana C. O. Neves

Prof. Kássio M. G. Lima

Analytical Methods

TECHNICAL NOTE

View Article Online
View Journal

Cite this: DOI: 10.1039/c5ay01135c

Estimation of Brazilian charcoal properties using attenuated total reflectance-Fourier transform infrared (ATR-FTIR) spectrometry coupled with multivariate analysis

Rafaela M. R. Bezerra,[a] Ana C. O. Neves,[b] Alexandre S. Pimenta[a] and Kássio M. G. Lima*[b]

Received 4th May 2015
Accepted 19th May 2015

DOI: 10.1039/c5ay01135c

www.rsc.org/methods

The aim of the present work was to estimate fixed-carbon, volatile matter and ash contents in Brazilian commercial charcoal by using attenuated total reflectance-Fourier transform infrared (ATR-FTIR) spectroscopy together with multivariate calibration methods. Several multivariate calibration techniques, including partial least squares (PLS), interval partial least squares (iPLS), and genetic algorithm (GA), were compared and validated by establishing significance testing. Charcoal samples ($n = 72$) were divided into calibration ($n = 52$) and validation sets ($n = 20$) by applying the classic Kennard–Stone (KS) selection algorithm to the ATR-FTIR spectra. For the fixed-carbon content, the results obtained using PLS-GA for the root mean square error of cross-validation (RMSECV) and prediction (RMSEP) were 3.77% and 4.29%, respectively. For volatile matter, RMSECV and RMSEP of 4.36% and 4.65% were achieved by the PLS model using seven latent variables (LV). Finally, for ash, RMSECV and RMSEP of 0.58% and 0.38% were achieved by the PLS model using eight latent variables (LV). A *t*-test and quantile–quantile (Q–Q) plot were performed to compare the results of the models with each other and with a reference method. These results suggest that ATR-FTIR spectroscopy and multivariate calibration can be used effectively to determine fixed-carbon, volatile matter content and ash content in Brazilian charcoal.

Introduction

Charcoal is the residue of solid non-agglomerating organic matter of vegetable or animal origin, which results from carbonization by heating in the absence of air at a temperature above 300 °C.[1,2] The characteristics of wood charcoal are effectively associated with the chemical structures formed during the heating process.[3,4] The structure of charcoal is believed to be closely related to that of activated carbon (AC), which is primarily composed of short stacks of graphene sheets rimmed with O-containing groups (–OH, $–CO_2H$, –O–, =O, –CHO, *etc.*) to form a microporous network.[5] Chemical (fixed-carbon, volatile matter, ash, sulfur and phosphorus) and physical (hardness, specific weight, yield and moisture) properties are greatly influenced by three factors – raw material type,[6] process characteristics, and post-treatment.[7]

Most of the techniques used for the analysis of chemical and physical properties in charcoal are time-consuming and expensive, while rapid methods that require little or no sample preparation are needed for large scale surveys. Alternative methods such as X-ray photoelectron spectroscopy,[8] emission scanning electron microscopy,[9] near Infrared spectroscopy,[10] nuclear magnetic resonance spectroscopy,[11] Raman spectroscopy[12] and infrared spectroscopy[13–15] are suitable to replace the usual physico-chemical analysis. Specifically, one recent development in FTIR techniques applied to coal is the incorporation of an attenuated total reflectance (ATR) crystal (~100 μm in diameter), in which the standard polished block samples can be used without further sample preparation.[16]

Furthermore, ATR-FTIR spectroscopy can distinguish components (macerals) of charcoal which have diverse chemical compositions and physical properties, quantifying the abundance of chemical functional groups.[13] Limited studies have applied ATR-FTIR spectra for qualitatively evaluating charcoal. For instance, Guo and Bustin[17] have used ATR-FTIR to establish the relationships between both temperature and duration of heating of charcoal formation, reflectance values and spectral characteristics of charcoals, such as coalification maturation. In addition, reflectance and FTIR spectra indicate that fungal-decayed wood is particularly susceptible to the formation of charcoal and thus inertinite. Labbé and colleagues[13] employed ATR-FTIR to investigate the chemical

[a]*Universidade Federal do Rio Grande do Norte, Unidade Acadêmica Especializada em Ciências Agrárias, Grupo de Pesquisa Florestas, Bioenergia e Meio Ambiente, 59072-970 – Natal, RN, Brazil*

[b]*Biological Chemistry and Chemometrics, Institute of Chemistry, Federal University of Rio Grande of Norte, Natal 59072-970, RN, Brazil. E-mail: kassiolima@gmail.com; Tel: +55 84 3342 2323*

structure of charcoal made from different maple species: sugar maple (*Acer saccharum*), red maple (*Acer rubrum*), and silver maple (*Acer saccharinum*). In the second part of the study, the authors investigated the effect of thermal treatments on the chemical structure of white oak to have a better understanding of the maturation process in toasted charred white oak barrels.

