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

Métodos espectroscópicos e classificação multivariada aplicada na diferenciação de microrganismos patógenos

Fernanda Saadna Lopes da Costa

Tese de Doutorado

Natal/RN, julho de 2019

Fernanda Saadna Lopes da Costa

MÉTODOS ESPECTROSCÓPICOS E CLASSIFICAÇÃO MULTIVARIADA APLICADOS NA DIFERENCIAÇÃO DE MICRORGANISMOS PATÓGENOS

Tese apresentada ao Programa de Pós-Graduação em Química da Universidade Federal do Rio Grande do Norte, como parte dos requisitos necessários para a obtenção do título de Doutora em Química.

Orientador: Prof. Dr. Kássio Michell Gomes de Lima

Natal – RN
2019

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI
Catalogação de Publicação na Fonte. UFRN - Biblioteca Setorial Prof. Francisco Gurgel De Azevedo - Instituto Química - IQ

Costa, Fernanda Saadna Lopes da.
Métodos espectroscópicos e classificação multivariada aplicados na diferenciação de microrganismos patógenos / Fernanda Saadna Lopes da Costa. - Natal: UFRN, 2019.
102f.: il.

Tese (Doutorado) - Universidade Federal do Rio Grande do Norte. Centro de Ciências Exatas e da Terra - CCET, Instituto de Química. Programa de Pós-Graduação em Química (PPGQ).
Orientador: Drº. Kássio Michell Gomes de Lima.

1. Cryptococcus - Tese. 2. Klebsiella sp - Tese. 3. Escherichia coli - Tese. 4. ATR-FTIR - Tese. 5. Fluorescência - Tese. 6. Análise Multivariada - Tese. I. Lima, Kássio Michell Gomes de. II. Título.

RN/UF/BSQ CDU 543(043.2)

Elaborado por FERNANDO CARDOSO DA SILVA - CRB-759/15

Fernanda Saadna Lopes da Costa

MÉTODOS ESPECTROSCÓPICOS E CLASSIFICAÇÃO MULTIVARIADA APLICADA NA DIFERENCIAÇÃO DE MICRORGANISMOS PATÓGENOS

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: 25 de julho de 2019

Comissão Examinadora:

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

Dr. Fabrício Gava Menezes – UFRN

Dr. Edgar Perin Moraes – UFRN

Dr. Josélio Maria Galvão de Araújo – UFRN

Dr. Sherlan Guimaraes Lemos – UFPB

Dr. Luciano Farias Almeida – UFPB

AGRADECIMENTOS

À Deus, primeiramente, por me conceder determinação, saúde e sabedoria para concluir esta jornada.

Ao meus pais, Ana Maria e Francisco Benjamim, por todo o amor, dedicação e apoio. Por nunca deixarem de acreditar em mim e por sempre se esforçarem para que alcançasse meus objetivos. Grande parte das minhas conquistas eu devo e ofereço a vocês.

Ao meu orientador Dr. Kássio Lima, por sua constante atenção, apoio, cobrança, pelos inúmeros conhecimentos transmitidos, e sobretudo, pela confiança que em mim foi depositada ao longo destes 9 anos.

À minha família que sempre acreditou e torceu por mim a cada novo desafio, em especial meus avós Maria Ester e Antônio de Assis.

Aos meus companheiros de CHEMOVECTOR e amigos, Camilo Morais, Heloiza Oliveira e Leomir Aires. Vocês foram fundamentais ao longo desses anos, me ensinaram, apoiaram e incentivaram a chegar aqui. Obrigada por todos os momentos partilhados, que trouxeram mais leveza a essa jornada.

Às Helenas: Aline Marques, Ana Carolina, Jábine Talitta, Karine Fonseca e Maria Raquel. Aos companheiros de república: Daniele Santos, Eduardo Vasconcelos e Rosângela Costa. As minhas amigas Gleyca Rocha, Lívia Borges, Luana Queiroz e Raiene Ohana. Meus primos Kauã Lopes, Lamonier Lopes e Tácylla Danyelle. Muito obrigada por todo apoio e compreensão que vocês tiveram ao longo desses anos, por sempre se fazerem presentes e me incentivar a alcançar meus objetivos.

À minha amiga Wendy Marina, que desde o ensino médio foi minha companheira de estudos, com quem tracei as principais metas da minha vida e quem sempre foi meu apoio, incentivo e “minha coragem” para seguir em frente.

À minha irmã Aline Marques, por me reerguer nos momentos difíceis, comemorar cada vitória e acreditar tanto em mim. Por não me deixar desistir e ser quem eu sei que sempre posso contar.

Ao meu marido Dogival Ferreira, por me incentivar, ensinar e estar ao meu lado nos bons e nos maus momentos. Obrigada pelo seu apoio, paciência e dedicação, por me fazer crer que tudo vai dar certo.

À CAPES pela bolsa concedida e financiamento do trabalho.

RESUMO

Este trabalho apresenta o desenvolvimento de métodos de classificação multivariada, aliados a técnicas espectroscópicas, como a espectroscopia na região do infravermelho médio e de fluorescência molecular, na detecção de microrganismos patógenos: fungos e bactérias. Os primeiros estudos, buscavam a diferenciação de *Cryptococcus neoformans* e *Cryptococcus gatti.* Estes fungos são os agentes etiológicos da criptococose, cujo tratamento adequado depende da detecção e diferenciação rápida e correta da espécie. Esta identificação é atualmente feita por técnicas clássicas e moleculares, que em sua maioria são trabalhosas e dispendiosas. Como método alternativo para discriminar *C. gattii* e *C. neoformans*, foi investigada inicialmente a espectroscopia no infravermelho médio por reflectância total atenuada, aliada a técnicas de classificação multivariada (PCA-LDA/QDA, GA-LDA/QDA, SPA-LDA/QDA), no qual o modelo GA-QDA obteve sensibilidade nas classes *C. neoformans* e *C. gatti* de 84,4% e 89,3%, respectivamente, utilizando apenas 17 números de onda. Em seguida, foi utilizada a espectroscopia de fluorescência em matriz excitação-emissão (EEM), combinada com métodos de classificação multivariada (UPCA-LDA/QDA, UGA-LDA/QDA, USPA-LDA/QDA, PARAFAC/PLS-DA, nPLS-DA). O modelo mais satisfatório foi o UGA-LDA, que utilizou apenas 5 comprimentos de onda, e apresentou sensibilidade de 88,9% em calibração e 100,0% de previsão para ambas as espécies, resultados que são comparáveis aos testes biológicos de rotina. O último estudo visava a diferenciação de bactérias sensíveis e multirresistentes do gênero *Klebsiella sp*. e *Escherichia coli*. Através da espectroscopia de fluorescência molecular e os métodos de classificação multivariada LDA, QDA e SVM acoplados a algoritmos de redução de dados PCA, GA e SPA. Dentre estes, os modelos que tiveram melhores desempenho para ambos os gêneros de bactéria, apresentaram taxas de sensibilidade e especificidade de 100%. Em comparação com os métodos clássicos, as metodologias propostas nestes estudos demonstram ser uma alternativa inovadora, mais rápida e barata para a identificação de microrganismos patógenos, como fungos e bactérias, abrindo a possibilidade de aplicação em laboratórios de diagnósticos de rotina.

Palavras-chave: FT-IR, Fluorescência, Análise Multivariada, *Cryptococcus, Klebsiella sp*. e *Escherichia coli*.

ABSTRACT

This paper presents the development of multivariate classification methods, combined with spectroscopic techniques, such as spectroscopy in the middle infrared region and molecular fluorescence, in the detection of pathogenic microorganisms: fungi and bacteria. The first studies sought the differentiation of *Cryptococcus neoformans* and *Cryptococcus gattii*. These fungi are the etiological agents of cryptococcosis, whose proper treatment depends on the rapid and correct detection and differentiation of the species. This identification is currently done by classical and molecular techniques, which are mostly laborious and expensive. As an alternative method to discriminate *C. gattii* and *C. neoformans*, we initially investigated medium infrared spectroscopy by attenuated total reflectance, together with multivariate classification techniques (PCA-LDA/QDA, GA-LDA/QDA, SPA-LDA/QDA), in which the GA-QDA model obtained sensitivity in classes *C. neoformans* and *C. gattii* of 84.4% and 89.3%, respectively, using only 17 wave numbers. Then, fluorescence spectroscopy in excitation-emission matrix (EEM) was used, combined with multivariate classification methods (UPCA-LDA/QDA, UGA-LDA/QDA, USPA-LDA/QDA, PARAFAC/PLS-DA, nPLS-DA). The most satisfactory model was the UGA-LDA, which used only 5 wavelengths, and showed sensitivity of 88.9% in calibration and 100.0% prediction for both species, results that are comparable to routine biological tests. The last study aimed to differentiate sensitive and multi-resistant bacteria of the genera *Klebsiella sp*. and *Escherichia coli*. Through molecular fluorescence spectroscopy and multivariate classification, methods LDA, QDA and SVM coupled with data reduction algorithms PCA, GA and SPA. Among these, the models with the best performance for both types of bacteria presented sensitivity and specificity rates of 100%. Compared to the classical methods, the methodologies proposed in these studies proved to be an innovative, faster and cheaper alternative for the identification of pathogenic microorganisms, such as fungi and bacteria, opening the possibility of application in routine diagnostic laboratories.

Keywords: FT-IR, Fluorescence, Multivariate Analysis, *Cryptococcus*, *Klebsiella sp*. and *Escherichia coli*.

LISTA DE ABREVIATURAS E SIGLAS

AIDS - Síndrome da Imunodeficiência Humana Adquirida (do inglês, *acquired immunodeficiency syndrome*)

ATR - Reflectância Total Atenuada (do inglês, *Attenuated Total Reflectance*)

BHI - Infusão Cérebro e Coração (*Brain Heart Infusion*)

CGB - Canavanina-glicina-azul de Bromotimol

DA - Análise de Discriminante (do inglês, *Discriminant Analysis*)

DNA - Ácido Desoxirribonucleico (do inglês, *deoxyribonucleic acid*)

EEM - Matriz de Excitação e Emissão (do inglês, *Excitation-Emission Matrix*)

FTIR - Espectroscopia de Infravermelho por Transformada de Fourier (do inglês, *Fourier Transform Infrared*)

GA - Algoritmo Genético (do inglês, *Genetic Algorithm*)

HIV - Vírus da Imunodeficiência Humana (do inglês, *Human immunodeficiency vírus*)

KS - *Kennard-Stone*

LCR - Líquido Cefalorraquidiano

LDA - Análise Discriminante Linear (do inglês, *Linear Discriminant Analysis*)

MBC - Concentrações Bactericidas Mínimas (do inglês, *Minimum Bactericidal Concentration*)

MIC - Concentrações Inibitórias Mínimas (do inglês, *Minimum Inhibitory Concentration*)

NIR - Espectroscopia no Infravermelho Próximo (do inglês, *Near-infrared spectroscopy*)

nPLS - *N-Partial Least Squares*

NPV - Valor Preditivo Negativo (do inglês, *Negative Predictive Value*)

PARAFAC - Análise de Fatores Paralelos (do inglês, *Parallel Factor Analysis*)

PC - Componentes Principais (do inglês, *Principal Components*)

PCA - Análise de Componentes Principais (do inglês, *Principal Component Analysis*)

PLS - Regressão por mínimos quadrados parciais (do inglês, *Partial Least Squares)*

PPV - Valor Preditivo Positivo (do inglês, *Positive Predictive Value*)

QDA - Análise Discriminante Quadrática (do inglês, *Quadratic Discriminant Analysis*)

SPA - Algoritmo de Projeções Sucessivas (do inglês, *Successive Projections Algorithm*)

SVM - Máquinas de Vetores Suporte (do inglês, *Support Vector Machines*)

UFC - Unidades formadoras de colônias

SUMÁRIO

CAPÍTULO 1 - INTRODUÇÃO GERAL

1 Organização da tese

Esta tese foi organizada em capítulos, que correspondem aos trabalhos desenvolvidos como primeira autora, em parceria com pesquisadores do Instituto de Medicina Tropical/UFRN, Hospital Gizelda Trigueiro/Natal/RN, Departamento de Microbiologia e Parasitologia/UFRN. Nos apêndices são apresentados os diferentes trabalhos realizados em colaborações, desenvolvidos ao longo deste doutoramento.

CAPÍTULO 2 – *"Attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy as a new technology for discrimination between Cryptococcus neoformans and Cryptococcus gattii."* (Publicado no periódico Analytical Methods, DOI: 10.1039/C6AY01893A). Relata uma aplicação de espectroscopia no infravermelho médio e algoritmos de classificação multivariada (PCA-LDA/QDA, SPA-LDA/QDA e GA-LDA/QDA) na diferenciação dos fungos *C. neoformans e C. gattii.* Como uma alternativa mais rápida para auxiliar o diagnóstico.

CAPÍTULO 3 – *"Comparison of multivariate classification algorithms using EEM fluorescence data to distinguish Cryptococcus neoformans and Cryptococcus gattii pathogenic fungi."* (Publicado no periódico Analytical Methods, DOI: 10.1039/c7ay00781g). Relata uma aplicação de espectroscopia de fluorescência molecular e algoritmos de classificação multivariada de primeira e segunda ordem (UPCA-LDA/QDA, UGA-LDA/QDA, USPA-LDA/QDA, PARAFAC-LDA/QDA, UPLS-DA e nPLS) na otimização da diferenciação dos fungos *C. neoformans e C. gattii.* A técnica foi escolhida visando um aumento na sensibilidade e especificidade na classificação entre as duas espécies de *Cryptococcus.*

CAPÍTULO 4 – *"Identification of resistance in Escherichia coli and Klebsiella pneumoniae using E. E. M. fluorescence Spectroscopy and multivariate analysis."* (Manuscrito em fase de escrita). Relata uma aplicação de espectroscopia de fluorescência molecular e algoritmos de classificação multivariada (2D-LDA, 2D-PCA-LDA, 2D-PCA-QDA, 2D-PCA-SVM, UPCA-LDA/QDA, UGA-LDA/QDA, USPA-

LDA/QDA) na identificação de resistência e sensibilidade em duas espécies de bactérias *E. coli e K. pneumoniae*. Como um método alternativo para a detecção de resistência bacteriana, que possa ser utilizado nas análises de rotina laboratoriais e tornar o tratamento mais eficaz.

CAPÍTULO 5 – Conclusões e Perspectivas: neste capítulo são apresentados resumidamente os principais resultados alcançados nesses estudos. As principais contribuições de cada metodologia proposta e as perspectivas para futuros trabalhos.

APÊNDICE A – *"Variable selection with a support vector machine for Discriminating Cryptococcus fungal species based on ATR-FTIR spectroscopy."* (Publicado no periódico Analytical Methods, DOI: 10.1039/c7ay00428a). Relata a aplicação de diferentes tipos de algoritmos de redução de dados e variáveis ao SVM com diferentes funções kernel na distinção de *C. gattii* e *C. neoformans* baseados na espectroscopia ATR-FTIR.

APÊNDICE B – *"On the synergy between silver nanoparticles and doxycycline towards the inhibition of Staphylococcus aureus growth."* (Publicado no periódico RSC Advances, DOI: 10.1039/c8ra02176g). Relata o estudo de discriminação das respostas metabólicas de bactérias *S. aureus* tratadas com NanoAg, doxiciclina e com o combinado NanoAg/doxiciclina com a finalidade de confirmar o efeito sinérgico da ação combinada entre a doxiciclina e as nanopartículas de prata.

APÊNDICE C – *"The Use of Near Infrared Spectroscopy and Multivariate Calibration for Determining the Active Principle of Olanzapine in a Pharmaceutical Formulation."* (Publicado no periódico Journal of the Brazilian Chemical Society, DOI: 10.21577/0103-5053.20160233). Relata a determinação quantitativa de olanzapina em formulação farmacêutica utilizando a espectroscopia de infravermelho próximo (NIR) combinada com regressão por mínimos quadrados parciais (PLS).

APÊNDICE D – *"A Multivariate Control Chart Approach for Calibration Transfer between NIR Spectrometers for Simultaneous Determination of Rifampicin and Isoniazid in Pharmaceutical Formulation."* (Publicado no periódico Current Analytical Chemistry, DOI: 10.2174/1573411014666171212141909). Relata uma abordagem de transferência de cartas de controle multivariadas entre dois espectrômetros de infravermelho próximo (NIR) usando Padronização Direta, para a determinação de teor de rifampicina e isoniazida em formulações farmacêuticas.

2 Introdução

Mesmo com todo o avanço na medicina, as doenças infecciosas ainda são uma grande ameaça, inclusive aos países desenvolvidos [1]. Estas doenças podem ser transmitidas ao homem através de vírus, protozoários, bactérias, fungos, dentre outros e podem se espalhar, direta ou indiretamente, de uma pessoa para outra. Nos últimos anos, tem sido observado em vários países, que os grandes centros urbanos vêm apresentando um aumento nos números de casos de enfermidades infectocontagiosas [2], o que torna este tema de grande relevância social.

A maioria dos patógenos que causam essas doenças possuem caráter oportunista, tendendo a acometer com frequência, pacientes recém transplantados e portadores de HIV (Vírus da Imunodeficiência Humana) [3,4], que constituem uma grande parcela da população. Segundo os dados apontados pelo último boletim apresentado pela UNAIDS (Organização das Nações Unidas para AIDS), em 2017 cerca de 21,7 milhões de pessoas possuíam HIV (UNAIDS, 2018) [5]. Nestes pacientes, a manifestação de doenças infecciosas agrava o caso clínico e compromete o tratamento.

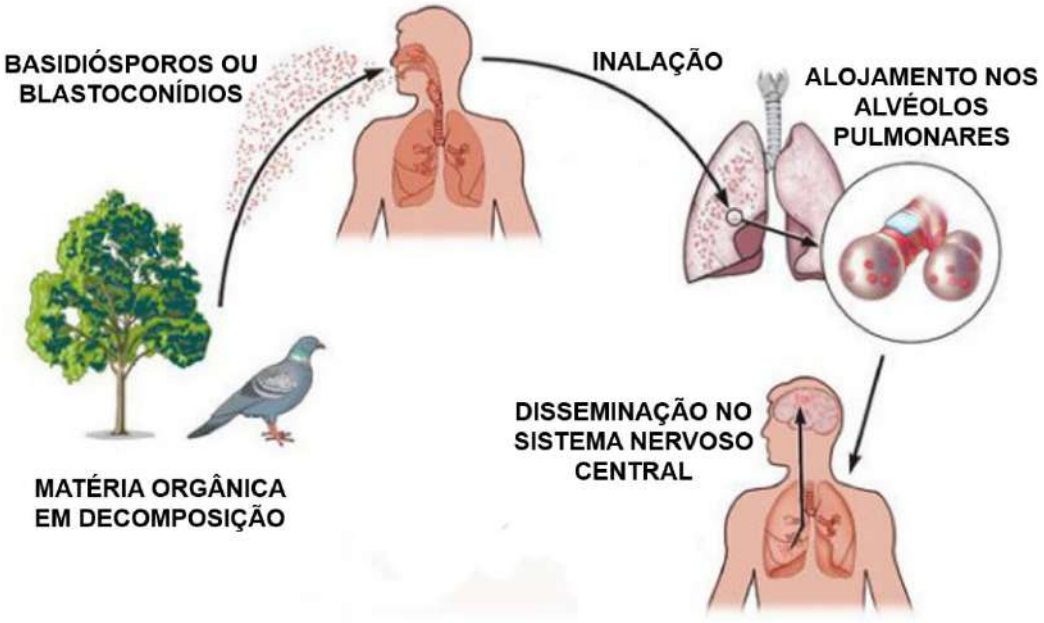
Dentre as doenças infecciosas mais comuns em pacientes portadores de HIV, temos a criptococose cujos principais agentes causadores são o *Cryptococcus neoformans* e *Cryptococcus gatti* [6]. São fungos que apresentam diferenças com relação ao habitat natural, à epidemiologia, às características fenotípicas, às manifestações clínicas e à resposta a terapia antifúngica [7]. Conforme a Tabela 1, na qual estão listadas as principais diferenças entre estas duas espécies de fungos, *C. neoformans* é um patógeno oportunista que geralmente acomete indivíduos imunocomprometidos, enquanto *C.gatti* está ganhando destaque como uma das principais causas de doenças em pacientes imunocompetentes [8].

Tabela 1: Principais diferenças entre *Cryptococcus neoformans* e *Cryptococcus gatti*

Cryptococcus gatti	***Cryptococcus neoformans***
Encontrado em meio campestre	Encontrado em grandes centros urbanos
Leveduras presentes em árvores em decomposição, nas regiões subtropicais e subtropicais	Leveduras presentes nas fezes dos pombos e outras aves
Acomete indivíduos imunocompetentes	Acomete principalmente indivíduos imunossuprimidos

A criptococose é uma micose que se tornou muito comum a partir dos anos 80, em virtude do surgimento da AIDS [9]. A forma mais comum é a criptococose pulmonar, que se não controlada, pode levar a meningite criptocócica potencialmente fatal ou meningoencefalite [10]. A infecção se dá pela inalação dos basidiósporos ou blastoconídios, advindos, em geral, de matéria orgânica em decomposição, devido a sua natureza saprofítica. Quando inalados, os propágulos instalam-se nos alvéolos pulmonares, assumindo então sua forma leveduriforme (Fig. 1) [11,12].

Figura 1: Representação do mecanismo de transmissão da criptococose.



Fonte: Adaptada da Referência [12].

Atualmente, o diagnóstico da criptococose é feito por meio da pesquisa direta, na qual o líquor ou liquido céfalorraquidiano (LCR) é analisado ao microscópio ótico, em solução de tinta da china e por meio do cultivo em ágar Sabouraud. A diferenciação entre estas espécies pode ser feita utilizando o teste de CGB (canavanina-glicina-azul de bromotimol). Porém, esse teste demanda muito tempo, além de não apresentar confiabilidade de 100%, e por isso não é empregado de forma rotineira no diagnóstico da criptococose [13].

Das bactérias que geralmente estão associadas ao cenário clínico, destacam-se a *Klebsiella pneumoniae* (*K. pneumoniae*) e a *Escherichia coli (E. Coli),* que são comumente fontes de infecções comunitárias e hospitalares. Ambas são bactérias gram-negativas, que fazem parte da família das Entereobactérias [14, 15]. *K.*

pneumoniae é a segunda causa mais comum de bacteremia gram-negativa, causa infecções oportunistas, como pneumonia, sepse e inflamação do trato urinário [16]. A *Escherichia coli*, que apesar de não ser tipicamente patogênica para humanos, pode causar várias doenças no trato gastrointestinal, no sistema renal e no sistema nervoso central [17].

O tratamento das doenças infecciosas causadas por bactérias enfrenta além da dificuldade de identificação do agente etiológico, um agravante que é a resistência bacteriana, que pode acontecer de forma natural ou adquirida. A resistência adquirida é a mais preocupante, uma vez que já foram descritas em praticamente todas as espécies de bactérias. Essa resistência está associada a capacidade que a bactéria tem de modificar sua estrutura celular e induzir a produção de substâncias capazes de neutralizar a ação de antibacterianos [18].

Nas enterobactérias, a principal resistência são aos carbapenêmicos [19]. Nas *K. pneumoniae* e *E. coli*, a susceptibilidade reduzida aos carbapenêmicos pode surgir por uma diminuição da permeabilidade da membrana externa devido à inativação ou expressão alterada de porinas em cepas que produzem β-lactamases com pelo menos alguma atividade hidrolítica contra os carbapenêmicos [20, 21].

A avaliação da resistência bacteriana é realizada através de testes nos quais uma cultura isolada é submetida a vários tipos de antibióticos. O perfil de sensibilidade bacteriana a antibióticos das cepas isoladas pode ser determinado pelo método de difusão em disco [22], Concentrações Inibitórias Mínimas (MIC) [23] ou Concentrações Bactericidas Mínimas (MBC) [24]. E o tempo de análise, em geral, leva de 24 a 48h.

Metodologias que sejam capazes de fornecer a identificação de fungos e avaliar a resistência bacteriana de forma rápida, precisa e com a confiabilidade dos métodos de referência, são cada vez mais necessários. Neste sentido, as técnicas espectroscópicas, como a fluorescência molecular e a do infravermelho médio, têm tido grande destaque nos últimos anos na área microbiológica, uma vez que, são técnicas rápidas, de custo relativamente baixo, não-destrutivas, que necessitam de pouca ou nenhuma manipulação da amostra, além da não utilizarem reagentes e/ou solventes. Se baseiam na interação da radiação com os grupos químicos que estão presentes nas amostras, através da medida da quantidade de radiação produzida ou absorvida pelas moléculas ou espécies atômicas de interesse [25].

Apesar das grandes vantagens das técnicas espectroscópicas, a seletividade pode ser prejudicada devido à grande sobreposição espectral ou presença de interferência da matriz. Para superar esses obstáculos, ferramentas quimiométricas podem ser utilizadas, para maximizar a extração de informações relevantes [26]. Os

métodos de reconhecimento de padrões supervisionados e não supervisionados são comumente aplicados para extrair características espectrais e desenvolver modelos de classificação [27]. Entre os métodos de monitoramento supervisionado, podemos destacar a Análise Discriminante Linear (LDA) [28], Análise Discriminante Quadrática (QDA) [29], acoplados a métodos de redução de dados, como a Mínimos Quadrados Parciais [30], Análise de Componentes Principais (PCA) [31], Análise de Fatores Paralelos (PARAFAC) [32], Algoritmo Genético (GA) [33] e o Algoritmo de Projeções Sucessivas (SPA) [34].

A detecção dos patógenos causadores dessas doenças é fundamental para garantir uma intervenção correta e eficaz ao tratamento do paciente, quanto mais rápido e de forma segura for feita a identificação, mais chances de sucesso tem o tratamento da enfermidade. Por esse motivo, há um grande interesse em metodologias que melhorem os métodos de detecção atuais, particularmente em questões relacionadas com a rapidez no diagnóstico.

2.1 Técnicas Instrumentais

2.1.1 Espectroscopia no infravermelho Médio

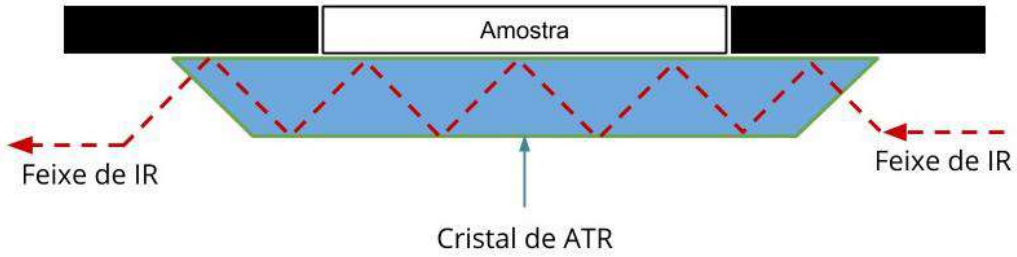
A espectroscopia no Infravermelho Médio é um método de análise largamente utilizada em muitas áreas da ciência e tecnologia [35]. A técnica consiste em incidir sobre uma amostra um feixe de radiação na faixa de 400 a 4000 cm^{-1}, de modo que as vibrações permitidas que provoquem alteração no dipolo da molécula, promoverão uma absorção de energia incidente em frequências específicas, gerando um espectro de infravermelho [25].

Os espectrômetros mais modernos, em sua grande maioria, são acoplados a um siatema óptico que associados à cálculos matemáticos, conhecidos como Transformada de Fourier, convertem o sinal analítico no domínio do tempo para um sinal no domínio das frequências [36]. A vantagem de usar a Transformada de Fourier com o infravermelho é que é possível obter dezenas de interferogramas para uma mesma amostra em menos de um segundo, e através do cálculo da média gerar um espectro com alta razão sinal/ruído [37].

A aplicação de acessórios de Reflectância Total Atenuada (ATR) no instrumento convencional de espectroscopia FTIR permite obter espectros de forma estável, robusta, não destrutiva e com mínima ou nenhuma preparação de amostras, uma vez que estas podem ser posicionadas diretamente sobre o cristal de ATR [38]. A

reflectância total é um caso especial de reflexão de uma onda eletromagnética em uma interface entre dois meios, como mostrado na Figura 2.

Figura 2: Processo de reflectância total atenuada



Fonte: Autor

Na espectroscopia ATR-FTIR, o raio infravermelho entra no cristal ATR a 45° em relação à superfície do cristal e é totalmente refletido na interface de cristal para amostra. A profundidade de penetração depende do comprimento de onda, dos índices de refração do cristal ATR e da amostra, bem como do ângulo da luz incidente. A onda evanescente é uma fração da luz que atingiu a amostra, que é atenuada nas regiões espectrais, onde a amostra absorve energia. Após um ou vários reflexos internos, o feixe IR sai do cristal ATR e é direcionado para o detector IR [39].

A espectroscopia FTIR é adequada para ser empregada em estudos biológicos, porque possui a capacidade de diferenciar substâncias orgânicas complexas, sendo utilizada para tanto para a classificação, quanto para quantificação [40]. No espectro de infravermelho médio, existe uma região chamada de impressão digital bioquímica [41], na qual as bandas espectrais podem ser correlacionadas com a presença/ausência de características estruturais das amostras estudadas [42]. Algumas estruturas presentes nas células podem ser caracterizadas por algumas absorções, correspondentes a ligações químicas que apresentam regiões de absorção e bandas espectrais específicas atribuídas, como por exemplo, lipídios (≈1750 cm^{-1}), carboidratos (≈1155 cm^{-1}), estrutura secundárias de proteínas (amida primária ≈1650 cm^{-1}; amida secundária ≈1550 cm^{-1}) e DNA/RNA (≈1225 cm^{-1}; ≈1080 cm^{-1}) [43,44].

Os dados fornecidos por técnicas como a espectroscopia FTIR carregam muitas informações. Diante desta necessidade de processar a grande quantidade de informações produzidas pelos instrumentos analíticos mais modernos, surgiu uma área da ciência, que faz uso de ferramentas estatísticas e matemáticas de análise multivariada que possibilita relacionar medidas de absorbância ou reflectância, por exemplo, com o estado ou propriedades físicas e químicas do sistema analisado [45]. A

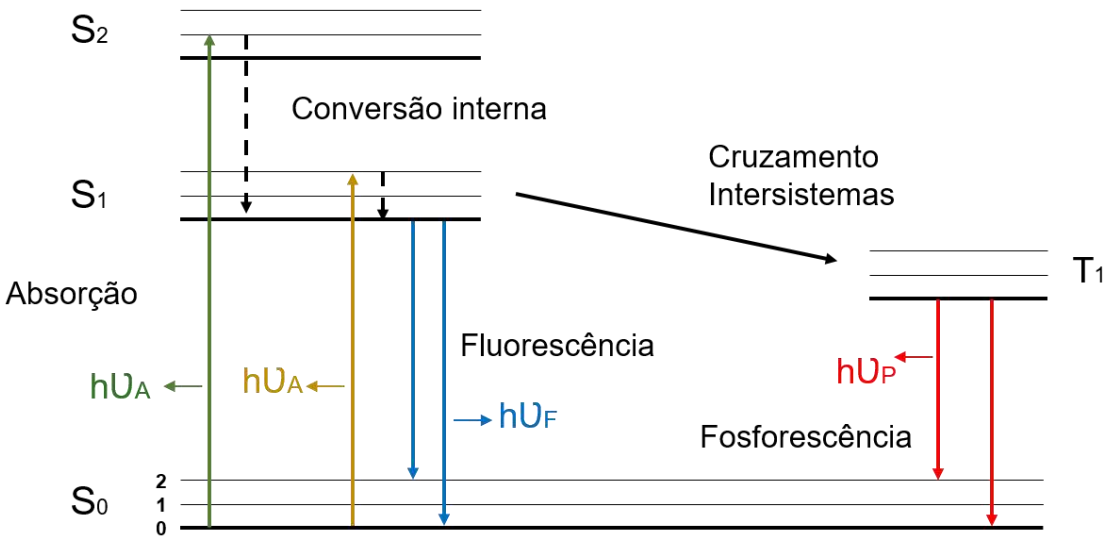
espectroscopia FTIR aliada a análise multivariada foi empregada com sucesso no diagnóstico da infecção pelo HPV [46], rastreio de câncer [47,48], identificação de fungos [49] e bactérias [50].

2.1.2 Espectroscopia de Fluorescência Molecular

A fotoluminescência é um fenômeno em que as moléculas ou átomos absorvem radiação eletromagnética, passando para um estado excitado e que ao retornarem ao estado fundamental, liberam o excesso de energia na forma de fótons. A fotoluminescência divide-se em fluorescência e fosforescência, dependendo da natureza do estado excitado [51].

Na fluorescência, o elétron envolvido na transição mantém sua orientação de spin no orbital excitado, se emparelhando com o elétron que está no orbital do estado fundamental, caracterizando o estado excitado singleto. A fosforescência ocorre com o estado excitado tripleto, quando o elétron no estado excitado, inverte a sua orientação de spin [52]. A Figura 3 apresenta o diagrama de Jablonsky que ilustra os processos de absorção e emissão de luz.

Figura 3: Diagrama de Jablonski ilustrando resumidamente o fenômeno da luminescência.



Fonte: Adaptada da Referência [51]

Em estados excitados podem ocorrer vários processos moleculares de desativação. Através de colisões, as moléculas podem rapidamente perder energia e caírem para níveis vibracionais menos energéticos. No processo de conversão interna, moléculas que ocupam um estado excitado maior de S_2 e S_1 relaxam até o nível

vibracional mais baixo de S_1. As transições que ocorrem de S_1 para S_0, constituem a fluorescência. O espectro de fluorescência é localizado em comprimentos de onda maiores (energia mais baixa) que o espectro de absorção, por causa da perda de energia no estado excitado devido à relaxação vibracional [51].

Para que a fluorescência ocorra é necessário que as moléculas tenham estruturas apropriadas e estar em um meio que favoreça a desativação radiativa. São potencialmente fluorescentes as moléculas com estruturas relativamente rígidas e ricas em elétrons π, como as moléculas aromáticas, contendo ou não heteroátomos em sua cadeia principal [53].

Entre as técnicas espectroscópicas, a fluorescência molecular tem grande destaque, pois é uma técnica analítica altamente sensível que permite medições em concentrações ambientais naturais, além de não ser destrutiva. Juntamente com isso, a fluorescência em matriz excitação-emissão (EEM, do inglês, *Excitation-Emission Matrix*) contém uma enorme quantidade de informações do analito, gerando um sinal abrangente contendo informações sobre todos os compostos fluorescentes dentro da amostra [54, 55]. Os componentes de sobreposição de fluorescência podem ser identificados de acordo com seus comprimentos de onda de excitação e emissão, que são capazes de identificar biomoléculas em amostras [56].

Apesar das grandes vantagens da espectroscopia de fluorescência EEM, a seletividade pode ser prejudicada devido à grande sobreposição espectral ou presença de interferência da matriz [57]. Neste sentido, as técnicas de análise de dados multivariadas podem decompor de forma confiável EEMs em componentes fluorescentes com variação independente, permitindo uma identificação mais precisa das substâncias [58], devido a possibilidade de classificar, sob certas condições, na presença de interferentes desconhecidos, que não estejam presentes no conjunto de treinamento, a chamada "vantagem de segunda ordem” [59, 60]. A combinação de análise multivariada e espectroscopia EEM foi bem-sucedida no estudo de hibridização de DNA [61], quantificação de colesterol [62] e rastreamento de câncer [63].

2.2 Análise Quimiométrica

2.2.1 Análise de Componentes Principais (PCA)

O método de compressão e extração de dados mais conhecido e utilizado é o PCA (do inglês, *Principal Component Analysis*) [31]. O PCA encontra uma combinação linear de variáveis, que descreve a maior variância dos dados. A matriz de dados original é decomposta em *scores*, matriz que contém a informação de como cada amostra se

correlaciona com as outras e nos *loadings*, que é a matriz que informa como cada variável se correlaciona com as demais. O PCA pode ser descrito com um número k de componentes e a variância residual, não contemplada pelo modelo está relacionada ao erro [64], conforme a equação 1:

$$\boldsymbol{X} = \boldsymbol{TP^t} + \boldsymbol{E} \qquad \text{Equação 1}$$

Onde $\boldsymbol{X}$ é a matriz espectral, $\boldsymbol{T}$ e $\boldsymbol{P}$ são as matrizes dos *scores* e *loadings*, respectivamente, e $\boldsymbol{E}$ é a matriz dos resíduos.

Os dados originais são resolvidos em componentes ortogonais, cuja combinação linear aproxima-se dos dados originais. O novo conjunto de dados, os autovetores, chamados de componentes principais (PC, do inglês, *Principal Components*), correspondem aos maiores autovalores da matriz de covariância, assim, representando a maior variação possível no conjunto de dados. A primeira PC representa variação máxima entre todas as combinações lineares e as demais PCs o máximo de variabilidade restante possível [65].

O PCA pode ser utilizado como uma ferramenta de análise exploratória [66] ou como redutor de dimensionalidade [67] para ser acoplado à algoritmos de classificação em métodos supervisionados.

2.2.2 Análise de Fatores Paralelos (PARAFAC)

O algoritmo de análise de fatores paralelos (PARAFAC, do inglês, *Parallel Factor Analysis*) é considerada uma generalização do PCA bilinear [68], que se aplica em sistemas, em que cada amostra é constituída por uma matriz de dados. O PARAFAC é usado para decompor dados trilineares com uma única solução, permitindo estimativas robustas de perfis de excitação e emissão presentes nos espectros e suas concentrações [69], a chamada “vantagem de segunda ordem” [70].

O PARAFAC pode ser utilizado em dados multidimensionais como os gerados pela fluorescência EEM. Cada amostra EEM constitui uma matriz de dimensão $\boldsymbol{J}$ x $\boldsymbol{K}$, onde $\boldsymbol{J}$ são os comprimentos de onda de emissão e $\boldsymbol{K}$ os comprimentos de onda de excitação. Ao agrupar $\boldsymbol{I}$ matrizes de amostras forma-se um tensor de dados $\boldsymbol{X}$, com as dimensões ($\boldsymbol{I}$ x $\boldsymbol{J}$ x $\boldsymbol{K}$) [71]. Através do PARAFAC, a decomposição de $\boldsymbol{X}$ é obtida minimizando a soma dos quadrados dos resíduos $\boldsymbol{e_{ijk}}$, conforme a Equação 2:

$$\boldsymbol{X_{ijk}} = \sum_{f=1}^{F} \boldsymbol{a_{if}} \boldsymbol{b_{jf}} \boldsymbol{c_{kf}} + \boldsymbol{e_{ijk}}$$ Equação 2

Em que $\boldsymbol{F}$ é o número de fatores, $\boldsymbol{e_{ijk}}$ é o erro residual que tem as mesmas dimensões da matriz $\boldsymbol{X}$, os vetores colunas $\boldsymbol{a_{if}}$, $\boldsymbol{b_{jf}}$ e $\boldsymbol{c_{kf}}$ estão contidos em matrizes de *scores* e *loadings* A, B e C, respectivamente [72].

Os dados decompostos provenientes do PARAFAC podem ser usados para construir os modelos de classificação, utilizando algum algoritmo de análise discriminante. Além disso, a seleção de recursos ou uma seleção reduzida pelo uso do pré-processamento em um conjunto de variáveis latentes podem melhorar a classificação [73].

2.2.3 Seleção de variáveis: Algoritmo de Projeções Sucessivas (SPA) e Algoritmo Genético (GA)

As técnicas instrumentais mais modernas possibilitam a aquisição de muitos dados por amostra, o que torna o tratamento desses dados muito complexo. Além de nem todas as variáveis serem relevantes para a construção dos modelos. Neste sentido, métodos de redução de dados que encontrem subconjuntos de variáveis mais relevantes, podem incrementar o desempenho dos modelos [74].

O algoritmo das Projeções Sucessivas (SPA, do inglês, *Successive Projections Algorithm*) é um método de seleção de variáveis, inicialmente desenvolvido para aplicação em métodos de regressão [34], tendo sido posteriormente ampliado para aplicação também em modelos de classificação [75]. O SPA seleciona as variáveis menos colineares, realizando uma série de projeções para redução do espaço de busca e selecionando aquelas que possuem o valor de projeção máxima, dentre todas as variáveis do espaço sub-ortogonal das variáveis previamente selecionadas [76].

Em comparação com outros algoritmos, o SPA apresenta vantagens em termos de simplicidade e facilidade de interpretação dos dados. Entretanto, apresenta algumas limitações, uma vez que é afetado pela baixa razão sinal/ruído e nesses casos, acabar selecionando amostras que não são relevantes [77].

O Algoritmo Genético (GA, do inglês, *Genetic Algorithm*) é uma técnica heurística popular de otimização probabilística que emprega o processo de pesquisa não-local, inspirado na teoria da seleção natural de Darwin [33], visando encontrar as variáveis que se encaixam melhor nas equações que serão passadas para o modelo da próxima geração [78].

Incialmente, o GA seleciona uma população aleatória de indivíduos, que terão suas aptidões avaliadas. Em seguida, são selecionados pares de indivíduos a serem cruzados, posteriormente submetidos a uma mutação. Os indivíduos são substituídos pela nova geração, em um ciclo que é repetido até que se encontre um conjunto de variáveis que promovam a melhor classificação dos dados, com uma função de menor custo [79].

As principais vantagens do GA é que ele independe da complexidade da estrutura do problema, e não se restringe a uma solução ótima local. Podendo ser utilizado tanto em modelos de regressão quanto de classificação [80]. Devido à natureza aleatória do GA, múltiplas execuções do algoritmo geram diferentes resultados afetando a reprodutibilidade. Além disso, em problemas com o espaço de busca suave, ou seja, com pouca correlação entre as variáveis, o GA pode não encontrar os melhores ótimos locais [81].

2.2.4 Análise discriminante

A análise discriminante é um método supervisionado em que funções de variáveis observadas são usadas para classificar as observações em grupos designados. A análise discriminante linear (LDA, do inglês, *Linear Discriminant Analysis*) é o método mais comumente utilizado [82]. No LDA, as densidades de probabilidade de classe condicional são consideradas como distribuições multivariadas normais, com diferentes vetores médios para cada classe, em que as matrizes de covariância são idênticas para todas os grupos [83].

A função limite para o LDA é mostrada na Equação 3 onde $\boldsymbol{k}$ é uma constante; $\boldsymbol{l_1}$ e $\boldsymbol{l_2}$ são os coeficientes lineares entre as observações ($\boldsymbol{x}$) das classes 1 e 2, respectivamente.

$$\boldsymbol{f_{LDA}(x_1, x_2) = k + l_1 x_1 + l_2 x_2} \qquad \text{Equação 3}$$

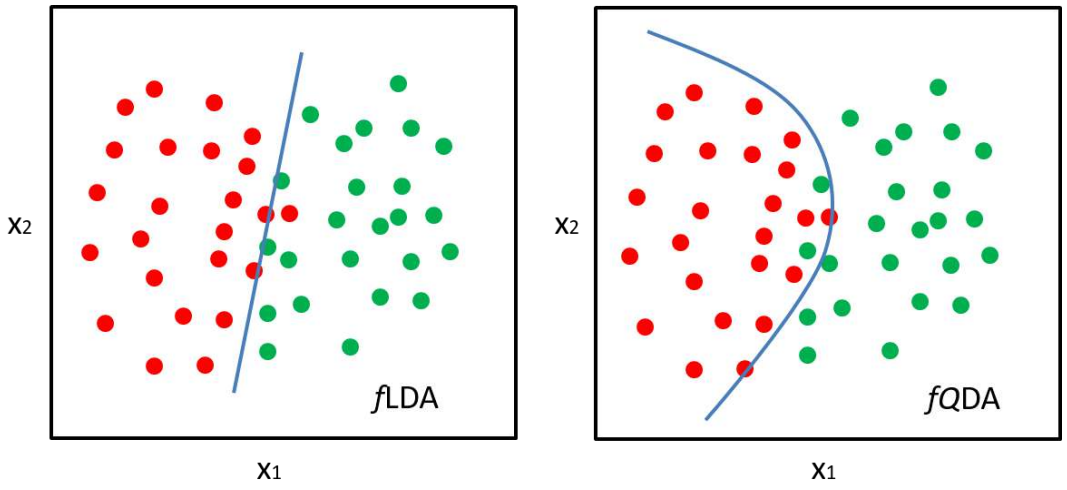
Nos casos em que as classes possuem covariância diferente é recomendado a análise discriminante quadrática (QDA, do inglês, *Quadratic Discriminant Analysis*) [69]. A classificação do QDA, difere-se do LDA, porque leva em consideração também os determinantes das matrizes de dispersão [29]. O QDA discrimina grupos que possuem matrizes de covariância de classes específicas significativamente diferentes e forma um modelo de variância separado para cada classe, enquanto as populações de classes representam distribuições normais multivariadas com a mesma média [84].

A função limite para o QDA é mostrada na Equação 4 onde $\boldsymbol{k}$ é uma constante; $\boldsymbol{l_1}$ e $\boldsymbol{l_2}$ são os coeficientes lineares entre as observações ($\boldsymbol{x}$) das classes 1 e 2, respectivamente; $\boldsymbol{q_1}$ e $\boldsymbol{q_2}$ são os coeficientes quadráticos entre as observações da classe 1 e 2, respectivamente.

$$f_{QDA}(x_1, x_2) = k + l_1 x_1 + l_2 x_2 + q_1 x_1^2 + (q_1 + q_2) x_1 x_2 + q_2 x_2^2 \qquad \text{Equação 4}$$

A Figura 4 ilustra as funções limite LDA e QDA para discriminação de dois conjuntos de dados, utilizando as Equações 3 e 4.

