UNIVERSIDADE FEDERAL DE SÃO PAULO

INSTITUTO DE CIÊNCIA E TECNOLOGIA

GRADUAÇÃO EM ENGENHARIA BIOMÉDICA

Desenvolvimento de um método para classificação de fungos do gênero Metarhizium utilizando dados de espectroscopia no infravermelho e técnica de machine learning

Aluno: Bruno João de Souza

Orientador: Prof. Dr. Thiago Martini Pereira

São José dos Campos, SP

2021

UNIVERSIDADE FEDERAL DE SÃO PAULO

INSTITUTO DE CIÊNCIA E TECNOLOGIA

GRADUAÇÃO EM ENGENHARIA BIOMÉDICA

Desenvolvimento de um método para classificação de fungos do gênero Metarhizium utilizando dados de espectroscopia no infravermelho e técnica de machine learning

Aluno: Bruno João de Souza

Orientador: Prof. Dr. Thiago Martini Pereira

Monografia apresentada ao Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo como requisito parcial para obtenção do título de Graduação em Engenharia Biomédica.

São José dos Campos, SP

2021

FICHA CATALOGRÁFICA

.

Elaborado por sistema de geração automática com os dados fornecidos pelo(a) autor(a).

João de Souza, Bruno
Desenvolvimento de um método para classificação de fungos do gênero Metarhizium utilizando dados de espectroscopia no infravermelho e técnica de machine learning/ Bruno João de Souza
Orientador(a) Thiago Martini Pereira-São José dos Campos, 2021.
62 p.

Trabalho de Conclusão de Curso-Engenharia Biomédica-Universidade Federal de São Paulo-Instituto de Ciência e Tecnologia, 2021.

1. IR Spectroscopy. 2. Machine Learning. 3. Bioespectroscopia. 4. Fungos. 5. Bioengenharia. I. Martini Pereira, Thiago, orientador(a). II. Título.

FOLHA DE APROVAÇÃO

BRUNO JOÃO DE SOUZA

Desenvolvimento de um método para classificação de fungos do gênero Metarhizium utilizando dados de espectroscopia no infravermelho e técnica de machine learning

Monografia apresentada ao Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo como requisito parcial para obtenção do título de Graduação em Engenharia Biomédica.

São José dos Campos,_____de_____de_____.

_____

Prof. Dr. Thiago Martini Pereira

UNIVERSIDADE FEDERAL DE SÃO PAULO

_____

Prof. Dr. Fábio Gava Aoki

UNIVERSIDADE FEDERAL DE SÃO PAULO

_____

Prof. Dr. Mateus Fernandes Réu Urban

UNIVERSIDADE FEDERAL DE SÃO PAULO

AGRADECIMENTOS

Aos meus pais **Lilian** e **Vilmar** pela base segura que me garantiram em casa, o que me permitiu focar em meus estudos e aprendizados acadêmicos de forma integral por grande parte da minha graduação e deste trabalho. Agradeço imensamente todo o amor, preocupação, paciência e cuidado que tiveram comigo e meus sentimentos durante as horas mais difíceis. Obrigado, mãe, por dedicar-se tanto a mim mesmo quando não era solicitado, mesmo quando estava enfrentando seus próprios desafios. Obrigado, pai, por garantir que eu teria todo o suporte em qualquer decisão, por ter me garantido oportunidades e por ter sido paciente com minhas conquistas.

Ao meu irmão **Vilmar Jr** pelo apoio e tempo dedicado a me ajudar com a universidade, estando sempre pronto para me auxiliar no que fosse necessário, tanto técnica quanto emocionalmente. Obrigado por me trazer a experiência de 5 anos de mais idade do que eu, me ajudando a evitar obstáculos que já havia enfrentado antes e me ajudando a traçar meus planos e objetivos claramente, especialmente neste trabalho.