However, the use of appropriate tools for multivariate calibration is mainly responsible for the advancement of the ATR-FTIR technique to achieve complete and fast characterization of charcoal or coal including partial least squares (PLS)[18] and methods based on the selection of variable intervals or spectral bands, such as iPLS (interval partial least squares),[19] GA (genetic algorithm)[20] and successive projection algorithm (SPA).[21] These last methods eliminate variables that do not directly correlate with the property of interest. They also eliminate potential interference and variables that generate a lower signal/noise ratio, which is indicative of low sensitivity.

Herein, we have attempted to make a comparison of the full-spectrum PLS (full-PLS), interval PLS (iPLS), successive projection algorithm (SPA), and genetic algorithm-PLS (GA-PLS), using ATR-FTIR for estimating the properties of charcoal made from wood such as volatile matter, fixed carbon content and ash. The topic of wavelength selection is of particular importance in ATR-FTIR spectra since they generally show relatively high sensitivity to small perturbations in the experimental conditions, as well as the physical and chemical properties of samples. To our knowledge, there is no report presenting ATR-FTIR-based calibrations for estimating the properties of charcoal made from wood, such as volatile matter, fixed-carbon and ash contents. Wavelength selection with interval based algorithms such as iPLS, SPA and GA is also not mentioned for charcoal analysis.

Materials and methods

Samples of charcoal

In this study, seventy-two commercial charcoal samples acquired from some locations in and around the Natal-Brazil region are the whole sample set. All samples were ground to a particle size of 40 mesh with a Wiley mill (Thomas Scientific, Philadelphia, PA).

Chemical properties

Fixed-carbon, volatile matter and ash contents were determined using the proximate chemical analysis of wood charcoal according to the procedure D-1762-84 of ASTM[22] and ABNT NBR 8112/83,[23] while the fixed carbon content was calculated following the equation of Anon.[24] To determine volatile matter, the furnace was then preheated to 950 °C with the vent port capped. Samples were introduced into the furnace as quickly as possible, rather than preheating crucibles by placing them on the outer edge of the furnace for 2 min and then on the edge of the furnace for 3 min, as described in the ASTM method. It should be noted that the furnace temperature would not rebound to 950 °C until approximately 8 min after the samples had been introduced. Samples were removed from the furnace after 10 min and placed on a refractory brick to cool until they could be safely transferred into desiccators, at a point above 200 °C. Covered crucibles were weighed after cooling to ambient temperature and the volatile matter content was calculated as follows:

$$\text{Volatile matter}\% = \frac{\text{weight}_{105\,^\circ\text{C dried}} - \text{weight}_{950\,^\circ\text{C devolatilized}}}{\text{weight}_{105\,^\circ\text{C dried}}} \times 100 \quad (1)$$

For determination of ash contents, covers were removed from the crucibles and the furnace vent port was connected to the fume hood exhaust. Following this, samples were placed in the furnace and the temperature was increased from 105 °C to 750 °C at 5 °C min^{-1}, and then held at 750 °C for 6 hours. The furnace was allowed to cool to 105 °C before the samples were transferred to desiccators. The ash content was determined by weight loss according to the following equation:

$$\text{Ash}\% = \frac{\text{weight}_{\text{residue after 750 °C}}}{\text{weight}_{105\,^\circ\text{C dried}}} \times 100 \quad (2)$$

Volatile and ash contents were used to calculate the fixed carbon content according to the following equation:

$$\text{Fixed carbon}\% = \frac{\text{weight}_{105\,^\circ\text{C dried}} - \text{weight}_{950\,^\circ\text{C devolatilized}} - \text{weight}_{\text{residue after 750 °C}}}{\text{weight}_{105\,^\circ\text{C dried}}} \times 100 \quad (3)$$

It should be noted that the so-called fixed carbon content is given as the mass residue, and is not strictly a C content. All analyses were done in duplicate. Further details of the samples, including chemical analysis of the individual charcoal and reference method used in each parameter, are summarized in Table 1.