Figura 4. Exemplo de funções limite LDA e QDA para discriminação de duas classes: (●) classe 1 e (●) classe 2.



Fonte: Autor

2.2.5 Máquinas de Vetores Suporte (SVM)

O algoritmo Máquinas de Vetores Suporte (SVM, do inglês, *Support Vector Machines*) é uma técnica de aprendizado de máquina, que pode ser utilizada para a construção de modelos de classificação [85]. O princípio básico do SVM consiste em encontrar um hiperplano ótimo de separação intermediário entre duas classes de dados, seguindo o princípio da margem máxima [86].

Entre os métodos existentes de classificação, O SVM fornece vantagens importantes, como generalização adequada para novas amostras, ausência de mínimos locais e uma representação que depende apenas de alguns parâmetros [87,88]. O SVM foi originalmente criado para lidar com classificações binárias [89], entretanto a maior

parte dos problemas reais requerem múltiplas classes. Para se utilizar um SVM em classificações de múltiplas classes, é necessário transformar o problema multiclasse em vários problemas de classes binárias. Por isso, o tempo de treinamento pode ser bem longo, dependendo do número de amostras e da dimensionalidade dos dados [90].

2.2.6 Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA)

Uma análise discriminante por mínimos quadrados parciais (PLS-DA, do inglês, *Partial Least Squares Discriminant Analysis*) é uma variação do método de regressão por mínimos quadrados parciais (PLSR), para utilização em modelos de classificação [91]. O PLS-DA se baseia na redução do tamanho dos dados originais (X) e substituí-los pela matriz de *scores* e *loadings*, maximizando a covariância entre a matriz espectral X e a matriz de resposta das classes (Y). O nível de redução é descrito pelo número de variáveis latentes significativas (VL) [92].

O PLS-DA apresenta algumas vantagens frente a outros algoritmos de redução de dimensionalidade, como por exemplo o PCA, que tenta descrever a variação máxima com o menor número possível de componentes, sem necessariamente levar à separação das duas classes entre si. Por ser um método supervisionado, o PLS-DA encontra a separação máxima entre cada classe mais efetivamente, uma vez que se baseia na designação de classes, o que permite encontrar padrões nos dados [93, 94].

3 Objetivos

3. 1 Objetivo geral

Avaliação da Espectroscopia no infravermelho médio e de fluorescência molecular em conjunto com métodos de classificação multivariada para rápida identificação de fungos e resistência de bactérias causadoras de doenças infecciosas.

3.2 Objetivos específicos

- Diferenciação dos fungos *Cryptococcus neoformans* e *Cryptococcus gattii* através de modelos de classificação multivariada, espectroscopia ATR-FTIR e fluorescência molecular.
- Comparação do desempenho de métodos de classificação de primeira e segunda ordem.
- Identificação de resistência em *Escherichia coli* e *Klebsiella pneumoniae*, por meio de espectroscopia de fluorescência molecular e classificação multivariada.

- Validação das metodologias propostas por meio do cálculo de figuras de mérito de classificação tais como sensibilidade, especificidade, PPV, NPV e index You.

4 Metodologia

4.1 Preparo das amostras

4.1.1 Fungos

Foram utilizadas culturas fúngicas *C. gatti* e *C. neoformans* da coleção particular do Laboratório de Micologia Médica e Ambiental (LAMEA-UFRN), do Hospital das Clínicas e Hospital Veterinário da UNESP, campus de Botucatu (SP), isolado de referência Fio Cruz e isolados recentes do Hospital Giselda Trigueiro (Natal – RN). Estes fungos foram cultivados em meio de cultura Ágar Sabouraud e incubados durante 48 horas a uma temperatura de 37°C, até o crescimento.

Depois do crescimento das amostras foi feito um repique nas placas de Ágar, para manter a viabilidade do fungo. Em 28 tubos de ensaio foram adiconados 1,0 mL de tampão fosfato (1 mol/L), que foram colocados em um autoclave para esterilização. Foi feita a transferência de parte do isolado original para estes tubos de ensaios. Em seguida, os tubos de ensaio foram armazenados na estufa a uma temperatura de aproximadamente 37°C durante quatro dias para o crescimento do fungo.

Após o crescimento, as células de levedura foram inativadas por meio de uma solução de paraformaldeído, para promover a biossegurança na manipulação. Depois lavados com solução salina e armazenados numa geladeira até a coleta dos espectros.

4.1.2 Bactérias

As amostras usadas foram: *E. coli* ATCC 25922 - Cepa padrão, *E. coli* CCHB NDM+, *E. coli* CCHB ampC 7018, *K. pneumoniae* ATCC 1003, *K pneumoniae* CCBH 4955 KPC e *K pneumoniae* CCBH 6633 resistente a carbapenêmicos. As cepas CCBH foram obtidas do Laboratório de Pesquisa em Infecção Hospitalar (LAPIH - Fiocruz/RJ). As cepas ATCC pertencem ao LABMIC/DMP – UFRN. Inicialmente as amostras puras foram repicadas no meio de cultura BHI, e então foram mantidas na estufa por 24 horas a 38°C, para que a bactérias se multiplicassem. A amostra foi então repicada em uma placa de petri, que também foi mantida na estufa por 24 horas. Finalmente, a massa bacteriana correspondente a aproximadamente 10^6 unidades de formação de colônia

(UFC) foi transferida de um meio de cultura para um tubo Falcon com 2 mL de solução tampão de fosfato (1mol/L).

4.2 Aquisição de dados

4.2.1 Espectroscopia ATR-FTIR

Espectros ATR-FTIR das amostras de *C. gatti* e *C. neoformans* foram medidos usando um espectrofotômetro modelo Bruker FT-IR, VERTEX 70 de laboratório LAMMEN. O instrumento foi configurado para executar um total de 16 varreduras, resolução de 4 cm^{-1}, na faixa de 400-4000 cm^{-1}, no modo de absorbância. Foram medidas 10 réplicas de cada amostra, para garantir a cobertura de toda a variabilidade amostral. Uma gota de 50 µL foi colocada diretamente no detector, e sobre a amostra, um pedaço de papel alumínio com a porção fluorescente virada para baixo, a fim de minimizar a propagação de radiação. O cristal ATR foi limpo com álcool a 70% v/v e um novo background foi coletado antes da leitura de uma nova amostra.

4.2.2 Espectroscopia de Fluorescência Molecular

4.2.2.1 Fungos

Os dados de fluorescência de excitação/emissão para as espécies de *C. gatti* e *C. neoformans* foram adquiridos na faixa de comprimento de onda de 220-320 nm para excitação e 250-900 nm para emissão, com passos de 10 e 1 nm para excitação e emissão, respectivamente. Foi utilizado um espectrofluorímetro RF-5301 Shimadzu com uma cubeta de quartzo de 0,5 mm. As larguras das fendas de excitação e emissão de monocromadores foram fixadas a 5 nm, a velocidade de varredura foi ajustada para o modo super (3000nm/min), o tubo fotomultiplicador foi ajustado para o nível médio e uma célula com uma sonda de reflectância de fibra óptica foi utilizada. Um total de 500µl de solução salina com células fúngicas foi adicionado à cubeta de fluorescência para leitura. Após cada leitura a cubeta foi lavada com uma solução de álcool a 70% e em seguida com água destilada para evitar contaminação entre amostras de fungos. A temperatura foi mantida a 25ºC (+/- 2ºC) ao longo dos experimentos.

4.2.2.2 Bactérias

Os dados de fluorescência de excitação/emissão para os dados de *E. coli* e *K. pneumoniae* foram adquiridos na faixa de comprimento de onda de 220-310 nm para excitação e 270-900 nm para emissão, com passos de 10 e 1 nm para excitação e emissão, respectivamente. Foi utilizado um espectrofluorímetro RF-5301 Shimadzu com uma cubeta de quartzo de 0,5 mm. As larguras das fendas de excitação e emissão de monocromadores foram configuradas para 3 e 5 nm, respectivamente. A velocidade de varredura foi ajustada para o modo super (3000nm/min), o tubo fotomultiplicador foi ajustado para o nível médio e uma célula com uma sonda de reflectância de fibra óptica foi utilizada. Foram coletadas cinco réplicas das concentrações de $1x10^6$ UFC/ml, $5x10^5$ UFC/ml, $1,3x10^5$ UFC/ml, $6,3x10^4$ UFC/ml e $3,1x10^4$ UFC/ml.

4.3 Análise Computacional

Os pré-processamentos espectrais e modelos de classificação multivariada foram feitos utilizando o software MATLAB R2011a (The Math-Works, Natick, EUA), e o pacote PLS-toolbox (Eigenvector Research, Inc., Wenatchee, WA, EUA). Os pré-processamentos espectrais realizados nos dados de espectroscopia FTIR foram: um corte na região 900-1800 cm^{-1} e normalização pelo pico da amida I (≈ 1650 cm^{-1}). Para os dados de fluorescência molecular, foi feita a remoção de dispersões Rayleigh e Raman usando o algoritmo 'EEMscat' e cortes nas matrizes de emissão.

Para a construção de modelos de classificação, as amostras foram divididas em conjuntos de calibração, validação e previsão usando o algoritmo de seleção de amostras Kennard-Stone (KS).

Os modelos de classificação de padrões utilizados foram: Análise de Discriminante Linear, Análise de Discriminante Quadrática e Máquinas de Vetores Suporte, acoplados a algoritmos de redução de dimensionalidade, como a Análise de Componentes Principais e dos algoritmos de seleção de variáveis como o Algoritmo Genético (GA) e Algoritmo das Projeções Sucessivas (SPA). Além disso, investigou-se os modelos PARAFAC, PLS-DA e nPLS-DA.

Por fim, para a validação do método analítico foram calculadas alguns parâmetros de desempenho analítico, tais como: sensibilidade, especificidade, VPP, VPN, LR (+), LR (-).

REFERÊNCIAS

[1] DE SOUSA MARQUES, Aline et al. Feature selection strategies for identification of Staphylococcus aureus recovered in blood cultures using FT-IR spectroscopy successive projections algorithm for variable selection: a case study. **Journal of microbiological methods**, v. 98, p. 26-30, 2014.

[2] NEIDERUD, Carl-Johan. How urbanization affects the epidemiology of emerging infectious diseases. **Infection ecology & epidemiology**, v. 5, n. 1, p. 27060, 2015.

[3] MAESTRALE, Caterina et al. Genetic and pathological characteristics of *Cryptococcus gattii* and *Cryptococcus neoformans var. neoformans* from meningoencephalitis in autochthonous goats and mouflons, Sardinia, Italy. **Veterinary microbiology**, v. 177, n. 3-4, p. 409-413, 2015.

[4] RODRIGUEZ-GONCER, Isabel et al. A case of pulmonary cryptococcoma due to *Cryptococcus gattii* in the United Kingdom. **Medical mycology case reports**, v. 21, p. 23-25, 2018.

[5] **UNAIDS: KNOWLEDGE IS POWER**. 2018. Disponível em: <www.unaids.org>.

[6] PEREIRA, Cristiane Bigatti et al. Antifungal activity of eicosanoic acids isolated from the endophytic fungus Mycosphaerella sp. against Cryptococcus neoformans and C. gattii. **Microbial pathogenesis**, v. 100, p. 205-212, 2016.

[7] GALIZA, Glauco JN et al. Características histomorfológicas e histoquímicas determinantes no diagnóstico da criptococose em animais de companhia. **Pesq. Vet. Bras**, v. 34, n. 3, p. 261-269, 2014.

[8] MACIEL, R.-A. et al. Corticosteroids for the management of severe intracranial hypertension in meningoencephalitis caused by Cryptococcus gattii: A case report and review. **Journal de mycologie medicale**, v. 27, n. 1, p. 109-112, 2017.

[9] LACAZ, Carlos da Silva et al. Tratado de micologia médica Lacaz. 2002.

[10] IKEDA-DANTSUJI, Yurika et al. Interferon-γ promotes phagocytosis of Cryptococcus neoformans but not Cryptococcus gattii by murine macrophages. **Journal of Infection and Chemotherapy**, v. 21, n. 12, p. 831-836, 2015.

[11] DE QUEIROZ, João Paulo Araújo Fernandes. Criptococose-uma revisão bibliográfica. **Acta Veterinaria Brasilica**, v. 2, n. 2, p. 32-38, 2008.

[12] Criptococose: causas, sintomas, tratamento e prevenção. 2018. Disponível em: <http://www.saude.gov.br>

[13] VIEILLE, Peggy et al. Isolation of *Cryptococcus gattii* VGIII from feline nasal injury. **Medical mycology case reports**, v. 22, p. 55-57, 2018.

[14] SUSANTO, Woen et al. Synthesis of the trisaccharide repeating unit of capsular polysaccharide from Klebsiella pneumoniae. **Tetrahedron letters**, v. 60, n. 3, p. 288-291, 2019.

[15] KUMAR, Harikesh; MANDAL, Pintu Kumar. Synthetic routes toward pentasaccharide repeating unit corresponding to the O-antigen of *Escherichia coli* O181. **Tetrahedron Letters**, v. 60, n. 12, p. 860-863, 2019.

[16] INAGAKI, Ayane et al. A case of Klebsiella pneumoniae spondylitis and bacteremia potentially due to inflammation around a fecalith. **Journal of Infection and Chemotherapy**, v. 25, n. 6, p. 470-472, 2019.

[17] LI, Xin et al. Disruption of blood-brain barrier by an *Escherichia coli* isolated from canine septicemia and meningoencephalitis. **Comparative Immunology, Microbiology and Infectious Diseases**, v. 63, p. 44-50, 2019.

[18] HOLANDA, Cecília Maria de Carvalho Xavier; MOTTA NETO, Renato; ARIMATEIA, Dayse Santos. **Manual de bacteriologia e de enteroparasitos.** Natal: EDUFRN, 2017. 134 p.

[19] MARKOVSKA, Rumyana et al. Multicentre investigation of carbapenemase-producing Klebsiella pneumoniae and Escherichia coli in Bulgarian hospitals–Interregional spread of ST11 NDM-1-producing K. pneumoniae. **Infection, Genetics and Evolution**, v. 69, p. 61-67, 2019.

[20] LÓPEZ-CAMACHO, Elena et al. Meropenem heteroresistance in clinical isolates of OXA-48–producing Klebsiella pneumoniae. **Diagnostic microbiology and infectious disease**, v. 93, n. 2, p. 162-166, 2019.

[21] JIMÉNEZ-GUERRA, Gemma et al. Extended-spectrum beta-lactamase-producing Escherichia coli and Klebsiella pneumoniae from urinary tract infections: Evolution of antimicrobial resistance and treatment options. **Medicina Clínica** (English Edition), v. 150, n. 7, p. 262-265, 2018.

[22] ALAM, Mohammad Zubair et al. Incidence and transferability of antibiotic resistance in the enteric bacteria isolated from hospital wastewater. **Brazilian Journal of Microbiology**, v. 44, n. 3, p. 799-806, 2013.

[23] OLORUNMOLA, Felix Oluwasola; KOLAWOLE, Deboye Oriade; LAMIKANRA, Adebayo. Antibiotic Resistance and Virulence Properties in *Escherichia coli* Strains from Cases of Urinary Tract Infections. **African journal of infectious diseases**, v. 7, n. 1, p. 1-7, 2013

[24] LEVISON, Matthew E.; LEVISON, Julie H. Pharmacokinetics and pharmacodynamics of antibacterial agents. **Infectious Disease Clinics**, v. 23, n. 4, p. 791-815, 2009.

[25] HERAUD, Philip; TOBIN, Mark J. The emergence of biospectroscopy in stem cell research. **Stem cell research**, v. 3, n. 1, p. 12-14, 2009.

[26] DA SILVA, Adenilton C. et al. Two-dimensional linear discriminant analysis for classification of three-way chemical data. **Analytica chimica acta**, v. 938, p. 53-62, 2016.

[27] RUBIO, L.; ORTIZ, M. C.; SARABIA, L. A. Identification and quantification of carbamate pesticides in dried lime tree flowers by means of excitation-emission molecular fluorescence and parallel factor analysis when quenching effect exists. **Analytica chimica acta**, v. 820, p. 9-22, 2014.

[28] LIU, Jianyu; YU, Guan; LIU, Yufeng. Graph-based sparse linear discriminant analysis for high-dimensional classification. **Journal of Multivariate Analysis**, v. 171, p. 250-269, 2019.

[29] BOSE, Smarajit et al. Generalized quadratic discriminant analysis. **Pattern Recognition**, v. 48, n. 8, p. 2676-2684, 2015.

[30] NESPECA, Maurílio Gustavo et al. Rapid and sensitive method for detecting adulterants in gasoline using ultra-fast gas chromatography and Partial Least Square Discriminant Analysis. **Fuel**, v. 215, p. 204-211, 2018.

[31] GUPTA, Ajay; BARBU, Adrian. Parameterized principal component analysis. **Pattern Recognition**, v. 78, p. 215-227, 2018.

[32] ROSA, Larissa Naida et al. Thermal rice oil degradation evaluated by UV–Vis-NIR and PARAFAC. **Food chemistry**, v. 273, p. 52-56, 2019.

[33] LIU, Peng et al. Deep Evolutionary Networks with Expedited Genetic Algorithm for Medical Image Denoising. **Medical Image Analysis**, 2019.

[34] MILANEZ, Karla Danielle Tavares Melo et al. Selection of robust variables for transfer of classification models employing the successive projections algorithm. **Analytica chimica acta**, v. 984, p. 76-85, 2017.

[35] ABBASIAN, Ali; EKBATANI, Shahed. Resin migration tracking via real-time monitoring FTIR-ATR in a self-stratifying system. **Progress in Organic Coatings**, v. 131, p. 159-164, 2019.

[36] SHRIVER, Duward F.; ATKINS, Peter William. **Química inorgânica**. Porto Alegre: Bookman, 2008.

[37] GOMPEL, Joe Van. **The Fundamentals Of Infrared Spectroscopy**. [*S. l.*], 2012. Disponível em: https://www.chemicalonline.com/doc/the-fundamentals-of-infrared-spectroscopy-0001. Acesso em: 13 jul. 2019.

[38] YANG, Xinhao et al. Pre-diabetes diagnosis based on ATR-FTIR spectroscopy combined with CART and XGBoots. **Optik**, v. 180, p. 189-198, 2019.

[39] KHAN, Shahid Ali et al. Fourier Transform Infrared Spectroscopy: Fundamentals and Application in Functional Groups and Nanomaterials Characterization. In: **Handbook of Materials Characterization**. Springer, Cham, 2018. p. 317-344.

[40] LUCARINI, M. et al. Determination of fatty acid content in meat and meat products: The FTIR-ATR approach. **Food chemistry**, v. 267, p. 223-230, 2018.

[41] WESTWORTH, Saori; ASHWATH, Nanjappa; COZZOLINO, Daniel. Application of FTIR-ATR spectroscopy to detect salinity response in Beauty Leaf Tree (Calophyllum inophyllum L). **Energy Procedia**, v. 160, p. 761-768, 2019.

[42] ABBASIAN, Ali; EKBATANI, Shahed. Resin migration tracking via real-time monitoring FTIR-ATR in a self-stratifying system. **Progress in Organic Coatings**, v. 131, p. 159-164, 2019.

[43] LIMA, Kássio MG et al. Classification of cervical cytology for human papilloma virus (HPV) infection using biospectroscopy and variable selection techniques. **Analytical Methods**, v. 6, n. 24, p. 9643-9652, 2014.

[44] SANTOS, Marfran CD et al. ATR-FTIR spectroscopy coupled with multivariate analysis techniques for the identification of DENV-3 in different concentrations in blood and serum: a new approach. **RSC Advances**, v. 7, n. 41, p. 25640-25649, 2017.

[45] SHARMA, Vishal; BHARDWAJ, Shweta; KUMAR, Raj. On the spectroscopic investigation of Kohl stains via ATR-FTIR and multivariate analysis: Application in forensic trace evidence. **Vibrational Spectroscopy**, v. 101, p. 81-91, 2019.

[46] RYMSZA, Taciana et al. Human papillomavirus detection using PCR and ATR-FTIR for cervical cancer screening. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 196, p. 238-246, 2018.

[47] FROST, Jonathan et al. Identification of cancer associated molecular changes in histologically benign vulval disease found in association with vulval squamous cell carcinoma using Fourier transform infrared spectroscopy. **Analytical Methods**, v. 8, n. 48, p. 8452-8460, 2016.

[48] GHOSH, Aritri et al. Chemometric analysis of integrated FTIR and Raman spectra obtained by non-invasive exfoliative cytology for the screening of oral cancer. **Analyst**, v. 144, n. 4, p. 1309-1325, 2019.

[49] CASTILHO, M. L. et al. The efficiency analysis of gold nanoprobes by FT-IR spectroscopy applied to the non-cross-linking colorimetric detection of Paracoccidioides brasiliensis. **Sensors and Actuators B: Chemical**, v. 215, p. 258-265, 2015.

[50] DE SOUSA MARQUES, Aline et al. Feature selection strategies for identification of Staphylococcus aureus recovered in blood cultures using FT-IR spectroscopy successive projections algorithm for variable selection: a case study. **Journal of microbiological methods**, v. 98, p. 26-30, 2014.

[51] LAKOWICZ, Joseph R. (Ed.). **Principles of fluorescence spectroscopy**. Springer Science & Business Media, 2013.

[52] SADAT, Azin; CORRADINI, Maria G.; JOYE, Iris J. Molecular spectroscopy to assess protein structures within cereal systems. **Current opinion in food science**, 2019.

[53] HASSOUN, Abdo et al. Fluorescence spectroscopy as a rapid and non-destructive method for monitoring quality and authenticity of fish and meat products: Impact of different preservation conditions. **LWT**, 2019.

[54] ACKOVIĆ, Lea Lenhardt et al. Modeling Food Fluorescence with PARAFAC. In: **Reviews in Fluorescence 2017**. Springer, Cham, 2018. p. 161-197.

[55] AMORELLO, Diana et al. An analytical method for monitoring micro-traces of landfill leachate in groundwater using fluorescence excitation–emission matrix spectroscopy. **Analytical Methods**, v. 8, n. 17, p. 3475-3480, 2016.

[56] ZHANG, Shihua et al. Assessing the stability in composting of penicillin mycelial dreg via parallel factor (PARAFAC) analysis of fluorescence excitation–emission matrix (EEM). **Chemical Engineering Journal**, v. 299, p. 167-176, 2016.

[57] RUBIO, L. et al. Determination of cochineal and erythrosine in cherries in syrup in the presence of quenching effect by means of excitation-emission fluorescence data and three-way PARAFAC decomposition. **Talanta**, v. 196, p. 153-162, 2019.

[58] HAMBLY, A. C. et al. Characterising organic matter in recirculating aquaculture systems with fluorescence EEM spectroscopy. **Water research**, v. 83, p. 112-120, 2015.

[59] OHAMMADI, Ghobad et al. Exploiting second-order advantage from mathematically modeled voltammetric data for simultaneous determination of multiple antiparkinson agents in the presence of uncalibrated interference. **Journal of the Taiwan Institute of Chemical Engineers**, v. 88, p. 49-61, 2018.

[60] SAJJADI, S. Maryam et al. Quantifying aflatoxins in peanuts using fluorescence spectroscopy coupled with multi-way methods: resurrecting second-order advantage in excitation–emission matrices with rank overlap problem. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 156, p. 63-69, 2016.

[61] EBRAHIMI, Sara; KOMPANY-ZAREH, Mohsen. Investigation of kinetics and thermodynamics of DNA hybridization by means of 2-D fluorescence spectroscopy and soft/hard modeling techniques. **Analytica chimica acta**, v. 906, p. 58-71, 2016.

[62] DE OLIVEIRA NEVES, Ana Carolina; TAULER, Romá; DE LIMA, Kássio Michell Gomes. Area correlation constraint for the MCR− ALS quantification of cholesterol using EEM fluorescence data: A new approach. **Analytica chimica acta**, v. 937, p. 21-28, 2016

[63] DE OLIVEIRA NEVES, Ana Carolina et al. The use of EEM fluorescence data and OPLS/UPLS-DA algorithm to discriminate between normal and cancer cell lines: a feasibility study. **Analyst**, v. 139, n. 10, p. 2423-2431, 2014.

[64] TRAN, Ngoc M. et al. Principal component analysis in an asymmetric norm. **Journal of Multivariate Analysis**, v. 171, p. 1-21, 2019.

[65] UDDIN, Md Nasir et al. Mapping of climate vulnerability of the coastal region of Bangladesh using principal component analysis. **Applied Geography**, v. 102, p. 47-57, 2019.

[66] DA SILVA FERNANDES, Rafael et al. Non-destructive detection of adulterated tablets of glibenclamide using NIR and solid-phase fluorescence spectroscopy and chemometric methods. **Journal of pharmaceutical and biomedical analysis**, v. 66, p. 85-90, 2012.

[67] SONG, Yongxing et al. A novel demodulation method for rotating machinery based on time-frequency analysis and principal component analysis. **Journal of Sound and Vibration**, v. 442, p. 645-656, 2019.

[68] ROSA, Larissa Naida et al. Thermal rice oil degradation evaluated by UV–Vis-NIR and PARAFAC. **Food chemistry**, v. 273, p. 52-56, 2019.

[69] GORDON, Devin A.; ZHAN, Zhonghao; BRUCKMAN, Laura S. Characterizing the weathering induced degradation of Poly (ethylene-terephthalate) using PARAFAC modeling of fluorescence spectra. **Polymer Degradation and Stability**, v. 161, p. 85-94, 2019.

[70] FARAHANI, Khalil Zarnousheh et al. Potentiality of PARAFAC approaches for simultaneous determination of N-acetylcysteine and acetaminophen based on the second-order data obtained from differential pulse voltammetry. **Talanta**, v. 192, p. 439-447, 2019.

[71] GOMES, Paulo RB et al. A nested-PARAFAC based approach for target localization in bistatic MIMO radar systems. **Digital Signal Processing**, v. 89, p. 40-48, 2019.

[72] NGUYEN, Viet-Dung; ABED-MERAIM, Karim; LINH-TRUNG, Nguyen. Second-order optimization based adaptive PARAFAC decomposition of three-way tensors. **Digital Signal Processing**, v. 63, p. 100-111, 2017.

[73] RINNAN, Åsmund; BOOKSH, Karl S.; BRO, Rasmus. First order Rayleigh scatter as a separate component in the decomposition of fluorescence landscapes. **Analytica Chimica Acta**, v. 537, n. 1-2, p. 349-358, 2005.

[74] LIU, Ke et al. A consensus successive projections algorithm–multiple linear regression method for analyzing near infrared spectra. **Analytica Chimica Acta,** v. 858, p. 16-23, 2015.

[75] DE ALMEIDA, Valber Elias et al. Vis-NIR spectrometric determination of Brix and sucrose in sugar production samples using kernel partial least squares with interval selection based on the successive projections algorithm. **Talanta**, v. 181, p. 38-43, 2018.

[76] MESQUITA, Diego PP et al. Building selective ensembles of randomization based neural networks with the successive projections algorithm. **Applied Soft Computing**, v. 70, p. 1135-1145, 2018.

[77] DE ARAÚJO GOMES, Adriano et al. The successive projections algorithm for interval selection in PLS. **Microchemical Journal**, v. 110, p. 202-208, 2013.

[78] SALGUEIRO, Rui; DE ALMEIDA, Ana; OLIVEIRA, Orlando. New genetic algorithm approach for the min-degree constrained minimum spanning tree. **European Journal of Operational Research**, v. 258, n. 3, p. 877-886, 2017.

[79] VILLACAMPA, Yolanda et al. A guided genetic algorithm for diagonalization of symmetric and Hermitian matrices. **Applied Soft Computing**, v. 75, p. 180-189, 2019.

[80] BAEK, Seung H.; PARK, Dong-Ho; BOZDOGAN, Hamparsum. Hybrid kernel density estimation for discriminant analysis with information complexity and genetic algorithm. **Knowledge-Based Systems**, v. 99, p. 79-91, 2016.

[81] TSURUTA, Jaime Hidehiko; NARCISO, Marcelo Gonçalves. **Um estudo sobre algoritmos genéticos**. Embrapa Informática Agropecuária, 2000.

[82] CAI, Wei et al. Network linear discriminant analysis. **Computational Statistics & Data Analysis**, v. 117, p. 32-44, 2018.

[83] CHEN, Lu-Hung; JIANG, Ci-Ren. Sensible functional linear discriminant analysis. **Computational Statistics & Data Analysis**, v. 126, p. 39-52, 2018.

[84] WANG, Xiumei et al. Quadratic discriminant analysis model for assessing the risk of cadmium pollution for paddy fields in a county in China. **Environmental pollution**, v. 236, p. 366-372, 2018.

[85] AOYAGI, Kenta et al. Simple method to construct process maps for additive manufacturing using a support vector machine. **Additive Manufacturing**, v. 27, p. 353-362, 2019.

[86] HOU, Qiuling et al. Discriminative information-based nonparallel support vector machine. **Signal Processing**, v. 162, p. 169-179, 2019.

[87] MALDONADO, Sebastián; PÉREZ, Juan; BRAVO, Cristián. Cost-based feature selection for Support Vector Machines: An application in credit scoring. **European Journal of Operational Research**, v. 261, n. 2, p. 656-665, 2017.

[88] TANG, Long; TIAN, Yingjie; PARDALOS, Panos M. A novel perspective on multiclass classification: Regular simplex support vector machine. **Information Sciences**, v. 480, p. 324-338, 2019.

[89] DE CARVALHO, Jhonnata B. et al. A combined Fourier analysis and support vector machine for EEG classification. **Chilean Journal of Statistics (ChJS)**, v. 10, n. 1, 2019.

[90] HUANG, Huajuan; WEI, Xiuxi; ZHOU, Yongquan. Twin support vector machines: A survey. **Neurocomputing**, v. 300, p. 34-43, 2018.

[91] LI, Hong-Dong; XU, Qing-Song; LIANG, Yi-Zeng. libPLS: An integrated library for partial least squares regression and linear discriminant analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 176, p. 34-43, 2018.

[92] SONG, Weiran et al. Nearest clusters based partial least squares discriminant analysis for the classification of spectral data. **Analytica chimica acta**, v. 1009, p. 27-38, 2018.

[93] LEE, Loong Chuen; LIONG, Choong-Yeun; JEMAIN, Abdul Aziz. Effects of data pre-processing methods on classification of ATR-FTIR spectra of pen inks using partial least squares-discriminant analysis (PLS-DA). **Chemometrics and Intelligent Laboratory Systems**, v. 182, p. 90-100, 2018.

[94] GÓRSKI, Łukasz et al. Voltammetric classification of ciders with PLS-DA. **Talanta**, v. 146, p. 231-236, 2016.

CAPÍTULO 2

Attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy as a new technology for discrimination between *Cryptococcus neoformans* and *Cryptococcus gattii*

Fernanda S. L. Costa

Priscila P. Silva — Camilo L. M. Morais

Thales D. Arantes — Eveline P. Milan

Raquel C. Theodoro — Kássio M. G. Lima

Anal. Methods, 2016, 8, 7107-7115.

Contribuição:

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

Analytical Methods

PAPER

View Article Online
View Journal | View Issue

Cite this: *Anal. Methods*, 2016, **8**, 7107

Attenuated total reflection Fourier transform-infrared (ATR-FTIR) spectroscopy as a new technology for discrimination between *Cryptococcus neoformans* and *Cryptococcus gattii*

Fernanda S. L. Costa,[a] Priscila P. Silva,[a] Camilo L. M. Morais,[a] Thales D. Arantes,[b] Eveline Pipolo Milan,[c] Raquel C. Theodoro*[b] and Kássio M. G. Lima*[a]

Systemic fungal infections are among the most difficult diseases to manage in humans, especially when the recognition of the correct species is required for a precise and successful treatment. This is the case for *Cryptococcus* species and its genotypes, which are the main cause of meningitides in immunocompromised patients. Attenuated total reflection Fourier transform-infrared (ATR-FTIR) spectroscopy with discriminant analysis was employed to distinguish between the pathogenic fungal species *Cryptococcus neoformans* and *Cryptococcus gattii* by determining which wavenumber–absorbance/intensity relationships might reveal biochemical differences. *Cryptococcus* inactivated colonies were applied to an ATR crystal, and vibrational spectra were obtained in the ATR mode. Twenty-eight *Cryptococcus* isolates, fourteen *C. neoformans* and fourteen *C. gattii* were investigated. Spectral categories were analyzed using principal component analysis (PCA), successive projection algorithm (SPA) and genetic algorithm (GA) followed by linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). Multivariate classification accuracy results were estimated based on sensitivity, specificity, positive (or precision) and negative predictive values, Youden index, and positive and negative likelihood ratios. Sensitivity for *C. neoformans* and *C. gattii* categories were 84.4% and 89.3%, respectively, using a QDA-LDA model with 17 wavenumbers with respect to their "fingerprints". Compared to classical methods for differentiation of *Cryptococcus* species, this new technology could represent an alternative and innovative tool for faster and cheaper fungal identification for routine diagnostic laboratories.

Received 4th July 2016
Accepted 5th September 2016

DOI: 10.1039/c6ay01893a

www.rsc.org/methods

Introduction

Cryptococcus genus is widely distributed in nature, but only two species, *Cryptococcus neoformans* and *Cryptococcus gattii*, are frequent human pathogens,[1] which infect the host on inhalation of viable propagules from environmental sources (usually pigeon feces and vegetal material).[2] Cryptococcosis may be opportunistic, most often caused by *C. neoformans*, or a primary disease affecting immunocompetent individuals, mainly caused by *C. gattii*. Meningoencephalitis is the most common clinical feature,[3] but there may also be pulmonary involvement, similar to tuberculosis and histoplasmosis.[4]

Conventional diagnosis of cryptococcosis is based on the direct visualization of encapsulated yeast with India ink or an isolated culture, which is fast compared to other more fastidious fungi. However, identifying the species in this case is as important as identifying the genus. This can be done by using molecular markers or a CGB (bromothymol blue canavanine-glycine) medium, in which *C. gattii* changes its color from yellow to blue.[5,6] However, despite its high sensitivity (about 93%), CGB identification is laborious. Its reading takes at least 48 hours and it can have a subjective interpretation, since some *C. gattii* isolates do not completely change from yellow to blue. In addition, other yeast species can also change the color of the medium.[7]

Several molecular techniques have been applied for the epidemiological study of cryptococcosis, including karyotyping, RAPD (random amplification of polymorphic DNA), RFLP (restriction fragment length polymorphism), AFLP (amplified fragment length polymorphism), PCR (polymerase chain reaction) using specific primer regions of minisatellite or microsatellite sequences, MLST (multi locus sequencing type) and PCR-RFLP of the *URA5* gene.[8–12] Thus, *C. neoformans* was divided into four genotypes: VNI (serotype A), VNII (serotype A), VNIII (serotype AD) and VNIV (serotype D). *C. gattii* was divided into VGI, VGII, VGIII and VGIV (serotypes B and C).

[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]*Tropical Medicine Institute, Federal University of Rio Grande do Norte, Natal 59072-970, RN, Brazil*

[c]*Giselda Trigueiro Hospital, Federal University of Rio Grande do Norte, Natal 59072-970, RN, Brazil*

Correct species identification and genotyping of *C. neoformans* and *C. gattii* isolates is of utmost importance, as there are differences in antifungal susceptibility among genotypes. Clinical data suggest that the response to antifungal therapy is less significant in infections caused by *C. gattii*, requiring more prolonged therapy.[13] Additionally, it was found that *C. gattii* VGII isolates are less susceptible to antifungal drugs (particularly azole), followed by VGI, VNI and VNIV isolates.[14–17]

Although PCR is considered a good alternative option for fungal molecular diagnostics due to its high specificity and sensitivity, it has some technical limitations due to protocol complexities, reagent costs and the choice of specific primers for each species. Accordingly, it is necessary to improve current detection methods, particularly for issues related to saving time, effort and cost.

One interesting alternative to molecular methods in microbiology is FTIR spectroscopy, which has been studied over the past decade for microbial identification.[18–20] However, the application of FTIR as bioanalytical spectroscopy to study fungal bacteria is a relatively new frontier. Infrared spectroscopy is a form of vibrational spectroscopy which provides "whole biochemical fingerprinting" by means of spectral features corresponding to a wide range of important functional groups (lipids $\sim$ 1750 cm^{-1}, carbohydrates $\sim$ 1155 cm^{-1}, primary amide $\sim$ 1650 cm^{-1}, secondary amide $\sim$ 1550 cm^{-1} and DNA/RNA $\sim$ 1225 cm^{-1}; $\sim$ 1080 cm^{-1}),[19] that together could provide important information about the biochemical constituents of each fungal pathogen and also would identify and discriminate fungal bacteria at a species or strain level.

Nonetheless, data analysis is a critical aspect of any diagnostic assay, particularly for IR spectroscopy. The major difficulties in the analysis of microbial species and subspecies are as follows: (i) spectrochemical studies generate complex datasets, giving rise to the challenges of extracting meaningful underlying variance within variables, and (ii) molecular examination of microbial species or subspecies generates complex spectral datasets and demands suitable data-handling tools in order to extract important discriminating information. To overcome these difficulties, many chemometric algorithms have been applied to IR data such as the following ones: (i) principal component analysis (PCA) for simplifying a dataset *via* linear transformations by choosing a new coordinate system, in which the first principal component (PC) describes the greatest variance within the dataset,[20] (ii) linear discriminant analysis (LDA) to reduce confounding factors of within-category heterogeneity, whilst maximizing discriminating biomarkers between each category,[21] (iii) quadratic discriminant analysis (QDA) in which each group is modeled by a separate normal density and a prior probability, and the classification of new observations is done by choosing the group that has the highest posterior probability,[22] and (iv) variable selection methods such as the successive projections algorithm (SPA)[23] and genetic algorithm (GA)[24] to improve the model performance compared to the full spectrum model, eliminating potential interfering variables that generate a lower signal/noise ratio.

The choice and development of the multivariate classification approaches are important to ensure reliable fungal detection using IR spectroscopy. For instance, multivariate classification quality features such as sensitivity, specificity, positive (or precision) and negative predictive values, Youden index, and positive and negative likelihood ratios should be calculated to ensure the validity of the results in accordance with International Guidelines.[25]

Taking into account the potential of IR spectroscopy in various biological assays and considering the difficulty in distinguishing between *C. neoformans* and *C. gattii*, this paper proposes a method of differentiation between these fungal pathogens using attenuated total reflection-FTIR (ATR-FTIR) spectroscopy coupled with multivariate classification techniques. Herein, we have attempted to evaluate the potential of a quicker, low cost method, which uses no reagents, as a new technology for identifying fungal pathogens. This method is based on biochemical intra-individual differences or "fingerprint" features with direct association between peaks and chemical bonds between *C. neoformans* and *C. gattii* with subsequent variable selection methods. In our study, sample preparation, spectroscopic measurement, data preprocessing, feature selection and analytical validation were addressed. To the best of our knowledge, this is the first paper that applies PCA-QDA, SPA-QDA and GA-QDA to differentiate fungi samples based on spectral data. Nevertheless, *C. neoformans* and *C. gattii* were never discriminated by IR spectroscopy using wavelength selection to elucidate the altered biochemical-microbial fingerprint.

Materials and methods

Sample preparation

Fungal cultures from the Veterinary Hospital – UNESP, campus Botucatu (SP), IMT/SP (Instituto de Medicina Tropical de São Paulo), UFPI (Universidade Federal do Piauí) and FioCruz mycological collection, as well as recently isolated fungi from Giselda Trigueiro Hospital (Natal/RN/Brazil) were used in this study (Table 1). The fungal cultures were sent to the Institute of Tropical Medicine of RN at UFRN for genotyping, under the approval of the ethics committee, number 51050415.6.0000.5537. These fungi were cultured on Sabouraud Agar and incubated for 48 hours at a temperature of 30 °C until satisfactory growth was obtained.

For genotyping, the DNA was extracted from a 2 days old sample, according to Trilles L. *et al.* (2008).[12] The amplification of the *URA5* gene was carried out in a 50 μL PCR containing 1× PCR buffer CG (Finnzymes), 3% DMSO, 0.2 mmol L^{-1} of each primer URA5-(5′-ATGTCCTCCCAAGCCCTCGACTCCG-3′) and SJO1-(5′-TTAAGACCTCTGAACACCGTACTC-3′)[10] and 1 unit of Phusion DNA polymerase (Finnzymes). Amplifications were performed in a thermocycler (Eppendorf) using the following cycle: 98 °C for 1 min, 40 cycles of 98 °C for 10 s, 61 °C for 30 s, 72 °C for 30 s and a final cycle at 72 °C for 10 min. The PCR product volume was concentrated to 12.5 μL in a concentrator (Eppendorf) and submitted to double digestion with Sau96I and HhaI restriction endonucleases. The digestion reaction was performed in a final volume of 15 μL, containing 12.5 μL of PCR product, 1.5 μL of 10× Cut Smart buffer, (New England BioLabs), 0.5 μL of HhaI (20 000 units per mL, New England

Table 1 Cryptococcus isolates used in this work

Isolate ID	Species	Genotype	Used for
BT14	*C. gattii*	VGI	Calibration
BT20	*C. gattii*	VGII	
HGT10	*C. gattii*	VGII	
FC3	*C. gattii*	VGIII	
FC9	*C. gattii*	VGIV	
PI1401	*C. gattii*	VGII	
CG751	*C. gattii*	VGII	
CN508	*C. gattii*	VGII	
BT2	*C. neoformans*	VNI	
BT3	*C. neoformans*	VNI	
FC5	*C. neoformans*	VNII	
HGT2	*C. neoformans*	VNII	
FC7	*C. neoformans*	VNIV	
HGT4	*C. neoformans*	VNII	
FC4	*C. neoformans*	VNIII	
BT29	*C. neoformans*	VNIV	
CN894	*C. gattii*	VGII	Validation
BT8	*C. gattii*	VGII	
FC6	*C. gattii*	VGII	
CN117	*C. neoformans*	VNIII	
HGT7	*C. neoformans*	VNI	
HGT5	*C. neoformans*	VNII	
FC1	*C. gattii*	VGII	Prediction
BT17	*C. gattii*	VGII	
CG606	*C. gattii*	VGII	
BT28	*C. neoformans*	VNIV	
HGT1	*C. neoformans*	VNI	
FC2	*C. neoformans*	VNIV	

BioLabs) and 0.5 μL of Sau96I (5000 units per mL, New England BioLabs) at 37 °C for 3 h. The fragments were separated by 3% agarose (GE healthcare) gel electrophoresis, stained with ethidium bromide. The genotype of each isolate was defined following the DNA band patterns described by Meyer, W. *et al.* (2003)[10] and Trilles, L. *et al.* (2008).[12]

For ATR-FTIR spectroscopy, some yeast colonies were placed in 1.0 mL of paraformaldehyde solution at 4% plus phosphate buffer (1 mol L^{-1}) v/v, and in 1.5 mL Eppendorf tubes for cell attachment to inactivate yeast cells for biosafety handling in the spectroscopy equipment. The final solution was added to 28 tubes with 28 different isolates of *Cryptococcus*. After 3 hours at room temperature, tubes with cells were placed under refrigeration at −20 °C until the next step. For spectral readings, the tubes were put at room temperature until defrosting, and then centrifuged for 10 minutes at 5000*g* for cell precipitation. The supernatant was removed and the cells were washed with 1.0 mL of sterile saline solution (0.95% w/v). The tubes were maintained at 4 °C until spectroscopic analysis.

ATR-FTIR spectroscopy

ATR-FTIR spectra [$n = 280$, 10 replicates of each one of the 28 samples (*C. neoformans* ($n = 14$) and *C. gattii* ($n = 14$)] were examined on a Bruker VERTEX 70 FTIR spectrometer (Bruker Optics Ltd., Coventry, UK) with a Helios ATR attachment containing a diamond crystal internal reflective element and a 45° incidence angle of the 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. Approximately 50 μL of each sample was applied to the ATR crystal immediately following the collection of each background. To ensure that no air bubbles were trapped on the crystal surface, a small piece of aluminum foil was placed on the sample, following the same strategy that was recently used by Cui *et al.* (2016).[22] The ATR crystal was cleaned with 70% v/v alcohol and a new background was collected prior to analysis of a new sample.

Chemometrics procedure and software

The data import, pre-treatment and construction of chemometric classification models (PCA-LDA, PCA-QDA, SPA-LDA, SPA-QDA, GA-LDA and GA-QDA) were implemented in MATLAB R2014a software (MathWorks, USA). Raw spectra were pre-processed by cutting between 1800 and 900 cm^{-1} (235 wavenumber at 4 cm^{-1} spectral resolution) and baseline-corrected. For PCA-LDA/QDA, SPA-LDA/QDA and GA-LDA/QDA models, the samples were divided into training (60%), validation (20%) and prediction sets (20%) by applying the classic Kennard–Stone (KS) uniform sampling algorithm to the IR spectra.[23] The KS algorithm was applied separately to each class to extract a representative set of objects from a given dataset by maximizing the minimal Euclidean distance between the already selected objects and the remaining objects. The training samples were used in the modelling procedure (including variable selection for LDA and QDA), whereas the prediction set was only used in the final classification evaluation. The optimum number of variables for SPA-LDA/QDA and GA-LDA/QDA was performed 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,

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

where $I(n)$ is the index of the true class for the nth validation object x_n.