Aos meus amigos **Eric** e **Bruna** que estiveram ao meu lado em centenas de desafios por mais de quatro anos na universidade. Agradeço imensamente por cada uma das horas que passamos juntos estudando, fazendo projetos e sendo simplesmente amigos inseparáveis. Obrigado por estar sempre a postos para ser a equipe dos sonhos para qualquer trabalho em grupo, para criticar e me fazer crescer e também para me abrir e me ajudar com meus problemas pessoais que, por muitas vezes, afetavam meu desempenho acadêmico e profissional.

Aos meus **colegas** da Graduação em Bacharelado em Ciência e Tecnologia e da Graduação em Engenharia Biomédica que estiveram sempre juntos para enfrentar cada novo desafio da universidade. Agradeço pelos grupos de estudos formados, pelas equipes prontas para solucionar qualquer lista de exercícios e por todos os eventos que serviram para espairecermos a cabeça. Obrigado por serem tão diferentes e tão unidos, me dando forças para continuar independente da dificuldade de cada etapa deste desafio.

Ao **Prof. Dr Thiago Martini Pereira** por toda a dedicação e paciência em lecionar pelo menos duas matérias das quais cursei. Agradeço imensamente pelo

cuidado e profissionalismo em me guiar como orientando neste presente trabalho, sempre com muita responsabilidade técnica e afetiva, equilibrando o dever e a capacidade em cada momento. Obrigado por ser tamanho porto seguro durante os dificílimos tempos de pandemia global de Coronavírus, sendo extremamente compreensivo e de grande ajuda sempre que necessário durante os termos de TCC cursados.

A todo o corpo de **funcionários** de limpeza e manutenção que sempre garantiram um ambiente pleno para todas as atividades e desafios do meio acadêmico. Obrigado pela dedicação em manter a limpeza e funcionamento do lugar que passei a grande maioria dos meus dias durante mais de cinco anos aprendendo para realizar este trabalho.

A **Deus** e todas as **Forças Espirituais**, independente de religião, as quais meu bem estar já foi solicitado e as quais já me apoiei para continuar nesta graduação. O equilíbrio entre fé e busca pelo conhecimento me garantiram a paz para seguir em frente e finalizar este trabalho.

RESUMO

A capacidade de identificar e classificar amostras de fungos é de grande interesse na medicina diagnóstica e mercado agroindustrial. No que se diz respeito a saúde pública, nos últimos anos, fungos do gênero *Aspergillus,* antes considerados inofensivos, passaram a contaminar seres humanos com o sistema imune inibido ou enfraquecido devido a medicação antibiótica ou tratamentos invasivos [1]. No caso da indústria de produção, fungos como o gênero *Metarhizium* vêm sendo usados como biopesticidas para enfrentamento de diversas pragas (espécies de insetos).Foram utilizados dados de espectroscopia no infravermelho por transformada de Fourier (FTIR) de seis linhagens de três espécies diferentes do fungo de gênero *Metarhizium* (filo *Ascomycota*), pertencentes ao USDA - ARSEF ("*United States Department of Agriculture – Agricultural Research Service Collection of Entomopathogenic Fungal Cultures"*). Foram três linhagens da espécie *Metarhizium acridum* (ARSEF 324, ARSEF 3391 e ARSEF 7486), uma da espécie *Metarhizium anisopliae* (ARSEF 5749) e duas da espécie *Metarhizium brunmeum* (ARSEF 1095 e ARSEF 5626). Os dados foram pré-processados em MATLAB® (2015a) com aplicação de *downsampling* (redução da taxa de amostragem), normalização e filtragem de suavização de Savitzky-Golay (S-G). Os dados representam a absorbância de espectros em função do número de onda, sendo avaliadas as bandas de 900 a 1350cm^{-1} e de 900 a 1800cm^{-1} com variação do grau de derivada e do tamanho da janela do filtro. A comparação foi realizada quanto a capacidade de classificação dos grupos com aplicação das técnicas combinadas de Análise de Componentes Principais (PCA) e Análise Discriminante Linear (LDA).Identificou-se que os parâmetros do filtro de Savitzky-Golay (grau de derivada e dimensão da janela) têm grande importância para construção de classificadores juntamente com aplicação de PCA para otimização da razão sinal-ruído (SNR), assim como a escolha das regiões do espectro, visto que a região de 900 a 1800cm^{-1} apresentou resultados com muita interferência de ruídos decorrentes de contaminante de vapor d'água.