ATR-FTIR spectra measurement

Spectral measurements were performed using a Bruker ALPHA FTIR spectrometer equipped with an ATR accessory. Spectra (8 cm^{-1} spectral resolution giving 4 cm^{-1} data spacing equivalent to 258 wavenumbers, co-added for 32 scans) were converted into absorbance by using Bruker OPUS software. For the infrared measurements, the powder of each sample was placed on the diamond crystal of an ATR accessory. The average value from two different measurements of each sample was properly stored, and the mean spectrum was then calculated for each sample, giving a total of 72 ATR-FTIR spectra. After each measurement, the ATR plate was washed with ethanol (70% v/v) and dried using tissue paper. Cleanliness of the ATR plate was verified by collecting an absorbance spectrum of the crystal

Table 1 Statistical results of full set sample Brazilian charcoal analysis (72 samples) and the reference method applied in each case

Property	Minimum	Maximum	Mean	S.D.	Reference method
Fixed carbon (%)	59.2000	87.5000	73.9042	6.6124	ASTM D 1762-84
Volatile matter (%)	11.5000	39	23.5028	6.5554	ASTM D 1762-84
Ash (%)	0.3000	8.7000	2.3681	1.7368	ASTM D 1762-84

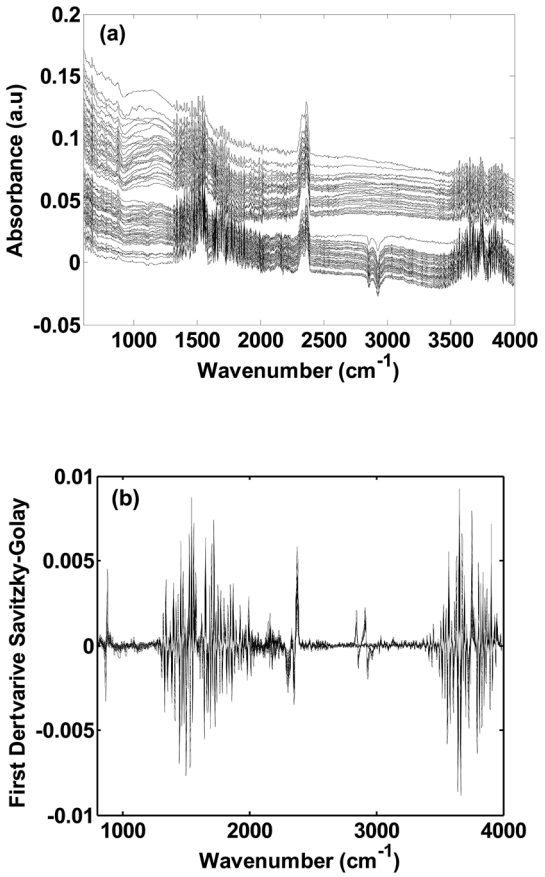


Fig. 1 (a) Original ATR-FTIR average spectra of 72 samples of commercial charcoals. (b) First derivative spectra of the original 72 samples of charcoal after pretreatment (Savitzky–Golay smoothing, MSC and a Savitzky–Golay derivative).

Table 2 Results for calibration and the external validation set for fixed-carbon content (%): root mean square error of cross-validation (RMSECV) and prediction (RMSEP), correlation coefficient for calibration set (r_c) and prediction set (r_p) and the number of used spectral variables (size). The number of latent variables in PLS, iPLS, PLS-SPA and PLS-GA models is shown in brackets

	Calibration		Prediction		
Models	r_c	RMSECV (%)	r_p	RMSEP (%)	Size
PLS (6)	0.65	3.92	0.63	4.37	1666
PLS (6)[a]	0.66	3.91	0.63	4.37	1656
PLS (8)[b]	0.57	4.18	0.77	3.99	1658
PLS (8)[c]	0.57	4.59	0.66	3.09	1658
PLS (5)[d]	0.54	4.37	0.71	3.67	1666
iPLS (5)	0.44	5.06	0.80	3.01	166
iPLS (5)	0.61	4.18	0.65	4.08	833
PLS-SPA (7)	0.72	3.50	0.60	4.56	20
PLS-GA (7)	0.68	3.77	0.66	4.29	407
PLS (7)[e]	0.62	4.04	0.79	3.35	1666
PLS (5)[f]	0.77	3.06	0.78	3.05	1666

[a] Smoothing (11 points). [b] First derivative (9 points). [c] Smoothing (5 points), first derivative (5 points) and MSC. [d] MSC. [e] One application of outlier detection. [f] Second application of outlier detection.

using the most recently collected background as a reference. Spectral measurements were done in an acclimatized room at a controlled temperature of 22 °C, and 60% relative air humidity.