In this definition, the numerator is the squared Mahalanobis distance between object x_n (of class index I_n) and the sample mean $m_{I(n)}$ of its true class. The denominator in eqn (2) corresponds to the squared Mahalanobis distance between object x_n and the center of the closest wrong class.

The QDA classification is made by applying the quadratic discriminant function (DF_Q) to the selected variables for the analyzed classes, according to eqn (3):

$$DF_Q = Q_{i1} - Q_{i2} \quad (3)$$

where Q_{i1} and Q_{i2} are the quadratic distance functions for classes 1 and 2, respectively.[24] This quadratic function is calculated for a given class k by the following equation:

$$Q_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}}\boldsymbol{\Sigma}_k{}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_k) + \log_e|\boldsymbol{\Sigma}_k| - 2\log_e \mu_k \quad (4)$$

where $\mathbf{x}_i$ is an unknown measurement vector for sample i, $\bar{\mathbf{x}}_k$ is the mean measurement vector of class k, $\boldsymbol{\Sigma}_k$ is the variance–covariance matrix of class k, and μ_k is the prior probability of class k.[25]

LDA classification follows a similar formulation, however in this case, the natural logarithm for the determination of the variance–covariance matrix ($\log_e|\boldsymbol{\Sigma}_k|$) is not taken into account, and the covariance matrix for LDA is based on a pooled covariance matrix.[25] In addition, LDA does not take into account different variance structures for each class, assuming that the studied classes have similar variance–covariance matrices. On the other hand, QDA forms a separated variance model for each class and does not assume that the classes have similar variance–covariance matrices.[26] Therefore, QDA is more suitable than LDA for building discriminant models when the analyzed classes have very different variance structures, such as biological media.

The GA routine was carried out during 40 generations with 80 chromosomes each. 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 realizations of the GA was employed. The LDA and QDA scores, loadings, and discriminant function (DF) or Fisher score values were obtained for each category.

Sensitivity (the confidence in a positive result for a sample of the label class was obtained), specificity (the confidence that a negative result for a sample of the non-label class was obtained), positive predictive value (PPV) (measures the proportion of the correctly assigned positive examples and its value varies between 0 and 1), negative predictive value (NPV) (measures the proportion of correctly assigned negative examples and its value varies between 0 and 1), Youden's index (YOU) (evaluates the classifier's ability to avoid failure), the likelihood ratio (LR+) (represents the ratio between the probability to predict an example as positive when it truly is positive, and the probability to predict an example as positive when it actually is not positive), the likelihood ratio (LR−) (represents the ratio between the probabilities to predict an example as negative when it is actually positive, and the probability to predict an example as negative when it truly is negative) were calculated as important quality standards in the test evaluation. The quality metrics used in this study for evaluating the classification results can be calculated with the following equations:

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

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

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \quad (7)$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}} \quad (8)$$

$$\mathrm{YOU} = \mathrm{SENS} - (1 - \mathrm{SPEC}) \quad (9)$$

$$\mathrm{LR}(+) = \frac{\mathrm{SENS}}{1-\mathrm{SPEC}} \quad (10)$$

$$\mathrm{LR}(-) = \frac{1-\mathrm{SENS}}{\mathrm{SPEC}} \quad (11)$$

where FN is a false negative, FP is a false positive, TP is a true positive and TN is a true negative. SENS is the sensibility and SPEC is the specificity.

Results

A typical representative average spectrum was obtained by ATR-FTIR for *C. neoformans* and *C. gattii* with a slight difference of glycogen intensities (1015, 1038, and 1045 cm^{-1}) and DNA/RNA wavenumbers (1123, 1126, and 1219 cm^{-1}) (Fig. 1). The maximum and minimum variances between the 10 replicates of the class *C. gattii* were from 1.05×10^{-8} to 1.91×10^{-5}, while those for the class *C. neoformans* ranged from 4.94×10^{-9} to 9.30×10^{-8}. The variance between the two classes was 5.48×10^{-11}. These spectra were normalized to the amide I (1650 cm^{-1}) absorbance band.

PCA-LDA/QDA models

The discriminant function (DF) plot of PCA-LDA on the average ATR-FTIR spectra reveals a degree of overlap between the classes (Fig. 2A). The PCA-LDA model was built from the calibration set using 3 PCs, together explaining 96.0% of the variance in the data. When PCA-QDA (also with 3 PCs and 96.0% for explained variance) was used to segregate the two categories, the DF plot was obtained, and there was a weak segregation from each category, as shown in Fig. 2B.

PCA-QDA presented better segregation than PCA-LDA between the classes, but a clear region of overlap existed. In

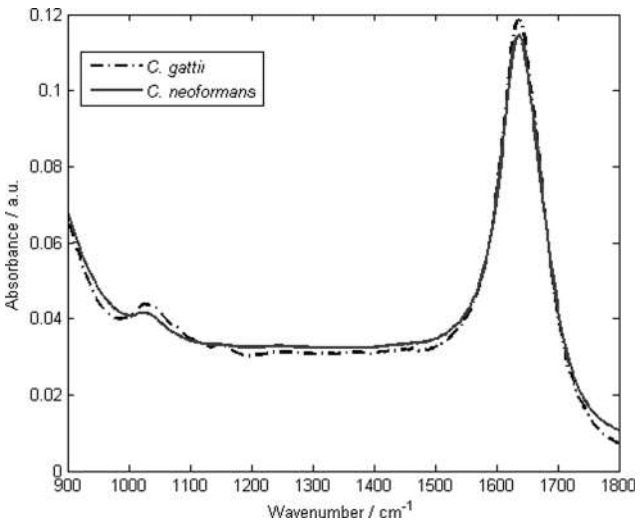


Fig. 1 Average spectrum for each original class (*C. gattii*, dash line; *C. neoformans*, solid line).

addition, even within a given category (*C. gattii*) we can see some inter-fungal variation, suggesting different genotypes.

SPA-LDA/QDA models

SPA-LDA was applied to the dataset to obtain the optimum number of variables by a minimum cost function *G*. As can be seen in Fig. 3A, the model built with only 2 variables (1043 cm^{-1} and 1683 cm^{-1}) slightly improved the segregation between *C. neoformans* and *C. gattii*, when compared with PCA-LDA. The coefficients of these functions were calculated using the training-set statistics (class means and pooled covariance matrix) for the 2 selected variables. The selected variables highlight discriminating differences at amide I (1683 cm^{-1}) and glycogen (1043 cm^{-1}). The SPA-QDA employed for comparison resulted in the selection of two variables, namely 1043 cm^{-1} and 1683 cm^{-1}. As can be seen in Fig. 3B, DF× samples do not fully discriminate among the categories investigated. The PCA-QDA, SPA-QDA results also show a variation within the category (*C. gattii*), suggesting differences in genotypes.

GA-LDA/QDA models

GA-LDA was applied to the data set and resulted in the selection of 17 variables, namely 946, 977, 991, 1074, 1180, 1186, 1259,

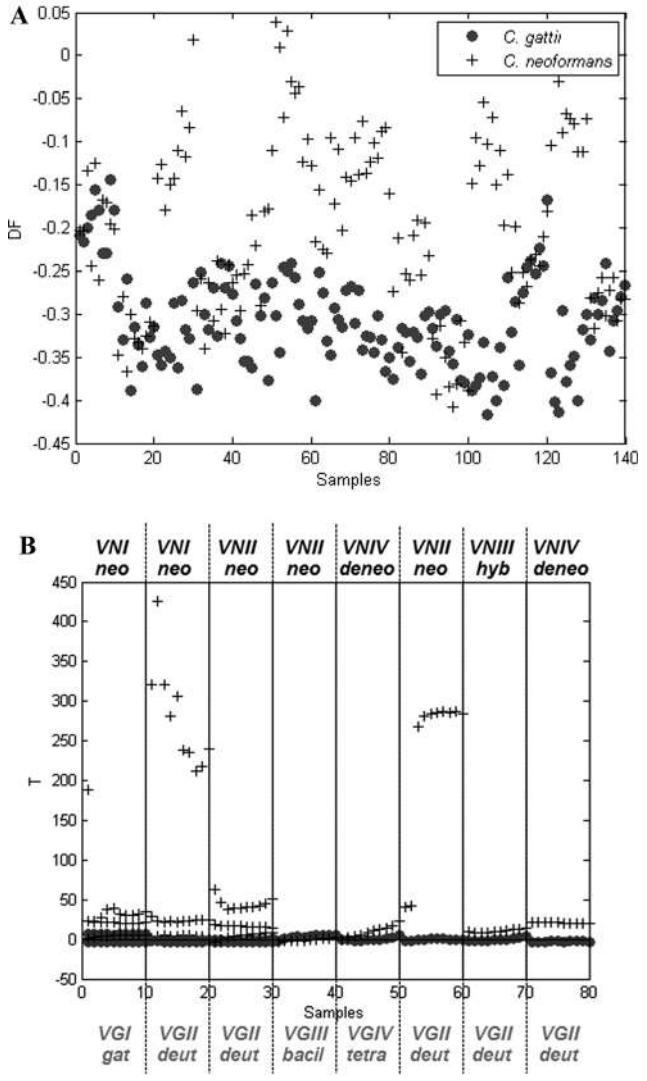


Fig. 2 (A) Discriminant function *versus* samples calculated using the PCA-LDA model from two categories (● = *C. gattii* and + = *C. neoformans*), (B) discriminant function *versus* samples calculated using the PCA-QDA model from two categories (● = *C. gattii* and + = *C. neoformans*), where neo, deneo and hyb refers to the cryptic species *C. neoformans* (VNI and VNII), *C. deneoformans* (VNIV) and the hybrid between *C. neoformans* and *C. deneoformans* (VNIII) respectively, while gat, deut, bacil and tetra refers to the cryptic species *C. gattii* (VGI), *C. deuterogattii* (VGII), *C. bacillosporus* (VGIII) and *C. tetragattii* (VGIV).

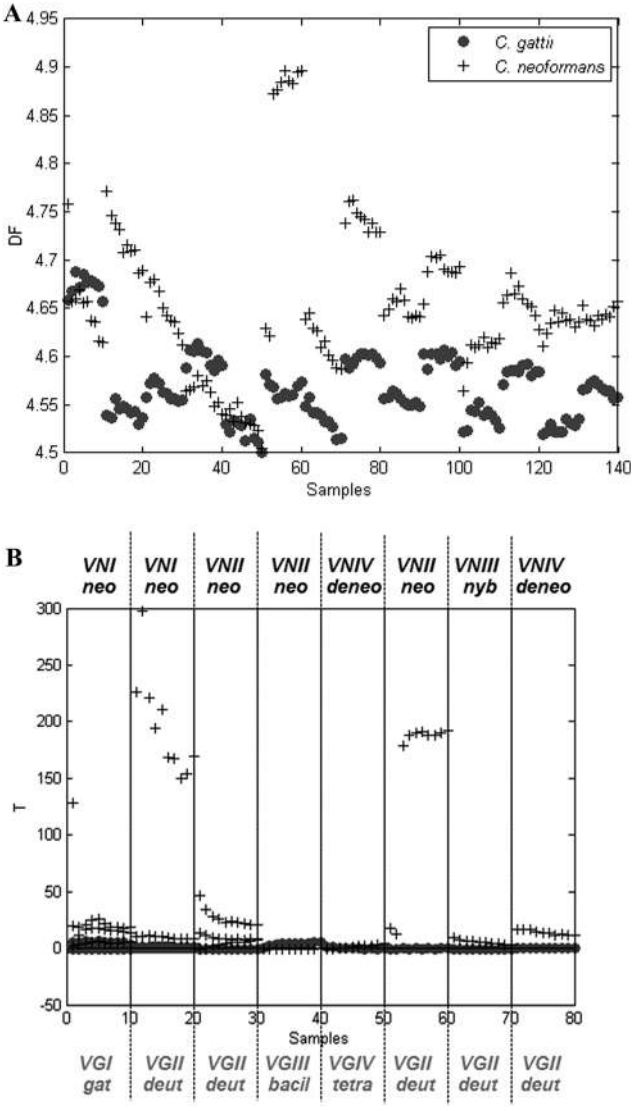


Fig. 3 (A) Discriminant function *versus* samples calculated using the SPA-LDA model from two categories (● = *C. gattii* and + = *C. neoformans*). (B) Discriminant function *versus* samples calculated by using the SPA-QDA model from two categories (● = *C. gattii* and + = *C. neoformans*), where neo, deneo and hyb refer to the cryptic species *C. neoformans* (VNI and VNII), *C. deneoformans* (VNIV) and the hybrid between *C. neoformans* and *C. deneoformans* (VNIII), respectively, while gat, deut, bacil and tetra refer to the cryptic species *C. gattii* (VGI), *C. deuterogattii* (VGII), *C. bacillosporus* (VGIII) and *C. tetragattii* (VGIV).

1276, 1309, 1340, 1371, 1442, 1512, 1525, 1670, 1702 and 1712 cm^{-1}. Using the 17 selected wavelengths, the Fisher scores for all the samples of the data set were obtained (Fig. 4A). There was clearly less overlap between *C. neoformans vs. C. gattii* compared to the other models (PCA-LDA/QDA and SPA-LDA/QDA). Examination of the selected wavenumbers following GA-LDA (Fig. 4A) indicated that the main biochemical alterations for segregation between the species were on glycogen, DNA/RNA, phosphate bands, amide I and amide II. Several selected wavenumbers appear to be of particular interest, namely the variables at 991 and 1074 cm^{-1}, representing carbohydrate bands. The variables at 1180 and 1259 cm^{-1} represent the spectral region of DNA/RNA. Finally, the variables at 1670, 1702 and 1712 cm^{-1} represent the fingerprint region for proteins. GA-QDA employed for comparison resulted in the selection of the same 17 wavenumbers selected for GA-LDA. However, as can be seen in Fig. 4B, there was a greater homogeneity effect among classes using only the 17 wavenumbers selected by GA in the QDA modeling with no misclassification obtained.

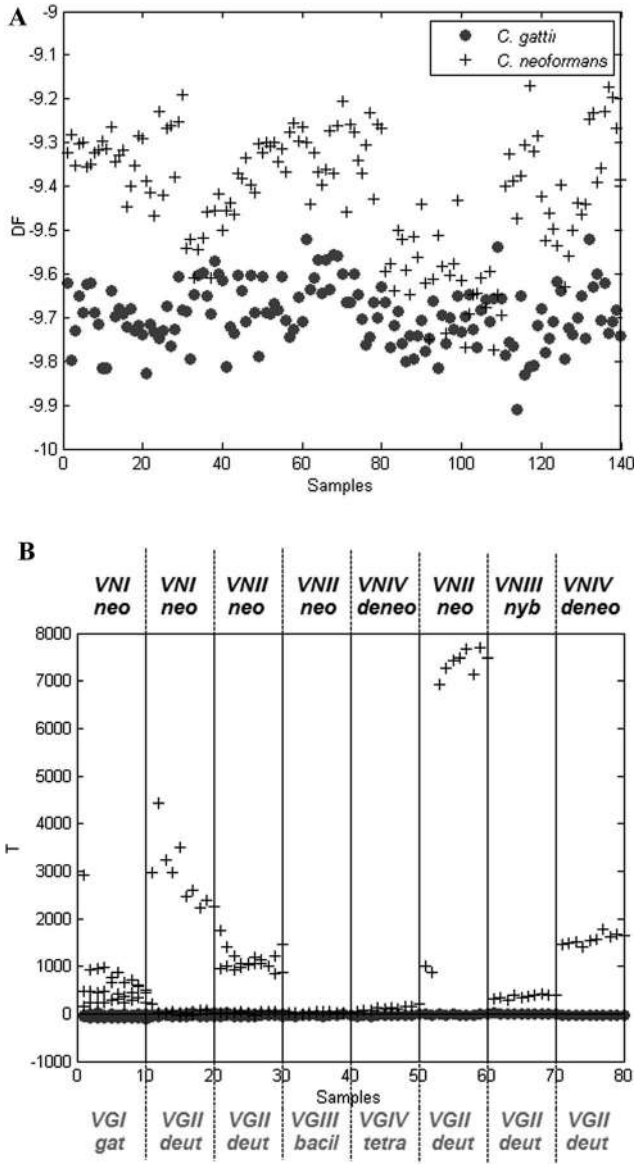


Fig. 4 (A) Discriminant function *versus* samples calculated by using the GA-LDA model from two categories (● = *C. gattii* and + = *C. neoformans*). (B) Discriminant function *versus* samples calculated by using the GA-QDA model from two categories (● = *C. gattii* and + = *C. neoformans*), where neo, deneo and hyb refer to the cryptic species *C. neoformans* (VNI and VNII), *C. deneoformans* (VNIV) and the hybrid between *C. neoformans* and *C. deneoformans* (VNIII), respectively, while gat, deut, bacil and tetra refer to the cryptic species *C. gattii* (VGI), *C. deuterogattii* (VGII), *C. bacillosporus* (VGIII) and *C. tetragattii* (VGIV).

Quality metrics

The empirical evaluation of classification in IR spectroscopy or new techniques for microbiology studies support the usefulness and effectiveness of the classification method. Assessment of the quality of the classification performance without focusing on a class is the most general way of comparing the quality of the classification results. Estimation of such metrics (sensitivity, specificity, positive, PPV, NPV, YOU, LR+ and LR−) were calculated as important quality standards in the test evaluation. Classification rates were determined using the best models.

The corresponding quality metrics achieved for PCA-LDA, SPA-LDA and GA-LDA models of each category are shown in Table 2. For the *C. gattii* category, all the rate classification values from all models using the LDA approach were well classified, showing that ATR-FTIR spectroscopy has the potential to detect and identify this category in a dataset. On the other hand, the sensitivity rates from PCA-LDA, SPA-LDA and GA-LDA for the *C. neoformans* category achieved scores of 33.3%, 70.0% and 20.8%, respectively, showing poor accuracy in comparison with the other category (*C. gattii*), where LDA was employed. Furthermore, the remaining quality metrics for the *C. neoformans* category using LDA models (PCA, SPA and GA) also achieved unsatisfactory results.

Table 3 presents the validation results for the optimized model (PCA-QDA, SPA-QDA and GA-QDA) of each category.

Table 2 Quality performance values from three classification methods (PCA-LDA, SPA-LDA and GA-LDA) by ATR-FTIR spectroscopy for each category

Stage performance features	PCA-LDA	SPA-LDA	GA-LDA
Cryptococcus gattii			
Sensitivity (%)	100.0	100.0	100.0
Specificity (%)	100.0	100.0	100.0
Positive predictive values (PPV)	100.0	100.0	100.0
Negative predictive values (NPV)	100.0	100.0	100.0
Youden index (YOU)	100.0	100.0	100.0
Positive likelihood ratios (LR+)	- - -	- - -	- - -
Negative likelihood ratios (LR−)	0.0	0.0	0.0
Cryptococcus neoformans			
Sensitivity (%)	33.3	70.0	20.8
Specificity (%)	33.3	70.0	30.6
Positive predictive values (PPV)	33.3	70.0	16.7
Negative predictive values (NPV)	33.3	70.0	36.7
Youden index (YOU)	−33.3	40.0	−48.6
Positive likelihood ratios (LR+)	0.5	2.3	0.3
Negative likelihood ratios (LR−)	2.0	0.4	2.6

Table 3 Quality performance values from three classification methods (PCA-QDA, SPA-QDA and GA-QDA) by ATR-FTIR spectroscopy for each category

Stage performance features	PCA-QDA	SPA-QDA	GA-QDA
Cryptococcus gattii			
Sensitivity (%)	3.5	96.6	89.3
Specificity (%)	6.5	93.6	84.4
Positive predictive values (PPV)	100.0	93.3	100.0
Negative predictive values (NPV)	6.7	96.7	90.0
Youden index (YOU)	−90.1	90.1	73.7
Positive likelihood ratios (LR+)	0.0	15.0	5.7
Negative likelihood ratios (LR−)	15.0	0.0	0.1
Cryptococcus neoformans			
Sensitivity (%)	50.0	74.4	84.4
Specificity (%)	93.6	95.2	89.3
Positive predictive values (PPV)	93.3	96.7	90.0
Negative predictive values (NPV)	50.9	66.7	83.3
Youden index (YOU)	43.6	69.6	73.7
Positive likelihood ratios (LR+)	7.8	15.6	7.9
Negative likelihood ratios (LR−)	0.5	0.3	0.2

According to the sensitivity results shown in Table 3, it is possible to see that the sensitivity rates from PCA-QDA, SPA-QDA and GA-QDA achieved scores of 3.5%, 96.6% and 89.3% for the *C. gattii* category, respectively. Although these results for sensitivity and other metrics using QDA models for *C. gattii* are below the ones achieved for LDA models (100%), the SPA-QDA and GA-QDA results can be considered satisfactory. For *C. neoformans* and using QDA models, the sensitivity rate from PCA-QDA, SPA-QDA and GA-QDA achieved scores of 50.0%, 74.4% and 84.4%, respectively, showing an interesting accuracy in comparison with LDA results, particularly for GA-QDA. GA-QDA for *C. neoformans* achieved the best classification rates when compared to other models (PCA and SPA) using the QDA approach.

Discussion

Making correct diagnoses is an urgent demand in the medical mycology field, since many molecular epidemiological studies have been pointing out some differences between close species or even genotypes concerning geographic distribution, clinical aspects and treatment response. This is the case for cryptococcosis; when it is caused by *C. gattii* VGII isolates, it is less responsive to antifungal drugs (especially azole), followed by VGI, VNI and VNIV isolates.[14–17] Furthermore, the molecular types are not equally distributed across the world. VNIV is more frequent in Europe, whereas *C. gattii* VGII is frequently found in the Americas and VGI is found in Oceania, Asia and Europe. Thus, for the *Cryptococcus* species, genotypes and geographic origins are important data that must be taken into consideration for choosing the correct treatment.[14]

Until now, the most successful methods for differentiating between *Cryptococcus* species were PCR, PCR-RFLP or sequence based methods, which are very laborious and expensive for routine applications. In this study, we applied ATR-FTIR spectral information coupled with multivariate classification techniques (PCA-LDA/QDA, SPA-LDA/QDA and GA-LDA/QDA) for the first time to distinguish between *C. neoformans* and *C. gattii*. In contrast to cryptococcosis diagnostic methods, metabolic fingerprinting revealed through ATR-FTIR spectral information coupled with multivariate classification techniques (PCA-LDA/QDA, SPA-LDA/QDA and GA-LDA/QDA) were able to quickly differentiate between *C. gattii* and *C. neoformans*. The main biochemical alterations for segregation between the species from the 17 variables used for the GA-LDA and GA-QDA models were on carbohydrates, DNA/RNA, phosphate bands and proteins.

C. neoformans and *C. gattii* are different in many aspects. There is a predilection of *C. neoformans* for CNS and *C. gattii* for the lungs, and as already mentioned *C. neoformans* mainly infects immunosuppressed patients (mostly those with HIV/AIDS), while *C. gattii* is considered a primary pathogen, infecting immunocompetent and apparently healthy individuals.[13,27] Essentially, this clinical pattern is common for the VGI and VGII *C. gattii* genotypes, which more frequently cause infections in immunocompetent individuals than the VGIII and VGIV genotypes.[28] The VGII genotype is responsible for almost all infections in the Pacific Northwest, including the outbreak on Vancouver Island,[29] and its clinical manifestation more often involves the lungs than the CNS.[30,31] According to Ma, H. *et al.* (2009),[32] the high virulence of this genotype is due to an unusual tubular mitochondrial morphology, as a consequence of mitochondrial fusions responsible for enhancing the repair of mitochondrial DNA damage from oxidative stress within the phagosome.

One important difference between *C. neoformans* and *C. gattii*, which might have contributed to the differences on carbohydrate bands of the spectra variables, is their polysaccharide capsule, which is composed of 90–95% GXM glucuronoxylomannan (GXM; composed of mannose, xylose and glucuronic acid), and 5% galactoxylomannan (GalXM). The capsule determines the serotypes of *C. neoformans* (serotypes A, D and AD) and the serotypes of *C. gattii* (serotypes B and C).[33,34]

Other biochemical differences between both pathogens have also been revealed, such as differences in creatinine, canavanine, and glycine metabolism, which have allowed the development of the CGB medium.[5] Further investigations in a rat model compared the metabolites released by *C. neoformans* and *C. gattii* using magnetic resonance spectroscopy. Although metabolites in *C. neoformans* were generally present in higher concentrations, two novel metabolites of acetoin and dihydroxyacetone were described in *C. gattii*, potentially giving less pro-inflammatory responses when compared to *C. neoformans*, which could facilitate fungal survival and local multiplication to form cryptococcomas.[35]

All these clinical and biochemical differences between *C. neoformans* and *C. gattii* are, certainly, a reflex of genomic and proteomic divergences between them, dating from 34 mya. Comparison of *C. neoformans* var. *grubii* and *C. gattii* VGI strain WM276 and VGIIa strain R265 revealed 85% to 87% identity. Similar to *C. neoformans*, *C. gattii* has a genome size around 18.4 Mb, divided into 14 chromosomes with conserved centromere locations, although some rearrangements such as inversions

and balanced translocations were observed.[36] One interesting observation was the loss of an iRNA mechanism in VGII isolates, in contrast to *C. neoformans* isolates.[37]

The losses and gains of certain genes could corroborate the spectral differences herein observed, not only between *C. neoformans* and *C. gattii*, but also between the genotypes within each species. Farrer, R. A. *et al.* (2015)[28] suggested the occurrence of some genes involved in the stress response and those in coding for the metal binding domains in VGI and VGII could contribute to their high virulence, while the loss of enolase and copper transporters in VGIII and VGIV genotypes might corroborate their low infection rates in immunocompetent individuals, including cases reported in HIV/AIDS patients.[38]

Phylogenetic studies based on MLST have challenged the existence of only two species, *C. neoformans* and *C. gattii*. As such, seven species have recently been proposed: *C. neoformans* (VNI and VNII), *C. deneoformans* (VNIV), *C. gattii* (VGI), *C. bacillosporus* (VGIII), *C. deuterogattii* (VGII), *C. tetragattii* (VGIV), and *C. decagattii* (VGIV and VGIIIc). The VNIII genotype would be a hybrid between *C. neoformans* and *C. deneoformans*.[39] These results show that the variability found within the genus may reflect more than one intraspecific polymorphism, but indicate the presence of different lineages that differ environmentally and clinically.

In our study, we clearly demonstrated the segregation between *C. gattii* and *C. neoformans* using our methodology with subsequent PCA-QDA, SPA-QDA and GA-QDA algorithms. The best model was GA-QDA, which successfully detected biochemical alterations for the fungi using only 17 wavenumbers, contrasting the traditional full-spectrum PCA model which presents some overlap between categories. It would be interesting to test this methodology with a larger number of isolates not only belonging to different *Cryptococcus* species, but also to different genotypes. As we can observe in Fig. 2B, 3B and 4B, there was some clustering for some genotypes or cryptic species of *C. neoformans*, however, it clearly does not separate the different cryptic species. Unfortunately, the number of isolates available for this work (most from Brazilian mycology collections) was limited. Increasing the amount of fungal samples from all over the globe from both species and from their respective genotypes would certainly improve the statistical significance and potentially distinguish even the very close genotypes.

In addition, our method was thoroughly validated in accordance with quality metrics, being considered as sensitive, specific, and suitable for use as an alternative methodology for fungal pathogenic determination, and potentially directly applicable to clinical samples. Generating a library of major fungal pathogens and evaluating more powerful multivariate classification methods (support vector machine – SVM) is required for this approach to become a standard classification method.

Acknowledgements

We would like to thank CAPES, the post-graduation programs PPGQ and PPGBQ and the Propesq (Pró-reitoria de pesquisa) from UFRN for the fellowships, and the financial support of the CNPq (Grants 305962/2014-4 and 475525/2013-2), as well as Professors Sandra de Moraes Gimines Bosco (UNESP/Brazil), Gilda del Negro (IMT/SP/Brazil) and Fernanda Fonseca (UFPI/Brazil) for the generous isolate supply.

References

1 R. Vilgalys and M. Hester, *J. Bacteriol.*, 1990, **172**, 4238–4246.
2 J. R. Köhler, A. Casadevall and J. Perfect, *Cold Spring Harbor Perspect. Med.*, 2015, **5**, a019273.
3 R. Rozenbaum and A. J. R. Gonçalves, *Clin. Infect. Dis.*, 1994, **18**, 369–380.
4 D. G. Campbell, *Am. Rev. Respir. Dis.*, 1966, **94**, 236–243.
5 K. J. Kwon-chung, I. Polacheck and J. E. Bennett, *J. Clin. Microbiol.*, 1982, **15**, 535–537.
6 K. H. Min and K. J. Kwon-Chung, *Zentralblatt für Bakteriol. Mikrobiol. und Hyg. Ser. A Med. Microbiol. Infect. Dis. Virol. Parasitol.*, 1986, **261**, 471–480.
7 L. McTaggart, S. E. Richardson, C. Seah, L. Hoang, A. Fothergill and S. X. Zhang, *J. Clin. Microbiol.*, 2011, **49**, 2522–2527.
8 Y. Yamamoto, S. Kohno, H. Koga, H. Kakeya, K. Tomono, M. Kaku, T. Yamazaki, M. Arisawa and K. Hara, *J. Clin. Microbiol.*, 1995, **33**, 3328–3332.
9 W. Meyer, D. M. Aanensen, T. Boekhout, M. Cogliati, M. R. Diaz, M. C. Esposto, M. Fisher, F. Gilgado, F. Hagen, S. Kaocharoen, A. P. Litvintseva, T. G. Mitchell, S. P. Simwami, L. Trilles, M. A. Viviani and J. Kwon-Chung, *Med. Mycol.*, 2009, **47**, 561–570.
10 W. Meyer, A. Castañeda, S. Jackson, M. Huynh and E. Castañeda, *Emerging Infect. Dis.*, 2003, **9**, 189–195.
11 W. Meyer and T. G. Mitchell, *Electrophoresis*, 1995, **16**, 1648–1656.
12 L. Trilles, M. d. S. Lazéra, B. Wanke, R. V. Oliveira, G. G. Barbosa, M. M. Nishikawa, B. P. Morales and W. Meyer, *Mem. Inst. Oswaldo Cruz*, 2008, **103**, 455–462.
13 T. C. Sorrell, *Med Mycol.*, 2001, **39**, 155–168.
14 L. Trilles, W. Meyer, B. Wanke, J. Guarro and M. Lazéra, *Med. Mycol.*, 2012, **50**, 328–332.
15 H. S. Chong, R. Dagg, R. Malik, S. Chen and D. Carter, *J. Clin. Microbiol.*, 2010, **48**, 4115–4120.
16 F. Hagen, M. T. Illnait-Zaragozi, K. H. Bartlett, D. Swinne, E. Geertsen, C. H. W. Klaassen, T. Boekhout and J. F. Meis, *Antimicrob. Agents Chemother.*, 2010, **54**, 5139–5145.
17 N. Iqbal, E. E. DeBess, R. Wohrle, B. Sun, R. J. Nett, A. M. Ahlquist, T. Chiller and S. R. Lockhart, *J. Clin. Microbiol.*, 2010, **48**, 539–544.
18 M. Beekes, P. Lasch and D. Naumann, *Vet. Microbiol.*, 2007, **123**, 305–319.
19 K. Becker, N. Al Laham, W. Fegeler, R. A. Proctor, G. Peters and C. von Eiff, *J. Clin. Microbiol.*, 2006, **44**, 3274–3278.
20 Y. Burgula, D. Khali, S. Kim, S. S. Krishnan, M. A. Cousin, J. P. Gore, B. L. Reuhs and L. J. Mauer, *J. Rapid Methods Autom. Microbiol.*, 2009, **15**, 146–175.
21 M. J. Walsh, S. W. Bruce, K. Pant, P. L. Carmichael, A. D. Scott and F. L. Martin, *Toxicology*, 2009, **258**, 33–38.

22 L. Cui, H. J. Butler, P. L. Martin-hirsch and F. L. Martin, *Anal. Methods*, 2016, 1–7.
23 R. Kennard and L. Stone, *Technometrics*, 1969, **11**, 137–148.
24 T. Næs, T. Isaksson, T. Fearn and T. Davies, *A User-friendly Guide to Multivariate Calibration and Classification*, NIR publications, Charlton, Chichester, UK, 2002.
25 W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D. L. Massart, S. Heuerding and F. Erni, *Anal. Chim. Acta*, 1996, **329**, 257–265.
26 S. J. Dixon and R. G. Brereton, *Chemom. Intell. Lab. Syst.*, 2009, **95**, 1–17.
27 B. Speed and D. Dunt, *Clin. Infect. Dis.*, 1995, **21**, 28–34.
28 R. A. Farrer, C. A. Desjardins, S. Sakthikumar, S. Gujja, S. Saif, Q. Zeng, Y. Chen, K. Voelz, J. Heitman, R. C. May, M. C. Fisher and C. A. Cuomo, *mBio*, 2015, **6**, 1–12.
29 E. J. Byrnes III, W. Li, Y. Lewit, H. Ma, K. Voelz, P. Ren, D. A. Carter, V. Chaturvedi, R. J. Bildfell, R. C. May and J. Heitman, *PLoS Pathog.*, 2010, **6**, 1–16.
30 J. R. Harris, S. R. Lockhart, E. Debess, N. Marsden-Haug, M. Goldoft, R. Wohrle, S. Lee, C. Smelser, B. Park and T. Chiller, *Clin. Infect. Dis.*, 2011, **53**, 1188–1195.
31 S. C. A. Chen, M. A. Slavin, C. H. Heath, E. Geoffrey Playford, K. Byth, D. Marriott, S. E. Kidd, N. Bak, B. Currie, K. Hajkowicz, T. M. Korman, W. J. H. McBride, W. Meyer, R. Murray and T. C. Sorrell, *Clin. Infect. Dis.*, 2012, **55**, 789–798.
32 H. Ma, F. Hagen, D. J. Stekel, S. A. Johnston, E. Sionov, R. Falk, I. Polacheck, T. Boekhout and R. C. May, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 12980–12985.
33 S. P. Franzot, I. F. Salkin and A. Casadevall, *J. Clin. Microbiol.*, 1999, **37**, 838–840.
34 K. J. Kwon-Chung, J. E. Bennett and J. C. Rhodes, *Antonie van Leeuwenhoek*, 1982, **48**, 25–38.
35 L. Wright, W. Bubb, J. Davidson, R. Santangelo, M. Krockenberger, U. Himmelreich and T. Sorrell, *Microbes Infect.*, 2002, **4**, 1427–1438.
36 C. A. D'Souza, J. W. Kronstad, G. Taylor, R. Warren, M. Yuen, G. Hu, W. H. Jung, A. Sham, S. E. Kidd, K. Tangen, N. Lee, T. Zeilmaker, J. Sawkins, G. McVicker, S. Shah, S. Gnerre, A. Griggs, Q. Zeng, K. Bartlett, W. Li, X. Wang, J. Heitman, J. E. Stajich, J. A. Fraser, W. Meyer, D. Carter, J. Schein, M. Krzywinsk, K. J. Kwon-Chung, A. Varma, J. Wang, R. Brunham, M. Fyfe, B. F. F. Ouellette, A. Siddiqui, M. Marra, S. Jones, R. Holt, B. W. Birren, J. E. Galagan and C. A. Cuomoe, *mBio*, 2011, **2**, 1–11.
37 R. B. Billmyre, S. Calo, M. Feretzaki, X. Wang and J. Heitman, *Chromosome Res.*, 2013, **21**, 561–572.
38 D. J. Springer, R. B. Billmyre, E. E. Filler, K. Voelz, R. Pursall, P. A. Mieczkowski, R. A. Larsen, F. S. Dietrich, R. C. May, S. G. Filler and J. Heitman, *PLoS Pathog.*, 2014, **10**, 1–19.
39 F. Hagen, K. Khayhan, B. Theelen, A. Kolecka, I. Polacheck, E. Sionov, R. Falk, S. Parnmen, H. T. Lumbsch and T. Boekhout, *Fungal Genet. Biol.*, 2015, **78**, 16–48.

CAPÍTULO 3

Comparison of multivariate classification algorithms using EEM fluorescence data to distinguish *Cryptococcus neoformans* and *Cryptococcus gattii* pathogenic fungi

Fernanda S. L. Costa

Priscila P. Silva

Camilo L. M. Morais

Thales D. Arantes

Raquel C. Theodoro

Kássio M. G. Lima

Anal. Methods, 2017, 9, 3968-3976.

Contribuição:

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

Analytical Methods

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/c7ay00781g

Comparison of multivariate classification algorithms using EEM fluorescence data to distinguish *Cryptococcus neoformans* and *Cryptococcus gattii* pathogenic fungi

Fernanda S. L. Costa,[a] Priscila P. Silva,[a] Camilo L. M. Morais,[a] Raquel C. Theodoro,[b] Thales D. Arantes[bc] and Kássio M. G. Lima *[a]

Cryptococcus neoformans and *Cryptococcus gattii* are the etiologic agents of cryptococcosis, whose suitable treatment depends on rapid and correct detection and differentiation of the *Cryptococcus* species. Currently, this identification is made by classical and molecular techniques; however most of them are considered laborious and expensive. As an alternative method to discriminate *C. gattii* and *C. neoformans*, excitation-emission matrix (EEM) fluorescence spectroscopy combined with multivariate classification methods, Unfolded Partial Least Squares Discriminant Analysis (UPLS-DA), multiway-Partial Least Squares Discriminant Analysis (nPLS-DA), Parallel Factor Analysis (PARAFAC), Principal Component Analysis (PCA), Successive Projection Algorithm (SPA) and Genetic Algorithm (GA), followed by Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) was herein investigated. This technique showed to be an innovative and low cost methodology which requires a small sample volume. Among the methods, the most successful model was UGA-LDA, which showed a sensitivity of 88.9% within only 5 selected wavelengths in calibration and 100.0% prediction for both classes of *C. neoformans* and *C. gattii*, equaling or surpassing some of the biological tests that are usually carried out to differentiate these fungi.

Received 23rd March 2017
Accepted 5th June 2017

DOI: 10.1039/c7ay00781g

rsc.li/methods

Introduction

Cryptococcosis is a systemic mycosis with global distribution, mainly affecting people in tropical and subtropical areas due to suitable climatic conditions for the saprophytic growth of its etiological yeast agents: *Cryptococcus neoformans* and *Cryptococcus gattii*,[1,2] being two complex species with four genotypes each. Differences between *C. neoformans* and *C. gattii* as well as among their genotypes may be associated with different clinical manifestations and treatment responses to antifungal drugs.[3] The infection occurs by inhalation of dehydrated yeasts (viable propagules) from the environment.[4] Cryptococcosis has variable clinical manifestations, such as pulmonary infection, local lymphatic infection and mainly involvement of the central nervous system. Most of the opportunistic cryptococcosis infections are caused by *C. neoformans*, while *C. gattii* commonly causes disease in immunocompetent individuals,[4] usually requiring a longer therapeutic period,[5–7] and a fast differentiation between *C. neoformans* and *C. gattii* is in demand.

Direct visualization of the agent in biological samples or culture diagnosis can be easy with cryptococcosis, but differentiating species and genotypes is an expensive challenge. Bromothymol blue canavanine-glycine (CGB) medium has high sensitivity (about 93%), but CGB identification is laborious and needs at least 48 hours for a subjective interpretation.[8] Associated with culture techniques, molecular methods such as Restriction Fragment Length Polymorphism (RFLP), Amplified Fragment Length Polymorphism (AFLP),[9] Polymerase Chain Reaction (PCR) and their variants can be used to discriminate *Cryptococcus* genotypes (VN-1, II, III and IV) from *C. neoformans* (VG-I, II, III and IV) and *C. gattii*.[10–14] All of these techniques are expensive and laborious for application in routine diagnosis of cryptococcosis.

For this reason, new analytical strategies that can distinguish fungal species with low cost and time should be developed. In this way, spectroscopic methods along with chemometric methods have increasingly been applied in biological analysis as efficient tools for clinical diagnosis due to their great potential for group classification. For instance, ATR-FTIR and EEM fluorescence have been applied to distinguish

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

[b]Tropical Medicine Institute, Federal University of Rio Grande do Norte, Natal 59072-970, RN-Brazil, Brazil

[c]Post-graduation Program in Biochemistry, Federal University of Rio Grande do Norte, Natal 59072-970, RN-Brazil, Brazil

cancer cells[15,16] and Raman for detecting phospholipids and proteins in blood.[17]

Among spectroscopic techniques, molecular fluorescence has great importance since it is a highly sensitive analytical technique which allows measurements in natural environmental concentrations, in addition of being non-destructive. Coupled with this, the matrix of excitation-emission (EEM) fluorescence contains a huge amount of analyte information, allowing for direct monitoring of fluorophores due to its particular sensitivity to a sample chemical environment.[18,19] Fluorescence overlapping components may be identified according to their wavelengths of excitation and emission, and this technique is able to identify biomolecules in samples.[20]

In spite of the great advantages of EEM fluorescence spectroscopy, selectivity may be impaired due to the large spectral overlap or presence of matrix interference. However, chemometric tools can maximize the extraction of relevant information.[21] Recognition methods of supervised and unsupervised standards are commonly applied to extract spectral characteristics and develop classification models.[22] Among the supervised monitoring methods, we can highlight partial least squares discriminant analysis (PLS-DA), which is an adaptation of PLS for pattern recognition where a code for each class is assigned during the calibration process.[23] It is a model that reduces the number of variables and may even be employed where the variability within groups is greater than the variability between groups. It can be constructed with its own 3D matrix (n-PLS-DA)[24] or built by unfolding the original data using the Unfolded Partial Least Squares (UPLS) method.[25]

Nevertheless, some sets of data using PLS do not provide satisfactory results. In these cases, it is recommended to use other methods such as linear discriminant analysis (LDA). In LDA, the densities of conditional class probability are considered as normal multivariate distributions with different mean vectors for different classes, wherein the scattering matrices are identical for all classes. When it is not possible to take the classes' dispersion matrices as equal, the result is to apply the quadratic discriminant analysis (QDA).[26] However, when compared with PLS-DA, the LDA and QDA methods present as a limitation the strong collinearity of the data. For bypassing this disadvantage, the number of training samples must be equal to or larger than the number of variables included in the LDA model. In this sense, dimensionality reduction methods are required before LDA and QDA for classification of spectral data.[27]

A widely used data reduction method is Principal Component Analysis (PCA).[28] PCA simplifies a dataset through linear transformations by choosing a new coordinate system. In this system, the first principal component (PC) describes the greatest variance within the dataset[29] and can be used to reduce data multidimensionality, maintaining the most relevant variability.

Parallel factor analysis (PARAFAC) models are a generalization of principal component analysis (PCA) for a set of data matrices.[30] They are used to decompose trilinear data with a single solution, enabling robust estimates of excitation and emission profiles present in the spectra and their concentrations, a property known as the advantage of the second order.[31] The decomposed data coming from PARAFAC can be used to build the LDA and QDA models. Also, resource selection or a selection reduced by the use of pre-processing into a set of latent variables may improve the classification, although there are some economic and technical constraints.[32]

A selection of a subset of great features allows for the construction of a model with high predictive capacity. Among variable selection strategies, the most commonly used are Genetic Algorithm (GA) and Successive Projections Algorithm (SPA). The GA is generally suitable for variable selection as a method that optimizes a set of data to be used for binders of an artificial evolutionary process. In order for it to be applied to a classification problem, it can be coupled with LDA resulting in response maximization and recognition capability estimated by cross-validation in the training set.[33] On the other hand, SPA is a variable selection method used to minimize multicollinearity problems in the original dataset. It comprises a first phase in which projection operations are performed with an array of descriptor values. These projections are used to generate descriptor subsets with less multicollinearity. In the next step, the best subset is selected to minimize a cost function associated with the average risk of classification error for a given set of validation by comparing the Mahalanobis distance of the sample relative to its true class and the nearest wrong class.[34]

Herein, this study aimed to use molecular fluorescence spectroscopy to discriminate between *C. neoformans* and *C. gattii* and compare the potential of scoring models built with second-order data (UPCA-LDA/QDA, UGA-LDA/QDA, USPA-LDA/QDA, PARAFAC-LDA/QDA, UPLS-DA and nPLS).

Experimental procedures

Sample preparation

The 28 isolates of *Cryptococcus* used in this study are from the Hospital of Clinics and Veterinary Hospital – UNESP, campus Botucatu (SP/Brazil), IMT/SP/Brazil (Instituto de Medicina Tropical de São Paulo), UFPI/Brazil (Universidade Federal do Piauí) and FioCruz/Brazil mycological collection. The fungal isolates from the hospitals were sent to the Institute of Tropical Medicine of RN at UFRN/Brazil for genotyping, under the approval of the ethics committee, number 51050415.6.0000.5537.