Palavras-chave: 1. FTIR 2. Machine Learning 3. Bioespectroscopia 4. Fungos 5. Bioengenharia.

ABSTRACT

The ability to identify and classify samples of fungi is of great interest in diagnostic medicine and agrobusiness. Regarding public health, in recent years, fungi within the genus *Aspergillus*, previously considered harmless, started to contaminate humans with inhibited or weakened immune system due to antibiotic medication or invasive treatments [1]. When considering industrial business, fungi within the genus *Metarhizium* have been used as biopesticides to fight various pests (species of insects). Fourier transform infrared spectroscopy (FTIR) data from six strains of three different species of the fungus of genus Metarhizium (phylum Ascomycota) were used, belonging to USDA - ARSEF ("United States Department of Agriculture – Agricultural Research Service Collection of Entomopathogenic Fungal Cultures"). There were three strains from the species *Metarhizium acridum* (ARSEF 324, ARSEF 3391 and ARSEF 7486), one strain of the species *Metarhizium anisopliae* (ARSEF 5749) and two strains of the species *Metarhizium brunmeum* (ARSEF 1095 and ARSEF 5626). The data was pre-processed in MATLAB® (2015a) with the application of downsampling (reduction of the sampling rate), normalization and Savitzky-Golay smoothing filtering (S-G). The data represents the absorbance of spectra as a function of the wave number, being analysed on the bands 900 to 1350cm^{-1} and 900 to 1800cm^{-1} with variation of the degree of differentiation and dimension of the filtering window. The comparison worked regarding the classification capacity of groups using the combined techniques of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). It was identified that the parameters of the Savitzky-Golay filter (degree of differentiation and window dimension) are of great importance for the construction of classifiers along with the application of PCA to optimize the signal-to-noise ratio (SNR), as well as the restriction of specific regions of the spectrum, such as 900 to 1800cm^{-1}, which showed results with high interference of noise from water vapor contaminants.

Keywords: 1. FTIR 2. Machine Learning 3. Biospectroscopy 4. Fungi 5. Bioengineering.

LISTA DE ABREVIATURAS

A. flavus – *Arpergillus flavus*

A. fumigatus – *Aspergillus fumigatus*

A. nidulans – *Aspergillus nidulans*

A. niger – *Aspergillus niger*

ARSEF - United States Department of Agriculture – Agricultural Research Service Collection of Entomopathogenic Fungal Cultures

DNA – Deoxyribonucleic Acid

FTIR – *Fourier Transform Infrared Spectroscopy*

IA – Inteligência Artificial

IR – Infrared

LDA – *Linear Discriminant Analysis*

M. acridum - *Metarhizium acridum*

M. anisopliae – *Metarhizium anisopliae*

M. brunmeum – *Metarhizium brunmeum*

PCA – *Principal Component Analysis*

PCR – Proteína-C Reativa

PDA – Potato dextrose Agar

S-G – *Savitzky-Golay Filter*

SNR – *Signal-to-Noise Ratio*

LISTA DE FIGURAS

Figura 1 - Espectro Eletromagnético. Observa-se que o infravermelho induz a vibração, enquanto micro-ondas levam a rotação [19]. 19

Figura 2 - Espectro biológico típico, mostrando a correspondência biomolecular dos picos [21]. 20

Figura 3 - Espectrômetro por Infravermelho [23]. 21

Figura 4 - Técnica de ATR, onde o feixe de IR interage com a interface ATR-amostral e tem comprimentos de onda absorvidos pela amostra, o que é mensurado pelo detector. [25]. 23