Chemometrics procedure and software

All the datasets were exported to MATLABversion7.12 (The Math-Works, Natick, USA). Data analysis was performed using the PLS-toolbox (Eigenvector Research, Inc., Wenatchee, WA, USA, version 7.8). Cross-validation was employed to optimize the number of PLS factors and to guide the selection process in PLS models. Before computing variable selections and calibrations, different preprocessing methods were used, including the multiplicative scattering correction (MSC), first and second derivative and smoothing Savitzky–Golay methods by varying the number of window points (3, 5, 7, 9 and 11 points) using a first-order polynomial. Mean centering was applied to all spectra before performing variable subset selection and calibration.

All the samples were divided into calibration and prediction sets using the SPXY algorithm.[25] Then, $n_{calibration} = 52$ and $n_{prediction} = 20$ samples were used. To verify the capability of the calibration models based on the selected region by different methods (full-PLS, iPLS, GA-PLS and SPA), each model mentioned above was used to predict the calibration dataset and the prediction dataset. The RMSEC (root mean square error of calibration), RMSEP (root mean square error of prediction) and correlation coefficients of each model for the calibration dataset (r_c) and prediction dataset (r_p) were taken into account. For an ideal model, correlation coefficients (r_c and r_p) should be close to 1 while RMSECV/RMSEP is close to 0. Furthermore, the root mean square error of both calibration and prediction samples was proposed for assessing the overall performance of the model. A smaller RMSECV/RMSEP value indicates better model quality. Additionally, we used *t*-paired statistic for a significant ($P < 0.05$) difference or trend in the concentration of each parameter with the reference method. If the *t* calculated is higher than the critical *t*-value at the 95% confidence level, there is evidence that the bias included in the multivariate model is significant. The quantile–quantile (Q–Q) plot

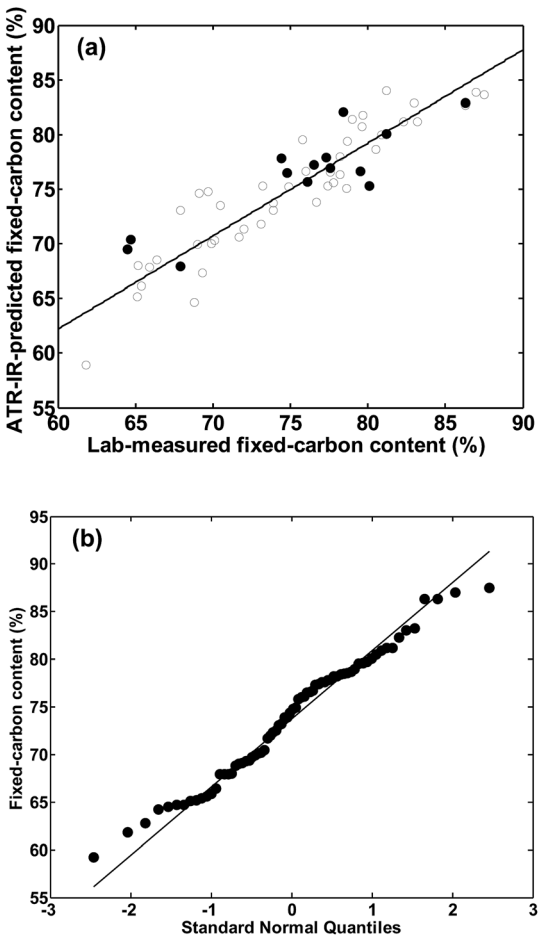


Fig. 2 (a) Predicted concentration *vs.* reference measured concentration of calibration and validation samples for the fixed-carbon content in commercial charcoals using the full PLS model after the outlier test, (○) calibration set and (●) prediction set. (b) Quantile–quantile (Q–Q) plot normal distribution for the fixed-carbon content.

compares the ordered distribution of a test sample with the quantiles of a standard normal distribution indicated by a straight line. If the sample is normally distributed, the points will lie along this line.[26]

Results and discussion

ATR-FTIR spectroscopic charcoal properties

The original spectra (calculated from the average between the two readings) giving a total of 72 ATR-FTIR spectra are shown in Fig. 1a. As can be seen in the ATR-FTIR spectra of charcoal containing information regarding its chemical composition and molecular structure, there were clear variations in the IR spectra of the charcoal samples. Although the direct interpretation of the IR spectrum is complicated, it is possible to assign some bands (inorganic structures) such as O–H stretching modes (3700–3600 cm^{-1}), aliphatic C–H stretching modes (2900–2800 cm^{-1}), Si–O stretching modes (1100–900 cm^{-1}), stretching modes of aromatic rings and carbonyl groups (1600 and 1400 cm^{-1}), C–O–C stretching (1030 cm^{-1}), aromatic C–H (900–700 cm^{-1}), and Si–O–Si and Si–O–Al bending modes (700–400 cm^{-1}).