As previously described,[35] the fungal isolates in culture on Sabouraud Agar with cloranfenicol (50 mg L^{-1}) were identified by PCR-RFLP of the *URA5* gene. For EEM fluorescence, 28 yeast colonies of different isolates of *Cryptococcus* were used, each one was placed in a 1.5 mL microtube, with 1.0 mL of 4% paraformaldehyde solution plus phosphate buffer (1 mol L^{-1}) v/v, for cell attachment to inactivate yeast cells for biosafety handling in the fluorescence equipment. The final solution of each of the microtubes was transferred to 28 different tubes. After 3 hours at room temperature, the tubes with cells were placed under refrigeration at −20 °C until the next step. For fluorescence reading, the tubes were put at room temperature until defrosted, and then centrifuged for 10 minutes at 5000*g* for cell precipitation. The supernatant was removed and the cells were washed with 1.0 mL of sterile saline solution (0.95% w/v). The tubes were maintained at 4 °C until fluorescence was recorded.

EEM fluorescence spectroscopy

The emission values obtained for calibration were acquired at an excitation of 240 nm and in the emission range from 250 to 900 nm (1 nm steps). For fungal culture samples, the excitation/emission fluorescence data were acquired in the wavelength range of 220–320 nm for excitation and 250–900 nm for emission, in steps of 10 and 1 nm for excitation and emission, respectively. A RF-5301 Shimadzu spectrofluorometer with a 0.5 mm quartz cuvette was used. The excitation and emission monochromator slit widths were fixed at 5 nm, the speed scan was set to super mode (3000 nm min^{-1}), the photomultiplier tube was set to the medium level and a cell with a fiber optic reflectance probe was used. A total of 500 μL of saline solution with fungal cells was added to the fluorescence cuvette for reading, and the cuvette was washed with distilled water after each mensuration in an alcohol solution at 70% and washed again with distilled water to avoid contamination between fungal samples. The temperature was maintained at 25 °C throughout the experiments.

Chemometric procedure and software

Spectral pre-processing and multivariate classification models were built using MATLAB R2011a software (The Math-Works, Natick, USA) and the PLS Toolbox 7.9.3 package (Eigenvector Research, Inc., Wenatchee, WA, USA). The spectral pre-processing was composed of a cut in the region of 801–900 nm in the emission range, and by removing Rayleigh and Raman scatterings using the 'EEMscat' algorithm.[36] Ranges of 220–320 nm for excitation and 250–800 nm for emission were used for model building, with steps of 10 and 1 nm used for excitation and emission, respectively. This resulted in a data matrix size of 11×551 for each sample. For the construction of classification models, the samples were divided into calibration (70%) and prediction (30%) sets using the Kennard–Stone (KS) sample selection algorithm.[37]

The following classification methods were utilized: UPCA-LDA/QDA; UGA-LDA/QDA; USPA-LDA/QDA; PARAFAC-LDA/QDA; UPLS-DA; and nPLS-DA.

The UPCA-LDA/QDA algorithm is based on the unfolding of the EEM matrices into row vectors. These vectors are organized into a matrix $\mathbf{X}$ containing n rows (samples) and k columns (variables). Then a regular PCA followed by LDA and QDA methods is used. The PCA decomposition of $\mathbf{X}$ takes the following form:[38]

$$\mathbf{X} = \mathbf{TP}^{\mathbf{T}} + \mathbf{E} \quad (1)$$

where $\mathbf{T}$ is the scores matrix; $\mathbf{P}$ is the loadings matrix; and $\mathbf{E}$ is the residual matrix. The LDA and QDA are applied to the PCA scores as follows:[39]

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

$$Q_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathbf{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) + \log_e |\boldsymbol{\Sigma}_k| - 2 \log_e \pi_k \quad (3)$$

in which L_{ik} is the LDA classification score; Q_{ik} is the QDA classification score; $\mathbf{x}_i$ is the vector containing the classification variables for sample i (*e.g.*, PCA scores for A components); $\bar{\mathbf{x}}_k$ is the mean vector of class k; $\boldsymbol{\Sigma}_k$ is the variance–covariance matrix of class k; Σ_{pooled} is the pooled covariance matrix; and π_k is the prior probability of class k. The $\boldsymbol{\Sigma}_k$, Σ_{pooled} and π_k are calculated as follows:

$$\boldsymbol{\Sigma}_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathbf{T}} \quad (4)$$

$$\Sigma_{\text{pooled}} = \frac{1}{n} \sum_{k=1}^{K} n_k \boldsymbol{\Sigma}_k \quad (5)$$

$$\pi_k = \frac{n_k}{n} \quad (6)$$

where n is the total number of objects in the training set; K is the number of classes; and n_k is the number of objects of class k.

The same procedure of LDA and QDA is applied to the selected variables by GA and SPA using the EEM matrices unfolded (UGA-LDA/QDA and USPA-LDA/QDA). The GA reduces the data into a few selected variables following an evolutionary process based on Darwin's theory, where the best set of variables (chromosomes) is selected according to a fitness function.[40] And the SPA reduces the original data in order to minimize their multi-collinearity according to the minimum of the cost function G.[41] The GA fitness is calculated as the inverse of the cost function G, which is determined as[16]

$$G = \frac{1}{N_V} \sum_{n=1}^{N_V} g_n \quad (7)$$

$$g_n = \frac{r^2(x_n, m_{I(n)})}{\min_{I(m) \neq I(n)} r^2(x_n, m_{I(m)})} \quad (8)$$

where the numerator is the squared Mahalanobis distance between object x_n of class index $I(n)$ and the sample mean $m_{I(n)}$ of its true class; and the denominator is the squared Mahalanobis distance between object x_n and the center of the closest wrong class.

PARAFAC is a method of decomposition of high-order data based on a trilinear system.[42] It decomposes the three-way data of EEM matrices $\underline{\mathbf{X}}$ by[43]

$$\underline{\mathbf{X}} = \mathbf{A}(\mathbf{C}|\otimes|\mathbf{B})^{\mathbf{T}} + \underline{\mathbf{E}} \quad (9)$$

where $\mathbf{A}$ is the PARAFAC scores matrix representing the sample direction; $\mathbf{B}$ is the PARAFAC loadings matrix representing the excitation direction; $\mathbf{C}$ is the PARAFAC loadings matrix representing the emission direction; and $\underline{\mathbf{E}}$ is the residual tensor. The symbol $|\otimes|$ represents the Khatri–Rao product.[44] For classification purposes, the PARAFAC scores are used in conjunction with LDA and QDA according to eqn (2) and (3), respectively.

UPLSA-DA and nPLS-DA are classification methods based on partial least squares (PLS). In UPLS-DA, the three-way EEM data are unfolded into a matrix $\mathbf{X}$ and a linear classification is made based on a criterion obtained by PLS. This is achieved by an interactive process involving eqn (10) and (11):[45]

$$\mathbf{X} = \mathbf{TP} + \mathbf{E} \quad (10)$$

$$\mathbf{c} = \mathbf{Tq} + \mathbf{f} \quad (11)$$

where **c** is the numerical representation of the label of each sample according its class membership; **T** is the scores matrix; **P** is the loadings matrix of **X**; **q** is the loadings matrix of **c**; and **E** and **f** are the residuals of the spectra and classes, respectively. nPLS-DA is the natural extension of PLS-DA to N-way structures.[46]

The UPCA-LDA/QDA models were built with 4 Principal Components (PCs), which included 89.0% of explained variance. UGA-LDA/QDA models were built from the best solution (in terms of the fitness value) resulting from the GA routine, which was carried out for 40 generations with 80 chromosomes each. Crossover and mutation probabilities were set to 60% and 10%, respectively. UPLS-DA and nPLS-DA models were built with 3 and 4 Latent Variables (LVs), which included 99.1% and 90.2% of explained variance, respectively. The PARAFAC model was previously developed and selected a two-factor score matrix, with 99.0% of explained variance.

Finally, the method was validated to determine whether it fulfilled its intended purpose. To do this, some figures of merit such as sensitivity, specificity, positive predicted value (PPV), negative predicted value (NPV), Youden's index (YOU), positive likelihood ratio (LR^+), and negative likelihood ratio (LR^-) were calculated to ensure the quality of the models.

Results and discussion

Fig. 1 presents the excitation/emission fluorescence spectra (EEM) of one sample of *C. gattii* (Fig. 1a) and *C. neoformans* (Fig. 1b), after removing Rayleigh and Raman scatterings (the excluded spectral regions were properly corrected by interpolation). As can be seen in Fig. 1a and b, the spectral profiles of both pathogen classes are very similar, making it difficult to distinguish between them. For this reason, it is necessary to use multivariate classification algorithms that maximize the difference between the two classes.

A total of 28 samples were used for building the models, divided into two groups: calibration (18 samples) and prediction (10 samples). Table 1 shows the results of PLS-DA classification models built using the EEM fluorescence data for differentiating *C. gattii* and *C. neoformans* pathogens.

The UPLS-DA model used the unfolded matrix with a size of 28 × 6061, and it was built with 3 latent variables, which contemplates 99.05% of the explained variance. For the prediction samples, this model correctly classified 100.0% of *C. gattii* samples and *C. neoformans* samples, with a RMSEP of 0.28. Also, it presented a satisfactory classification for the calibration samples for both classes, with a RMSEC of 0.30. The nPLS-DA model used a 3D-matrix with a size of 28 × 11 × 551. The model was built using only 4 latent variables, which contemplated 99% of the explained variance. In the calibration set, the classification was more satisfactory for *C. gattii*, with a RMSEC of 0.38, while in the prediction set there was a 100% correct classification rate for both pathogen classes, with a RMSEP of 0.31. As already mentioned in the literature,[20] in some cases the PLS-DA models do not provide a good classification rate. The above results confirm that, as seen in Table 1, the correct classification rate in calibration was low, especially for the *C. gattii* class. This lower result in calibration suggests that the model is not well fitted due to the small number of samples used.

Therefore, classification models using LDA and QDA were employed in an attempt to maximize the difference between the classes. The UPCA-LDA model built with 4 principal components (PCs) showed a higher classification capacity for the *C.*

Table 1 Correct classification rates obtained for classification models (UPLS-DA and nPLS-DA) between *C. gattii* and *C. neoformans*

Model	Class	Calibration	Prediction
UPLS-DA (3)[a]	*C. gattii*	66.7	100.0
	C. neoformans	88.9	100.0
nPLS-DA (4)[a]	*C. gattii*	66.7	100.0
	C. neoformans	88.9	100.0

[a] Number of latent variables.

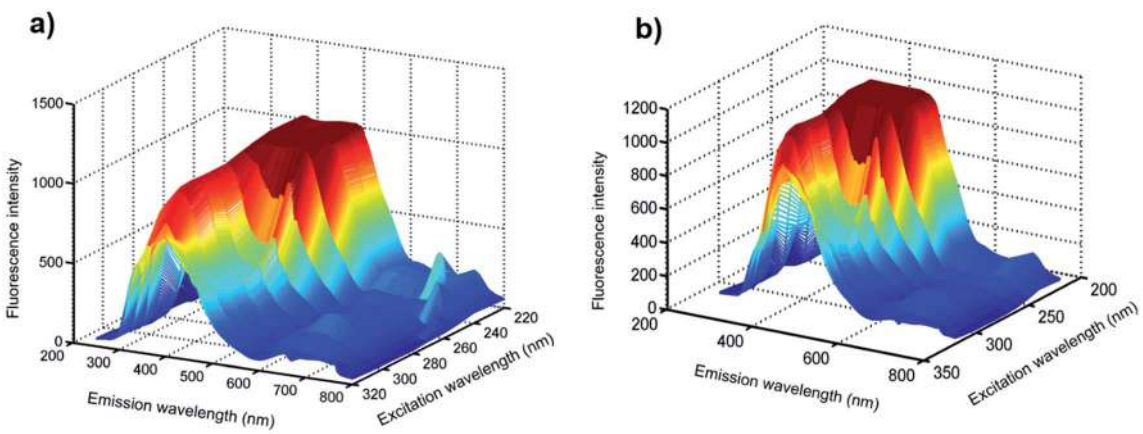


Fig. 1 Excitation–emission molecular fluorescence spectra obtained for *C. gattii* (a) and *C. neoformans* (b). The Rayleigh and Raman scatterings have been removed from the spectra.

Table 2 Correct classification rates obtained for classification models (PARAFAC-LDA/QDA, UPCA-LDA/QDA, USPA-LDA/QDA and UGA-LDA/QDA) between *C. gattii* and *C. neoformans*

Model	Class	Calibration	Prediction
PARAFAC-LDA (2)[a]	*C. gattii*	55.6	60.0
	C. neoformans	88.9	100.0
PARAFAC-QDA (2)[a]	*C. gattii*	77.8	60.0
	C. neoformans	66.7	80.0
UPCA-LDA (4)[b]	*C. gattii*	77.8	100.0
	C. neoformans	88.9	60.0
UPCA-QDA (4)[b]	*C. gattii*	100.0	100.0
	C. neoformans	66.7	20.0
USPA-LDA (2)[c]	*C. gattii*	66.7	100.0
	C. neoformans	77.8	100.0
USPA-QDA (2)[c]	*C. gattii*	66.7	100.0
	C. neoformans	77.8	100.0
UGA-LDA (5)[c]	*C. gattii*	88.9	100.0
	C. neoformans	88.9	100.0
UGA-QDA (11)[c]	*C. gattii*	55.6	40.0
	C. neoformans	100.0	100.0

[a] Number of parallel factors. [b] Number of principal components. [c] Number of selected variables.

gattii class. The same was observed in the UPCA-QDA model also built with 4 PCs, in which the percentage of correct classification was even lower in the calibration and the prediction set compared to the previous algorithm.

A PARAFAC model was previously constructed using two components that explained 99.0% of the data variance. The two-factor scores matrix was submitted to the LDA and QDA routines, in order to obtain the classification models. PARAFAC-LDA showed a better classification for *C. neoformans*, especially in the prediction set, in which 100.0% of the samples were correctly identified. However, for *C. gatti* the correct classification rate was very low both in the calibration and in the prediction set. PARAFAC-QDA, in comparison with the previous model, provided similar results for *C. gatti*, but the performance for *C. neoformans* was not satisfactory and lower than that of the LDA, as can be seen in Table 2.

The first variable selection model based on the successive projections algorithm (SPA) obtained a 100% classification rate for the two classes, selecting 2 variables for classification (Fig. 2a). Both models better classified the *C. neoformans* group in calibration.

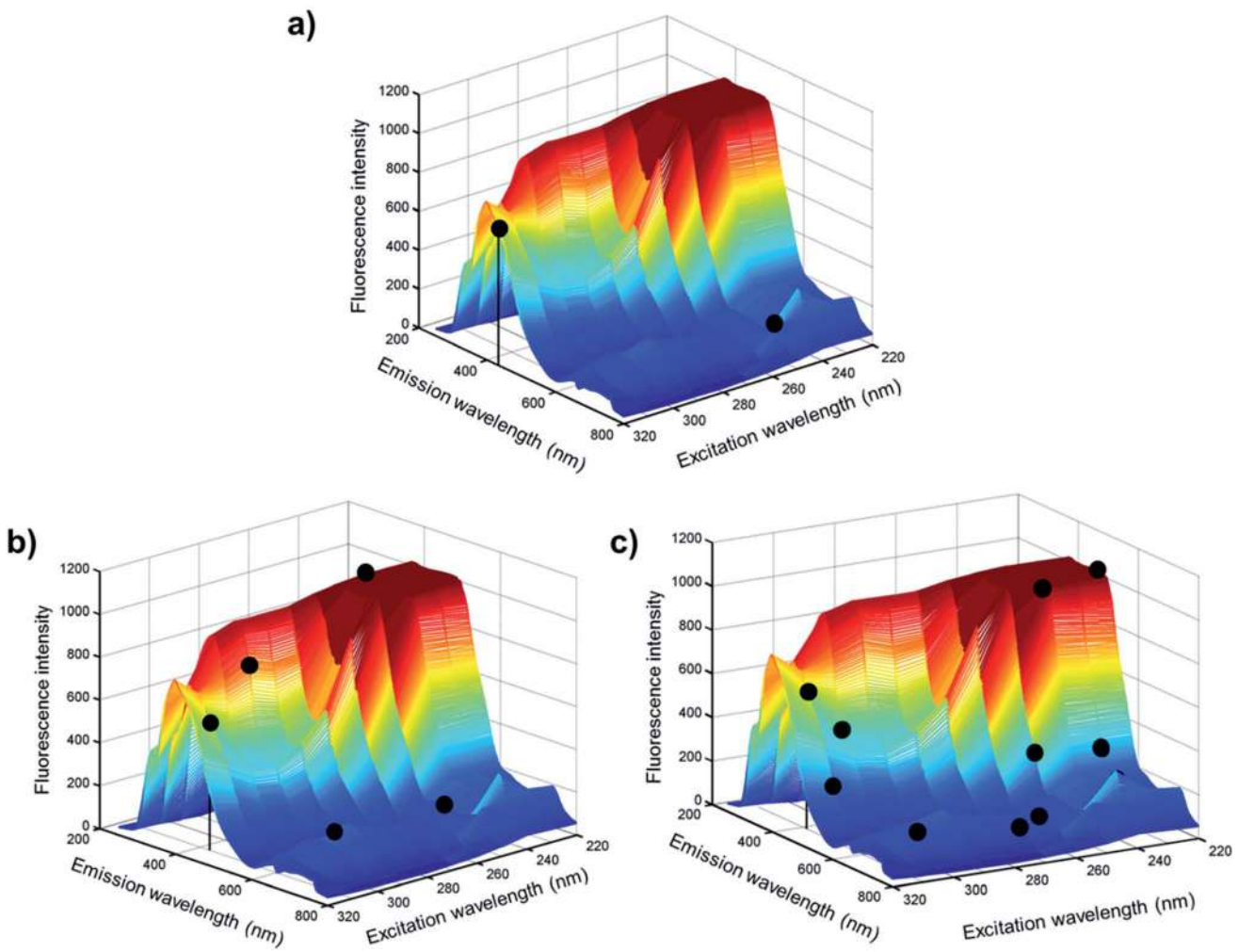


Fig. 2 Excitation–emission molecular fluorescence mean spectra with the variables selected (●) by the (a) USPA-LDA/QDA model, (b) UGA-LDA model and (c) UGA-QDA model.

The UGA-LDA model selected 5 variables (Fig. 2b). As seen in Table 2, the classification rate was 88.9%, which corresponds to only one misclassified sample for the calibration set of the two pathogen classes. The UGA-QDA model was built with only 11 variables (Fig. 2c), and had a satisfactory rating only for the *C. neoformans* class, while only 50.0% of the samples were classified correctly for *C. gattii* in calibration and prediction sets. Using the 5 selected wavelengths, Fisher scores for all the data set samples were obtained (Fig. 3). These results corroborate with the literature[24] since GA includes an optimization process in which many combinations of features and their interactions are considered, as GA has the advantage of enabling efficient searches for clusters with high data size and complexity.

Table 3 presents the validation results of the optimized models (UPLS-DA, nPLS-DA and PARAFAC-LDA/QDA) for each classification category. The results in Table 3 show that the sensitivity rate of UPLS-DA, nPLS-DA, PARAFAC-LDA and PARAFAC-QDA achieved scores of 100.0%, 100.0%, 50.0% and 75.0% for the *C. gattii* category, respectively. Among these models, the most satisfactory models for *C. gattii* classification were UPLS-DA and nPLS-DA. For the classification of *C. neoformans* using the UPLS-DA, nPLS-DA, PARAFAC-LDA and PARAFAC-QDA models, the sensitivity rate achieved values of 100.0%, 100.0%, 50.0% and 66.7%, respectively, showing perfect accuracy only for UPLS-DA and nPLS-DA models. The similar performance of UPLS-DA and nPLS-DA demonstrates that the unfolding procedure does not influence the classification performance in this case.

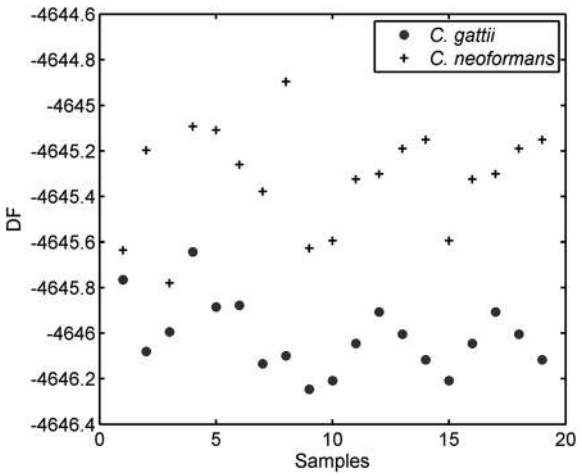


Fig. 3 Discriminant function *versus* samples calculated by using the UGA-LDA model from the two categories (● *C. gattii* and + *C. neoformans*).

In addition, the results in Table 3 show that the QDA had superior classification performance than LDA using PARAFAC as a dimensionality reduction technique. This probably occurred because although both LDA and QDA are based on a Mahalanobis distance calculation, the QDA algorithm forms a separated variance model for each class, not assuming similar class variance–covariance matrices as LDA does.[47]

Table 4 presents the validation results of the optimized models (UPCA-LDA/QDA, USPA-LDA/QDA and UGA-LDA/QDA) for each classification category. According to Table 4, the sensitivity of UPCA-LDA, USPA-LDA and UGA-LDA for class *C. gattii* was 100.0% for all models. Other parameters such as the specificity, NPV and PV were all equal to 100% for class *C. gattii*, corroborating with the sensitivity results. For the *C. neoformans* class, the sensitivity of UPCA-LDA, USPA-LDA and UGA-LDA was 60.0%, 100.0% and 100.0%, respectively, thus demonstrating that the UPCA-LDA is more satisfactory for *C. gattii* classification.

The validation results of the optimized models UPCA-QDA, USPA-QDA and UGA-QDA (Table 4) showed that the sensitivity rates for *C. gattii* were 55.56%, 100.0% and 100%, respectively, while for *C. neoformans* they were equal to 100.0%, 100.0% and 83.33%. Although presenting 100% sensitivity for the *C.*

Table 3 Quality performance values of three classification methods (UPLS-DA, nPLS-DA and PARAFAC-LDA/QDA) by molecular fluorescence spectroscopy for each category

Stage performance features	UPLS-DA	nPLS-DA	PARAFAC-LDA	PARAFAC-QDA
C. gattii				
Sensitivity (%)	100.0	100.0	50.0	75.0
Specificity (%)	100.0	100.0	50.0	66.7
Positive predictive values (PPV)	100.0	100.0	60.0	60.0
Negative predictive values (NPV)	100.0	100.0	40.0	80.0
Youden index (YOU)	100.0	100.0	0.0	41.7
Positive likelihood ratios (LR^+)	—	—	1.0	2.3
Negative likelihood ratios (LR^-)	0.0	0.0	1.0	0.38
C. neoformans				
Sensitivity (%)	100.0	100.0	50.0	66.7
Specificity (%)	100.0	100.0	—	75.0
Positive predictive values (PPV)	100.0	100.0	100.0	80.0
Negative predictive values (NPV)	100.0	100.0	0.0	60.0
Youden index (YOU)	100.0	100.0	—	41.7
Positive likelihood ratios (LR^+)	—	—	—	2.7
Negative likelihood ratios (LR^-)	0.0	0.0	—	0.4

Table 4 Quality performance values of three classification methods (UPCA-LDA/QDA, USPA-LDA/QDA and UGA-LDA/QDA) by molecular fluorescence spectroscopy for each category

Stage performance features	UPCA-LDA	UPCA-QDA	USPA-LDA	USPA-QDA	UGA-LDA	UGA-QDA
C. gattii						
Sensitivity (%)	100.0	55.6	100.0	100.0	100.0	100.0
Specificity (%)	100.0	100.0	100.0	100.0	100.0	62.5
Positive predictive values (PPV)	100.0	100.0	100.0	100.0	100.0	40.0
Negative predictive values (NPV)	100.0	20.0	100.0	100.0	100.0	100.0
Youden index (YOU)	100.0	55.6	100.0	100.0	100.0	62.5
Positive likelihood ratios (LR^+)	—	—	—	—	—	2.67
Negative likelihood ratios (LR^-)	0.0	0.4	0.0	0.0	0.0	0.0
C. neoformans						
Sensitivity (%)	60.0	100.0	100.0	100.0	100.0	83.3
Specificity (%)	60.0	55.6	100.0	100.0	100.0	100.0
Positive predictive values (PPV)	60.0	20.0	100.0	100.0	100.0	100.0
Negative predictive values (NPV)	60.0	100.0	100.0	100.0	100.0	80.0
Youden index (YOU)	20.0	55.6	100.0	100.0	100.0	83.3
Positive likelihood ratios (LR^+)	1.5	2.3	—	—	—	—
Negative likelihood ratios (LR^-)	0.7	0.0	0.0	0.0	0.0	0.2

neoformans class, the UPCA-QDA model presented low specificity, showing that this model could not satisfactorily discriminate between the two classes of pathogens. With the UGA-QDA, the sensitivity for *C. gattii* was 100.0% with a specificity and PPV of 62.5% and 40%, respectively. In addition, it presented a sensitivity rate of 83.3% for the *C. neoformans* class.

The results found here were very satisfactory for discriminating *C. neoformans* and *C. gattii* classes, especially for UGA-LDA, which showed an accuracy of 88.9% in calibration with 5 selected wavelengths and 100.0% in the prediction set for both *C. neoformans* and *C. gattii*. The validation results of the UGA-LDA optimized model confirmed the applicability potential of this method for the classification of the two pathogens. The sensitivity and specificity of 100% for both classes show how this model is able to correctly identify the individuals belonging and not belonging to each class. The obtained PPV and NPV values of 100% suggest that the method was done correctly. Youden's index (YOU) of 100% indicates the model's ability to avoid failure. Also, the values LR^+ and LR^- corroborate the indication that the UGA-LDA gave a satisfactory rating for both fungal classes.

According to the literature, the differentiation into two species is due the composition of the capsular polysaccharide structure, the biochemical properties, the reservoir and the immunological state.[48,49] *C. gattii* and *C. neoformans* present important differences in the nucleotide composition of their RNAs.[50,51] The satisfactory classification of the models based on the EEM fluorescence data is probably due to the different luminescence patterns for each class, since they have different chromophores (nucleotides) in their RNAs. Marbumrung (2012) used fluorescence spectroscopy and PCA to discriminate 10 nucleotides in 2 different solvents and obtained accuracies of almost 100% for both.[52] Another study that corroborates with this hypothesis is that of Cekan and collaborators (2012), in which the fluorescence technique was used to study conformation and nucleotide dynamics in DNA and the results showed that there is an intrinsic relationship between the nucleotide and the intensity of the fluorescence, which was proven by Electron Paramagnetic Resonance (EPR) spectroscopy.[53]

This study shows the power of these chemometric algorithms associated with molecular fluorescence spectroscopy to discriminate *Cryptococcus neoformans* and *Cryptococcus gattii*, where the sensitivity found (100%) was higher than other values reported in the literature for discriminating these species based on ATR-FTIR spectroscopy, for instance, where sensitivity using the GA-QDA algorithm (17 wavenumbers) for *C. neoformans* and *C. gattii* categories was reported to be 84.4% and 89.3%, respectively.[35]

Conclusion

This study demonstrates that EEM fluorescence spectroscopy in combination with multivariate analysis has the potential to differentiate between *C. gattii* and *C. neoformans* cultures, with a possible potential use directly for biological samples from patients with cryptococcosis which would make the diagnosis faster with a species-specific accuracy, reducing procedural costs. Among the models used, UGA-LDA was the most satisfactory, since its sensitivity values for both classes (*C. neoformans* and *C. gattii*) were greater than or equal to those of classical identification methods for these fungi. Thus, we can conclude that this study puts forth an efficient and low-cost method, with analyses which can be carried out more quickly with a very small amount of sample. We not only believe that the use of this method in biological samples is suitable for the diagnosis and epidemiological studies of cryptococcosis, but also believe that it should be developed for other systemic fungal diseases as a promising diagnostic tool in medical mycology.

Acknowledgements

Fernanda S. L. Costa and Camilo L. M. Morais would like to thank CAPES/PPGQ/UFRN for financial and scientific support. Priscila P. Silva acknowledges PROPESQ/UFRN for financial support. Thales D. Arantes would like to thank CAPES/PNPD/PPgBQ for financial support and IMT/RN for providing laboratory infrastructure. Kássio M. G. Lima thanks the CNPq (305962/2014-0).

References

1 R. Vilgalys and M. Hester, *J. Bacteriol.*, 1990, **172**, 4238–4246.
2 J. R. Köhler, A. Casadevall and J. Perfect, *Cold Spring Harbor Perspect. Med.*, 2015, **5**, a019273.
3 M. I. Butler and R. T. M. Poulter, *Fungal Genet. Biol.*, 2005, **42**, 452–463.
4 C. S. Lacaz, E. Porto, J. E. C. Martins, E. M. Heins-Vaccari and N. T. de Melo, *Tratado de Micologia Médica Lacaz, Sarvier Editora de Livros Médicos Ltda*, São Paulo, Brazil, 2002.
5 R. Rozenbaum and A. J. R. Gonçalves, *Clin. Infect. Dis.*, 1994, **18**, 369–380.
6 D. G. Campbell, *Am. Rev. Respir. Dis.*, 1966, **94**, 236–243.
7 T. Sorrell, *Med. Mycol.*, 2001, **39**, 155–168.
8 L. McTaggart, S. E. Richardson, C. Seah, L. Hoang, A. Fothergill and S. X. Zhang, *J. Clin. Microbiol.*, 2011, **49**, 2522–2527.
9 Y. Yamamoto, S. Kohno, H. Koga, H. Kakeya, K. Tomono, M. Kaku, T. Yamazaki, M. Arisawa and K. Hara, *J. Clin. Microbiol.*, 1995, **33**, 3328–3332.
10 W. Meyer, D. M. Aanensen, T. Boekhout, M. Cogliati, M. R. Diaz, M. C. Esposto, M. Fisher, F. Gilgado, F. Hagen, S. Kaocharoen, A. P. Litvintseva, T. G. Mitchell, S. P. Simwami, L. Trilles, M. A. Viviani and J. Kwon-Chung, *Med. Mycol.*, 2009, **47**, 561–570.
11 W. Meyer, A. Castañeda, S. Jackson, M. Huynh and E. Castañeda, *Emerging Infect. Dis.*, 2003, **9**, 189–195.
12 W. Meyer and T. G. Mitchell, *Electrophoresis*, 1995, **16**, 1648–1656.
13 L. Trilles, M. dos S. Lazéra, B. Wanke, R. V. Oliveira, G. G. Barbosa, M. M. Nishikawa, B. P. Morales and W. Meyer, *Mem. Inst. Oswaldo Cruz*, 2008, **103**, 455–462.
14 L. Trilles, W. Meyer, B. Wanke, J. Guarro and M. Lazéra, *Med. Mycol.*, 2012, **50**, 328–332.
15 A. C. D. O. Neves, R. F. de Araújo, A. L. C. S. L. Oliveira, A. A. de Araújo and K. M. G. de Lima, *Analyst*, 2014, **139**, 2423–2431.
16 G. Theophilou, K. M. G. Lima, P. L. Martin-Hirsch, H. F. Stringfellow and F. L. Martin, *Analyst*, 2016, **141**, 585–594.
17 J. Depciuch, M. Sowa-Kućma, G. Nowak, D. Dudek, M. Siwek, K. Styczeń and M. Parlińska-Wojtan, *J. Pharm. Biomed. Anal.*, 2016, **131**, 287–296.
18 M. C. G. Antunes and J. C. G. E. da Silva, *Anal. Chim. Acta*, 2005, **546**, 52–59.
19 M. C. G. Antunes, C. C. C. Pereira and J. C. G. E. da Silva, *Anal. Chim. Acta*, 2007, **595**, 9–18.
20 S. Zhang, Z. Chen, Q. Wen and J. Zheng, *Chem. Eng. J.*, 2016, **299**, 167–176.
21 M. C. Ortiz, L. A. Sarabia, M. S. Sánchez and D. Giménez, *Anal. Chim. Acta*, 2009, **642**, 193–205.
22 L. Rubio, M. C. Ortiz and L. A. Sarabia, *Anal. Chim. Acta*, 2014, **820**, 9–22.
23 B. Dejaegher, L. Dhooghe, M. Goodarzi, S. Apers, L. Pieters and Y. Vander Heyden, *Anal. Chim. Acta*, 2011, **705**, 98–110.
24 A. de A. Gomes, M. R. Alcaraz, H. C. Goicoechea and M. C. U. Araújo, *Anal. Chim. Acta*, 2014, **811**, 13–22.
25 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.
26 S. Bose, A. Pal, R. Saharay and J. Nayak, *Pattern Recognition*, 2015, **48**, 2676–2684.
27 A. L. B. Brito, L. R. Brito, F. A. Honorato, M. J. C. Pontes and L. F. B. L. Pontes, *Food Res. Int.*, 2013, **51**, 924–928.
28 S. Fernandes and J. Bala, *International Journal of Signal Processing Systems*, 2013, **1**, 1–6.
29 L. Gan, W. Lv, X. Zhang and X. Meng, *Phys. Procedia*, 2012, **24**, 1689–1695.
30 A. Muñoz de la Peña, N. M. Díez, D. B. Gil, A. C. Olivieri and G. M. Escandar, *Anal. Chim. Acta*, 2006, **569**, 250–259.
31 J. M. M. Leitão, H. Gonçalves, C. Mendonça and J. C. G. E. da Silva, *Anal. Chim. Acta*, 2008, **628**, 143–154.
32 H. Yoshida, R. Leardi, K. Funatsu and K. Varmuza, *Anal. Chim. Acta*, 2001, **446**, 485–494.
33 E. B. Huerta, B. Duval and J. K. Hao, *Neurocomputing*, 2010, **73**, 2375–2383.
34 M. Goodarzi, W. Saeys, M. C. U. de Araujo, R. K. H. Galvão and Y. V. Heyden, *Eur. J. Pharm. Sci.*, 2014, **51**, 189–195.
35 F. S. L. Costa, P. P. Silva, C. L. M. Morais, T. D. Arantes, E. P. Milan, R. C. Theodoro and K. M. G. Lima, *Anal. Methods*, 2016, **8**, 7107–7115.
36 M. Bahram, R. Bro, C. Stedmon and A. Afkhami, *J. Chemom.*, 2006, **20**, 99–105.
37 R. Kennard and L. Stone, *Technometrics*, 1969, **11**, 137–148.
38 R. Bro and A. K. Smilde, *Anal. Methods*, 2014, **6**, 2812–2831.
39 W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D. L. Massart, S. Heuerding and F. Erni, *Anal. Chim. Acta*, 1996, **329**, 257–265.
40 J. McCall, *J. Comput. Appl. Math.*, 2005, **184**, 205–222.
41 S. F. C. Soares, A. A. Gomes, A. R. Galvão Filho, M. C. U. Araujo and R. K. H. Galvão, *TrAC, Trends Anal. Chem.*, 2013, **42**, 84–98.
42 R. Bro, *Chemom. Intell. Lab. Syst.*, 1997, **38**, 149–171.
43 M. M. Sena, M. G. Trevisan and R. J. Poppi, *Quim. Nova*, 2005, **28**, 910–920.
44 S. Liu, *Linear Algebra and its Applications*, 1999, **289**, 267–277.
45 R. G. Brereton and G. R. Lloyd, *J. Chemom.*, 2014, **28**, 213–225.
46 R. Bro, *J. Chemom.*, 1996, **10**, 47–61.
47 S. J. Dixon and R. G. Brereton, *Chemom. Intell. Lab. Syst.*, 2009, **95**, 1–17.
48 M. Chan, D. Lye, M. K. Win, A. Chow and T. Barkham, *Int. J. Infect. Dis.*, 2014, **26**, 110–115.

49 M. E. Cattana, M. A. Sosa, M. Fernández, F. Rojas, M. Mangiaterra and G. Giusiano, *Revista Iberoamericana de Micología*, 2014, **31**, 188–192.
50 V. Rivera, M. Gaviria, C. Muñoz-Cadavid, L. Cano and T. Naranjo, *Braz. J. Infect. Dis.*, 2015, **19**, 563–570.
51 F. Hagen, K. Khayhan, B. Theelen, A. Kolecka, I. Polacheck, E. Sionov, R. Falk, S. Parnmen, H. T. Lumbsch and T. Boekhout, *Fungal Genetics and Biology*, 2015, **78**, 16–48.
52 S. Marbumrung, K. Wongravee, V. Ruangpornvisuti, G. Tumcharern, T. Tuntulani and B. Tomapatanaget, *Sens. Actuators, B*, 2012, **171**, 969–975.
53 P. Cekan and S. T. Sigurdsson, *Biochem. Biophys. Res. Commun.*, 2012, **420**, 656–661.

CAPÍTULO 4

Identification of resistance in *Escherichia coli* and *Klebsiella pneumoniae* using E. E. M. fluorescence Spectroscopy and multivariate analysis

Fernanda S. L. Costa

Caio C. R. Bezerra — Renato M. Neto

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

Manuscrito submetido à Scientifc Reports**.**

Contribuição:

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

Identification of resistance in *Escherichia coli* and *Klebsiella pneumoniae* using excitation-emission matrix fluorescence spectroscopy and multivariate analysis

Fernanda S. L. Costa[1], Caio C. R. Bezerra[2], Renato M. Neto[2], Camilo L. M. Morais[3], Kássio M. G. Lima[1]

[1]Institute of Chemistry, Biological Chemistry and Chemometrics, Federal University of Rio Grande do Norte, Natal 59072-970, RN-Brazil

[2] Laboratory of Mycobateria, Department of Microbiology and Parasitology, Federal University of Rio Grande do Norte, Natal 59072-970, RN-Brazil

[3]Lancashire Teaching Hospitals NHS Trust, Fulwood, Preston PR2 9HT, United Kingdom

ABSTRACT

Klebsiella pneumoniae and *Escherichia coli* are part of the Enterobacteriaceae family, being common sources of community hospital infections and having high antimicrobial resistance. This resistance profile has become the main problem of public health infections. Determining whether a bacterium has resistance is critical to the correct treatment of the patient. Currently the method for determination of bacterial resistance used in laboratory routine is the antibiogram, whose time to obtain the results can vary from 1 to 3 days. An alternative method to perform this determination faster is excitation-emission matrix (EEM) fluorescence spectroscopy combined with multivariate classification methods. In this paper, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Support Vector Machines (SVM), coupled with dimensionality reduction and variable selection algorithms: Principal Component Analysis (PCA), Genetic Algorithm (GA), and the Successive Projections Algorithm (SPA) were used. The most satisfactory models achieved sensitivity and specificity rates of 100% for all classes, both for *E. coli* and for *K. pneumoniae*. This finding demonstrates that the proposed methodology has promising potential in routine analyzes, streamlining the results and increasing the chances of treatment efficiency.

Introduction

The Enterobacteriaceae family is one of the most clinically prominent bacteria groups. One of the main gram-negative pathogen is *Klebsiella pneumoniae* (*K. pneumoniae*), which causes opportunistic infections, such as pneumonia, sepsis and inflammation of the urinary tract[1]. Another gram-negative that compose the entereobacteriaceae family is *Escherichia coli*, which are not typically pathogenic to humans and have the ability to cause several diseases in different sites including gastrointestinal tract, the renal system and the central nervous system[2,3].

Antibiotic therapy induces the selection of resistant bacteria[4], which generate environmental and health hazards, and economical risk. Over the last decades, several bacterial strains have become progressively resistant to antimicrobial agents[5]. Bacteria may have natural or acquired resistance. Among the genetic variations that confer resistance in bacteria, the main ones are extended spectrum betalactamases[6] (ESBL), AmpC production, Carbapenemases production[7], KPC group and MBL group[5].

Currently, the standard detection method is culture-based, which is time-consuming and labor intensive, providing a slow detection[8]. Other methods can be used to obtain faster results, such as low cytometry[9], electrochemical detection[10], and polymerase chain reaction (PCR)[11]. Near infrared (NIR)[12], Raman[13] and Fourier transform infrared (FTIR) spectroscopy[14] have been also reported for these applications.

To identify if a strain of bacteria have resistance is necessary a test where an isolated culture is submitted at several types of antibiotics. The antibiotic sensitivity behavior of the isolated strains can be determined by disc diffusion method[15], such as Minimal Inhibitory Concentrations (MIC)[16] or Minimal Bactericidal Concentrations (MBC)[17].

Fluorescence spectroscopy has already been used in the detection[18], structural investigation[19,20] and in the construction of a DNA biosensor for *E. coli*[21]. Chemometric methods such as Linear Discriminant Analysis (LDA)[22], Quadratic Discriminant Analysis (QDA)[23] and Support Vector Machines (SVM)[24], coupled with the dimensionality reduction algorithm: Principal Component Analysis (PCA)[25,26]; and variable selection algorithms: Genetic Algorithm (GA)[27] and Successive Projections Algorithm (SPA)[28], tend to enhance the spectroscopic techniques[29,30,31].

This paper brings a new perspective for the differentiation of sensitive and resistant bacteria of *E. coli* and *K. pneumoniae* species using excitation-emission fluorescence spectroscopy allied to multivariate classification methods.

Results and discussion

Klebsiella pneumoniae samples belonged to three groups, which were named as: Control (ATCC 10031 - sensitive samples), resistant 1 (CCBH 6633 - samples that show resistance to carbapenems) and resistant 2 (CCBH 4955 KPC - samples resistant to carbapenems, cephalosporins, penicillin). Fig. 1 presents the mean excitation-emission fluorescence matrix (EEM) of *Klebsiella pneumoniae:* control (Fig. 1a), carbapenems resistant (Fig. 1b) and KPC (Fig. 1c), after removing Rayleigh and Raman scatterings (the excluded spectral regions were properly corrected by interpolation) and truncation done in the emission matrix.

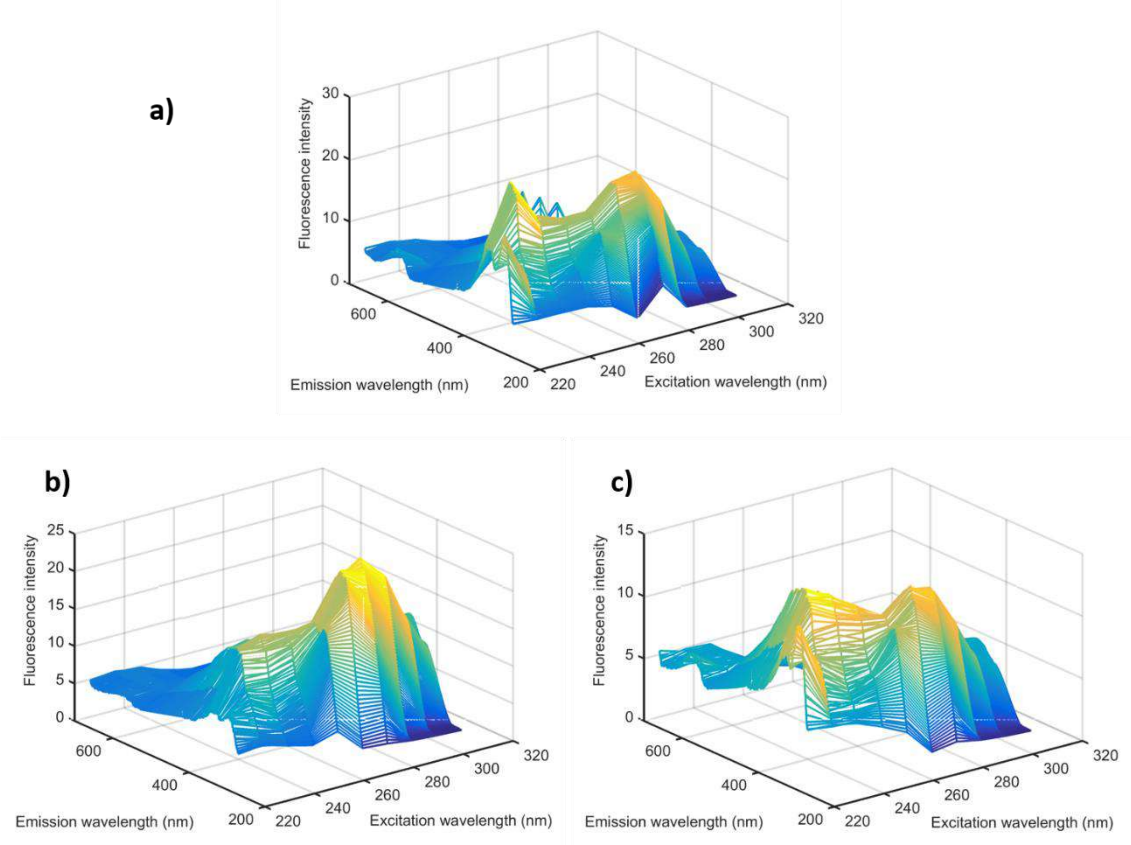


Figure 1: Excitation–emission molecular fluorescence matrix obtained for *Klebsiella pneumoniae:* sensitive (a), carbapenems resistant (b) and KPC (c). The Rayleigh and Raman scatterings were removed from the spectra.