Table 3 Results for calibration and the external validation set for volatile matter content (%): root mean square error of cross-validation (RMSECV) and prediction (RMSEP), correlation coefficient for calibration set (r_c) and prediction set (r_p) and the number of used spectral variables (size). The number of latent variables in PLS, iPLS, PLS-SPA and PLS-GA models is shown in brackets

	Calibration		Prediction		
Models	r_c	RMSECV (%)	r_p	RMSEP (%)	Size
PLS (7)	0.57	4.36	0.55	4.65	1666
PLS (6)[a]	0.56	4.46	0.56	4.50	1656
PLS (9)[b]	0.61	3.98	0.66	4.33	1656
PLS (8)[c]	0.44	5.09	0.72	3.10	1658
PLS (5)[d]	0.47	4.67	0.81	3.18	1666
iPLS (5)	0.56	4.42	0.53	4.59	166
iPLS (6)	0.53	4.57	0.58	4.22	833
SPA (7)	0.59	4.28	0.53	4.75	20
GA (7)	0.60	4.23	0.54	4.75	403
PLS (5)[e]	0.65	3.87	0.74	3.43	1666
PLS (6)[f]	0.70	3.51	0.85	2.83	1666

[a] Smoothing (11 points). [b] First derivative (9 points). [c] Smoothing (5 points), first derivative (5 points) and MSC. [d] MSC. [e] One application of outlier detection. [f] Second application of outlier detection.

In addition, Fig. 1a shows baseline shifts and bias present in the spectra, and undesirable features which need to be removed using some pre-treatments, such as smoothing (first-order), multiplicative scattering correction (MSC) and first- and second-order derivatives (Savitzky–Golay). Fig. 1b shows 72 ATR-FTIR spectra obtained during the pretreatment stage utilizing Savitzky–Golay smoothing (with a window of 5 points), MSC and the first derivative of the Savitzky–Golay polynomial (with a window of 5 points). Mean centering was also applied to all spectra before performing variable subset selection and calibration.

Fixed-carbon content

The fixed carbon content in the charcoal specimen was determined following eqn (3),[24] as the difference between 100 and the sum of moisture content, volatile matter and ash content. In other words, the carbon content can be estimated as a difference; all the other constituents are deducted from 100 as percentages and the remainder is assumed to be the fixed carbon. The results obtained for the calibration and predicted models in the ATR-FTIR region for the fixed-carbon content of commercial charcoal are displayed in Table 2. In addition to the full PLS models, the results of the iPLS, PLS-SPA and PLS-GA models are shown. Only the best results from the tested pre-processing techniques are presented. The number of latent variables (LV) calculated for each model corresponded to the first minimal residual variance. As can be shown in Table 2, the performance of the full PLS model is slightly better than that of the iPLS, GA and SPA models for the fixed-carbon content. The correlation coefficient for the prediction set ranged from 0.60 to

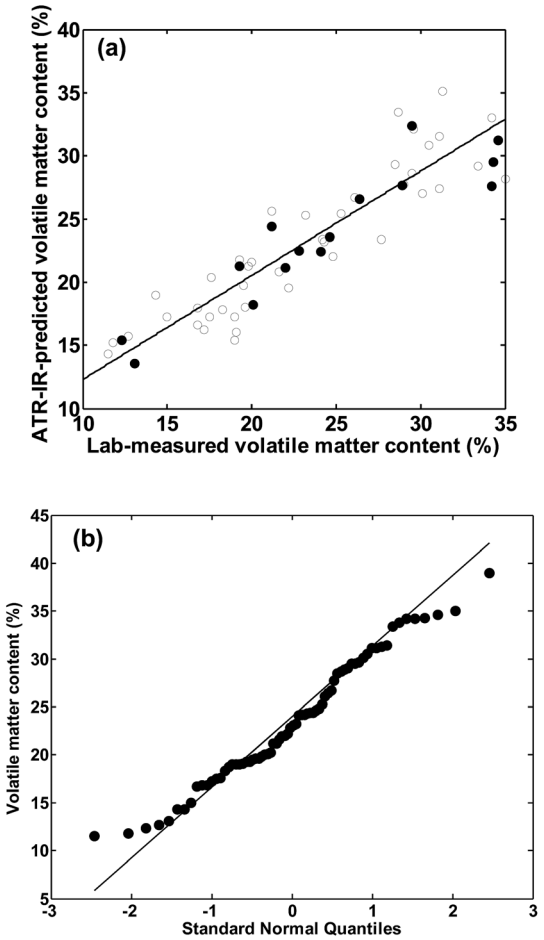


Fig. 3 (a) Predicted concentration *vs.* reference measured concentration of calibration and validation samples for the volatile matter content in commercial charcoals using the full PLS model after the outlier test, (○) calibration set and (●) prediction set. (b) Quantile–quantile (Q–Q) plot normal distribution for the volatile matter content.