The *E. coli* samples were composed of three groups, named control, resistant 1 and resistant 2. The control group was formed by sensitive *E. coli* samples (ATCC 25922). Resistance class 1 was composed of CCHB NDM samples, which have an enzyme called New Delhi metallo betalactamase, which attribute resistance to all beta-lactams, especially carbapenems. The resistant class 2 was formed by CCHB ampC

7018, which shows a type of beta-lactamase that causes hydrolysis of penicillins, monobactams, cephalosporins and cefoxitin. The EEM data obtained for *Escherichia coli*: sensitive (Fig. 2a), NDM (Fig. 2b) and ampC (Fig. 2c) are presents in Fig. 2, after spectral pre-processing.

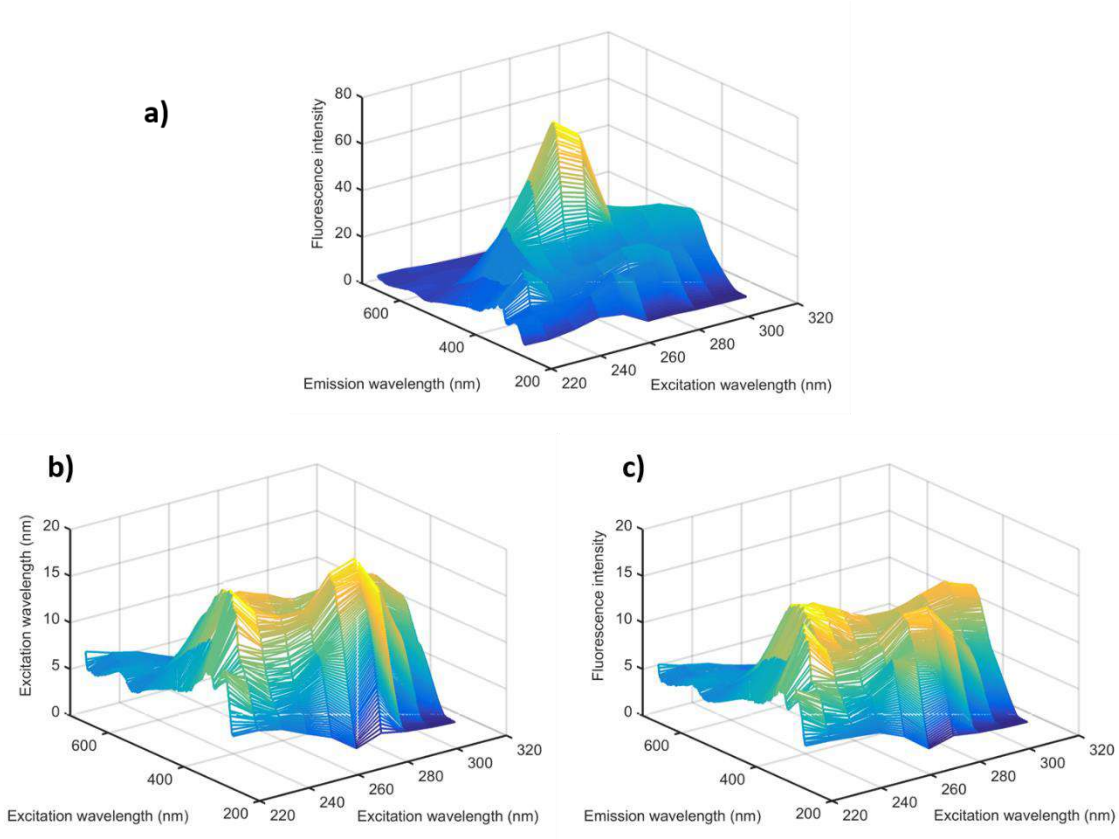


Figure 2: Excitation–emission molecular fluorescence matrix obtained for sensitive *Escherichia coli*: sensitive (a), NDM (b) and ampC (c). The Rayleigh and Raman scatterings were removed from the spectra.

As depicted in Figure 1 and 2, it is very difficult to distinguish the classes of sensitive and resistant bacteria only by their spectral profiles due to the great similarity between them. An exploratory analysis was performed using PCA with the unfolded data after spectral pre-processing. Figure 3 shows the PCA scores for *Klebsiella pneumoniae* data, built with 3 principal components (PCs).

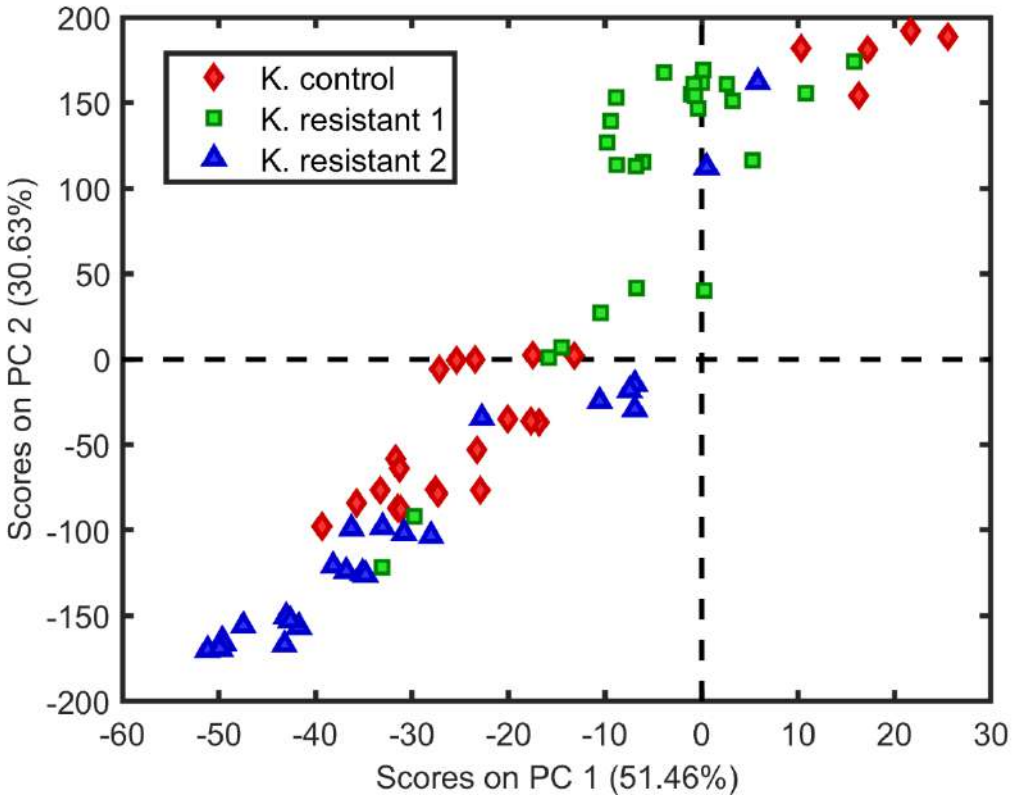


Figure 3: Scores on the first principal component *versus* the second principal component for classes *Klebsiella pneumoniae:* sensitive (♦), carbapenems resistant (■) and KPC (▲).

It can be observed that in the first component, which explains 51.5% of the explained variance, the control samples do not present separation in relation to the resistant *Klebsiella* samples. The second PC explains 30.6% of the data variance and also fails to distinguish between control and resistant classes. For the *Escherichia coli* spectra, we also constructed a PCA using 4 PCs, where the scores are shown in Fig 4.

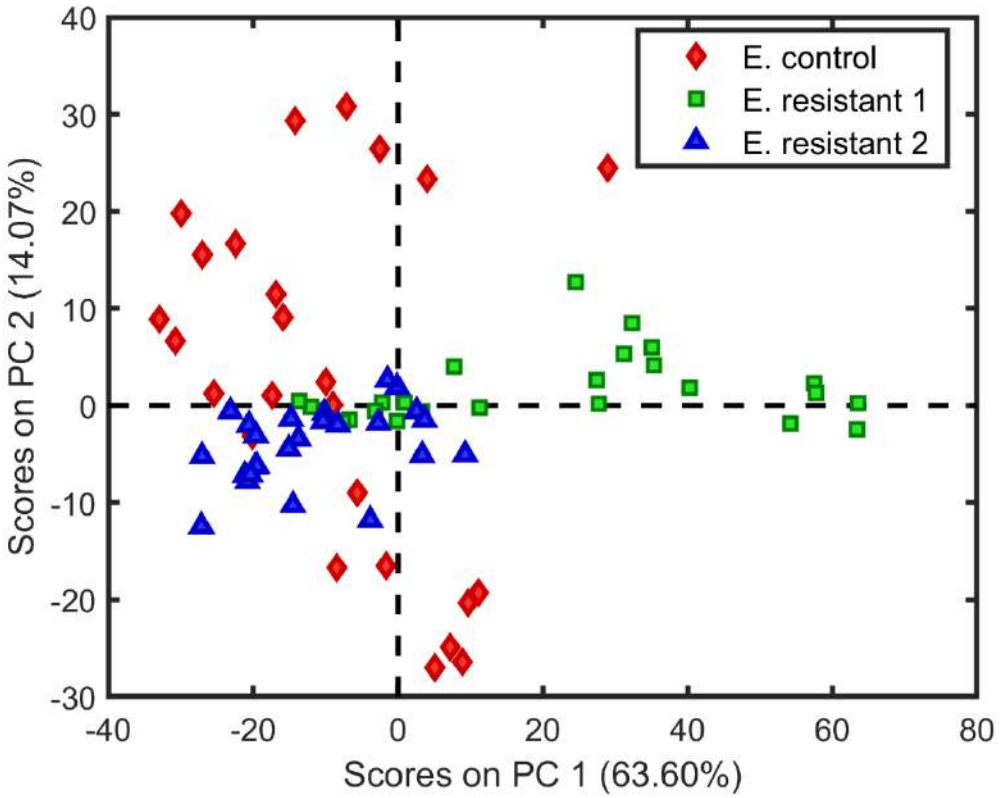


Figure 4: Scores on the first principal component *versus* the second principal component for classes *Escherichia coli*: sensitive (♦), NDM (■) and ampC (▲).

In Fig. 4, it is not possible to identify a separation between the three classes. Projecting the scores for the first PC, which explains 63.6% of the data variance, it is possible to observe a segregation between part of resistance group 1, in relation to the others samples. However, projecting in the second PC, which explains 14.1% of the data variance the data, the three classes cannot be distinguished. PCA results support that it is necessary to use multivariate classification algorithms that maximize the difference between the sensitive and resistant classes. A total of 75 samples were used for building the models, divided into three groups: calibration (45 samples), validation (15 samples) and prediction (15 samples). Table 1 shows the results of classification models built using the EEM fluorescence data for differentiating sensitive *Klebsiella pneumoniae* and resistants *Klebsiella pneumoniae*.

Model	Class	Calibration	Prediction
2D-LDA	Control	100.0	100.0
	Resistant 1 + 2	100.0	100.0
2D-PCA-LDA (5)[a]	Control	37.5	62.5
	Resistant 1 + 2	56.5	81.2
2D-PCA-QDA (5)[a]	Control	100.0	93.7
	Resistant 1 + 2	100.0	100.0
2D-PCA-SVM (5)[a]	Control	100.0	100.0
	Resistant 1 + 2	93.8	93.7
2D-LDA	Control	100.0	60.0
	Resistant 1	100.0	100.0
	Resistant 2	100.0	100.0
UPCA-QDA (4)[a]	Control	100.0	100.0
	Resistant 1	100.0	100.0
	Resistant 2	100.0	100.0
USPA-QDA (2)[b]	Control	100.0	100.0
	Resistant 1	93.3	100.0
	Resistant 2	100.0	80.0
UGA-QDA (7)[b]	Control	100.0	100.0
	Resistant 1	100.0	100.0
	Resistant 2	100.0	100.0
UPCA-SVM (4)[a]	Control	100.0	60.0
	Resistant 1	100.0	100.0
	Resistant 2	100.0	100.0
USPA-SVM (2)[b]	Control	73.3	100.0
	Resistant 1	80.0	100.0
	Resistant 2	86.7	80.0
UGA-SVM (12)[b]	Control	100.0	100.0
	Resistant 1	100.0	100.0
	Resistant 2	100.0	100.0

[a] Number of principal components. [b] Number of selected variables.

Table 1: Results obtained for classification models (2D-LDA, 2D-PCA-LDA, 2D-PCA-QDA, 2D-PCA-SVM, UPCA-QDA/SVM, USPA-QDA/SVM and UGA-QDA/SVM) for sensitive *Klebsiella pneumoniae* and resistant.

Initially, models were constructed comparing the class of *Klebsiella* sensitive and that of resistant. For built this last group, samples of two resistant classes are combined. Among these models, the ones that presented the most satisfactory results were 2D-LDA and 2D-PCA-QDA, which obtained 100.0% calibration accuracy and classification rates above 93% in all classes in the prediction set. Models were constructed using the three classes of samples, applying QDA and SVM, coupled to dimensionality reduction algorithms (PCA, SPA and GA) in the unfolded data. With the exception of the USPA-QDA, UPCA-SVM and USPA-SVM models, all others presented satisfactory results, with 100% accuracy, both in calibration and in prediction, for the three classes.

Model	Class	Calibration	Prediction
2D-LDA	Control	100.0	87.5
	Resistant 1 + 2	100.0	100.0
2D-PCA-LDA (3)[a]	Control	100.0	100.0
	Resistant 1 + 2	100.0	100.0
2D-PCA-QDA (5)[a]	Control	100.0	100.0
	Resistant 1 + 2	100.0	100.0
2D-PCA-SVM (5)[a]	Control	93.7	100.0
	Resistant 1 + 2	100.0	100.0
2D-LDA	Control	80.0	60.0
	Resistant 1	80.0	80.0
	Resistant 2	100.	100.0
UPCA-QDA (4)[a]	Control	100.0	100.0
	Resistant 1	100.0	100.0
	Resistant 2	100.0	100.0
USPA-QDA (2)[b]	Control	100.0	100.0
	Resistant 1	100.0	100.0
	Resistant 2	100.0	80.0
UGA-QDA (7)[b]	Control	100.0	100.0
	Resistant 1	100.0	100.0
	Resistant 2	100.0	100.0
UPCA-SVM (4)[a]	Control	93.3	60.0
	Resistant 1	100.0	100.0
	Resistant 2	100.0	100.0
USPA-SVM (2)[b]	Control	100.0	100.0
	Resistant 1	100.0	100.0
	Resistant 2	100.0	80.0
UGA-SVM (5)[b]	Control	100.0	100.0
	Resistant 1	100.0	100.0
	Resistant 2	100.0	100.0

[a] Number of principal components. [b] Number of selected variables.

Table 2: Results obtained for classification models (2D-LDA, 2D-PCA-LDA-2D, 2D-PCA-QDA, 2D-PCA-SVM, UPCA-QDA/SVM, USPA-QDA/SVM and UGA-QDA/SVM) for sensitive *Escherichia coli* and resistant.

The same strategy was applied to the *E. coli* samples, the results are shown on Table 2. The first models were created with only two classes: *E. coli* sensitive and the combined resistant samples. The results were satisfactory, mainly for 2D-PCA-LDA and 2D-PCA-QDA, which obtained 100.0% accuracy in both classes, both in calibration and in prediction. The models constructed with the three classes presented satisfactory results in the classification. Unfolded models (UPCA-QDA and UGA-QDA) also resulted in 100.0% accuracy in calibration and prediction of the three classes in this comparison.

	Stage performance features							
	UPCA-QDA			UGA-SVM			2D-LDA	
	Cont.	Res. 1	Res. 2	Cont.	Res. 1	Res. 2	Cont.	Res. 1+2
Accuracy	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Sensitivity	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Specificity	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
F-score	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 3: Quality performance values for the three classification methods (UPCA-QDA, UGA-SVM and 2D-LDA with 2 classes) by molecular fluorescence spectroscopy for each category of *Klebsiella pneumoniae.*

Table 3 presents the validation results of the optimized models (UPCA-QDA, UGA-SVM and 2D-LDA) for each classification category of *Klebsiella pneumoniae*. The models that considered three classes (UPCA-QDA, UGA-SVM) showed promising results, with 100.0% sensitivity and specificity rates. Another notable result is the 2D-LDA model, built with only two classes, achieved similar results, with the same 100.0% sensitivity and specificity rates. The parameters accuracy and F-score were all equal to 100.0%, showing that those models are valid to distinguish between different groups of *Klebsiella pneumoniae* bacteria.

	Stage performance features							
	UPCA-QDA			UGA-SVM			2D-PCA-QDA	
	Cont.	Res. 1	Res. 2	Cont.	Res. 1	Res. 2	Cont.	Res. 1+2
Accuracy	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Sensitivity	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Specificity	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
F-score	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 4: Quality performance values of three classification methods (UPCA-QDA, UGA-SVM and 2D-PCA-QDA) by molecular fluorescence spectroscopy for each category of *Escherichia coli.*

The validation results of the optimized models UPCA-QDA, UGA-SVM and 2D-PCA-QDA for the *E. coli* are illustrated in Table 4. The sensitivity and specificity rates for these models are 100.0% for all the analyzed classes. The accuracy and F-score values also reinforce the model efficiency.

According to the literature, bacterial resistance is usually associated with the ability of bacteria to modify their cellular structure and induce them to produce

substances that neutralize the action of antibacterial agents. Satisfactory results from the models using EEM fluorescence data, for the *E. coli* and *K. pneumoniae* bacteria, demonstrate the sensitivity of the technique in detecting variations in the nuclear content of the cells and in the structure of the membranes itself. As reported by Opačić *et al.*[19], who used fluorescence spectroscopy on structural investigation of the transmembrane C domain of the mannitol permease from *Escherichia coli*, the results showed that the technique was capable to differentiated the structure of EII^{mtl} from structure of a IIC protein transporting diacetylchitobiose. Additionally, Romantsov *et. al*.[20] used dynamic data obtained by fluorescence correlation spectroscopy to extract structural information on isolated nucleoids, besides the evaluation of the characteristic size of the structural units in terms of the DNA length and estimation of their spatial dimensions.

Methods

Sample preparation

The samples used were: *E. coli* ATCC 25922 - Standard strain, *E. coli* CCHB NDM +, *E. coli* CCHB ampC 7018, *K. pneumoniae* ATCC 1003, *K. pneumoniae* CCBH 4955 KPC and *K. pneumoniae* CCBH 6633 resistant to Carbapenems. The CCBH strains were obtained from the Laboratory of Hospital Infection (LAPIH - Fiocruz/RJ). The ATCC strains belong to LABMIC / DMP – UFRN. Initially the pure samples were pealed in a BHI broth, then kept in the oven for 24 hours at 38 °C, so that the bacteria multiplied. The sample was then pealed on a petri dish, which was also kept in the oven for 24 hours. Finally, a bacterial mass corresponding to approximately 10^6 colony forming units (CFU) was transferred from culture medium to falcon tube with 2 mL of phosphate buffer solution (1 mol L^{-1}).

EEM Fluorescence spectroscopy

The excitation/emission fluorescence data were acquired in the wavelength range of 220–310 nm for excitation and 270–900 nm for emission, with steps of 10 and

1 nm for excitation and emission, respectively. A RF-5301 Shimadzu spectrofluorometer with a 0.5 mm quartz cuvette was used. The excitation and emission slits were set at 3 and 5 nm, respectively, the speed scan was set to super mode; the photomultiplier tube was set to the medium level and a cell with a fiber optic reflectance probe was used. Five replicates of the concentrations at $1x10^6$ UFC/mL, $5x10^5$ UFC/mL, $1,3x10^5$ UFC/mL, $6,3x10^4$ UFC/mL and $3,1x10^4$ UFC/mL were performed.

Data analysis

Chemometrics procedure and software

Spectral pre-processing and multivariate classification models were built using MATLAB R2011a software (The MathWorks, Natick, USA), and the PLS Toolbox 7.9.3 package (Eigenvector Research, Inc., Wenatchee, USA). A spectral range between 220-310 nm for excitation and 270–900 nm for emission was used for model construction, with steps of 10 and 1 nm used for excitation and emission, respectively. This resulted in a data matrix size of 10×651 for each sample. The spectral pre-processing was composed by a cut in the region of 270-659 nm in the emission range, and by removing Rayleigh and Raman scatterings using the 'EEMscat' algorithm.[32]

The following classification methods were utilized: two-dimensional linear discriminant analysis (2D-LDA)[33], two-dimensional principal component analysis with linear discriminant analysis (2D-PCA-LDA)[34], quadratic discriminant analysis (2D-PCA-QDA)[34], and support vector machines (2D-PCA-SVM)[34]. In addition to these, first-order classification using LDA, QDA and SVM were used in conjunction with the output from the dimensionality reduction algorithms: PCA, GA and SPA.

For the construction of classification models, the samples were divided into calibration (60%), validation (20%) and prediction (20%) sets using the Kennard-Stone (KS) sample selection algorithm[35]. The proposed models were evaluated by calculating some quality parameters such as accuracy, sensitivity, specificity and F-score.

To statistically evaluate the classification models, calculations of sensitivity and specificity were performed using the test samples as important quality measures of model accuracy. Both parameters have a maximum value of 100 and a minimum of 0, and are obtained as follows:

$$\text{Sensitivity}\ (\%) = \frac{\text{TP}}{\text{TP+FN}} \times 100 \quad (1)$$

$$\text{Specificity}\ (\%) = \frac{\text{TN}}{\text{TN+FP}} \times 100 \quad (2)$$

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

Also, the models were evaluated using the area under the curve (AUC) and F-score. The AUC is the area under the receiver operating characteristics conditions (ROC) curve, and the F-score is a measurement of the model accuracy defined by:

$$F - score = \frac{2 \times SENS \times SPEC}{SENS + SPEC} \quad (3)$$

where SENS stands for sensitivity; and SPEC stands for specificity.

Conclusion

The present study demonstrates the ability of EEM fluorescence spectroscopy associated with multivariate classification in differentiating classes of susceptible and resistant bacteria of the species *E. coli* and *K. pneumoniae*. The most satisfactory models for the classification of *K. pneumoniae* were UPCA-QDA, UGA-SVM and 2D-LDA, which presented 100% accuracy rates for all classes. For the *E. coli* data, the UPCA-QDA, UGA-SVM and 2D-PCA-QDA models were the best, having 100% predictive performance for the classification of all groups. All these models obtained a sensitivity and specificity rate of 100%. This paper suggest a new alternative in the detection of bacterial resistance, through a methodology that is faster than traditional methods of analysis, simplifying the diagnosis, and increasing the chances of recovery of the patients.

References

1. Susanto, W., Kong, K.-H., Hua, K.-F., Wu, S.-H. & Lam, Y. Synthesis of the trisaccharide repeating unit of capsular polysaccharide from klebsiella pneumoniae. *Tetrahedron Lett.* **60**, 288–291 (2019).

2. Kumar, H. & Mandal, P. K. Synthetic routes toward pentasaccharide repeating unit corresponding to the o-antigen of escherichia coli o181. *Tetrahedron Lett.* **60**, 860–863 (2019).

3. Li, X. *et al.* Disruption of blood-brain barrier by an escherichia coli isolated from canine septicemia and meningoencephali- tis. *Comp. Immunol. Microbiol. Infect. Dis.* **63**, 44–50 (2019).

4. Mukherjee, P. *et al.* Studies on formulation of a combination heat killed immunogen from diarrheagenic escherichia coli and vibrio cholerae in ritard model. *Microbes infection* (2019).

5. Alharbi, N. S. *et al.* Prevalence of escherichia coli strains resistance to antibiotics in wound infections and raw milk. *Saudi J. Biol. Sci.* (2018).

6. Rodrigues, C. *et al.* Description of klebsiella africanensis sp. nov., klebsiella variicola subsp. tropicalensis subsp. nov. and klebsiella variicola subsp. variicola subsp. nov. *Res. microbiology* (2019).

7. López-Camacho, E. *et al.* Meropenem heteroresistance in clinical isolates of oxa-48–producing klebsiella pneumoniae. *Diagn. microbiology infectious disease* **93**, 162–166 (2019).

8. Li, J., Li, B. & Liu, M. One-step synthesis of mannose-modified polyethyleneimine copolymer particles as fluorescent probes for the detection of escherichia coli. *Sensors Actuators B: Chem.* **280**, 171–176 (2019).

9. Wang, X. *et al.* A metabolomics-based method for studying the effect of yfcc gene in escherichia coli on metabolism. *Anal. biochemistry* **451**, 48–55 (2014).

10. Ya-li, F. *et al.* Isolation and characterization of an electrochemically active and cyanide-degrading bacterium isolated from a microbial fuel cell. *RSC Adv.* **4**, 36458–36463 (2014).

11. Fonseca, E. L. *et al.* A one-step multiplex pcr to identify klebsiella pneumoniae, klebsiella variicola, and klebsiella quasipneumoniae in the clinical routine. *Diagn. microbiology infectious disease* **87**, 315–317 (2017).

12. Siripatrawan, U., Makino, Y., Kawagoe, Y. & Oshita, S. Near infrared spectroscopy integrated with chemometrics for rapid detection of e. coli atcc 25922 and e. coli k12. *Sensors Actuators B: Chem.* **148**, 366–370 (2010).

13. Kusić, D., Rösch, P. & Popp, J. Fast label-free detection of legionella spp. in biofilms by applying immunomagnetic beads and Raman spectroscopy. *Syst. applied*

microbiology **39**, 132–140 (2016).

14. Dieckmann, R. *et al.* Rapid characterisation of klebsiella oxytoca isolates from contaminated liquid hand soap using mass spectrometry, FTIR and Raman spectroscopy. *Faraday discussions* **187**, 353–375 (2016).

15. Alam, M. Z., Aqil, F., Ahmad, I. & Ahmad, S. Incidence and transferability of antibiotic resistance in the enteric bacteria isolated from hospital wastewater. *Braz. J. Microbiol.* **44**, 799–806 (2013).

16. Olorunmola, F. O., Kolawole, D. O. & Lamikanra, A. Antibiotic resistance and virulence properties in escherichia coli strains from cases of urinary tract infections. *Afr. journal infectious diseases* **7**, 1–7 (2013).

17. Levison, M. E. & Levison, J. H. Pharmacokinetics and pharmacodynamics of antibacterial agents. *Infect. Dis. Clin.* **23**, 791–815 (2009).

18. Siripatrawan, U., Makino, Y., Kawagoe, Y. & Oshita, S. Rapid detection of escherichia coli contamination in packaged fresh spinach using hyperspectral imaging. *Talanta* **85**, 276–281 (2011).

19. Opačić, M., Hesp, B. H., Fusetti, F., Dijkstra, B. W. & Broos, J. Structural investigation of the transmembrane c domain of the mannitol permease from escherichia coli using 5-ftrp fluorescence spectroscopy. *Biochimica et Biophys. Acta (BBA)-Biomembranes* **1818**, 861–868 (2012).

20. Romantsov, T., Fishov, I. & Krichevsky, O. Internal structure and dynamics of isolated escherichia coli nucleoids assessed by fluorescence correlation spectroscopy. *Biophys. journal* **92**, 2875–2884 (2007).

21. Sun, J. *et al.* Dna biosensor-based on fluorescence detection of e. coli o157: H7 by au@ ag nanorods. *Biosens. Bioelectron.* **70**, 239–245 (2015).

22. Liu, J., Yu, G. & Liu, Y. Graph-based sparse linear discriminant analysis for high-dimensional classification. *J. Multivar. Analysis* **171**, 250–269 (2019).

23. Gaynanova, I. & Wang, T. Sparse quadratic classification rules via linear dimension reduction. *J. Multivar. Analysis* **169**, 278–299 (2019).

24. Tang, L., Tian, Y. & Pardalos, P. M. A novel perspective on multiclass classification: Regular simplex support vector machine. *Inf. Sci.* **480**, 324–338 (2019).

25. Tran, N. M., Burdejová, P., Ospienko, M. & Härdle, W. K. Principal component analysis in an asymmetric norm. *J. Multivar. Analysis* **171**, 1–21 (2019).

26. Feng, Y., Zhao, T., Wang, M. & Owen, D. Characterising particle packings by principal component analysis. *Comput. Methods Appl. Mech. Eng.* **340**, 70–89 (2018).

27. Islam, M. L., Shatabda, S., Rashid, M. A., Khan, M. G. & Rahman, M. S. Protein structure prediction from inaccurate and sparse nmr data using an enhanced genetic algorithm. *Comput. biology chemistry* **79**, 6–15 (2019).

28. Milanez, K. D. T. M., Nóbrega, T. C. A., Nascimento, D. S., Galvão, R. K. H. & Pontes, M. J. C. Selection of robust variables for transfer of classification models employing the successive projections algorithm. *Anal. chimica acta* **984**, 76–85 (2017).

29. Morais, C. L., Lima, K. M. & Martin, F. L. Uncertainty estimation and misclassification probability for classification models based on discriminant analysis and support vector machines. *Anal. chimica acta* **1063**, 40–46 (2019).

30. Costa, F. S. *et al.* Attenuated total reflection fourier transform-infrared (ATR-FTIR) spectroscopy as a new technology for discrimination between cryptococcus neoformans and cryptococcus gattii. *Anal. Methods* **8**, 7107–7115 (2016).

31. Silva, H. F. *et al.* On the synergy between silver nanoparticles and doxycycline towards the inhibition of staphylococcus aureus growth. *RSC advances* **8**, 23578–23584 (2018).

32. Bahram, M., Bro, R., Stedmon, C. & Afkhami, A. Handling of rayleigh and raman scatter for parafac modeling of fluorescence data using interpolation. *A J. Chemom. Soc.* **20**, 99–105 (2006).

33. da Silva, A. C. *et al.* Two-dimensional linear discriminant analysis for classification of three-way chemical data. *Anal. chimica acta* **938**, 53–62 (2016).

34. Morais, C. L. & Lima, K. M. Comparing unfolded and two-dimensional discriminant analysis and support vector machines for classification of EEM data. *Chemom. Intell. Lab. Syst.* **170**, 1–12 (2017).

35. Kennard, R. W. & Stone, L. A. Computer aided design of experiments. *Technometrics* **11**, 137–148 (1969).

Acknowledgements

F.S.L.C. would like to thank CAPES/PPGQ/UFRN for financial and scientific support. C.C.R.B. would like to thank CNPQ for financial support. K.M.G.L. would like to thank to CNPq (Grant: 303733/2017−9) for financial support.

Authors contribution

F.S.L.C. was responsible for the construction of chemometric models and wrote the manuscript. C.L.M.M. revised the manuscript and multivariate analysis. C.C.R.B and R.M.N were responsible for experimental section. K.M.G.L. supervised the project and revised the manuscript.

Additional information

Competing financial interests: All authors declare no competing financial interest.

CAPÍTULO 5 - CONCLUSÃO E PERSPECTIVAS

Nesta tese, foi proposta a aplicação de técnicas espectroscópicas em conjunto com análise multivariada para a diferenciação de fungos *Cryptococcus gattii* e *Cryptococcus neoformans* e para a identificação de resistência bacteriana, em espécies de *Escherichia coli* e *Klebsiella pneumoniae.*

O primeiro estudo demonstrou que a espectroscopia FTIR em combinação com análise multivariada tem o potencial para distinguir *C.gatti* e *C. neoformans*, através da identificação de biomarcadores, que podem facilitar a discriminação entre as duas espécies, reduzindo o tempo de diagnóstico. O que é de importância fundamental para o tratamento adequado dos pacientes, especialmente para imunocomprometidos. Além disso, pode reduzir os custos dos serviços de saúde. Entre os modelos utilizados, o GA-QDA mostrou resultados mais satisfatórios, apresentando sensibilidade para ambas as classes igual ou superior a ensaios convencionais. Desta maneira, a metodologia proposta apresenta uma alternativa eficiente, de rapidez na análise, que exige uma quantidade pequena de amostra. Característica fundamental em análises biológicas, uma vez que o volume de fluidos biológicos é normalmente limitado, como neste caso o líquor.

A espectroscopia de fluorescência EEM em combinação com análise multivariada também demonstrou grande potencial para diferenciar as culturas de *C. gatti* e *C. neoformans*, com resultados ainda mais promissores que os encontrados com a espectroscopia FTIR. Entre os modelos utilizados, o UGA-LDA foi a mais satisfatório, pois apresentou valores de sensibilidade para ambas as classes (*C. neoformans* e *C. gatti*) iguais e em alguns casos, até maiores que alguns dos métodos clássicos de identificação desses fungos. A metodologia proposta tem grande potencial de ser empregada diretamente nas amostras biológicas de pacientes com criptococose, o que tornaria o diagnóstico mais rápido, reduzindo os custos processuais e mantendo os padrões de sensibilidade e especificidade das análises de rotina.

A técnica de florescência EEM associada à algoritmos de classificação multivariada também foi empregada na identificação de resistência bacteriana para espécies *E. coli* e *K. pneumoniae*. Os resultados foram bastante promissores. Para a classificação de *K. pneumoniae,* os melhores modelos foram UGA-SVM e 2D-LDA, que apresentaram 100% de acertos em todas as classes, além de sensibilidade e especificidade também de 100%. Para os dados de *E. coli*, os modelos UPCA-QDA, UGA-SVM e PCA-QDA_2D, apresentaram 100% de desempenho na classificação de todos os grupos. Todos esses modelos obtiveram uma taxa de sensibilidade e especificidade de 100%. Este estudo sugere uma nova alternativa na detecção de

resistência bacteriana, por meio de uma metodologia mais rápida que os métodos tradicionais de análise. O que pode conferir mais agilidade ao diagnóstico e aumentar as chances de recuperação dos pacientes.

Os estudos propostos são indicativos de que técnicas espectroscópicas e a classificação multivariada podem ser ferramentas poderosas nas análises clínicas. Podendo ser desenvolvidos novos modelos para diagnóstico de outras doenças fúngicas sistêmicas, se tornando uma ferramenta promissora na micologia médica. Assim, como utilizadas para a construção de um banco de dados de identificação de espécies e de resistência bacteriana. Auxiliando os profissionais da área da saúde na identificação do diagnóstico e possibilitando mais chances de cura aos pacientes, sobretudo no caso dos imunocomprometidos.

APÊNDICE A

Variable selection with a support vector machine for discriminating *Cryptococcus* fungal species based on ATR-FTIR spectroscopy

Camilo L. M. Morais

Fernanda S. L. Costa Kássio M. G. Lima

Anal. Methods, 2017, 9, 2964-2970.

Contribuição:

- Realizei a aquisição espectral;
- Ajudei na escrita da primeira versão do manuscrito.

Fernanda S. L. Costa **Prof. Kássio M. G. Lima**

Analytical Methods

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/c7ay00428a

Variable selection with a support vector machine for discriminating *Cryptococcus* fungal species based on ATR-FTIR spectroscopy

Camilo L. M. Morais, Fernanda S. L. Costa and Kássio M. G. Lima *

Received 17th February 2017
Accepted 11th April 2017

DOI: 10.1039/c7ay00428a

rsc.li/methods

Variable selection with supervised classification is currently an important tool for discriminating biological samples. In this paper, 15 supervised classification algorithms based on a support vector machine (SVM) were applied to discriminate *Cryptococcus neoformans* and *Cryptococcus gattii* fungal species using ATR-FTIR spectroscopy. These two fungal species of the *Cryptococcus* genus are the etiological agents of Cryptococcosis, which is an opportunistic or primary fungal infection with global distribution. This disease is potentially fatal, especially for immunocompromised patients, like those suffering from AIDS. The multivariate classification algorithms tested were based on principal component analysis (PCA), successive projections algorithm (SPA) and genetic algorithm (GA) as data reduction and variable selection methods, being coupled to a SVM with different kernel functions (linear, quadratic, 3rd order polynomial, radial basis function, and multilayer perceptron). Some of these algorithms achieved very successful classification rates for discriminating fungal species, with accuracy, sensitivity, and specificity equal to 100% using both SPA-SVM-polynomial and GA-SVM-polynomial algorithms. These results show the potential of such techniques coupled to ATR-FTIR spectroscopy as a rapid and non-destructive tool for classifying these fungal species.

Introduction

Cryptococcosis is an opportunistic fungal infection caused by inhaling basidiospores[1] or dissected yeasts present in the environment, causing an infection of the central nervous system which affects immunocompromised individuals, including AIDS patients and organ transplant recipients or other patients receiving immunosuppressive drugs.[2,3] This disease affects the respiratory tract of the host causing severe pneumonia and respiratory insufficiency and is responsible for the majority of worldwide deaths from HIV-related fungal infections.[1,3]

The main etiologic agents of Cryptococcosis in humans are two species, namely *Cryptococcus neoformans* (serotypes A, D and AD) and *Cryptococcus gattii* (serotypes B and C), which differ in their epidemiology, host range, virulence, antifungal susceptibility and geographic distribution.[1] *Cryptococcus gattii* is a primary pathogen which infects immunocompetent and healthy individuals, having predilection for the lungs.[4] On the other hand, *Cryptococcus neoformans* has predilection for the central nervous system and mainly infects immunosuppressed patients mostly having HIV/AIDS.[5] *Cryptococcus gattii* is responsible for many infection cases in the Pacific Northwest of the United States.[6] This high virulence occurs due to an unusual tubular mitochondrial morphology caused by mitochondrial fusions to enhance the repair of mitochondrial DNA damage from oxidative stress within the phagosome.[4] In addition, *Cryptococcus gattii* has two metabolites of acetoin and dihydroxyacetone which potentially produce less pro-inflammatory response than those of *Cryptococcus neoformans*. This facilitates fungal survival and local multiplication causing more cryptococcomas.[4] There are some morphological features that are specifically associated with each of the two species such as texture, pigmentation produced by their colonies, and yeast form.[1,7] However, it is still more reliable to distinguish them by their growth phenotype on certain media formulations based on their biochemical differences.[8]

Cryptococcosis is a treatable disease, however its effects are devastating to the patients, resulting in death or central nervous system dysfunction unless the condition is diagnosed and treated at the time of onset.[1] Currently, the techniques used in the identification of these pathogens are direct microscopic examination and molecular methods such as DNA hybridization and PCR-based methods (particularly nested, multiplex and real time PCR).[9,10] These methods provide both high sensitivity and specificity; however, most have some limitations that may hinder the final diagnosis, further requiring several days to detect and identify the microorganisms.[10]

In order to improve the ability to properly control fungal infections in humans, early identification of the pathogen is necessary, since they have different responses to antifungal

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

treatments.[11] In this sense, Fourier transform infrared spectroscopy (FTIR) has been standing out in the past few years in the microbiological area,[12,13] because it provides a large amount of information about typical absorption bands for each functional group, providing a spectroscopic fingerprint of the total biochemical and structural composition unique for each molecule.[14] The mid-IR region at 1800–900 cm^{-1} contains the fundamental vibrational modes of key chemical bonds of intracellular mechanisms corresponding to the biochemical fingerprint of the material under study, therefore being called the biofingerprint region.[14] In addition, FT-IR has the advantages of being rapid and non-destructive, using small sample sizes, and requiring an easy sample preparation.[14]

In attenuated total reflection – Fourier transform infrared (ATR-FTIR), the ATR module enhances the signal by passing the IR beam through the sample, taking advantage of several internal reflections with the crystal.[15,16] Such reflections generate an evanescent wave that penetrates the material to a depth between 0.5 and 2 μm.[16] ATR-FTIR has been very effective in analyzing biological samples, as demonstrated in analyzing diverse types of cancer,[17] insects,[18,19] and bacteria;[20,21] as well as to monitor plant health in a controlled[22] and natural[23] environment.

Good computation tools are required to follow the advances in spectroscopy techniques applied to biological samples. These tools enable building classification models for screening and diagnosis methods, which is a common task in biospectroscopy applications.[14] A very powerful multivariate classification technique is the support vector machine (SVM).[24] SVMs are binary classifiers that work by finding a classification hyperplane which separates two classes or objects providing the largest margin of separation.[25] A key advantage of SVMs over most other classical classification methods is that an SVM is capable of classifying nonlinearly separable data.[25] This makes its performance superior to linear-dependent classification methods, such as linear discriminant analysis (LDA).[25] The kernel function is responsible for transforming the data into a different feature space (linear, quadratic, and polynomial, among others) changing the classification ability of SVMs.[26] SVM algorithm applications in biological data include classifying low-grade cervical cytology;[27] breast cancer diagnosis;[28] ovarian cancer identification;[29] analysis of dengue infection;[30] and classifying *Candida* fungi.[31]

Data reduction and variable selection methods can be coupled with the SVM algorithm in order to speed up computational analysis. A common method of data reduction is principal component analysis (PCA).[32] PCA reduces the original data to a few principal components (PCs) having most of the original explained variance;[32] and the scores on each PC can be used as classification variables for the SVM. Among the variable selection methods, successive projections algorithm (SPA)[33] and genetic algorithm (GA)[34] have found many applications in biological data.[17,18] SPA reduces the original data to few variables by minimizing its collinearity,[33] while GA reduces the data following an evolutionary process where the fittest set of variables is chosen.[34] Both algorithms maintain the original data dimension, being consequently used as a tool to search for specific molecular fragments, also called biomarkers.[14]

In this paper, we have applied different types of algorithms based on PCA-SVM, SPA-SVM, and GA-SVM with different kernel functions (linear, quadratic, 3rd order polynomial, radial basis function, and multilayer perceptron) as a rapid and non-destructive method to discriminate *Cryptococcus gattii* and *Cryptococcus neoformans* fungal species based on ATR-FTIR spectroscopy. In addition, a tentative assignment of possible biomarkers involved in differentiating these fungal species is performed.

Methods

Sample preparation

In this study, 28 isolated samples from UFPI (Universidade Federal do Piauí); IMT/SP (Instituto de Medicina Tropical de São Paulo), Veterinary Hospital-UNESP, campus Botucatu (SP), FioCruz mycological collection and recently isolated fungus from Giselda Trigueiro Hospital (Natal/RN/Brazil) were used. Genotyping of the isolated fungus in culture on Sabouraud Agar with Chloramphenicol (50 mg L^{-1}) was done at the Institute of Tropical Medicine of RN at UFRN, using PCR-RFLP of the URA5 gene as previously described,[4] under approval of the ethics committee, number 51050415.6.0000.5537.

These fungi were incubated for 48 hours at a temperature of 30 °C until satisfactory growth is achieved. Yeast cells were inactivated for biosafety handling in the spectroscopy equipment by placing some yeast colonies in 1.0 mL of paraformaldehyde solution at 4% plus phosphate buffer (1 mol L^{-1}) v/v, and in 1.5 mL eppendorf tubes for cell attachment to inactivate yeast cells. The final solution was added to 28 tubes with 28 different *Cryptococcus* isolates. After 3 hours at room temperature, the tubes with cells were placed under refrigeration at −20 °C until the next step. For spectra reading, the tubes were put at room temperature until defrosted, and then centrifuged for 10 minutes at 5000*g* for cell precipitation. The supernatant was removed and the cells were washed with 1.0 mL of sterile saline solution (0.95% w/v). The tubes were maintained at 4 °C until spectroscopy reading.

ATR-FTIR spectroscopy

The ATR-FTIR measurements ($n = 280$, 10 replicates of each one of the 28 *C. neoformans* ($n = 14$) and *C. gattii* ($n = 14$) samples) were recorded using a Bruker VERTEX 70 FTIR spectrometer (Bruker Optics Ltd., UK) with Helios ATR attachment containing a diamond crystal internal reflective element and a 45 incidence angle of IR beam. The ATR-FTIR spectra of fungal samples were acquired in the range of 400–4000 cm^{-1} with a resolution of 4 cm^{-1}. Each spectrum was collected at 16 scans in the absorbance mode. Approximately 50 μL of each sample was applied to the ATR crystal immediately following collection of each background. A small piece of aluminum foil was placed on the sample to ensure that no air bubbles were trapped on the crystal surface and to improve the signal-to-noise ratio of the spectra.[35] The ATR crystal was cleaned with 70% v/v alcohol and a new background was collected prior to the analysis of a new sample

and compared to the first background to ensure no interference in the sample signal.

Computational analysis

Computational analysis was performed within a Matlab R2012b environment (MathWorks, USA) by using PLS Toolbox version 7.9.3 (Eigenvector Research, Inc., USA) and homemade algorithms. Raw spectral data were pre-processed by cutting the region of 1800–900 cm^{-1}, followed by normalization to the amide I peak (~1650 cm^{-1})[15] and baseline correction.

Samples for training (n = 196), validation (n = 42), and prediction (n = 42) sets were selected using the Kennard–Stone uniform sampling selection algorithm.[36] The training set was used to build the classification models, and the validation set to evaluate its internal performance. The prediction set was only used in the final classification evaluation.

The pre-processed spectra were utilized in the classification algorithms as follows: first, data reduction was performed by means of PCA, SPA, and GA; utilizing PCA, the scores on the first PCs were utilized as classification variables for the SVM; whereas during SPA and GA, the selected variables having the lowest average risk of miss classification G were utilized as classification variables for the SVM. The G cost function is calculated in the validation set as[18]

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

where N_V is the number of validation samples; and g_n is defined as,

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

In eqn (2), the numerator is the squared Mahalanobis distance between the object x_n (of class index I_n) and the sample mean $m_{I(n)}$ of its true class; whereas the denominator is the squared Mahalanobis distance between the object x_n and the mean $m_{I(m)}$ of the closest wrong class. GA was performed through 80 generations, having 160 chromosomes each. Crossover and mutation probability were set to 60% and 10%, respectively. The algorithm was repeated three times and the best result was chosen.

Thereafter, the PCA-SVM, SPA-SVM, and GA-SVM models were constructed. Different types of SVM kernels were utilized: linear (L), quadratic (Q), 3rd order polynomial (P), radial basis function (RBF), and multilayer perceptron (MPL). Such kernels transform the data into a feature space and are responsible for the SVM classification ability.[26] These kernels are calculated as follows:[26,37]

Linear,

$$K(\mathbf{x}_i,\mathbf{z}_j) = \mathbf{x}_i^T\mathbf{z}_j \quad (3)$$

Quadratic,

$$K(\mathbf{x}_i,\mathbf{z}_j) = (\tau + \mathbf{x}_i^T\mathbf{z}_j)^2,\ \tau \geq 0 \quad (4)$$

3rd order polynomial,

$$K(\mathbf{x}_i,\mathbf{z}_j) = (\tau + \mathbf{x}_i^T\mathbf{z}_j)^3,\ \tau \geq 0 \quad (5)$$

Radial basis function (RBF),

$$K(\mathbf{x}_i,\mathbf{z}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{z}_j\|^2) \quad (6)$$

Multilayer perceptron (MLP),

$$K(\mathbf{x}_i,\mathbf{z}_j) = \tan h(\kappa_1\mathbf{x}_i^T\mathbf{z}_j + \kappa_2) \quad (7)$$

where $\mathbf{x}_i$ and $\mathbf{z}_j$ are sample measurement vectors; τ is a constant; γ is the parameter that determines the RBF width; and κ_1 and κ_2 are constants. The SVM classifier takes the form of:

$$f(x) = \mathrm{sign}\left(\sum_{i=1}^{N_{SV}} \alpha_i y_i K(\mathbf{x}_i,\mathbf{z}_j) + b\right) \quad (8)$$

where N_{SV} is the number of support vectors; α_i is the Lagrange multiplier; y_i is the class membership ($\pm$1); $K(\mathbf{x}_i,\mathbf{z}_j)$ is the kernel function; and b is the bias parameter.[26,37]

By using these distinct types of kernel functions, 15 algorithms were utilized for classifying the fungal species: PCA-SVM-L, PCA-SVM-Q, PCA-SVM-P, PCA-SVM-RBF, PCA-SVM-MLP, SPA-SVM-L, SPA-SVM-Q, SPA-SVM-P, SPA-SVM-RBF, SPA-SVM-MLP, GA-SVM-L, GA-SVM-Q, GA-SVM-P, GA-SVM-RBF, and GA-SVM-MLP. In the RBF kernel, the γ parameter was set to 1; and in the MLP kernel, the κ_1 and κ_2 were respectively set to 1 and -1. The τ parameter was set to 0 for quadratic and 3rd order polynomial kernels.

Statistical validation

The models were statistically evaluated according to accuracy, sensitivity, specificity, F-score, and G-score. Accuracy is related to the percentage of correct classification achieved by the model; sensitivity measures the proportion of positive results that are correctly identified; specificity measures the proportion of negative results that are correctly identified; F-score represents the weighted average of the precision and sensitivity; and G-score accounts for the model precision and sensitivity without the influence of positive and negative class sizes.[38,39] These parameters were calculated as follows:[38,39]

$$\text{Accuracy}\ (\%) = 100 - \left(\frac{1}{N}\sum_{h=1}^{H} y_h^*\right) \times 100 \quad (9)$$

$$\text{Sensitivity}\ (\%) = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100 \quad (10)$$

$$\text{Specificity}\ (\%) = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \times 100 \quad (11)$$

$$F\text{-score} = \frac{2 \times \text{sensitivity} \times \text{specificity}}{\text{sensitivity} + \text{specificity}} \quad (12)$$

$$G\text{-score} = \sqrt{\text{sensitivity} \times \text{specificity}} \quad (13)$$

where N is the total number of samples; H is the total number of classes; y_h^* is the number of samples incorrectly classified in the h class; TP is the number of true positives; TN is the number of true negatives; FP is the number of false positives; and FN is the number of false negatives.

Results and discussion

Cryptococcus gattii (*C. gattii*) and *Cryptococcus neoformans* (*C. neoformans*) fungals samples were acquired by ATR-FTIR spectroscopy in the region of 3200–800 cm^{-1}. The raw spectra were preprocessed by cutting the spectra at 1800–900 cm^{-1} corresponding to the biological fingerprint region; followed by normalization to the amide I peak (~1650 cm^{-1}) and baseline correction. The preprocessed spectra are shown in Fig. 1.

The difference in the between-mean spectrum of *C. gattii* and *C. neoformans* is shown in Fig. 2a. In this figure, it is possible to observe that the large difference between the class' spectra is in the amide I region (~1650 cm^{-1}), where there is an absorbance difference close to -4.5×10^{-3} (−4.6%). The negative signal implies that this band is more intense in the *C. neoformans* class. A less intense difference between the class-mean is observed at ~1035 cm^{-1}, corresponding to glycogen bands.[40] In addition, the spectral difference close to 900 cm^{-1} increases due to phosphodiester and protein phosphorylation absorptions.[40,41]

In order to classify these fungal species, the SVM was used as a classification technique based on PCA as data reduction; and SPA and GA as variable selection methods. The PCA model applied to these data reduced the 468 variables (as wavenumbers inside the 1800–900 cm^{-1} range) to only 3 PCs, accounting for 99.98% of explained cumulative variance. Fig. 2b shows the PCA loadings on PC1, PC2 and PC3. In this figure, the loadings on PC1 which account for the largest variance from the original data (99.32% of explained variance) have higher coefficients in the amide I peak region (~1650 cm^{-1}), coinciding with the largest between-mean spectrum difference depicted in Fig. 2a. The loadings on PC2 (0.51% of explained variance) have higher coefficients in the phosphodiester and protein phosphorylation region (~900 cm^{-1}). The loadings on PC3 (0.15% of explained variance) show high coefficients in the glycogen region (~1035 cm^{-1}). These bands evidenced by PCA loadings are most important for class differentiation in the PCA-SVM-based models, which were built using five types of kernel functions: linear (PCA-SVM-L), quadratic (PCA-SVM-Q), 3rd order polynomial (PCA-SVM-P), RBF (PCA-SVM-RBF), and MLP (PCA-SVM-MLP).

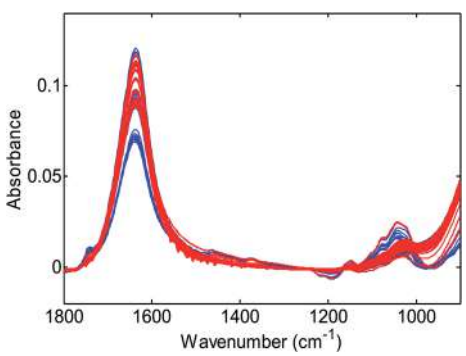


Fig. 1 Pre-processed spectra of *C. gattii* (blue color) and *C. neoformans* (red color) classes.

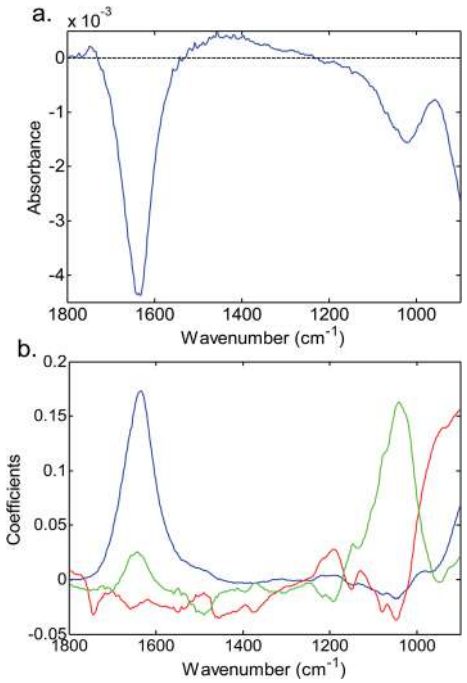


Fig. 2 (a) Difference between mean spectra of *C. gattii* and *C. neoformans* classes. (b) PCA loadings on PC1 (blue color), PC2 (red color), and PC3 (green color).

In addition to PCA, SPA and GA were applied to reduce the number of variables and be further used with SVM classifiers. The accuracy for each SVM-based algorithm in training, validation, and prediction set is shown in Table 1.

The most accurate PCA-SVM algorithm in the prediction set was composed of MLP kernel (PCA-SVM-MPL), which had 85.7% accuracy, whereas the most accurate for SPA-SVM and

Table 1 Accuracy (%) for SVM-based algorithms in training, validation, and prediction set

	Training	Validation	Prediction
PCA-SVM-L	85.7	88.1	78.6
PCA-SVM-Q	84.7	88.1	78.6
PCA-SVM-P	91.3	83.3	76.2
PCA-SVM-RBF	84.2	88.1	78.6
PCA-SVM-MLP	83.2	85.7	85.7
SPA-SVM-L	85.7	88.1	78.6
SPA-SVM-Q	92.9	92.9	92.9
SPA-SVM-P	98.0	100	100
SPA-SVM-RBF	93.4	92.9	90.5
SPA-SVM-MLP	77.0	76.2	83.3
GA-SVM-L	91.3	90.5	81.0
GA-SVM-Q	98.0	95.2	95.2
GA-SVM-P	99.5	97.6	100
GA-SVM-RBF	96.4	95.2	97.6
GA-SVM-MLP	72.4	66.7	71.4

GA-SVM had 3rd polynomial kernel (SPA-SVM-P and GA-SVM-P) with an accuracy of 100%. The classification performance by means of sensitivity, specificity, *F*-score, and *G*-score for PCA-SVMs, SPA-SVMs, and GA-SVMs models is shown in Fig. 3.

As shown in Fig. 3, the PCA-SVM algorithm with the best classification performance was PCA-SVM-MPL, achieving sensitivity, specificity, *F*-score, and *G*-score equal to 85.7%. For variable selection, the best algorithms were SPA-SVM-P and GA-SVM-P, achieving sensitivity, specificity, *F*-score, and *G*-score equal to 100%. These classification rates of 100% show the model's ability to correct the classification of all samples in which both positive and negative results were correctly identified. The selected variables by SPA-SVM are shown in Table 2. The percentage of absorbance variation (ΔA) between the classes at each selected wavenumber is also shown in this table.

Nine original wavenumbers were selected from 468 by SPA-SVM algorithms as classification variables. From the selected wavenumbers, absorbance at 1635 cm^{-1} had the most intense variation between the *C. gattii* and *C. neoformans* classes, with a variation of −4.4% (Table 2). This absorption is characteristic of the amide I β-sheet structure or proportions of β-sheet secondary structures.[40] The selected wavenumber at 906 cm^{-1} had the second largest ΔA (−2.4%). This wavenumber is in the phosphodiester region, composed of stretching of collagen and glycogen bands. The wavenumbers at 1443 cm^{-1} and 1745 cm^{-1} are respectively associated with the CH bending and symmetric stretching vibration of polysaccharides.[40] The polysaccharide capsules composed of 90–95% glucuronoxylomannan (GXM) and 5% galactoxylomannan (GalXM) determine the serotypes of *C. gattii* (serotypes B and C) and *C. neoformans* (serotypes A, D and AD) fungi,[4] therefore being important for class differentiation. The less intense ΔA for the selected wavenumbers by the SPA-SVM algorithm was found at 1541 cm^{-1}, a band of amide II absorption (N–H bending coupled to C–N stretching),[40] which is characteristic of proteins predominantly in β-sheet conformation.[42]

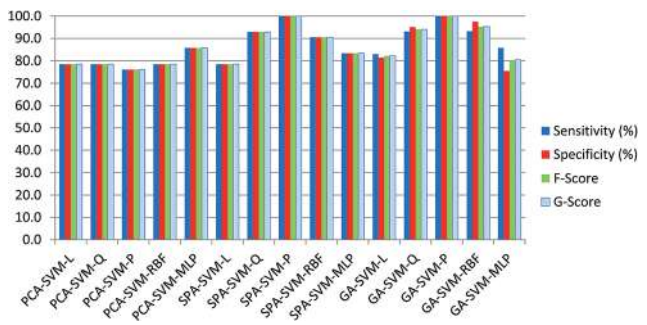


Fig. 3 Classification performance parameters (sensitivity, specificity, *F*-score, and *G*-score) for all SVM-based algorithms applied to discriminate *C. gattii* and *C. neoformans* classes.

Table 2 Selected variables by SPA-SVM-based algorithms and tentative assignment of possible biomarkers

Wavenumber (cm^{-1})	Tentative biomarker assignment[a]	ΔA[b] (%)
~906	Phosphodiester	−2.4
~964	C–C, C–O deoxyribose	−0.8
~999	Ring ν(C–C)–δ(C–H)	−1.3
~1041	Glycogen	−1.3
~1086	$\nu_s(PO_2^-)$ DNA/RNA	−0.7
~1443	δ(CH) polysaccharides	+0.4
~1541	Amide II	+0.03
~1635	Amide I	−4.4
~1745	ν_s(C=O) polysaccharides	+0.2

[a] ν = stretching vibration; δ = bending vibration; ν_s = symmetric stretching vibration. [b] Positive signal (+) indicates higher absorbance in the *C. gattii* class; negative signal (−) indicates higher absorbance in the *C. neoformans* class.

The variables selected by the GA-SVM algorithm are shown in Table 3. In this case, GA-SVM selected 12 wavenumbers as classification variables. Similar to Table 2, most of them have negative ΔA values. These negative ΔA values show that most selected wavenumbers have more intense absorption bands in the *C. neoformans* class. The higher absorbance in this class could be due to *C. neoformans* generally having a higher concentration of metabolites than *C. gattii*,[43] therefore increasing its absorption.

The higher ΔA for the selected wavenumbers by the GA-SVM algorithm (Table 3) is at 912 cm^{-1} (−2.2%). This value is close to the value obtained by the SPA-SVM algorithm at 906 cm^{-1} as shown in Table 2, and represents the phosphodiester region. The second largest ΔA was found at 991 cm^{-1} (ΔA = −1.2%), being assigned as the vibration of C–O in ribose.[40] This region is also a characteristic of other carbohydrate molecules,[15] therefore its signal could have contributions from more than one biomarker.

Amide I absorption was identified at 1697 cm^{-1} with ΔA of −1.0%. This band is a characteristic of high frequency vibration of an antiparallel amide I β-sheet (in-plane C=O stretching

Table 3 Selected variables by GA-based algorithms and tentative assignment of possible biomarkers

Wavenumber (cm^{-1})	Tentative biomarker assignment[a]	ΔA[b] (%)
~912	Phosphodiester	−2.2
~955	$\nu_s(PO_4^{3-})$	−0.8
~978	OCH_3 polysaccharides	−0.9
~991	C–O ribose	−1.2
~1070	$\nu_s(PO_2^-)$ DNA/RNA	−1.0
~1147	C–O oligosaccharides	−0.3
~1248	$\nu_{as}(PO_2^-)$ DNA	+0.05
~1278	Amide III	+0.1
~1323	Amide III	+0.2
~1508	Amide II	+0.3
~1697	Amide I	−1.0
~1734	ν_s(C=O) lipids	+0.1

[a] ν_s = symmetric stretching vibration; ν_{as} = asymmetric stretching. [b] Positive signal (+) indicates higher absorbance in the *C. gattii* class; negative signal (−) indicates higher absorbance in the *C. neoformans* class.

weakly coupled to C–N stretching and in-plane N–H bond bending).[40] Other vibrations of almost the same ΔA was found at 955 cm^{-1} (symmetric stretching of PO_4^{3-}), 978 cm^{-1} (OCH_3 vibration in polysaccharides), and 1070 cm^{-1} (symmetric PO_2^- stretching in DNA/RNA).[40] In this case, the influence of the polysaccharide capsules and nucleic acid contributions in the fungal species discrimination is clear. Amide II and amide III had very small ΔA contributions (+0.1–0.3%). Amide II absorption at 1508 cm^{-1} can be caused by N–H bending coupled to C–N stretching of amide II; whereas amide III absorptions at 1278 cm^{-1} and 1323 cm^{-1} are associated with vibration modes of collagen proteins in amide III. The lower ΔA for GA-SVM algorithms was found at 1248 cm^{-1} ($\Delta A = +0.05$) and corresponds to asymmetric PO_2^- stretching in DNA.[40]

The results shown here corroborate to the development of a rapid and non-destructive method for classifying *C. gattii* and *C. neoformans* fungal species with high accuracy, sensitivity, and specificity by using ATR-FTIR spectroscopy coupled with SVM-based techniques. The non-destructive nature of ATR-FTIR spectroscopy enables to reuse the samples in further studies, including genotyping by PCR-based methods. In addition, variable selection techniques (SPA and GA) can help to identify possible biomarkers responsible for class differentiation.

Furthermore, this research can be translated to real-world continuous monitoring by using these techniques to analyze cerebrospinal fluid of infected patients.[1] ATR-FTIR spectroscopy combined with chemometric techniques could be used to reduce the volume of the fluid utilized in the analysis, since the procedure to extract this fluid is quite invasive; as well as to reduce the cost, since the actual detection of both fungi follows genotyping procedures using molecular methods. In addition, this study could be used as a support to try the detection of both fungi in serum, which would reduce drastically the invasiveness of the procedure, allied to the advantages of using FTIR spectroscopy reported before.

Conclusion

PCA, SPA and GA were coupled to SVM classifiers to discriminate *C. gattii* and *C. neoformans* fungal species. Five different types of SVM kernels (linear, quadratic, 3rd order polynomial, RBF and MLP) were evaluated by means of quality metrics such as accuracy, sensitivity and specificity providing high classification rates. SPA-SVM and GA-SVM algorithms with 3rd order polynomial kernels (SPA-SVM-P and GA-SVM-P) achieved classification rates of 100% in accuracy, sensitivity, specificity, *F*-score, and *G*-score, showing these models to have the ability to provide reliable class differentiation. The SPA-SVM algorithm was highly influenced by amide I (1635 cm^{-1}) and phosphodiester (906 cm^{-1}) vibrations. In addition, the GA-SVM algorithm had higher influences of C–O ribose (991 cm^{-1}) and phosphodiester (912 cm^{-1}) vibrations. This report supports the development of an alternative method to classify *C. gattii* and *C. neoformans* fungal species using ATR-FTIR spectroscopy, which could be translated to real applications using cerebrospinal fluid in the future, for example. This could speed up the analysis of these fungi, thereby increasing its analytical frequency, reducing possible costs with reagents, and providing non-destructive data acquisition.

Acknowledgements

Camilo L. M. Morais and Fernanda S. L. Costa would like to thank CAPES/PPGQ/UFRN for their fellowship. Kássio M. G. Lima would like to acknowledge the CNPq grant (305962/2014-0) for financial support. In addition, the authors acknowledge PPGBQ/UFRN, as well as Professors Sandra de Moraes Gimines Bosco (UNESP/Brazil), Gilda del Negro (IMT/SP/Brazil), Fernanda Fonseca (UFPI/Brazil), Eveline P. Milan (Giselda Trigueiro Hospital/UFRN/Brazil), Thales D. Arantes (IMT/UFRN/Brazil), and Raquel C. Theodoro (IMT/UFRN/Brazil) for providing isolated fungus supplies.

References

1 E. K. Maziarz and J. R. Perfect, *Infect. Dis. Clin. North Am.*, 2016, **30**, 179–206.
2 L. Guazzelli, O. McCabe and S. Oscarson, *Carbohydr. Res.*, 2016, **433**, 5–13.
3 S. Samantaray, J. N. Correia, M. Garelnabi, K. Voelz, R. C. May and R. A. Hall, *Int. J. Antimicrob. Agents*, 2016, **48**, 69–77.
4 F. S. L. Costa, P. P. Silva, C. L. M. Morais, T. D. Arantes, E. P. Milan, R. C. Theodoro and K. M. G. Lima, *Anal. Methods*, 2016, **8**, 7107–7115.
5 X. Lin, *Infect., Genet. Evol.*, 2009, **9**, 401–416.
6 J. R. Harris, S. R. Lockhart, E. Debess, N. Marsden-Haug, M. Goldoft, R. Wohrle, S. Lee, C. Smelser, B. Park and T. Chiller, *Clin. Infect. Dis.*, 2011, **53**, 1188–1195.
7 K. J. Kwon-Chung and A. Varma, *FEMS Yeast Res.*, 2006, **6**, 574–587.
8 C. Maestrale, M. Masia, D. Pintus, S. Lollai, T. R. Kozel, M. A. Gates-Hollingsworth, M. G. Cancedda, P. Cabras, S. Pirino, V. D'Ascenzo and C. Ligios, *Vet. Microbiol.*, 2015, **177**, 409–413.
9 N. E. Nnadi, I. B. Enweani, M. Cogliati, G. M. Ayanbimpe, M. O. Okolo, E. Kim, M. Z. Sabitu, G. Criseo, O. Romeo and F. Scordino, *J. Mycol. Med.*, 2016, **26**, 306–311.
10 V. Rivera, M. Gaviria, C. Muñoz-Cadavid, L. Cano and T. Naranjo, *Braz. J. Infect. Dis.*, 2015, **19**, 563–570.
11 F. Sangalli-Leite, L. Scorzoni, A. C. A. de P. e Silva, J. de F. da Silva, H. C. de Oliveira, J. de L. Singulani, F. P. Gullo, R. M. da Silva, L. O. Regasini, D. H. S. da Silva, V. da S. Bolzani, A. M. Fusco-Almeida and M. J. S. Mendes-Giannini, *Int. J. Antimicrob. Agents*, 2016, **48**, 504–511.
12 C. B. Fígoli, R. Rojo, L. A. Gasoni, G. Kikot, M. Leguizamón, R. R. Gamba, A. Bosch and T. M. Alconada, *Int. J. Food Microbiol.*, 2017, **244**, 36–42.
13 N. Branan and T. A. Wells, *Vib. Spectrosc.*, 2007, **44**, 192–196.
14 J. Trevisan, P. P. Angelov, P. L. Carmichael, A. D. Scott and F. L. Martin, *Analyst*, 2012, **137**, 3202–3215.
15 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.
16 F. Zaera, *Chem. Soc. Rev.*, 2014, **43**, 7624–7663.
17 L. F. S. Siqueira and K. M. G. Lima, *Analyst*, 2016, **141**, 4833–4847.
18 T. C. Baia, R. A. Gama, L. A. S. de Lima and K. M. G. Lima, *Anal. Methods*, 2016, **8**, 968–972.
19 M. Boulet-Audet, F. Vollrath and C. Holland, *J. Exp. Biol.*, 2015, **218**, 3138–3149.
20 R. G. Saraiva, J. A. Lopes, J. Machado, P. Gameiro and M. J. Feio, *J. Biophotonics*, 2014, **7**, 392–400.
21 D. Naumann, V. Fijala, H. Labischinski and P. Giesbrecht, *J. Mol. Struct.*, 1988, **174**, 165–170.
22 H. J. Butler, M. R. McAinsh, S. Adams and F. L. Martin, *Anal. Methods*, 2015, **7**, 4059–4070.
23 J. Ord, H. J. Butler, M. R. McAinsh and F. L. Martin, *Analyst*, 2016, **141**, 2896–2903.
24 C. Cortes and V. Vapnik, *Mach. Learn.*, 1995, **20**, 273–297.
25 P. D. B. Harrington, *Anal. Chem.*, 2015, **87**, 11065–11071.
26 S. J. Dixon and R. G. Brereton, *Chemom. Intell. Lab. Syst.*, 2009, **95**, 1–17.
27 J. G. Kelly, P. P. Angelov, J. Trevisan, A. Vlachopoulou, E. Paraskevaidis, P. L. Martin-Hirsch and F. L. Martin, *Anal. Bioanal. Chem.*, 2010, **398**, 2191–2201.
28 M. Sattlecker, R. Baker, N. Stone and C. Bessant, *Chemom. Intell. Lab. Syst.*, 2011, **107**, 363–370.
29 G. L. Owens, K. Gajjar, J. Trevisan, S. W. Fogarty, S. E. Taylor, B. Da Gama-Rose, P. L. Martin-Hirsch and F. L. Martin, *J. Biophotonics*, 2014, **7**, 200–209.
30 S. Khan, R. Ullah, A. Khan, N. Wahab, M. Bilal and M. Ahmed, *Biomed. Opt. Express*, 2016, **7**, 2249–2256.
31 E. Pranckeviciene, R. Somorjai, R. Baumgartner and M. Jeon, *Artif. Intell. Med.*, 2005, **35**, 215–226.
32 R. Bro and A. K. Smilde, *Anal. Methods*, 2014, **6**, 2812–2831.
33 S. F. C. Soares, A. A. Gomes, A. R. Galvão Filho, M. C. U. Araujo and R. K. H. Galvão, *TrAC, Trends Anal. Chem.*, 2013, **42**, 84–98.
34 J. McCall, *J. Comput. Appl. Math.*, 2005, **184**, 205–222.
35 L. Cui, H. J. Butler, P. L. Martin-Hirsch and F. L. Martin, *Anal. Methods*, 2016, **8**, 481–487.
36 R. W. Kennard and L. A. Stone, *Technometrics*, 1969, **11**, 137–148.
37 J. Luts, F. Ojeda, R. Van De Plas, B. De Moor, S. Van Huffel and J. A. K. Suykens, *Anal. Chim. Acta*, 2010, **665**, 129–145.
38 K. S. Parikh and T. P. Shah, *Procedia Technol.*, 2016, **23**, 369–375.
39 L. C. de Carvalho, C. L. M. de Morais, K. M. G. de Lima, L. C. Cunha Junior, P. A. M. Nascimento, J. B. de Faria and G. H. A. Teixeira, *Anal. Methods*, 2016, **8**, 5658–5666.
40 Z. Movasaghi, S. Rehman and I. ur Rehman, *Appl. Spectrosc. Rev.*, 2008, **43**, 134–179.
41 J. G. Kelly, J. Trevisan, A. D. Scott, P. L. Carmichael, H. M. Pollock, P. L. Martin-Hirsch and F. L. Martin, *J. Proteome Res.*, 2011, **10**, 1437–1448.
42 D. E. Halliwell, C. L. M. Morais, K. M. G. Lima, J. Trevisan, M. R. F. Siggel-King, T. Craig, J. Ingham, D. S. Martin, K. A. Heys, M. Kyrgiou, A. Mitra, E. Paraskevaidis, G. Theophilou, P. L. Martin-Hirsch, A. Cricenti, M. Luce, P. Weightman and F. L. Martin, *Sci. Rep.*, 2016, **6**, 29494.
43 L. Wright, W. Bubb, J. Davidson, R. Santangelo, M. Krockenberger, U. Himmelreich and T. Sorrell, *Microbes Infect.*, 2002, **4**, 1427–1438.

APÊNDICE B

On the synergy between silver nanoparticles and doxycycline towards the inhibition of Staphylococcus aureus growth

Heloiza F. O. Silva

Rayane P. de Lima **Fernanda S. L. da Costa**

Edgar P. Moraes Maria C. N. Melo

Celso Sant'Anna Mateus Eugênio

Luiz H. S. Gasparotto

RSC Adv., 2018, 8, 23578.

Contribuição:

- Realizei a aquisição espectral;
- Ajudei na construção dos modelos de classificação multivariados;
- Ajudei na escrita da primeira versão do manuscrito.

RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2018, **8**, 23578

On the synergy between silver nanoparticles and doxycycline towards the inhibition of *Staphylococcus aureus* growth

Heloiza F. O. Silva,[a] Rayane P. de Lima,[a] Fernanda S. L. da Costa,[a] Edgar P. Moraes,[a] Maria C. N. Melo,[b] Celso Sant'Anna,[c] Mateus Eugênio[c] and Luiz H. S. Gasparotto*[a]

Received 12th March 2018
Accepted 21st June 2018

DOI: 10.1039/c8ra02176g

rsc.li/rsc-advances

In a previous paper (*RSC Adv.*, 2015, **5**, 66886–66893), we showed that the combination of silver nanoparticles (NanoAg) with doxycycline (DO) culminated in an increased bactericidal activity towards *E. coli*. Herein we further investigated the metabolic changes that occurred on *Staphylococcus aureus* upon exposure to NanoAg with the help of attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FTIR) coupled with multivariate data analysis. It has been discovered that the combination of DO with NanoAg produced metabolic changes in *S. aureus* that were not simply the overlap of the treatments with DO and NanoAg separately. Our results suggest that DO and NanoAg act synergistically to impede protein synthesis by the bacteria.

1 Introduction

It is widely known that the indiscriminate administration of antibiotics has rendered pathogens resistant to a variety of broad-spectrum antibiotics.[1] In order to circumvent this issue, nanoscience has worked in conjunction with biology and medicine to develop more efficient bactericidal agents.[2] As examples, silver nanoparticles have been used against *E. coli*[3,4] and gold nanoparticles for killing *S. aureus*.[5] Some researchers have attempted to combine nanoparticles with antibiotics to generate more potent antimicrobial agents.[6] Due to their large surface-area-to-volume ratio and biocompatibility, inorganic nanoparticles are considered ideal candidates for carrying large amounts of antibiotics without compromising their activity. Li *et al.*[7] demonstrated that the combination of silver nanoparticles with amoxicillin produced stronger bactericidal effect towards *Escherichia coli* in comparison to the administration of the components separately. Our group[8] showed that the conjugation of polyvinylpyrrolidone (PVP)-capped silver nanoparticles (NanoAg) with doxycycline (DO) yielded quite a potent agent for the inhibition of *E. coli*. An interesting question that follows is what biological changes occur upon contacting bacteria with NanoAg and DO-modified NanoAg. With that information in hand, it would be possible to fashion NanoAg with superior biocidal activities against a broader range of pathogens.

Silver nanoparticles may act *via* four main routes:[9] (1) adhesion to the microbial cell membrane causing damage and altering transport activity; (2) penetration inside the cell leading to organelle (ribosomes, DNA, RNA) dysfunction; (3) oxidation of proteins, lipids and DNA bases through oxidative stress; (4) alteration of cell signaling. Thus, as multiple factors are altered simultaneously, it is logical to measure the metabolism directly instead of selecting a single marker at a time. To that end, infrared spectroscopy (FT-IR) emerges as an interesting technique for metabolic fingerprinting,[10] owing to its capability to examine proteins, carbohydrates, lipids, amino acids and fatty acids concurrently. Coupled with multivariate data analysis,[10] FT-IR renders metabolic fingerprinting an excellent tool to discriminate between groups of related biological samples, in addition to being rapid and non-destructive.

In the present study, we exposed *S. aureus* to silver nanoparticles modified with doxycycline (DO, a member of the tetracycline group) and employed FT-IR coupled with multivariate data analysis to access variations of the *S. aureus* metabolism. DO-functionalized NanoAg caused the greatest alteration in the metabolism of *S. aureus* in comparison to that of bacteria treated with DO and NanoAg separately, which made possible the discrimination of bacteria subjected to those different treatments. These results corroborate nicely our previous work[8] in which we showed that the combination of DO with NanoAg delivered an increased antimicrobial activity towards *E. coli*.

2 Experimental section

2.1 Chemicals and reagents

Sodium hydroxide, glycerol, silver nitrate, polyvinylpyrrolidone (PVP; molecular weight = 10 000), and doxycycline hyclate (>98%) were obtained from Sigma-Aldrich Chemical Co (MO,

[a]*Biological Chemistry and Chemometrics Research Group, Institute of Chemistry, Federal University of Rio Grande do Norte, Natal 59072-970, RN, Brasil. E-mail: lhgasparotto@ufrnet.br; Tel: +55 84 33422323*

[b]*Laboratory of Medical Bacteriology, Center of Biosciences, Federal University of Rio Grande do Norte, Natal 59078-970, RN, Brasil*

[c]*Laboratory of Biotechnology – Labio, National Institute of Metrology, Quality and Technology – Inmetro, Duque de Caxias 25250-020, RJ, Brazil*

USA). *Staphylococcus aureus* (ATCC® 25923™) was cultivated in laboratory.

2.2 Production and characterization of NanoAg

NanoAg were produced according to a reported method.[11] Briefly, all glassware was cleaned thoroughly with a $KMnO_4$ + NaOH solution and piranha solution. The following aqueous stock solutions were then produced: 50 mmol L^{-1} $AgNO_3$, 100 g L^{-1} PVP and a solution containing 1.0 mol L^{-1} NaOH + 1.0 mol L^{-1} glycerol. In a beaker, determined volumes of the PVP and $AgNO_3$ solutions were dissolved in water to yield a 5 ml solution. In a separate beaker, a known volume of the NaOH + glycerol solution was mixed with water to generate another 5 ml solution. The glycerol–NaOH solution was poured into the $AgNO_3$–PVP one to yield the following final concentrations: 0.10 mol L^{-1} glycerol and NaOH, 10.0 g L^{-1} PVP and 1.0 mmol L^{-1} $AgNO_3$. The NanoAg colloidal solutions had then their pH adjusted to 7 by addition of diluted HCl.

UV-VIS was performed with an Ocean Optics USB-650 Tide spectrophotometer. FTIR in ATR mode was carried out with a Bruker Vertex 70 spectrophotometer and Transmission Electron Microscopy (TEM) images were acquired with a FEI Tecnai G^2 Spirit BioTWIN microscope operating at 120 kV.

2.3 Conjugation of DO with NanoAg

Conjugation of NanoAg with DO was achieved by simple incubation according to a previous protocol.[8] Five milliliters of a 200 μg ml^{-1} doxycycline stock solution were added to the same volume of the NanoAg colloidal solution, generating a 10 ml DO–NanoAg solution. The final concentrations of NanoAg and DO in the conjugate were 0.2×10^{-9} mol L^{-1} and 0.2×10^{-3} mol L^{-1}, respectively. All the above-mentioned techniques were employed to characterize the NanoAg–antibiotic complex.

2.4 Exposition of *S. aureus* to DO, NanoAg and DO–NanoAg

Staphylococcus aureus (strain ATCC® 25923™) was cultured in Brain-Heart-Infusion (BHI) agar medium on a Petri dish at 37 °C for 24 h. A microbiological strain suspension was standardized at 0.5 McFarland which is equivalent to 1.0×10^8 CFU ml^{-1} in 0.9% sterile saline medium. The suspension was then swabbed onto another Petri dish containing sterile Müller–Hinton agar medium and allowed to grow at 37 °C for 12 h. Afterwards, 1.5 ml of the antimicrobial agents (NanoAg, DO or the DO–NanoAg conjugate) was applied on the bacterial colony with the aid of a micropipette. The plates were incubated for further 12 h at 37 °C.

2.5 ATR-FTIR analysis

Bacteria were gently scraped off the Petri dish with a sterile metal handle, placed on the sample holder of the ATR-FTIR equipment, and covered with a piece of aluminum foil. The latter enhances the FTIR signal without interference due to its featureless background signal.[12] FTIR spectra were acquired in quintuplicate from each sample of the following groups (24 samples per group): control (*S. aureus* without any treatment), DO (*S. aureus* treated with doxycycline), NanoAg (*S. aureus* treated with silver nanoparticles), and DO–NanoAg (*S. aureus* treated with the conjugate), adding up to a total of 480 spectra. Measurements were conducted on a Bruker VERTEX 70 FTIR spectrometer (Bruker Optics Ltd., Coventry, UK) with a Helios ATR attachment containing a diamond crystal internal reflective element and a 45° incidence angle of the IR beam. Each spectrum was a result of 16 scans at a spectral resolution of 4 cm^{-1}. After each acquisition the sample holder was cleaned with 70% alcohol (v/v).

2.6 Chemometric procedure

Data import, pre-treatment and chemometric procedures were carried out with MATLAB R2014a software (MathWorks, USA) with the PLS-toolbox version 7.5.2 (Eigenvector Research, Inc., Wenatchee, WA). Raw spectra were pre-processed by selecting the range of 1800 cm^{-1} to 900 cm^{-1} (468 wavenumbers at 4 cm^{-1} spectral resolution) and mean-centering. PCA model was constructed with 96 samples (24 samples of each class: control, DO, DO + NanoAg, NanoAg), using 4 PCs, that explained 97.6% of total variance. PLS-DA models were made for each two classes of treatments (control, DO, DO + NanoAg, NanoAg).

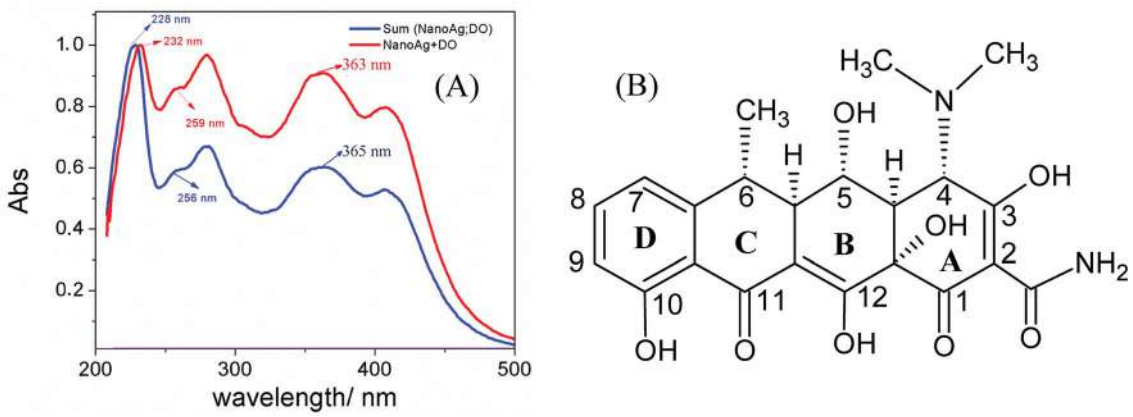


Fig. 1 (A) UV-VIS spectra of DO mixed with AgNPs (red curve) and the mathematical combination of the DO and NanoAg pure spectra. (B) Chemical structure of doxycycline.

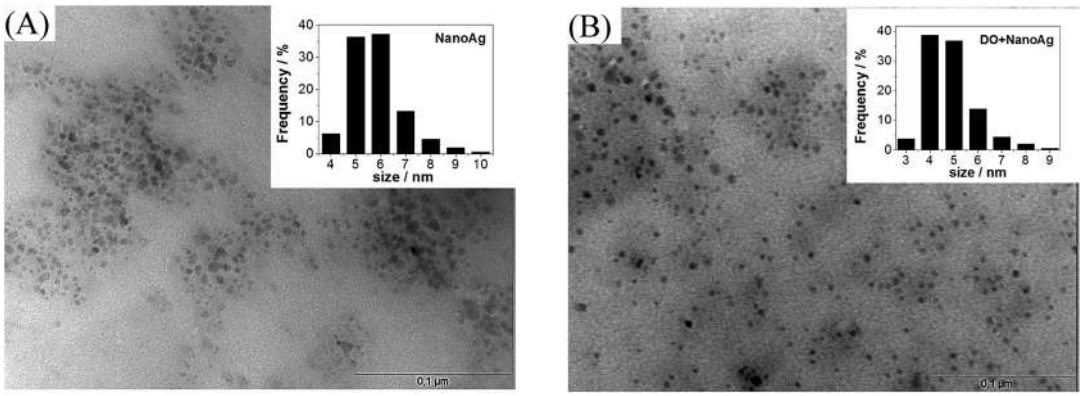


Fig. 2 TEM images of (A) NanoAg and (B) NanoAg mixed with doxycycline.

Using the algorithm Kennard–Stone (KS), separately to each class, the samples were divided into training/validation (70%) and prediction sets (30%). The model performance was evaluated by figures of merit: sensitivity, specificity and confusion matrix.

3 Results and discussion

3.1 Synthesis and characterization of NanoAg

In this study, NanoAg were produced with glycerol in alkaline medium as reducing agent at room temperature. Since glycerol is produced nowadays as a byproduct of the biodiesel fabrication, its supply has surpassed the current demand making glycerol a relatively inexpensive chemical.[8] Due to its biodegradability under aerobic conditions, non-toxicity, and low price, glycerol has become more attractive for generating nanoparticles than established reducing chemicals such as formamide, sodium borohydride and hydrazine. Fig. 1A suggests that NanoAg and DO interact to some extent. The UV-VIS spectrum of the mathematical combination of pure DO and NanoAg spectra (blue curve) shows a maximum at 410 nm corresponding to the characteristic Surface Plasmon Resonance (SPR) of PVP-stabilized spherical NanoAg, a peak at 365 nm due to the π-electron system located in the BCD chromophore (see Fig. 1B), and absorptions below 300 nm due to a combined contribution of the BCD system with the tricarbonylmethane keto–enol system comprised in ring A.[13] The mixing of DO with NanoAg (red curve) causes all DO absorptions to shift, implying an interaction of that system with the NanoAg. In a previous study,[8] we deeply investigated the interaction between DO and NanoAg *via* FTIR, showing that the capping agent, the PVP, was of paramount importance in augmenting the DO concentration around the nanoparticles. In addition to the chemical interaction, DO is kept in the vicinity of the particle due to the PVP-shell structure that encapsulates the DO.[14]

TEM images of NanoAg (Fig. 2B and C) showed that the conjugation with DO had practically no impact on both shape and size distribution of NanoAg, which is an important result since it can rule out size and shape effects on bacteriological experiments.

3.2 Exposition of *S. aureus* to DO, NanoAg and DO–NanoAg

3.2.1 Sample analysis and calibration/training dataset. The objective of the present study was to apply the ATR-FTIR spectroscopy in conjunction with PCA and PLS-DA to evaluate the metabolic response of *S. aureus* after treatment with NanoAg, DO and DO + NanoAg. As mentioned earlier, in a previous study we discovered that the combination of DO with NanoAg delivered a conjugate with enhanced growth inhibition properties against *E. coli* compared to the constituents administered separately.[8] This result prompted us to investigate the metabolic impact of NanoAg, DO and DO + NanoAg on *S. aureus*, a simpler microorganism in terms of cell wall complexity.[15] Fig. 3 presents average pre-treated spectra for each class acquired in the "bio-fingerprint" range of 900–1800 cm^{-1}. As noticed, it is not straightforward to distinguish the spectra visually, probably because the metabolic alterations upon treatment are minute.

The spectra were then subjected to the unsupervised Principal Components Analysis (PCA) classification model, followed by the supervised classification of Partial Least Squares Discriminant Analysis (PLS-DA) for the binary classification, as shown in Fig. 4.

The plot of the PCA discrimination function with the mean FTIR-ATR spectra (Fig. 4A) revealed a degree of segregation between the classes, meaning that the methodology allowed for the detection of variables that differentiate the groups which were then compared in pairs *via* PLS-DA: control *vs.* DO, control *vs.* NanoAg, and control *vs.* DO + NanoAg. This method indicated the wavenumbers whose changes were statistically

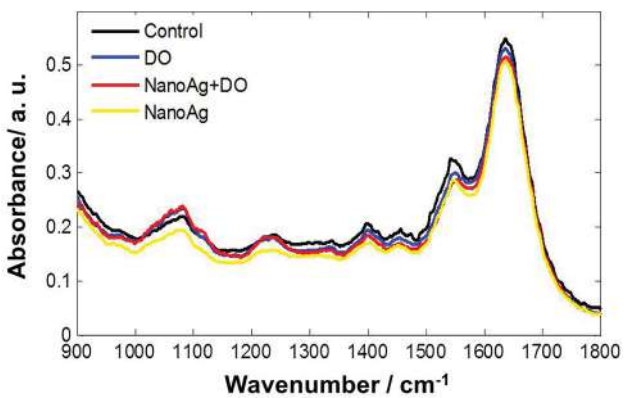


Fig. 3 Average spectrum for each original class control, DO, DO + NanoAg and NanoAg.

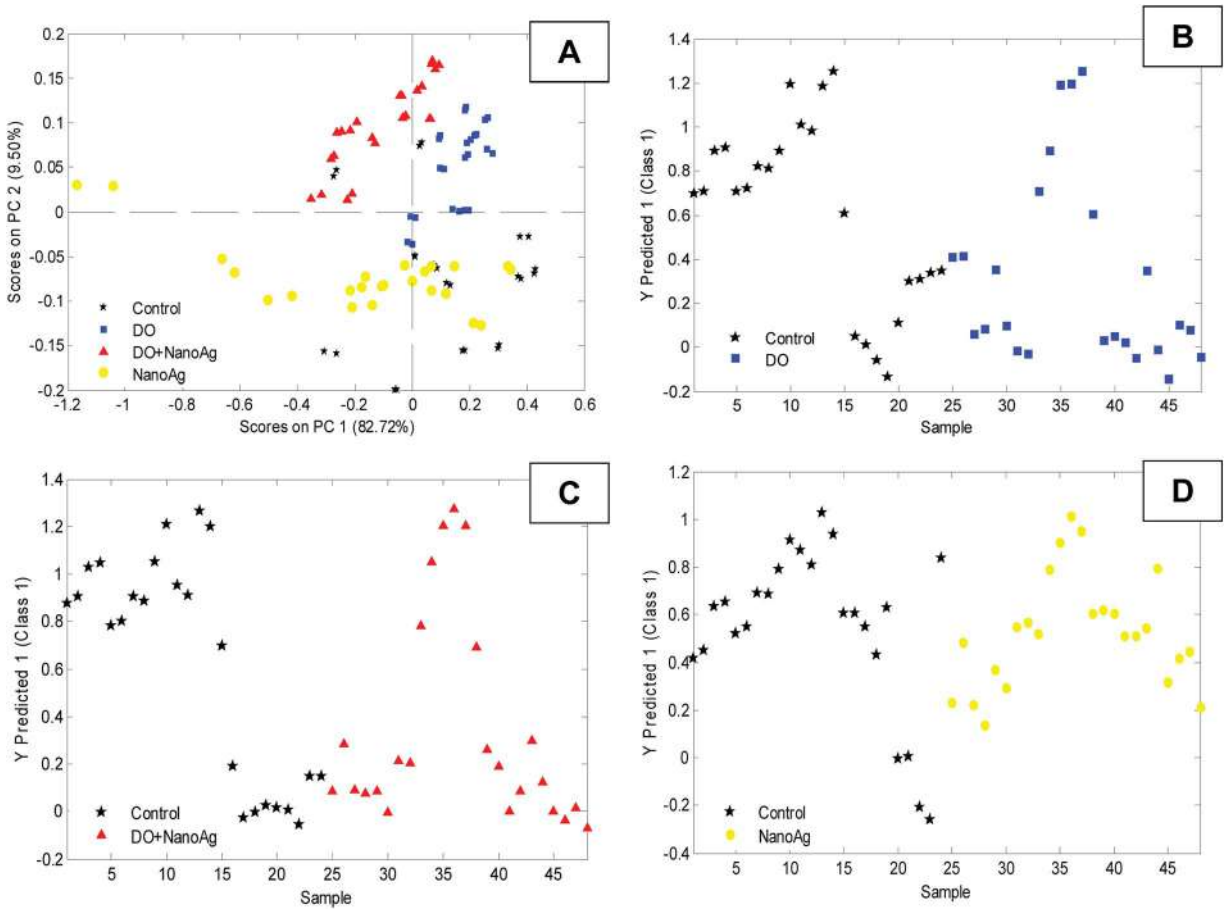


Fig. 4 Multivariate data analysis of selected variables in the samples. (A) Principal Component Analysis (PCA) of variables by the four classes and (B–D) PLS-DA by pairs.

significant for each group. Fig. 4B–D shows the PLS-DA plots for the three comparisons. For the determination of the best models, a confusion matrix (Table 1) was compiled with the values of true positive, true negative and type I and II errors.