Table 4 Results for calibration and the external validation set for ash content (%): root mean square error of cross-validation (RMSECV) and prediction (RMSEP), correlation coefficient for calibration set (r_c) and prediction set (r_p) and the number of used spectral variables (size). The number of latent variables in PLS, iPLS, PLS-SPA and PLS-GA models is shown in brackets

	Calibration		Prediction		
Models	r_c	RMSECV (%)	r_p	RMSEP (%)	Size
PLS (7)	0.72	1.00	0.34	0.65	1666
PLS (8)[a]	0.73	0.97	0.20	0.76	1662
PLS (3)[b]	0.62	1.11	0.79	0.81	1664
PLS (3)[c]	0.64	1.10	0.88	0.71	1658
PLS (6)[d]	0.75	0.96	0.65	0.46	1662
PLS (5)[e]	0.73	1.00	0.38	0.63	1666
iPLS (3)	0.65	1.13	0.65	0.57	166
iPLS (5)	0.53	1.33	0.01	1.42	831
SPA (5)	0.68	1.09	0.34	0.65	20
GA (9)	0.89	0.63	0.67	0.48	413
PLS (8)[f]	0.83	0.77	0.74	0.41	1662
PLS (8)[g]	0.89	0.58	0.75	0.38	1662

[a] Smoothing (5 points). [b] First derivative (3 points). [c] Smoothing (5 points), first derivative (5 points) and MSC. [d] Smoothing (5 points) and MSC. [e] MSC. [f] One application of outlier detection. [g] Second application of outlier detection.

0.80. In addition, it was also observed that models with wavelength selection in the ATR-FTIR spectral region (iPLS, GA and SPA) achieved RMSEP values between 3.01 and 4.56 (%). The number of LV used for the PLS, iPLS, SPA and GA models using ATR-FTIR spectra for the fixed-carbon content varied between 5 and 8. The calibration set was optimized by the exclusion of the samples that presented leverage, non-modeled residuals in the parameter (fixed-carbon content). Five outliers were excluded from the calibration set, and the best PLS model for the fixed-carbon content achieved RMSECV and RMSEP of 3.06 and 3.05, respectively. In addition, the correlation coefficients for the calibration and validation set for this model were 0.77 and 0.78, respectively, using 5 latent variables.

These results are corroborated by the graph of predicted *versus* reference values obtained by full PLS using 1666 spectral variables and a correlation coefficient of 0.78 for the prediction set using 5 LV, as shown in Fig. 2a. Moreover, to obtain a better insight for this model, *t*-test (ASTM E1655-00)[27] and normal ($P < 0.05$, quantile–quantile (Q–Q) plot) were calculated. The results showed that the bias included in the model was not significant, since the obtained *t* value of 0.67 for the fixed-carbon content was lower that the critical value of 2.14 with 95% of confidence. The Q–Q plot is an excellent graphical test of the normality of a sample and is commonly used for that purpose. The full PLS model for the fixed-carbon content was subjected to the Q–Q plot univariate normality test, and indicated a univariate normal data distribution as shown in Fig. 2b. It was therefore concluded that the dataset could be multivariate normally distributed.

Volatile matter content

Volatile matter was extracted by pre-heating the specimen in a tube furnace for 2 min at 300 °C, and then heating for 3 min at 500 °C and for 6 min at 950 °C. The volatile matter content was determined as a proportion of the oven-dry weight of the charcoal specimen. Table 3 displays the results for the analysis of the volatile matter of the charcoals. As can be seen for PLS models, better values were obtained for the RMSEP with smoothed data and MSC treatment compared to the models obtained with original raw data. For this parameter, the variable selection using the iPLS, the GA and SPA algorithms produced inferior results to those of full PLS. The best model found for this parameter was achieved using full PLS after exclusion of the outliers. When 1666 spectral variables were used to build the full PLS (6) model, we found a correlation coefficient of 0.85 for the prediction set. The plot of laboratory-determined volatile matter *versus* ATR-FTIR-predicted volatile matter is given in Fig. 3a, using 6 VL. We tested the presence of relevant bias with the prediction results for the prediction samples using the full PLS of the *t*-test suggested by ASTM E1655-00. The results showed that the bias included in the model was not significant