It is possible to observe in Table 1 that of the three models the treatments with DO and DO + NanoAg presented the smallest errors, being classified 100% in their classes. The NanoAg model presented a 50% accuracy rate, corroborating the PCA discrimination function, showing that it is difficult to find a standard for differentiation because it presents a variance within the very high class. The values of the parameters of quality: sensitivity, specificity, RMSEC, RMSECV and RMSEP were also taken into account for the three best models. All the best models used 2 latent variables (see Table 2). These values show that there was good classification, especially for the control *versus* DO (sensibility: control 75% and DO 100% – specificity control 100% and DO 75%) and control *versus* DO + NanoAg (sensibility: control 75% and DO + NanoAg 100% – specificity control 100% and DO + NanoAg 75%) pairs that presented the highest values of sensitivity and specificity. These results confirm the potential of FITR-ATR spectroscopy to detect and identify groups with different metabolic responses of *S. aureus* after exposure to the three antimicrobials DO, NanoAg and DO + NanoAg.

Table 3 was constructed with the aid of the most significant variables present in the loadings generated in each of the three models. From these variables, the wave numbers responsible for the discrimination were recovered. Literature data were used to assign the characteristic group to each wave number retrieved.

Control vs. DO. In order to investigate the control and DO samples, as observed in Table 3, six variables were selected for PLS-DA (1647 cm^{-1}, 1631 cm^{-1}, 1547 cm^{-1}, 1543 cm^{-1}, 1400 cm^{-1} and 1080 cm^{-1}). It is interesting to note that 83.3% of the wave numbers responsible for the discrimination between the two classes are related to protein: amide I in 1647

Table 1 Confusion table for actual and predicted groups

	Actual (%)		
Predicted (%)		Control	DO
	Control	75.0	0
	DO	25.0	100
		Control	DO + NanoAg
	Control	75.0	0
	DO + NanoAg	25.0	100
		Control	NanoAg
	Control	100	50.0
	NanoAg	0	50.0

Table 2 Quality performance values from PLS-DA method (2 latent variables) by ATR-FTIR spectroscopy for each category of the three models

	Models PLS-DA (2 LVs)		
Accuracy (%)	Control *vs.* DO	Control *vs.* NanoAg	Control *vs.* DO + NanoAg
Calibration			
Sensibility (%)	93.8	87.5	93.8
Specificity (%)	100	62.5	100
Prediction			
Sensibility (%)	75.0	100	75.0
Specificity (%)	100	50.0	100
RMSEC	0.284	0.388	0.201
RMSECV	0.339	0.443	0.222
RMSEP	0.388	0.412	0.318

and 1631 cm^{-1}, amide II in 1547 and 1543 cm^{-1} and amino acid in 1400 cm^{-1}. Note that only the wavelength 1647 cm^{-1} showed a significant change ($p < 0.05$) in absorbance upon exposure to DO. This corroborates with the literature, since this antibiotic belongs to the class of tetracyclines, whose mechanism of action is the interference in the binding of the tRNA by blocking the adhesion of aminoacyl-t-RNA, to the mRNA-ribosome complex; in other words, DO interacts with the 30S portion of the ribosome thereby impairing protein synthesis. It should therefore be noted that tetracycline are classified as bacteriostatic, *i.e.* their interaction occurs reversibly.[16,17] To this we can attribute the absence of significant change ($p < 0.05$) in the absorbances of the other wave numbers related to the proteins. An interesting result was the presence of the band at 1080 cm^{-1}, characteristic of polysaccharides. Zmantar *et al.*[18] showed that *S. aureus* ATCC 25923, the same strain used in the present study, produces biofilm after stress and Cerca *et al.*[19] stated that the proteins encoded by intercellular adhesin genes (icaADBC) synthesize polysaccharide, which contributes to the formation of this biofilm in *S. aureus*. In contrast to the results obtained in Table 3, *S. aureus* showed a significant increase in mean absorbance ($p < 0.05$) when the DO band was treated with 1080 cm^{-1}, a characteristic of polysaccharides.

Control vs. NanoAg. From the investigation between the control and NanoAg samples, five variables were selected for SPA-LDA (1641 cm^{-1}, 1635 cm^{-1}, 1547 cm^{-1}, 1543 cm^{-1} and 1086 cm^{-1}). Of these wave numbers 80% are related to proteins as well, namely 1641 cm^{-1} and 1635 cm^{-1} for amide I, while the absorptions at 1547 and 1543 cm^{-1} are related to amide II. Although NanoAg has an extensive list of studies involving its effect against bacteria, the mechanism of action is not clearly known.[9,20] Some authors attribute such difficulty to the fact that

Table 3 Infrared band assignments of the Gram-positive *S. aureus* and average absorbances of the control, DO, NanoAg and DO + NanoAg classes presented in the regions (variables) used in the discrimination by PLS-DA[a]

Model	Wavelength	Abscontrol	Abstreated	Assignment	Literature
Control *versus* DO	~1647 cm^{-1}	0.51 (0.02)	0.51 (0.01)	Stretching of C=O in amide (amide I) of structural proteins.	28–30
	~1631 cm^{-1}	0.53 (0.02)	0.52 (0.01)	Stretching of C=O in amide (amide I) of structural proteins.	28–30
	~1547 cm^{-1}	**0.31 (0.02)***	**0.30 (0.01)***	N–H bending and C–N stretching in amide (amide II) of structural proteins.	27–30
	~1543 cm^{-1}	0.30 (0.02)	0.30 (0.01)	N–H bending and C–N stretching in amide (amide II) of structural proteins.	27–30
	~1400 cm^{-1}	0.19 (0.02)	0.19 (0.01)	$-COO^-$ symmetric stretching of amino acid side chains and fatty acids	27–29 and 31
	~1080 cm^{-1}	**0.21 (0.01)***	**0.23 (0.01)***	C–O–C. C–O of various polysaccharides	27–29 and 31
Control *versus* DO + NanoAg	~1635 cm^{-1}	**0.54 (0.02)***	**0.50 (0.01)***	β-pleated sheet structures (amide I) of structural proteins.	27 and 29
	~1630 cm^{-1}	**0.53 (0.02)***	**0.49 (0.01)***	Stretching of C=O in amide (amide I) of structural proteins.	28–30
	~1543 cm^{-1}	**0.30 (0.02)***	**0.27 (0.01)***	N–H bending and C–N stretching in amide (amide II) of structural proteins.	27–30
	~1539 cm^{-1}	**0.29 (0.02)***	**0.25 (0.01)***	N–H bending and C–N stretching in amide (amide II) of structural proteins.	27–30
	~1398 cm^{-1}	**0.19 (0.02)***	**0.17 (0.01)***	$-COO^-$ symmetric stretching of amino acid side chains and fatty acids	27–29 and 31
	~1078 cm^{-1}	**0.21 (0.01)***	**0.22 (0.01)***	C–O–C. C–O of various polysaccharides	27–29 and 31
Control *versus* NanoAg	~1641 cm^{-1}	**0.53 (0.02)***	**0.50 (0.05)***	Stretching of C=O in amide (amide I)	28–30
	~1635 cm^{-1}	**0.54 (0.02)***	**0.50 (0.05)***	β-pleated sheet structures (amide I)	27 and 29
	~1547 cm^{-1}	**0.31 (0.02)***	**0.28 (0.03)***	N–H bending and C–N stretching in amide (amide II)	27–30
	~1543 cm^{-1}	**0.30 (0.02)***	**0.28 (0.03)***	N–H bending and C–N stretching in amide (amide II)	27–30
	~1086 cm^{-1}	**0.21 (0.01)***	**0.19 (0.02)***	C–O–C. C–O of various polysaccharides	27–29 and 31

[a] *Different averages ($p < 0.05$).

the antibacterial action is strongly dependent on the physical–chemical parameters such as size, shape, surface charge, concentration and colloidal state.[21–26] Despite many factors, Dakal *et al.*[9] found that the antimicrobial action of NanoAg in general is linked to at least four distinct mechanisms. According to them, the NanoAg act (A) inducing cellular toxicity through the oxidative stress caused by the generation of reactive oxygen species (ROS) and free radicals, (B) adhering to the surface of the wall and cell membrane, (C) interfering in the modulation or (D) damaging intracellular structures (mitochondria, vacuoles, ribosomes) and biomolecules (proteins, lipids and DNA) after the endocytosis of NanoAg. Based on the selected variables (most related to proteins), we can suggest that the path of action of NanoAg synthesized and administered in our work was D, since after the treatment of *S. aureus* with NanoAg the mean absorbances suffered a significant decrease ($p < 0.05$), as shown in Fig. 5. The pathway suggested is that it acts more expressively in the inhibition of protein synthesis. In agreement with the expression in Table 3, the variable related to the wave number 1086 cm^{-1} was selected which is characteristic of polysaccharides that predominate in biofilm expressed by *S. aureus* after stress.[17,18] However, it was unusual to note that this treatment with free NanoAg of DO showed a significant decrease ($p < 0.05$) in mean absorbance suggesting that action of the nanoparticles did not allow *S. aureus* to effectively express its biofilm or even that it expressed, but the damage to DNA and polysaccharides was more expressive.

Control vs. DO + NanoAg. Finally, the comparison between the Control and DO + NanoAg samples allowed for the selection of six variables for SPA-LDA (1635, 1630, 1543, 1539, 1398 and 1078 cm^{-1}). Of these, 80% wave numbers are protein related, as observed in Table 3. The wavenumber at 1635 and 1630 cm^{-1} are assigned to amide I, 1543 and 1539 cm^{-1} assigned to amide II and 1398 cm^{-1} attributed to amino acid. At first, it is possible to observe that the treatment with the DO + NanoAg conjugate allowed for the selection of the variables observed both in the DO treatment and in the treatment with NanoAg. Interestingly, the DO + NanoAg conjugate caused a significant increase in the mean absorbance ($p < 0.001$) at the 1078 cm^{-1} wavenumber attributed to polysaccharides. Thus, the defense by biofilm expression was presented by *S. aureus* after treatment with the DO + NanoAg conjugate. Another important point was the significant decrease in mean absorbances ($p < 0.05$) attributed to protein expression exhibited after treatment with this system, which did not occur after treatment with DO. Based on what has been presented so far, we can conclude that the conjugate caused expressive metabolic responses, even though the two starting constituents, DO and NanoAg, were at half their original concentrations. The combination of DO and NanoAg boosts the inhibition of protein synthesis.

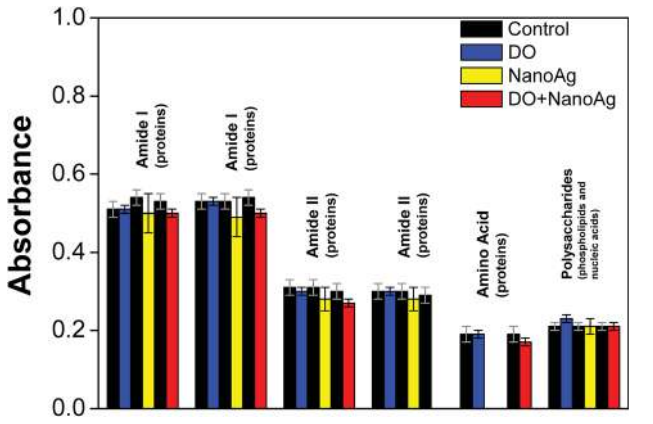


Fig. 5 Mean average absorbances of the control, DO, NanoAg and DO + NanoAg classes presented in the six regions used in the discrimination by PLS-DA.

4 Conclusions

Herein we have shown that FT-IR coupled with multivariate analysis is an excellent tool to discriminate bacteria that have been treated with DO, NanoAg and DO + NanoAg. From PCA analysis it is clear that, although both DO and NanoAg affect protein synthesis, their combination promotes biological changes in *S. aureus* sufficient for discrimination among the classes.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors are grateful to CNPq (grant 442087/2014-4).

References

1 A. J. Alanis, Resistance to Antibiotics: Are We in the Post-Antibiotic Era?, *Arch. Med. Res.*, 2005, **36**(6), 697–705.
2 A. J. Huh and Y. J. Kwon, "Nanoantibiotics": A new paradigm for treating infectious diseases using nanomaterials in the antibiotics resistant era, *J. Controlled Release*, 2011, **156**(2), 128–145.
3 I. Sondi and B. Salopek-Sondi, Silver nanoparticles as antimicrobial agent: a case study on *E. coli* as a model for Gram-negative bacteria, *J. Colloid Interface Sci.*, 2004, **275**(1), 177–182.
4 S. K. Rastogi, V. J. Rutledge, C. Gibson, D. A. Newcombe, J. R. Branen and A. L. Branen, Ag colloids and Ag clusters over EDAPTMS-coated silica nanoparticles: synthesis, characterization, and antibacterial activity against Escherichia coli, *Nanomedicine*, 2011, **7**, 305–314.
5 V. P. Zharov, K. E. Mercer, E. N. Galitovskaya and M. S. Smeltzer, Photothermal Nanotherapeutics and Nanodiagnostics for Selective Killing of Bacteria Targeted with Gold Nanoparticles, *Biophys. J.*, 2006, **90**(2), 619–627.
6 A. N. Brown, K. Smith, T. A. Samuels, J. Lu, S. O. Obare and M. E. Scott, Nanoparticles Functionalized with Ampicillin Destroy Multiple-Antibiotic-Resistant Isolates of Pseudomonas aeruginosa and Enterobacter aerogenes and Methicillin-Resistant *Staphylococcus aureus*, *Appl. Environ. Microbiol.*, 2012, **78**(8), 2768–2774.

7 P. Li, J. Li, C. Wu, Q. Wu and J. Li, Synergistic antibacterial effects of β-lactam antibiotic combined with silver nanoparticles, *Nanotechnology*, 2005, **16**(9), 1912.

8 H. F. O. Silva, K. M. G. Lima, M. B. Cardoso, J. F. A. Oliveira, M. C. N. Melo, C. Sant'Anna, M. Eugenio and L. H. S. Gasparotto, Doxycycline conjugated with polyvinylpyrrolidone-encapsulated silver nanoparticles: a polymer's malevolent touch against Escherichia coli, *RSC Adv.*, 2015, **5**(82), 66886–66893.

9 T. C. Dakal, A. Kumar, R. S. Majumdar and V. Yadav, Mechanistic Basis of Antimicrobial Actions of Silver Nanoparticles, *Front Microbiol*, 2016, **7**, 1831.

10 D. I. Ellis and R. Goodacre, Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy, *Analyst*, 2006, **131**(8), 875–885.

11 J. F. Gomes, A. C. Garcia, E. B. Ferreira, C. Pires, V. L. Oliveira, G. Tremiliosi-Filho and L. H. S. Gasparotto, New insights into the formation mechanism of Ag, Au and AgAu nanoparticles in aqueous alkaline media: alkoxides from alcohols, aldehydes and ketones as universal reducing agents, *Phys. Chem. Chem. Phys.*, 2015, **17**(33), 21683–21693.

12 L. Cui, H. J. Butler, P. L. Martin-Hirsch and F. L. Martin, Aluminium foil as a potential substrate for ATR-FTIR, transflection FTIR or Raman spectrochemical analysis of biological specimens, *Anal. Methods*, 2016, **8**(3), 481–487.

13 S. Schneider, M. O. Schmitt, G. Brehm, M. Reiher, P. Matousek and M. Towrie, Fluorescence kinetics of aqueous solutions of tetracycline and its complexes with Mg^{+2} and Ca^{+2}, *Photochem. Photobiol. Sci.*, 2003, **2**(1), 1107–1117.

14 M. Behera and S. Ram, Inquiring the mechanism of formation, encapsulation, and stabilization of gold nanoparticles by poly(vinyl pyrrolidone) molecules in 1-butanol, *Appl. Nanosci.*, 2014, **4**(1), 247–254.

15 K. D. Young, The Selective Value of Bacterial Shape, *Microbiol. Mol. Biol. Rev.*, 2006, **70**(3), 660–703.

16 E. C. Pereira-Maia, P. P. Silva, W. B. Almeida, H. F. Santos, B. L. Marcial, R. Ruggiero and W. Guerra, Tetraciclinas e Glicilciclinas: Uma Visão Geral, *Quim. Nova*, 2010, **33**(3), 700–706.

17 I. Chopra and M. Roberts, Tetracycline Antibiotics: Mode of Action, Applications, Molecular Biology, and Epidemiology of Bacterial Resistance, *Microbiol. Mol. Biol. Rev.*, 2001, **65**(2), 232–260.

18 T. Zmantar, B. Kouidhi, H. Miladi, K. Mahdouani and A. Bakhrouf, A Microtiter plate assay for *Staphylococcus aureus* biofilm quantification at various pH levels and hydrogen peroxide supplementation, *New Microbiol.*, 2010, **33**(1), 137–145.

19 N. Cerca, J. L. Brooks and K. K. Jefferson, Regulation of the Intercellular Adhesin Locus Regulator (icaR) by SarA, B, and IcaR in *Staphylococcus aureus*, *J. Bacteriol.*, 2008, **190**(19), 6530–6533.

20 S. Prabhu and E. K. Poulose, Silver nanoparticles: mechanism of antimicrobial action, synthesis, medical applications, and toxicity effects, *Int. Nano Lett.*, 2012, **2**(32), 2–10.

21 S. Pal, Y. K. Tak and J. M. Song, Does the Antibacterial Activity of Silver Nanoparticles Depend on the Shape of the Nanoparticle? A Study of the Gram-Negative Bacterium Escherichia coli, *Appl. Environ. Microbiol.*, 2007, **73**(6), 1712–1720.

22 R. Bhattacharya and P. Mukherjee, Biological properties of "naked" metal nanoparticles, *Adv. Drug Delivery Rev.*, 2008, **60**(1), 1289–1306.

23 M. K. Rai, S. D. Deshmukh, A. P. Ingle and A. K. Gade, Silver nanoparticles: the powerful nanoweapon against multidrug-resistant bacteria, *J. Appl. Microbiol.*, 2012, **112**(1), 841–852.

24 M. R. Nateghi and H. Hajimirzababa, Effect of silver nanoparticles morphologies on antimicrobial properties of cotton fabrics, *J. Text. Inst.*, 2014, **105**(1), 806–813.

25 A. Abbaszadegan, Y. Ghahramani, A. Gholami, B. Hemmateenejad, S. Dorostkar, M. Nabavizadeh and H. Sharghi, The Effect of Charge at the Surface of Silver Nanoparticles on Antimicrobial Activity against Gram-Positive and Gram-Negative Bacteria: A Preliminary Study, *J. Nanomater.*, 2015, **2015**(1), 1–8.

26 F. Zhang, J. A. Smolen, S. Zhang, R. Li, P. N. Shah, S. Cho, H. Wang, J. E. Raymond, C. L. Cannon and K. L. Wooley, Degradable polyphosphoester-based silver-loaded nanoparticles as therapeutics for bacterial lung infections, *Nanoscale*, 2015, **7**, 2265–2270.

27 K. Maquelin, C. Kirschner, L. P. Choo-Smith, N. V. D. Braak, H. Ph Endtz, D. Naumann and G. J. Puppels, Identification of medically relevant microorganisms by vibrational spectroscopy, *J. Microbiol. Methods*, 2002, **51**(1), 255–271.

28 W. Jiang, A. Saxena, B. Song, B. B. Ward, T. J. Beveridge and S. C. B. Myneni, Elucidation of Functional Groups on Gram-Positive and Gram-Negative Bacterial Surfaces Using Infrared Spectroscopy, *Langmuir*, 2004, **20**, 11433–11442.

29 D. Naumann, Infrared Spectroscopy in Microbiology, *Encyclopedia of Analytical Chemistry*, ed. R. A. Meyers, John Wiley & Sons Ltd, Chichester, 2000, pp. 102–131.

30 Y. Burgula, D. Khali, S. Kim, S. S. Krishnan, M. A. Cousin, J. P. Gore, B. L. Reuhs and L. J. Mauer, Review Of Mid-Infrared Fourier Transform-Infrared Spectroscopy Applications For Bacterial Detection, *J. Rapid Methods Autom. Microbiol.*, 2007, **15**(1), 146–175.

31 B. Buszewski, E. Dziubakiewicz, P. Pomastowski, K. Hrynkiewicz, J. Ploszaj-Pyrek, E. Talik, M. Kramer and K. Albert, Assignment of functional groups in Gram-positive bacteria. Analytical Method Development and Validation: A Concise Review, *J. Anal. Bioanal. Tech.*, 2015, **6**(1), 1–8.

APÊNDICE C

The Use of Near Infrared Spectroscopy and Multivariate Calibration for Determining the Active Principle of Olanzapine in a Pharmaceutical Formulation

Marcelo V. P. Amorim

Fernanda S. L. Costa Cícero F. S. Aragão

Kássio M. G. Lima

J. Braz. Chem. Soc., 2016, 00, 1-7.

Contribuição:

- Ajudei na escrita do artigo;
- Ajudei na construção dos modelos de regressão PLS;
- Ajudei na escrita da primeira versão do manuscrito.

Fernanda S. L. Costa **Prof. Kássio M. G. Lima**

http://dx.doi.org/10.21577/0103-5053.20160233

J. Braz. Chem. Soc., Vol. 00, No. 00, 1-7, 2016.
Printed in Brazil - ©2016 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

Short Report

The Use of Near Infrared Spectroscopy and Multivariate Calibration for Determining the Active Principle of Olanzapine in a Pharmaceutical Formulation

Marcelo V. P. Amorim,[a,b] Fernanda S. L. Costa,[c] Cícero F. S. Aragão[b] and Kássio M. G. Lima*[*,c]

[a]Núcleo de Pesquisa em Alimentos e Medicamentos, [b]Laboratório de Controle de Qualidade, Departamento de Farmácia and [c]Grupo de Química Biológica e Quimiometria, Instituto de Química, Universidade Federal do Rio Grande do Norte, 59072-970 Natal-RN, Brazil

The aim of this study was to quantitatively determine the olanzapine in a pharmaceutical formulation for assessing the potentiality of near infrared spectroscopy (NIR) combined with partial least squares (PLS) regression. The method was developed with samples based on a commercial formulation containing olanzapine and seven excipients. Laboratory and commercial samples (n = 27 and 18, respectively) were used by defining the calibration and prediction sets. The method was validated in the range from 1.0 to 12.5 of olanzapine *per* 100 mg of powder (average mass 210 mg), by accuracy, precision, linearity, analytical sensitivity, limit of detection and quantification. The multivariate model developed for olanzapine was based on PLS and the determination coefficient (r_c and r_p), with the root mean square error of calibration and prediction being 0.95, 0.93, 3.2×10^{-3} and 4.0×10^{-3}% m/m, respectively. The proposed NIR method is an effective alternative for quantification of olanzapine in the pharmaceutical industry.

Keywords: near infrared spectroscopy, olanzapine, partial least squares regression, figure of merit, HPLC

Introduction

Atypical antipsychotics are a group of antipsychotic drugs used to treat psychiatric conditions. Some atypical antipsychotics[1] have received regulatory approval for schizophrenia, bipolar disorder, autism, and as an adjunct in major depressive disorder. The first-line psychiatric treatment for schizophrenia and bipolar disorder is antipsychotic medication which includes olanzapine.[2] Olanzapine (Figure 1) is a synthetic derivative of thienobenzodiazepine with antipsychotic, antinausea and antiemetic activities.[3]

Several analytical methods have been described for the quantification of olanzapine in biological fluids, pharmaceutical formulations and tissues such as high performance liquid chromatography (HPLC) with ultraviolet[4,5] or electrochemical detection,[6] liquid chromatography/electrospray ionization tandem mass spectrometry (LC-ESI-MS/MS)[7] and mass spectrometry imaging (MSI) using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS).[8] Although these cited analytical methods indicate the effectiveness of liquid chromatography owing to its reliability, accuracy, reproducibility of results and sensitive analytical method for the determination of olanzapine in various studies, they are time consuming and require experienced personnel to perform the analysis. Furthermore, they are also destructive methods involving sample preparations. For all of these reasons, the search for new analytical techniques is of fundamental importance, especially those which lower both analysis time and cost.

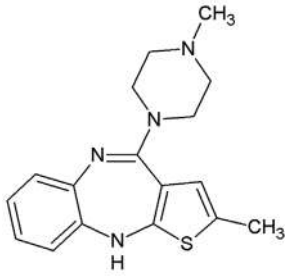


Figure 1. Chemical structure of olanzapine.

On the other hand, near infrared spectroscopy (NIRS) has been developed and proven to be a powerful tool for the pharmaceutical industry due to some characteristics such as being a fast and non-destructive

*e-mail: kassiolima@gmail.com

method, requiring minimal or no sample preparation and its high precision. Also, no reagents are required and no waste is produced, in contrast with traditional analytical methods (liquid chromatography, for example). Hertrampf *et al.*[9] employed NIR spectroscopy coupled with multivariate models to analyze tablets containing two different active pharmaceutical ingredients (API) (bisoprolol, hydrochlorothiazide) in different commercially available dosages. Two pharmaceutical excipients (lactose monohydrate and microcrystalline cellulose) and one API (acetaminophen) were used, and investigated using NIRS and partial least squares (PLS) by Sánchez-Paternina *et al*.[10] For pharmaceutical industry examples using NIRS technique, we can cite special interest in the identification of raw materials and finished products,[11] reaction monitoring in blending processes,[12] determination of active principles,[13] dissolution testing,[14] hardness testing[15] and polymorphs.[16]

The use of appropriate mathematical and statistical methods (i.e., chemometrics) is largely responsible for the advancement of the NIR technique, including multivariate calibration techniques such as partial least squares (PLS),[17] principal component regression (PCR),[18] artificial neural networks (ANN)[19] and least squares-support vector machine (LS-SVM).[20] The main advantages of using the multivariate calibration techniques listed above is that fast, cheap, or non-destructive analytical measurements (such as NIRS) can be used to estimate sample properties (for example, physicochemical parameters of pharmaceutical formulations) which would otherwise require time-consuming, expensive or destructive testing (such as liquid chromatography). Additionally, the establishment of validation procedures for multivariate calibration is very important because it is the first step for recognizing these methods for official analysis, especially in pharmaceutical legislation. Validation occurs via determination of several parameters, known as the figures-of-merit (FOM).[21] According to ANVISA (RE 899/2003),[22] validating a pharmaceutical analysis method is done by following the parameters of: sensibility, selectivity, accuracy, precision, linearity, range, limit of detection (LOD), quantification (LOQ) and robustness. Brazilian pharmacopeia[23] and the European Medicines Agency (EMEA)[24] have also adopted guidelines for validating methodology which employs NIR spectroscopy using established chemometrics tools, and evaluating parameters such as specificity, linearity, range, accuracy, precision and robustness.

Herein, we have attempted to quantitatively determine the active principle of olanzapine in different pharmaceutical excipients using NIRS and multivariate calibration. Nevertheless, olanzapine content has never been calibrated by NIR spectroscopy, or any other rapid technique. In addition, data pre-processing methods were evaluated to determine the most suitable method for analyzing the data type. Finally, the best performing models were validated by calculating the FOM obtained from the analyses, which included selectivity, sensitivity, analytical sensitivity, precision, accuracy, limit of detection and limit of quantification.

Experimental

Sample preparation and mixture design

The pharmaceutical preparation studied was a powder mixture with antipsychotic action containing olanzapine as the active principle and seven excipients (lactose, microcrystalline cellulose, poloxamer, crospovidone, silicon dioxide, magnesium stearate and coating mixture). All compounds (active principle and excipients) were supplied by the Center for Food and Drug Research of the Federal University of Rio Grande do Norte (NUPLAM/UFRN), Brazil. In this work, olanzapine from NUPLAM/UFRN (Brazil) and EMS sigma pharma (State of São Paulo, Brazil) was used to correspond to form II (polymorphic).

Laboratory and commercial samples were weighed, crushed and individually placed in the same vials in variable proportions to span a concentration range (1.0 to 12.5 mg *per* 100 mg of powder (average mass 210 mg)) of nominal content in the active principle and ± 5% for excipients. Laboratory samples were made by individually weighing all excipients (including the coating powder mixture) and active principle, according to its mass used in the master formula. Commercial samples provided for the study were weighed, crushed and individually placed into the vials. From there, they were also homogenized using a Tube Mixer for 5 minutes to ensure the same concentration of active principle *per* milligram of powder. Commercial samples (2.5, 5.0 and 10.0 mg of olanzapine also *per* 100 mg of powder (average mass 210 mg)) were obtained from EMS sigma pharma (State of São Paulo, Brazil). The concentration of the active principle within laboratory samples was between 0.0047 to 0.0595% m/m and for the commercial samples it was 0.0119 to 0.047% m/m.

The ternary mixtures were selected according to a D-Optimal solution[25] (Modde software version 4.0, MKS Data Analytic Solutions, Umeå, Sweden) totaling twenty-seven experiments, covering all corners at the center point of the mixture space. D-Optimal design was employed to select the concentration levels of olanzapine and excipients in the laboratory samples of calibration and external validation sets in order to build the multivariate

model. At the center point, all constituents in the mixture had nominal values. Six additional mixtures were made in order to achieve nearly equidistant steps in mass fraction for calibration and validation. Figure 2 illustrates the mixing ratios of the powder mixtures. One gram of every laboratory sample was homogenized in a Tube Mixer from BIOMATIC (Rio Grande do Sul, Brazil).

NIR spectroscopy

NIR spectra were collected in diffuse reflection mode via FT-NIR spectrometer (MPA, Bruker Optics, Ettlingen, Germany) equipped with an integrating sphere. Each measured spectrum (in triplicate) was the average of 32 scans obtained with a resolution of 16 cm^{-1} and over the range of 900-2500 nm. The background spectrum was recorded using a gold coated slide. Spectral measurements were done in an acclimatized room under controlled temperature of 22 °C, and 60% relative air humidity.

HPLC analysis

After NIR analysis, the samples were subject to reference analysis using HPLC. The API olanzapine was determined by performing isocratic analysis by using an HPLC instrument from HITACHI equipped with pump (5160), auto-injector (5260), column oven (5310), iodine array detector (5430), all from Hitachi (Tokyo, Japan), column Xterra® (Waters), 150 × 4.6 mm × 5 µm at 25 ºC. For each analysis, the mobile phase used was in proportion 64:17:19 v/v of citrate buffer pH 5.9, acetonitrile and methanol, respectively.

The HPLC procedure used as reference to determine the API (olanzapine) in production tablets was as follows: each different concentration (2.5, 5.0 and 10.0 mg) tablet was weighed, dissolved in hydrochloric acid 0.1 N, sonicated for 10 min, diluted to 25 mL (2.5 mg), 50 mL (5.0 mg) and 100 mL (10.0 mg) with the same acid. An aliquot of 15 µL was injected at HPLC to obtain the chromatogram at 260 nm. The API in each sample, in milligrams of API *per* gram of tablets was used as reference datum.

Chemometrics procedure and software

All calculations (models and pretreatments) were performed using the MATLAB version 6.5 (The Math-Works, Natick, USA), specifically the PLS-toolbox (Eigenvector Research, Inc. ,Wenatchee, WA, USA, version 6.01). The calculated NIR spectra was log 1/R transformed in the first step, followed by the average spectra for each sample. Different pretreatments such as Smoothing Savitzky-Golay (SGS) (7 window points) followed by MSC (multiplicative scatter correction) and first-order derivative Savitzky-Golay (7 window points) were applied on the spectra in order to minimize undesirable features such as spectral offset, noise, baseline and scattering.[14,26]

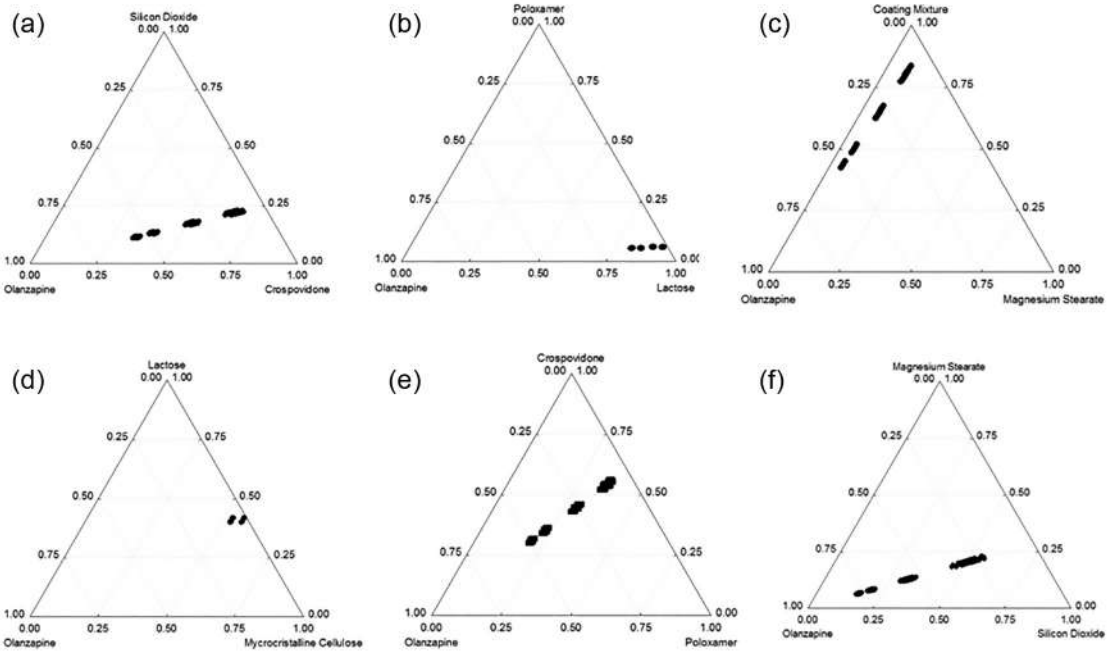


Figure 2. Ternary mixture design for NIR calibration measurements according to a D-Optimal design. (a) Olanzapine-crospovidone-silicon dioxide; (b) olanzapine-poloxamer-lactose; (c) olanzapine-coating mixture-magnesium stearate; (d) olanzapine-microcrystalline cellulose-lactose; (e) olanzapine-crospovidone poloxamer; (f) olanzapine-magnesium stearate-silicon dioxide.

A PLS-regression model was developed and validated by leave-one-out full-cross-validation. To avoid overfitting, test-set validated calibrations were used (75 and 25% of the spectra of laboratory and commercial samples, respectively, were applied in the calibration set and, 25 and 75% of the spectra of laboratory and commercial samples, respectively, were applied in the validation set) through the classic Kennard-Stone (KS) selection algorithm.[27] The following quality parameters were used to evaluate the calibration models, respectively: RMSEC (root mean square error of calibration), RMSEP (root mean square error of prediction), correlation coefficients of each model for calibration data set (r_c) and prediction data set (r_p). An elliptical joint confidence region (EJCR) was calculated to evaluate the slope, intercept the reference regression, and to predict values at a 95% confidence interval.

Finally, validating an analytical method entails determining whether it fulfills its intended purpose. To do this, some figures of merit were determined such as sensitivity (fraction of analytical signal that is due to the increase of the concentration of a particular analyte at unitary concentration), selectivity (indicates the portion of the instrumental signal that is used for the multivariate calibration model), analytical sensitivity (ratio between the sensitivity and the instrumental noise), precision (degree of scatter between a series of measurements for the same sample under prescribed conditions), accuracy (closeness of agreement between the reference value and the value found by the calibration model, generally expressed as the root mean square error of the prediction samples (RMSEP)), limit of detection (minimum detectable value of net signal (or concentration) for which the probabilities of false negatives (β) and false positives (α) are 0.05) and limit of quantification (signal or analyte concentration value that will produce estimates having a specified relative standard deviation). The quality metrics[28] used in this study for evaluating the figures of merit results can be calculated following the equations:

$$\text{Sensitivity} = \mathbf{S}_{\mathbf{k,j}}^{\mathbf{nas}} = \frac{\hat{\mathbf{X}}_{\mathbf{A,k}}^{\mathbf{nas}}}{y_i} \quad (1)$$

$$\text{Selectivity} = \hat{SEL}_{k,un} = \frac{\hat{nas}_{k,un}}{\left\|\mathbf{x}_{\mathbf{k,un}}\right\|} \quad (2)$$

$$\text{Analytical sensitivity} = \gamma = \frac{\hat{SEN}}{\left\|\sigma x\right\|} \quad (3)$$

$$\text{Precision} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\hat{y}_{i,j} - \hat{\bar{y}}_i\right)^2}{n(m-1)}} \quad (4)$$

$$\text{Accuracy} = \text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \quad (5)$$

$$\text{Limit of detection} = \text{LOD} = 3\delta x\left\|b_k\right\| = 3\delta x\frac{1}{\hat{SEN}} \quad (6)$$

$$\text{Limit of quantification} = \text{LOQ} = 10\delta x\left\|b_k\right\| = 10\delta x\frac{1}{\hat{SEN}} \quad (7)$$

where the vector of sensitivities $\mathbf{S}_{\mathbf{k}}^{\mathbf{nas}}$ must be the same for all calibration samples, $\hat{\mathbf{X}}_{\mathbf{A,k}}^{\mathbf{nas}}$ is the vector for the net analyte signal for the k analyte and y_i is the reference value of the sample i. $\mathbf{x}_{\mathbf{k,un}}$ is the Euclidean norm of the original vector of the instrument responses. δx is an estimate for the instrumental noise, calculated as the standard deviation of 15 blank samples. n is the number of samples and m the number of replicates.

Results and Discussion

The objective of this work was to develop a methodology to determine the active principle of olanzapine in a mixture of seven pharmaceutical excipients (lactose, microcrystalline cellulose, poloxamer, crospovidone, silicon dioxide, magnesium stearate and coating mixture) in laboratory samples using a simple, rapid and non-destructive method. The raw NIR spectra (27 laboratory samples and 18 commercial samples) show the main effect of variations on NIR-spectra (baseline offset and overlapping peak). The spectrum for the pharmaceutical preparation was highly similar to that for all excipients, being consistent with the low concentrations of the active principle. The best models obtained during the pretreatment stage utilized Savitzky-Golay smoothing (with a window of 7 points), MSC and the first derivative of the Savitzky-Golay polynomial (with a window of 7 points), as can be seen in Figure 3.

A PLS-regression model was developed for active principle and validated by leave-one-out full-cross-validation and the optimal number of PLS factors chosen like the minimum in the graph of residual variance *versus* the number of factors. PLS is a mathematical method that is able to describe the covariance between multidimensional NIR spectral data and response variables by means of a small number of latent variables or PLS factors. Six latent variables were found to sufficiently describe the variance in the spectra (99%). The performed calibration models achieved low RMSEC (3.2×10^{-3}% m/m), RMSEP (4.0×10^{-3}% m/m) and high regression coefficients for calibration ($r_c = 0.95$) and prediction ($r_p = 0.93$). Figure 4 shows the relationship between the predicted and reference

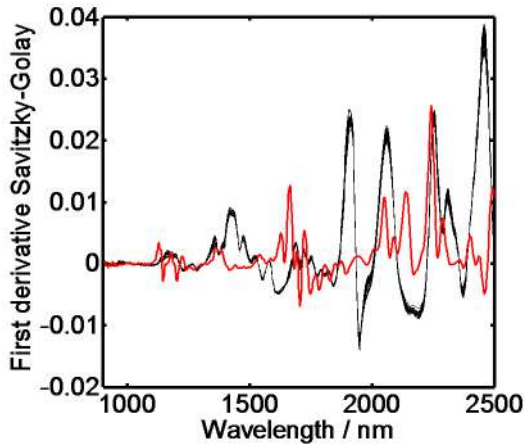


Figure 3. NIR derivative spectra of the active principle of olanzapine (red line) and original 27 laboratory samples and 18 commercial samples after pretreatment [(Smoothing, MSC and a Savitzky-Golay first derivative, black line)].

values of the laboratory samples in the calibration and validation sets. The diagonal black line represents ideal results, where the closer the points plot to the diagonal, the better the fit to the model. All the calculated concentrations including samples of both calibration and test sets were close to the real values.

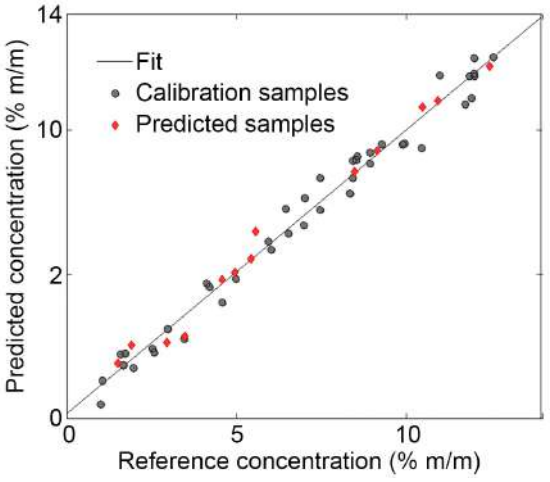


Figure 4. Predicted *versus* reference concentration from calibration and validation samples for olanzapine using the PLS model. (●) calibration set; (◆) validation set.

In order to gain further insight into the accuracy of the methods, linear regression analysis of nominal *versus* found concentration values was applied. The estimated intercept and slope were compared with their ideal values of 0 and 1 using the EJCR test. EJCR calculations are a convenient means to ascertain if bias exists in determining both parameters when using the PLS model. As can be seen in Figure 5, the point (a = 0, b = 1) was inside the EJCR, therefore it can be concluded that constant and proportional bias are absent.

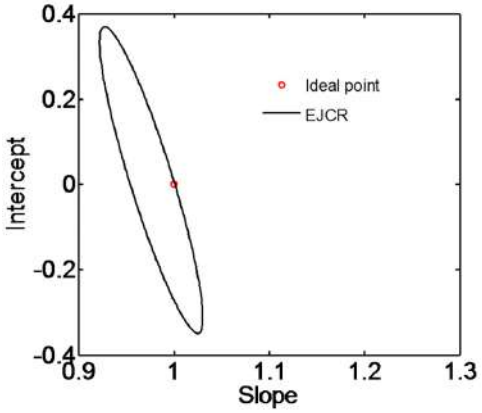


Figure 5. Elliptical joint confidence region for the regression slope and intercept of predicted *versus* reference concentration of olanzapine using an external validation set by PLS model.

Based on the comparison analysis above, PLS model with smoothing, MSC and first derivative spectral pretreatments were applied to predict the olanzapine of 12 unknown samples (laboratory samples, n = 1-6, commercial samples, n = 7-12) after similar spectral pretreatment to the calibration ones, as is shown in Table 1. To compare the methods between conventional (HPLC) measurement and PLS algorithm, the paired *t*-test method was applied. The paired *t*-test revealed no significant statistical difference between the two methods (NIR and HPLC) at a 95% confidence level ($p = 0.05$ and $t = 1.06$). The repeatability of the chromatographic method was followed as described by ANVISA[22] and assessed by the injection of the standard preparation at the sample concentration of all samples (100 ppm) in six replicates, according to the HPLC analysis. The HPLC method presented a precision with RSD 0.042%.

Table 1. Comparison results with reference method for commercial samples by NIR and HPLC

Sample		Concentration / % (m/m)	
		Predicted (NIR)	Reference (HPLC)
1	laboratory	0.034	0.033
2		0.044	0.047
3		0.036	0.045
4		0.054	0.060
5		0.042	0.043
6		0.013	0.012
7	commercial	0.028	0.023
8		0.021	0.023
9		0.023	0.023
10		0.025	0.023
11		0.023	0.024
12		0.022	0.024

NIR: near infrared spectroscopy; HPLC: high performance liquid chromatography.