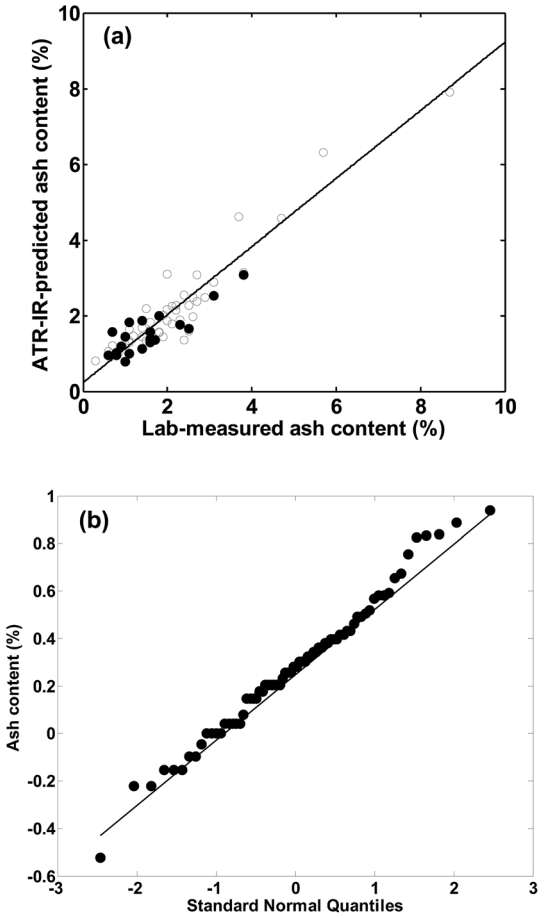


Fig. 4 (a) Predicted concentration *vs.* reference measured concentration of calibration and validation samples for the ash content in commercial charcoals using the full PLS model after the outlier test, (○) calibration set and (●) prediction set. (b) Quantile–quantile (Q–Q) plot normal distribution for the ash content.

($t_{calculated}$ = 0.94, $t_{critical}$ = 2.14, 95% confidence level). The volatile matter content was also subjected to the Q–Q plot univariate normality test, and indicated a univariate normal data distribution as shown in Fig. 3b.

Ash content

The ash content was calculated as a proportion of the oven-dry weight of the residue to the oven-dry weight of the charcoal specimen. Table 4 presents the model statistics of the ATR-FTIR models for the ash content. As can be seen in Table 4, the performance of the full PLS model was better than the wavelength selection models (iPLS, GA and SPA) for the ash content. The correlation coefficient for the prediction set ranged from 0.01 to 0.88. The number of LV used for the PLS, iPLS, SPA and GA models using ATR-FTIR spectra for the ash content varied between 3 and 9. The best model found for the ash content was achieved using the full PLS after exclusion of the outliers and smoothing (5 points window) and MSC as preprocessing methods. When 1662 spectral variables were used to build the full PLS (8) model, a correlation coefficient of 0.75 for the prediction set was achieved. These results are corroborated by the graph of predicted *versus* reference values obtained by full PLS, as shown in Fig. 4a.

The model (full PLS) was not significantly different using prediction samples for ash content when compared with the reference values according to a paired *t*-test ($t_{calculated}$ = 0.39, $t_{critical}$ = 2.09, 95% confidence level). Finally, the full PLS model for the ash content was subjected to the Q–Q plot univariate normality test, and indicated a univariate normal data distribution as shown in Fig. 4b. It was therefore concluded that the dataset could be multivariate normally distributed.

Conclusions

In this study, we demonstrated ATR-FTIR based on full PLS and wavelength variable models (iSPA, GA and SPA) for estimating the fixed-carbon, volatile matter and ash contents of commercial charcoal. It can be concluded that ATR-FTIR is a very promising technique for the non-destructive quantification of important parameters in charcoals. An advantage of the ATR method applied to charcoal is that standard polished block samples can be used without further sample preparation. For instance, the full PLS models developed for each parameter can be useful for monitoring charcoal quality in steel industries. These models were validated by cross-validations and independent statistic tests. The findings presented in this paper provide a detailed analytical view of real ATR-FTIR data and they could be applied to other spectral signals as well.

Acknowledgements

The authors would like to acknowledge the financial support from the PPGQ/UFRN/CAPES for a fellowship for Ana Carolina de Oliveira Neves. We are grateful to Fabio Godoy (Bruker Optics Ltd.) for the excellent technical assistance in handling the Bruker ALPHA FT-IR spectrometer. K. M. G. Lima acknowledges the CNPq/CAPES project (Grant 070/2012 and 442087/2014-4) and FAPERN (Grant 005/2012) for financial support.

References

1 W. Emrich, in *Solar Energy R&D in the European Community*, Springer-Science+Business Media, 1985, p. 278.
2 M. J. Antal, *Ind. Eng. Chem. Res.*, 2003, **42**, 1619–1640.
3 P. L. Ascough, M. I. Bird, P. Wormald, C. E. Snape and D. Apperley, *Geochim. Cosmochim. Acta*, 2008, **72**, 6090–6102.
4 M. Somerville and S. Jahanshahi, *Renewable Energy*, 2015, **80**, 471–478.
5 J. Bourke, M. Manley-Harris, C. Fushimi, K. Dowaki, T. Nunoura and M. J. Antal, *Ind. Eng. Chem. Res.*, 2007, **46**, 5954–5967.
6 A. L. Missio, B. D. Mattos, D. A. Gatto and E. A. de Lima, *J. Wood Chem. Technol.*, 2013, **34**, 191–201.
7 L. Wang, Y. Xin, Z. Zhou, X. Xu and H. Sun, *J. Hazard. Mater.*, 2013, **244–245**, 268–275.
8 K. Nishimiya, T. Hata, Y. Imamura and S. Ishihara, *J. Wood Sci.*, 1998, **44**, 56–61.

9 E. Hobley, G. R. Willgoose, S. Frisia and G. Jacobsen, *Eur. J. Soil Sci.*, 2014, **65**, 751–762.
10 C. Andrade, P. Trugilho, P. Gherardi Hein, J. Lima and A. Napoli, *J. Near Infrared Spectrosc.*, 2012, **20**, 657.
11 A. V. McBeath, R. J. Smernik, M. P. W. Schneider, M. W. I. Schmidt and E. L. Plant, *Org. Geochem.*, 2011, **42**, 1194–1202.
12 O. Francioso, S. Sanchez-Cortes, S. Bonora, M. L. Roldán and G. Certini, *J. Mol. Struct.*, 2011, **994**, 155–162.
13 N. Labbé, D. Harper, T. Rials and T. Elder, *J. Agric. Food Chem.*, 2006, **54**, 3492–3497.
14 B. Pizzo, E. Pecoraro, A. Alves, N. Macchioni and J. C. Rodrigues, *Talanta*, 2015, **131**, 14–20.
15 H. Li, Y. Yang, S. Yang, A. Chen and D. Yang, *J. Spectrosc.*, 2014, **2014**, 1–7.
16 J. Thomasson, C. Coin, H. Kahraman and P. M. Fredericks, *Fuel*, 2000, **79**, 685–691.
17 Y. Guo and R. Bustin, *Int. J. Coal Geol.*, 1998, **37**, 29–53.
18 S. Wold and M. Sjostrom, *Chemom. Intell. Lab. Syst.*, 2001, **58**, 109–130.
19 L. Nørgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck and S. B. Engelsen, *Appl. Spectrosc.*, 2000, **54**, 413–419.
20 M. Ferrand, B. Huquet, S. Barbey, F. Barillet, F. Faucon, H. Larroque, O. Leray, J. M. Trommenschlager and M. Brochard, *Chemom. Intell. Lab. Syst.*, 2011, **106**, 183–189.
21 S. F. C. Soares, A. A. Gomes, A. R. G. F. Filho, M. C. U. Araujo and R. K. H. Galvão, *Trends Anal. Chem.*, 2013, **42**, 84–98.
22 American Society for Testing and Materials—ASTM, *Standard Methods for Chemical Analysis of Wood Charcoal, D1762-84*, American Society for Testing and Materials, Philadelphia, PA, USA, 1989.
23 *Associação brasileira de normas técnicas - NBR 8112/83 Carvão vegetal – análise imediata.*, 1983.
24 FAO, *Simple technologies for charcoal making*, FAO Forestry Paper 41, FAO, Rome, Italy, 1987.
25 R. K. H. Galvão, M. C. U. Araujo, G. E. José, M. J. C. Pontes, E. C. Silva and T. C. B. Saldanha, *Talanta*, 2005, **67**, 736–740.
26 A. R. Henderson, *Clin. Chim. Acta*, 2006, **366**, 112–129.
27 ASTM International, *Annual Book of ASTM Standards, Standard Practices for Infrared Multivariate Quantitative Analysis - E1655-00*, West Conshohocken, Pennsylvania, USA, 2000.