New analytical methods must be validated prior to use by the pharmaceutical industry. The proposed NIR method was validated in accordance with ICH guidelines by assessing its selectivity, sensitivity, analytical sensitivity, precision, accuracy, limit of detection and limit of quantification. Table 2 presents the FOM assessed for the optimized model. Accuracy values represented by RMSEC and RMSEP indicated the estimated multivariate model values exhibited acceptable agreement with the reference method. Precision at a level of repeatability was assessed by analyzing five samples/ten replicates *per* sample, with measurements recorded on the same day, through an estimate of the relative standard deviation (RSD). The method was considered precise, with a repeatability RSD value of 4.02%. Trueness was estimated through absolute error parameters, such as a RMSEP of 4.0×10^{-3}% m/m. Trueness and precision results corroborated that the method can be considered accurate. Considering accuracy and linearity studies, the analytical working range was defined from 1.0 to 12.5% for olanzapine. Acceptable results were observed for sensitivity and sensibility to the evaluated parameters, considering the analytical range of the model. The results estimated for LOD and LOQ values might be optimistic.

Table 2. FOM (figures of merit) for the best performing PLS model: olanzapine in % m/m

FOM	Olanzapine content / % (m/m)
RMSEC	3.2×10^{-3}
RMSECV	4.4×10^{-3}
RMSEP	4.0×10^{-3}
Bias	-1.87×10^{-6}
r_c	0.95
r_p	0.93
Slope	0.952
Intercept	0.0014
Precision	4.02
SEN	0.1834
SEL	1.7988×10^{-4}
LOD	0.000619
LOQ	0.002

RMSEC: root mean square error of calibration; RMSECV: root mean square error of cross validation; RMSEP: root mean square error of prediction; r_c: calibration samples correlation coefficient; r_p: prediction samples correlation coefficient; SEN: sensibility; SEL: selectivity; LOD: limit of detection; LOQ: limit of quantification.

Conclusions

A NIR method was developed that allows for pharmaceutically determining olanzapine accurately and precisely in commercial drug products with minimal sample treatment. According to the results, PLS is presented as a good regression method to be used together with pretreatment steps that must be performed initially on the sample spectra, ensuring the construction of good calibration models and consistent prediction results. The NIR method was compared with the conventional (HPLC) method for tablet samples; no difference was found at 95% confidence interval. The values for accuracy, precision, and other figures of merit exhibited promising results, indicating that the model developed by NIR spectroscopy for olanzapine can be used as an alternative methodology for pharmaceutical purposes.

Acknowledgments

Marcelo V. P. Amorim would like to thank the Center for Food and Drug Research (NUPLAM/UFRN) for all materials and equipment, and PPgDITM-UFRN and PPGQ-UFRN for scientific support. K. M. G. Lima acknowledges the CNPq (grant 305962/2014-4) for financial support. F. S. L. Costa would like to acknowledge the financial support from the PPGQ/UFRN/CAPES for a fellowship.

References

1. Seeman, P.; *Can. J. Psychiatry* **2002**, *47*, 27.
2. Citrome, L.; Holt, R. I. G.; Walker, D. J.; Hoffman, V. P.; *Clin. Drug Invest.* **2011**, *31*, 455.
3. Navari, R. M.; *Eur. J. Pharmacol.* **2014**, *722*, 180.
4. Boulton, D. W.; Markowitz, J. S.; DeVane, C. L.; *J. Chromatogr. B: Biomed. Sci. Appl.* **2001**, *759*, 319.
5. D'Arrigo, C.; Migliardi, G.; Santoro, V.; Spina, E.; *Ther. Drug Monit.* **2006**, *28*, 388.
6. Kasper, S. C.; Mattiuz, E. L.; Swanson, S. P.; Chiu, J. A.; Johnson, J. T.; Garner, C. O.; *J. Chromatogr. B: Biomed. Sci. Appl*. **1999**, *726*, 203.
7. Nirogi, R. V. S.; Kandikere, V. N.; Shukla, M.; Mudigonda, K.; Maurya, S.; Boosi, R.; Yerramilli, A.; *J. Pharm. Biomed. Anal.* **2006**, *41*, 935.
8. Hamm, G.; Bonnel, D.; Legouffe, R.; Pamelard, F.; Delbos, J.-M.; Bouzom, F.; Stauber, J.; *J. Proteomics* **2012**, *75*, 4952.
9. Hertrampf, A.; Sousa, R. M.; Menezes, J. C.; Herdling, T.; *J. Pharm. Biomed. Anal.* **2016**, *124*, 246.
10. Sánchez-Paternina, A.; Román-Ospino, A. D.; Martínez, M.; Mercado, J.; Alonso, C.; Romañach, R. J.; *J. Pharm. Biomed. Anal.* **2016**, *123*, 120.
11. Alvarenga, L.; Ferreira, D.; Altekruse, D.; Menezes, J. C.; Lochmann, D.; *J. Pharm. Biomed. Anal.* **2008**, *48*, 62.
12. Blanco, M.; Bañó, R. G.; Bertran, E.; *Talanta* **2002**, *56*, 203.
13. Boyer, C.; Gaudin, K.; Kauss, T.; Gaubert, A.; Boudis, A.;

Verschelden, J.; Franc, M.; Roussille, J.; Boucher, J.; Olliaro, P.; White, N. J.; Millet, P.; Dubost, J.-P.; *J. Pharm. Biomed. Anal.* **2012**, *67-68*, 10.
14. Neves, A. C. O.; Soares, G. M.; de Morais, S. C.; da Costa, F. S. L.; Porto, D. L.; de Lima, K. M. G.; *J. Pharm. Biomed. Anal.* **2012**, *57*, 115.
15. Blanco, M.; Alcalá, M.; *Anal. Chim. Acta* **2006**, *557*, 353.
16. Blanco, M.; Valdés, D.; Llorente, I.; Bayod, M.; *J. Pharm. Sci.* **2005**, *94*, 1336.
17. Bodson, C.; Rozet, E.; Ziemons, E.; Evrard, B.; Hubert, P.; Dellatre, L.; *J. Pharm. Biomed. Anal.* **2007**, *45*, 356.
18. Xie, Y.-L.; Kalivas, J. H.; *Anal. Chim. Acta* **1997**, *348*, 29.
19. Makino, Y.; Ichimura, M.; Oshita, S.; Kawagoe, Y.; Yamanaka, H.; *Food Chem.* **2010**, *121*, 533.
20. Shao, Y.; Zhao, C.; Bao, Y.; He, Y.; *Food Bioprocess Technol.* **2012**, *5*, 100.
21. Thompson, M.; Ellison, S. L. R.; Wood, R.; *Pure Appl. Chem.* **2002**, *74*, 835.
22. Agência Nacional de Vigilância Sanitária (ANVISA); Resolução No. 899, *Guia para Validação de Métodos Analíticos e Bioanalíticos*; Brasília, Brasil, 2003.
23. *Farmacopeia Brasileira*, 5ª ed.; Agência Nacional de Vigilância Sanitária (ANVISA): Brasília, 2010.
24. EMEA/CHMP/CVMP/QWP; *Guideline on the Use of Near Infrared Spectroscopy by the Pharmaceutical Industry and the Data Requirements for New Submissions and Variations*, Agency of European Union: London, UK, 2009.
25. El-Hagrasy, A. S.; D'Amico, F.; Drennen, J. K.; *J. Pharm. Sci.* **2006**, *95*, 392.
26. Mariani, N. C. T.; da Costa, R. C.; de Lima, K. M. G.; Nardini, V.; Cunha Jr., L. C.; Teixeira, G. H. D. A.; *Food Chem.* **2014**, *159*, 458.
27. Kennard, R.; Stone, L.; *Technometrics* **1969**, *11*, 137.
28. Braga, J. W. B.; Poppi, R. J.; *Quim. Nova* **2004**, *27*, 1004.

Submitted: May 24, 2016
Published online: August 12, 2016

APÊNDICE D

A Multivariate Control Chart Approach for Calibration Transfer between NIR Spectrometers for Simultaneous Determination of Rifampicin and Isoniazid in Pharmaceutical Formulation

Eduardo W. V. Andrade

Camilo L. M. Morais **Fernanda S. L. Costa**

Kássio M. G. Lima

Current Analytical Chemistry, 2018, 14, 488-494.

Contribuição:

- Ajudei na escrita do manuscrito;
- Ajudei na construção das cartas de controle.

Send Orders for Reprints to reprints@benthamscience.ae

Current Analytical Chemistry, **2018**, *14,* **000-000**

RESEARCH ARTICLE

A Multivariate Control Chart Approach for Calibration Transfer between NIR Spectrometers for Simultaneous Determination of Rifampicin and Isoniazid in Pharmaceutical Formulation

Eduardo Wagner Vasconcelos de Andrade, Camilo de Lelis Medeiros de Morais, Fernanda Saadna Lopes da Costa and Kássio Michell Gomes de Lima*

Biological Chemistry and Chemometrics, Institute of Chemistry, Federal University of Rio Grande do Norte, Natal, Brazil

ARTICLE HISTORY

Received: June 29, 2017
Revised: November 13, 2017
Accepted: November 14, 2017

DOI:
10.2174/1573411014666171212141909

Abstract: ***Background:*** Multivariate transfer techniques have become a widely accepted concept over the past few years, since they avoid full recalibration procedures when instruments are changed to analyze a specific sample.

Objective: This paper reports a multivariate control chart transfer approach between two near infrared (NIR) spectrometers for simultaneous determination of rifampicin and isoniazid in pharmaceutical formulation using direct standardization (DS).

Method: The control charts are based on the calculation of net analyte signal (NAS) models and the transfer samples are selected by the Kennard-Stone (KS) algorithm. Three control charts (NAS, interference and residual) transferred on both the master and slave instruments were measured.

Results: As a result, a classification model for rifampicin and isoniazid developed on a primary instrument has been successfully transferred to a secondary instrument. The spectral differences after the standardization procedure were considerably reduced and errors values found in the charts for both analytes were comparable with the errors obtained for the original chart models.

Conclusion: The proposed approach appears to be a valid alternative to the commonly used transfer of multivariate calibration models in simultaneous determination of isoniazid and rifampicin in pharmaceutical formulation.

Keywords: Multivariate control chart, NAS, isoniazid, rifampicin, direct standardization, NIR.

1. INTRODUCTION

Multivariate calibration transfer techniques (also known as instrumental standardization) have become a widely accepted concept over the past few years mainly due to avoiding the use of time-consuming complete recalibration procedures [1-5]. Usually, the instrument standardization procedure for multivariate calibration transfer involves two steps: (i) a set of standardization samples are measured on both instruments to evaluate their different responses; (ii) standardization parameters are computed with standardization samples and used for spectra transfer [6, 7].

Direct standardization (DS) [8], piecewise direct standardization (PDS) [9], orthogonal signal correction (OSC) [3], reverse standardization (RS) [10], piecewise reverse standardization (PRS) [11], slope and bias correction (SBC) [1], orthogonal projections to latent structures (O-PLS) [12] and model updating (MU) [4] are examples that have been successfully applied to various calibration transfer problems. These methods relatively correct differences between data collected by two instruments where the entire spectra from the new (secondary) instrument are transformed by relating its spectral variables (e.g., wavelengths) to resemble the spectral data from the original (primary) instrument used to build a prior calibration model [13].

On the other hand, there are many situations in which the simultaneous monitoring or control of two or more related quality–process characteristics is necessary. Multivariate control charts based on principal component analysis [14-16], partial least squares [17], multivariate exponential weighted moving average [18], multivariate cumulative sum [19] and Bayesian probability [20] are some examples for building an empirical model of a set of measurements achieved under normal operating conditions (NOC).

An interesting approach for quality multivariate control chart is based on net analyte signal (NAS) [21-24]. This method is carried out by the decomposition of a sample spectrum into a vector that is unique for the analyte; a vector re-

*Address correspondence to Prof. Dr. Kássio M. G. Lima at Biological Chemistry and Chemometrics, Institute of Chemistry, Federal University of Rio Grande do Norte, Natal 59072-970, Brazil; Tel: +55 84 3342 2323; E-mail: kassiolima@gmail.com

1573-4110/18 $58.00+.00

lated to the other compounds in the sample (interfering constituents); and a remaining residual vector [25]. Thereafter, the statistical limits for the NAS control charts are derived from the NAS value for each of the NOC spectra calculated [26].

This paper investigates a multivariate control chart transfer approach between two near infrared (NIR) spectrometers (primary and secondary) for simultaneous determination of rifampicin and isoniazid in pharmaceutical formulation, which are important drugs used in tuberculosis treatment [27,28]. Three control charts (NAS, interference and residual) transfer using NAS and DS between the primary and secondary instruments were developed. These control charts were built with the spectral data before and after calibration transfer, in which the classification rates were evaluated before and after DS according to the upper and low limits on these charts.

2. MATERIALS AND METHOD

2.1. Samples

The pharmaceutical preparation studied contained isoniazid (99.29%, Amsal Quality Control Laboratory, India) and rifampicin (98.87%, Sanofi Aventis, Italy) as the active principles and four excipients (magnesium stearate, sodium starch glycolate, talc and amide). All compounds (active principles and excipients) were supplied by the Center for Food and Drug Research of the Federal University of Rio Grande do Norte (NUPLAM/UFRN) – Brazil. The capsules produced at UFRN were available in one absolute active pharmaceutical ingredient (API) content per dose. The capsules were uncoated, thus permitting diffuse reflectance.

Laboratory samples were prepared by a D-optimal experimental design using MODDE® 4.0 (MKS Data Analytics Solutions, Umeå, Sweden). D-optimal design is performed when the classical symmetrical design cannot be used because the shape of the experimental region is irregular or the number of experiments selected by a classical design is too large. A total of 120 pharmaceutical formulation samples were generated to efficiently represent the design space for the large number of possible combinations of these substances and to build the NAS charts. These samples were weighed on an analytical scale with a total weight accuracy of 0.012 g. Then, the samples were mixed for 3 min and vortexed for 1 min before NIR analysis.

Next, the samples were used to design news samples varying API concentrations (isoniazid and rifampicin) and to provide a variable matrix from which NAS charts could be derived providing an independent set of samples that could be used to check the accuracy of the control charts for each instrument. Samples containing only the excipients (blank samples) were also prepared. The samples (blank, in control and out-of-control) used for NAS, interference and residual charts were distributed as follows:

i) 20 blank samples: 10 samples for isoniazid and 10 samples for rifampicin;

ii) 10 samples in control (isoniazid, rifampicin, magnesium stearate, microcrystalline cellulose, talc and starch);

iii) 20 samples in control (2.5% of the nominal content of each active substance);

iv) 20 samples in control (5.0% of the nominal content of each active substance);

v) 20 samples out-of-control (8.0% of the nominal content of each active substance);

vi) 20 samples out-of-control (12.0% of the nominal content of each active substance);

vii) 10 samples out-of-control (16.0% of the nominal content of each active substance).

2.2. Instruments

The primary (master) instrument used was an Antaris MX Fourier Transform NIR spectrophotometer (Thermo Fisher Scientific Inc., USA) equipped with a transflectance optical fiber probe being positioned onto the sample surface (less than 1 cm and at 90° from the surface). The transflectance probe was washed with ethanol (70% v/v) and dried using tissue paper after each sample. The spectrum of a polytetrafluoroethylene sample was used as the background. The NIR spectra were obtained over a range of 1000–2400 nm, and were recorded with a spectral resolution of 1 nm, with 32 scans co-added. The measurement time was 26 s (32 scans) per spectrum. A Fourier Transform NIR MPA spectrometer was used as the secondary (slave) instrument (Bruker Optics, Germany) equipped with an integrating sphere via diffuse reflection mode. Each measured spectrum (in triplicate) was the average of 32 scans obtained with a resolution of 2 nm and over the range of 1000–2400 nm. The background spectrum was recorded using a gold coated slide. Spectral measurements for both instruments were done in an acclimatized room under controlled temperature of 22°C and 60% relative air humidity.

2.3. Data Analysis

The data import, pre-processing, and construction of multivariate control charts were implemented in MATLAB® version 7.12.0 (MathWorks Inc., USA) using an in-house developed algorithm. Different preprocessing methods were tested, including baseline correction; multiplicative scatter correction (MSC); variance scaling; derivative; and Savitzky-Golay smoothing using first- and second-order polynomial functions varying the number of window points (7, 11 and 15). However, the best pre-processing were baseline correction and MSC for isoniazid charts; and baseline correction for rifampicin charts. These pre-processing were the same for both equipment. The technique chosen for selection of transfer samples was the classic Kennard Stone (KS) algorithm [29].

3. THEORY

Fundamentally, in order to build a multivariate control chart based on NAS, an out-of-control indicator is required for diagnostic and corrective measures. In this sense, two steps are required: 1) (diagnostic) discovery which measurement variables contribute to the out-of-control signal and 2) (corrective) determining what occurs in the process that disturbs the behavior of these variables.

The first step to perform before any standardization method is to select the standardization samples to transfer, which is commonly obtained by using sample selection techniques, such as KS algorithm [29] or leverage [9]. The number of transfer samples is evaluated by an arbitrary cost function, which for calibration models is usually the root-mean-squared error of prediction [13]. In our case, for classification purpose, this cost function was calculated as the classification rate of the NAS control charts.

The DS is a multivariate standardization method employed to correct relatively large differences between data collected by two instruments [9]. In this method, the entire spectra from the new (secondary) instrument are transformed by relating its spectral variables (e.g., wavelengths) to resemble the spectral data from the original (primary) instrument used to build a prior calibration model [13]. The linear relationship between the primary and secondary response is described by the transformation matrix F according to Eq. (01) [4]:

$$S_1 = S_2F \quad (01)$$

where S_1 and S_1 are the data matrices of the standardization samples for the primary and secondary instruments, respectively.

Thus, the transformation matrix is estimated in a least-squares sense according to Eq. (02) [30]:

$$F = S_2^+ S_1 \quad (02)$$

where S_2^+ is the pseudo-inverse of S_2. S_2 must contain independent rows (samples) or columns (variables) for the pseudo-inverse calculation to be feasible (Eq. (03)).

$$S_2^+ = (S^TS)\ ^1S^T \quad (03)$$

After the calculation of F, the projection of the response vector for a new sample x from the secondary instrument on the original space from the primary instrument is estimated according to Eq. (04) [4]:

$$\hat{X}^T = X^TF \quad (04)$$

where $\hat{x}$ is the standardized response vector for x.

In order to solve possible problems related to different background information in both instruments, the standardization process was performed using the background correction method [30] where the data matrices of standardization samples from the primary and secondary instruments relate to each other by the transformation matrix calculated with the background correction F_b and an additive background correction vector b_s according to Eq. (05):

$$S_1 = S_2F_b + 1b_s^T \quad (05)$$

where b_s is obtained using Eq. (06):

$$b_s = s_{1m} - F_b^Ts_{2m} \quad (06)$$

in which S_{1m} is the mean vector of matrix S_1 and S_{2m} is the mean vector of matrix S_2.

Multivariate control charts based on NAS provide multivariate product quality monitoring and they are carried out in two stages: (i) model building and (ii) calculation of statistical limits [26]. The first stage consists on the decomposition of a sample spectrum r into three vectors: a vector r_{NAS} that is unique for the analyte; a vector r_{INT} that is related to the other compounds in the sample (interfering constituents); and a residual vector r_{res} [25]:

$$r = r_{NAS} + r_{INT} + r_{res} \quad (07)$$

In the second stage, the statistical limits of the NAS control charts are derived from the NAS value for each of the NOC spectra calculated as follows:

$$nas_{NOC} = R_{NOC}^Tb_k \quad (08)$$

where nas_{NOC} is a vector with the NAS value of the individual NOC spectra; R_{NOC} is the set of NOC spectra used to set the control limits for the NAS chart; and b_k is the orthogonal part of the model spectra used to define the NAS direction on the interference space [26]. The NAS values are assumed to follow a normal distribution, which can be verified by statistical normality tests such as QQ plot [31]. Its mean and standard deviation are computed for statistical limits (95% confidence limits called the upper and lower warning lines, and 99.7% confidence limits called the upper and lower action lines) that are plotted in the NAS control chart [26]. The classification rate was calculated based on the 2-sigma (95%) confidence interval; so that any sample outside this limit would be considered out of control. The 3-sigma (99.7%) confidence interval was not used for classification evaluation but it represents the limit with the largest probability for a sample be identified as out of control.

The interference chart is based on projecting the RNOC matrix on the interference space. The projected "under control" spectra occupy a restricted region on the interference space, wherein the pharmaceutical formulation is constructed with placebo and blank samples. The validation of these control charts is made by using "in-control" and "out-of-control" samples, based on the concentration of the active pharmaceutical ingredient [25, 26]. The residual charts are obtained after calculation of NAS and interference vector, in which its control limits are estimated based on Q-statistics by fitting a chi-squared distribution to the reference distribution obtained from NOC data [25, 26]. The Q-statistics is the first type of statistical calculation recommend to test significance of an individual observation vector [32]. It is calculated for a sample vector following a chi-squared distribution [26]:

$$Q_{NOC} \sim g\chi_h^2 \quad (09)$$

where Q_{NOC} contains the Q-statistics of the NOC spectra; *g* represents the weight to account for the magnitude; and X_h^2 is the chi-squared distribution to the reference distribution obtained from NOC data, where the parameter *h* denotes the degrees of freedom. Using this statistics, if we have a situation where the residual vector of a new sample is not only random noise then the observation will have a large Q-statistics and flag in the residual chart [26]. In addition, the chi-squared distribution is very adequate for large samples sets [33], therefore, being very suitable for industrial or routine applications. On the other hand, statistics such as F dis-

tribution was not used due to the lack of sensitivity to detect normality of distribution in the residuals [33].

The control chart transfer procedure is summarized on the flowchart in Fig. (**1**). In this, the spectral data is separated into three sets: Xcal(M) corresponding to the calibration samples of the primary (master) instrument; Xcal(S) corresponding to the calibration samples of the secondary (slave) instrument; and Xpred(S) corresponding to the prediction samples of the secondary instrument to be analyzed using the control chart of the primary instrument. The spectral resolution and the number of calibration samples used for transferring data from both instruments must be equal or otherwise all algebraic operations will not be possible since the matrices sizes would be different. Therefore, an algorithm for correct spectral resolution was inserted on the transfer approach. This algorithm resizes the spectra of higher resolution to match with the spectra of lower resolution keeping its shape constant; in other words, it "compresses" the largest spectra. The transfer samples are selected by the calibration indexes found employing KS algorithm on the Xcal(M) data.

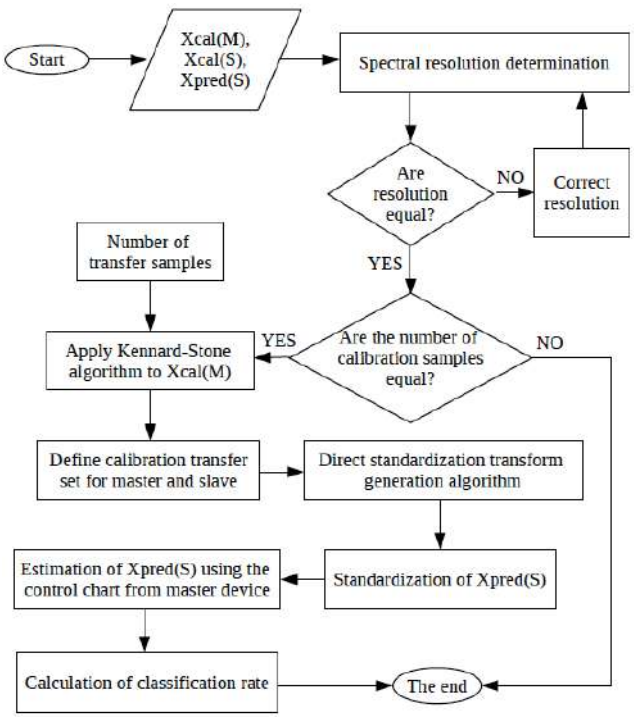


Fig. (1). Flowchart for control chart transfer procedure using DS.

The DS transferring is performed by combining the transfer samples from both instruments. Then, the prediction set from the secondary instrument (Xpred(S)) is standardized by using the transformation matrix with additive background correction (Eq. (05)). At the end, the standardized Xpred(S) is analyzed by the NAS-based primary control charts and the classification rate is calculated. This parameter is used as the cost function to define the ideal number of transfer samples. After the model is optimized with the ideal number of samples to transfer, all external prediction samples from the secondary instrument are standardized and predicted using the primary control chart.

4. RESULTS AND DISCUSSIONS

Fig. (**2**) shows the raw NIR spectra of a pharmaceutical formulation sample acquired on the two instruments employed in this study. They are the averages of triplicate measurements for each sample recorded in the region from 1000 to 2400 nm.

As mentioned before, the primary and secondary instruments were from different manufacturers and different measurement procedures were employed with each. As can be seen in Fig. (**2**), there are resulting spectral differences between master and slave measurements. Some preprocessing methods needed to be applied to reduce instrumental noise and light scattering that can affect the baseline. The performance of each preprocessing method was evaluated according to their correct classification rate (predicted sample index equal to the correct class index) and incorrect classification rate (predicted sample index different from the correct class index) using a calibration and validation set. The best prediction rates were obtained using baseline correction combined with MSC for isoniazid control chart; and baseline correction for rifampicin control chart (see Fig. **3**). The spectral differences between both instruments motivate the use of control chart transfer techniques.

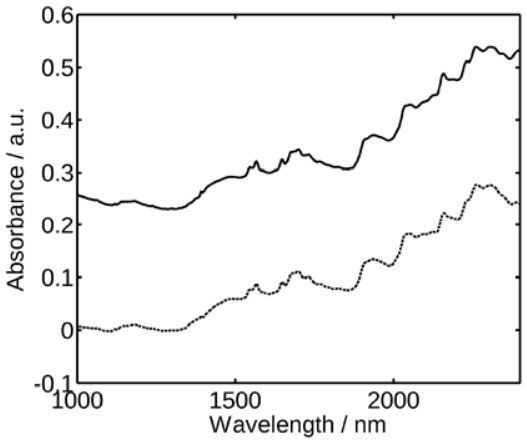


Fig. (2). Spectra of a representative pharmaceutical formulation sample acquired on two NIR instruments: dashed line represents the NIR spectrophotometer equipped with transflectance optical fiber probe (primary); and the continuous line represents the NIR spectrophotometer equipped with an integrating sphere (secondary).

For isoniazid, the control charts (NAS, interference and residual) constructed for master instrument achieved the following correct classification rates: 92% (NAS chart); 100% (interference chart); and 100% (residual chart). When the isoniazid model was directly applied to the slave instrument, the following correct classification rates were achieved: 71% (NAS chart); 100% (interference chart); and 100% (residual chart). The prediction accuracy was particularly poor for the slave instrument because of the major spectral differences between this instrument and the master. Fig. (**4**) and Fig. (**5**) show the control charts developed for isoniazid using the master and slave instrument, respectively. These results justify the application of multivariate control chart transfer to the acquired data, as without a transfer technique, the correct classification rate may be completely different when the NIR spectra from an instrument is validated into another.

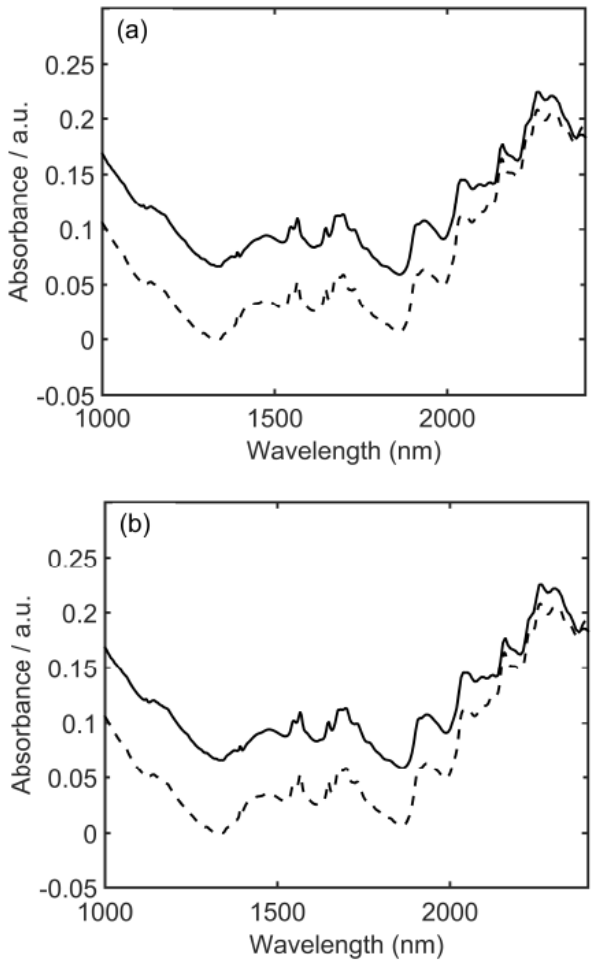


Fig. (3). NIR spectra after application of: **a**) baseline to rifampicin; **b**) baseline and MSC to isoniazid. Dashed line – primary instrument; and continuous line – secondary instrument.

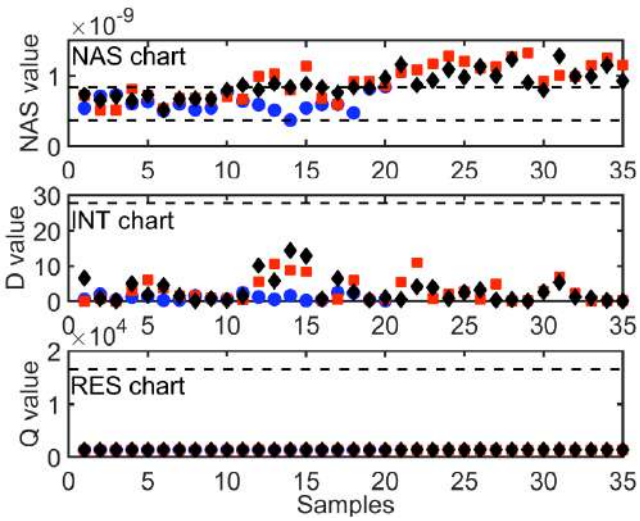


Fig. (4). Control charts for isoniazid using master instrument: (●) calibration, (■) validation and (◆) prediction. NAS: net analyte signal; INT: interference; RES: residual.

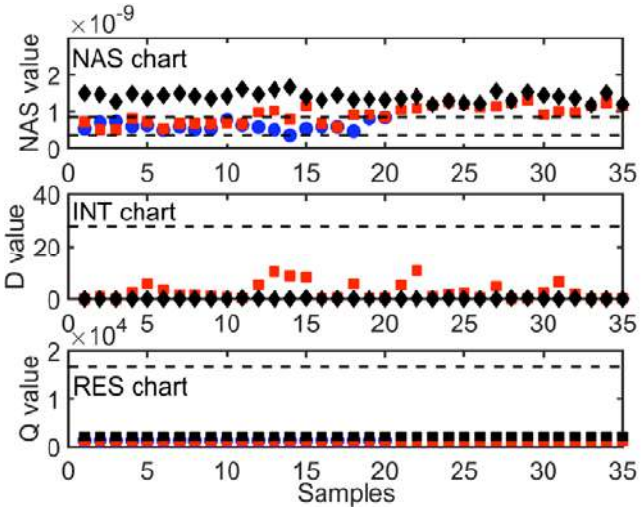


Fig. (5). Control charts for isoniazid using slave instrument: (●) calibration, (■) validation and (◆) prediction. NAS: net analyte signal; INT: interference; RES: residual.

The prediction performance of the multivariate control charts (NAS, interference and residual) for isoniazid in the master and slave instruments after DS multivariate control chart transfer procedure were calculated by using 20 transfer samples selected by KS algorithm. The standardization improved the correct classification for NAS chart from 71% (without DS) to 92% (after DS); and maintained the same correct classification rates for interference (100%) and residual (100%) charts. These correct classification rates are satisfactory, mainly for NAS chart, considering the simplicity of the multivariate control chart transfer used, and showing their importance to avoid a full recalibration step.

For rifampicin, the control charts (NAS, interference and residual) built for the master instrument achieved the following correct classification rates: 86% (NAS chart); 99% (interference chart); and 73% (residual chart). When the isoniazid model was directly applied to the slave instrument, the following correct classification rates were achieved: 71% (NAS chart); 99% (interference chart); and 60% (residual chart). The prediction accuracy was particularly poor for the slave instrument because of the major spectral differences between this instrument and the master. However, the standardization improved the correct classifications for NAS chart (86%), interference chart (99%) and residual chart (73%) using 11 transfer samples selected by KS algorithm. Fig. (**6**) and Fig. (**7**) show the control charts developed for rifampicin using the master and slave instruments, respectively.

These classification values demonstrate that after multivariate transfer the response obtained with the secondary instrument gave the same results observed with the primary instrument despite the differences of resolution, equipment and probe. Therefore, the standardization methodology shown herein was a successful case for rinfampicin and isoniazid determination using NAS control charts constructed with different NIR spectrometers, which can avoid a full recalibration when analyzing these samples with different NIR equipment and shows its potential to further applications.

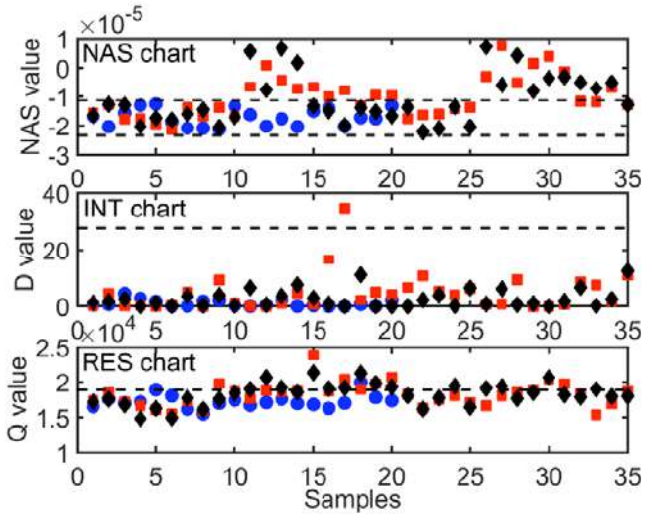


Fig. (6). Control charts for rifampicin using master instrument: (●) calibration, (■) validation and (◆) prediction. NAS: net analyte signal; INT: interference; RES: residual.

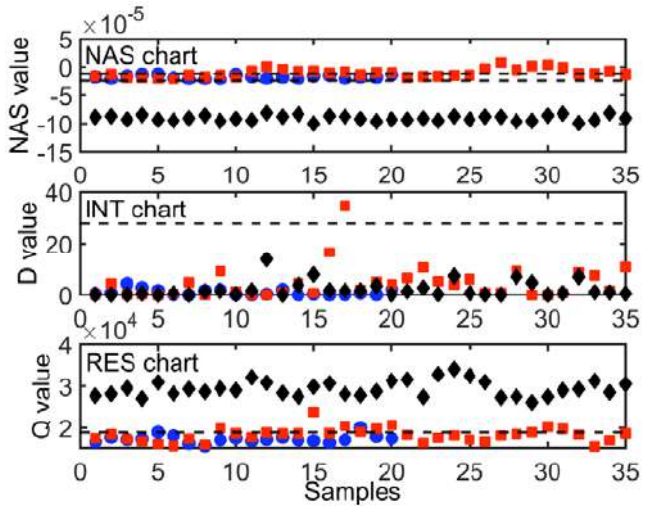


Fig. (7). Control charts for rifampicin using slave instrument: (●) calibration, (■) validation and (◆) prediction. NAS: net analyte signal; INT: interference; RES: residual.

CONCLUSION

This paper presents a multivariate control chart transfer approach between two NIR spectrometers for simultaneous determination of rifampicin and isoniazid in pharmaceutical formulation using DS. The study reported herein supports the usefulness and effectiveness of this approach for simultaneous determination of isoniazid and rifampicin using NIR spectroscopy. The results (in terms of correct classification) demonstrated that the direct application of the master instrument to the control charts (NAS, interference and residual) acquired on a slave instrument may lead to poor predictions, making the use of multivariate control chart transfer necessary.

LIST OF ABBREVIATIONS

API	=	Active pharmaceutical ingredient
DS	=	Direct standardization
KS	=	Kennard-Stone
MSC	=	Multiplicative scatter correction
MU	=	Model updating
NAS	=	Net analyte signal
NIR	=	Near infrared
NOC	=	Normal operating conditions
O-PLS	=	Orthogonal projections to latent structures
OSC	=	Orthogonal signal correction
PDS	=	Piecewise direct standardization
PRS	=	Piecewise reverse standardization
RS	=	Reverse standardization
SBC	=	Slope and bias correction

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are base of this research.

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Eduardo W. V. Andrade thanks PROPESQ and PIBIT/NUPLAM/UFRN for financial support. Camilo L. M. Morais and Fernanda S. L. Costa would like to acknowledge PPGQ/UFRN and CAPES for financial support. K.M.G. Lima acknowledges the CNPq/CAPES project (Grant 070/2012 and 305962/201-4) and FAPERN (PPP 005/2012) for financial support.

REFERENCES

[1] Abdelkader, M.F.; Cooper, J.B.; Larkin, C.M. Calibration transfer of partial least squares jet fuel property models using a segmented virtual standards slope-bias correction method. *Chemom. Intell. Lab. Syst.*, **2012**, 110, 64-73.

[2] Honorato, F.A.; Galvão, R.K.H.; Pimentel, M.F.; Neto, B.B.; Araújo, M.C.U.; de Carvalho, F.R. Robust modeling for multivariate calibration transfer by the successive projections algorithm. *Chemom. Intell. Lab. Syst.*, **2005**, 76, 65-72.

[3] Greensill, C.V.; Wolfs, P.J.; Spiegelman, C.H.; Walsh, K.B. Calibration Transfer between PDA-Based NIR Spectrometers in the NIR Assessment of Melon Soluble Solids Content. *Appl. Spectrosc.*, **2001**, 55, 647-653.

[4] Feudale, R.N.; Woody, N.A.; Tan, H.; Myles, A.J.; Brown, S.D.; Ferré, J. Transfer of multivariate calibration models: a review. *Chemom. Intell. Lab. Syst.*, **2002**, 64, 181-192.

[5] Panchuk, V.; Kirsanov, D.; Oleneva, E.; Semenov, V.; Legin, A. Calibration transfer between different analytical methods. *Talanta*, **2017**, 170, 457-463.

[6] Lin, J.; Lo, S.-C.; Brown, C.W. Calibration transfer from a scanning near-IR spectrophotometer to a FT-near-IR spectrophotometer. *Anal. Chim. Acta*, **1997**, 349, 263-269.

[7] Wülfert, F.; Kok, W.T.; de Noord, O.E.; Smilde, A.K. Correction of Temperature-Induced Spectral Variation by Continuous Piecewise Direct Standardization. *Anal. Chem.*, **2000**, 72, 1639-1644.

[8] Zamora-Rojas, E.; Pérez-Marín, D.; De Pedro-Sanz, E.; Guerrero-Ginel, J.E.; Garrido-Varo, A. Handheld NIRS analysis for routine meat quality control: Database transfer from at-line instruments. *Chemom. Intell. Lab. Syst.*, **2012**, 114, 30-35.

[9] Wang, Y.; Veltkamp, D.J.; Kowalski, B.R. Multivariate Instrument Standardization. *Anal. Chem.*, **1991**, 63, 2750-2756.

[10] de Noord, O.E. Multivariate calibration standardization. *Chemom. Intell. Lab. Syst.*, **1994**, 25, 85-97.

[11] Lima, F.S.G.; Borges, L.E.P. Evaluation of standardisation methods of near infrared calibration models. *J. Near Infrared Spectrosc.*, **2002**, 10, 269-278.

[12] Rodrigues, R.R.T.; Rocha, J.T.C.; Oliveira, L.M.S.L.; Dias, J.C.M.; Müller, E.I.; Castro, E.V.R.; Filgueiras, P.R. Evaluation of calibration transfer methods using the ATR-FTIR technique to predict density of crude oil. *Chemom. Intell. Lab. Syst.*, **2017**, 166, 7-13.

[13] Pereira, C.F.; Pimentel, M.F.; Galvão, R.K.H.; Honorato, F.A.; Stragevitch, L.; Martins, M.N. A comparative study of calibration transfer methods for determination of gasoline quality parameters in three different near infrared spectrometers. *Anal. Chim. Acta*, **2008**, 611, 41-47.

[14] Clavaud, M.; Roggo, Y.; Von Daeniken, R.; Liebler, A.; Schwabe, J.-O. Chemometrics and in-line near infrared spectroscopic monitoring of a biopharmaceutical Chinese hamster ovary cell culture: Prediction of multiple cultivation variables. *Talanta*, **2013**, 111, 28-38.

[15] Alcalà, M.; Blanco, M.; Bautista, M.; González, J.M. On-Line Monitoring of A Granulation Process By NIR Spectroscopy. *J. Pharm. Sci.*, **2010**, 99, 336-345.

[16] Tôrres, A.R.; Grangeiro Jr, S.; Fragoso, W.D. Vibrational spectroscopy and multivariate control charts: A new strategy for monitoring the stability of captopril in the pharmaceutical industry. *Microchem. J.*, **2017**, 133, 279-285.

[17] Kourti, T.; Nomikos, P.; MacGregor, J.F. Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *J. Proc. Cont.*, **1995**, 5, 277-284.

[18] Zou, C.; Tsung, F.; Wang, Z. Monitoring General Linear Profiles Using Multivariate Exponentially Weighted Moving Average Schemes. *Technometrics*, **2007**, 49, 395-408.

[19] Bodnar, O.; Schmid, W. CUSUM charts for monitoring the mean of a multivariate Gaussian process. *J. Stat. Plan. Inference.*, **2011**, 141, 2055-2070.

[20] Tian, Y.; Du, W.; Makis, V. Improved cost-optimal Bayesian control chart based auto-correlated chemical process monitoring. *Chem. Eng. Res. Des.*, **2017**, 123, 63-75.

[21] Rocha, W.F.C.; Rosa, A.L.; Martins, J.A.; Poppi, R.J. Multivariate control charts based on net analyte signal and near infrared spectroscopy for quality monitoring of Nimesulide in pharmaceutical formulations. *J. Mol. Struct.*, **2010**, 982, 73-78.

[22] Rocha, W.F.C.; Poppi, R.J. Multivariate control charts based on net analyte signal (NAS) for characterization of the polymorphic composition of Piroxicam using near infrared spectroscopy. *Microchem. J.*, **2010**, 96, 21-26.

[23] Skibsted, E.T.S.; Boelens, H.F.M.; Westerhuis, J.A.; Witte, D.T.; Smilde, A.K. Simple assessment of homogeneity in pharmaceutical mixing processes using a near-infrared reflectance probe and control charts. *J. Pharm. Biomed. Anal.*, **2006**, 41, 26-35.

[24] Sitoe, B.V.; Máquina, A.D.V.; Santana, F.B.; Gontijo, L.C.; Santos, D.Q.; Borges Neto, W. Monitoring of biodiesel content and adulterant presence in methyl and ethyl biodiesels of jatropha in blends with mineral diesel using MIR spectrometry and multivariate control charts. *Fuel*, **2017**, 191, 290-299.

[25] Costa, F.S.L.; Pedroza, R.H.P.; Porto, D.L.; Amorim, M.V.P.; Lima, K.M.G. Multivariate Control Charts for Simultaneous Quality Monitoring of Isoniazid and Rifampicin in a Pharmaceutical Formulation Using a Portable Near Infrared Spectrometer. *J. Braz. Chem. Soc.*, **2015**, 26, 64-73.

[26] Skibsted, E.T.S.; Boelens, H.F.M.; Westerhuis, J.A.; Smilde, A.K.; Broad, N.W.; Rees, D.R.; Witte, D.T. Net Analyte Signal Based Statistical Quality Control. *Anal. Chem.*, **2005**, 77, 7103-7114.

[27] Wang, P.; Pradhan, K.; Zhong, X.; Ma, X. Isoniazid metabolism and hepatotoxicity. *Acta Pharm. Sin. B*, **2016**, 6, 384-392.

[28] Campbell, E.A.; Korzheva, N.; Mustaev, A.; Murakami, K.; Nair, S.; Goldfarb, A.; Darst, S.A. Structural Mechanism for Rifampicin Inhibition of Bacterial RNA Polymerase. *Cell*, **2001**, 104, 901-912.

[29] Kennard, R.W.; Stone, L.A. Computer Aided Design of Experiments. *Technometrics*, **1969**, 11, 137-148.

[30] Wang, Z.; Dean, T.; Kowalski, B. Additive Background Correction in Multivariate Instrument Standardization. *Anal. Chem.*, **1995**, 67, 2379-2385.

[31] Wilk, M.B.; Gnanadesikan, R. Probability plotting methods for the analysis of data. *Biometrika*, **1968**, 55, 1-17.

[32] Jackson, J.E.; Mudholkar, G.S. Control procedures for residuals associated with principal component analysis. *Technometrics*, **1979**, 21, 341-349.

[33] Miller, J.N.; Miller, J.C. *Statistics and Chemometrics for Analytical Chemistry*, 6th ed.; Pearson Education Limited: Essex, **2010**.