Figura 5 - Espectros A e B de absorção simulado (a) e suas respectivas derivadas de segundo grau (b). É possível identificar os pontos de mínimo paralelos aos picos do espectro, caracterizando a separação dos picos de absorção [28]. 25

Figura 6 - Demonstração hipotética da interpolação de pontos de um sinal amostrado [30]. 26

Figura 7 - Visão esquemática da Análise Discriminante Linear de Fisher (LDA) [36]. 29

Figura 8 - Espectros aleatórios médios gerados por soma de gaussianas [39]..... 33

Figura 9 - Loading Plot dos espectros médicos demonstrados na Figura 2 [39]. 34

Figura 10 - Classificação Taxonômica de seis linhagens distribuídas em 3 espécies do fungo de gênero Metarhizium (3 linhagens da espécie M. acridum, 1 linhagem da espécie M. anisopliae e 2 linhagens da espécie M. brunmeum). Tal gênero pertencente ao filo Ascomyota. **Fonte:** Próprio autor. 37

Figura 11 - Espectro médio completo das seis linhagens de fungo do gênero Metarhizium utilizadas neste trabalho. O intervalo (I) define a região de interesse, onde concentra-se mais informação bioquímica relevantes para classificação dos fungos. O intervalo (II) representa uma região mais ruidosa. 42

Figura 12 - Espectro médio normalizado filtrado pelo método de Savitzky-Golay com derivação de segundo grau e janela de dimensão 11, sendo (a) o espectro de 900 a 1350cm^{-1} e (b) o espectro de 900 a 1800cm^{-1}.....44

Figura 13 - PCA - scatter plot do espectro médio normalizado filtrado pelo método de Savitzky-Golay com derivação de segundo grau e janela de dimensão 11, sendo (a) o espectro de 900 a 1350cm^{-1} e (b) o espectro de 900 a 1800cm^{-1} (que apresenta uma rotação de 90º. 45

Figura 14 - PCA - Loading plot do espectro médio normalizado filtrado pelo método de Savitzky-Golay com derivação de segundo grau e janela de dimensão 11, sendo (a) o espectro de 900 a 1350cm^{-1} e (b) espectro de 900 a 1800cm^{-1}.....46

Figura 15 - Espectro médio normalizado das seis linhagens do gênero Metarhizium, comparando a variação de janela e derivada. (a) – 1º Derivada com janela de 11 pontos; (b) - 1º Derivada com janela de 13 pontos; (c) – 1º Derivada com janela de 15 pontos; (d) – 2º Derivada com janela de 11 pontos; (e) - 2º Derivada com janela de 13 pontos; (f) – 2º Derivada com janela de 15 pontos.....48

Figura 16 - Espectro médio analisado após Análise de Componentes Principais das PCs 1 e 2, contendo as seis linhagens do fungo do gênero Metarhizium analisadas. (a) – 1º Derivada com janela de 11 pontos; (b) - 1º Derivada com janela de 13 pontos; (c) – 1º Derivada com janela de 15 pontos; (d) – 2º Derivada com janela de 11 pontos; (e) - 2º Derivada com janela de 13 pontos; (f) – 2º Derivada com janela de 15 pontos.....49

Figura 17 - Loading Plots com a variação de dimensão das janelas e graus de derivada. (a) – 1º Derivada com janela de 11 pontos; (b) - 1º Derivada com janela de 13 pontos; (c) – 1º Derivada com janela de 15 pontos; (d) – 2º Derivada com janela de 11 pontos; (e) - 2º Derivada com janela de 13 pontos; (f) – 2º Derivada com janela de 15 pontos. Duas regiões relevantes para comparação estão destacadas nos loading de segundo grau de derivada, identificados pelos algarismos I e II.....50

SUMÁRIO

1. INTRODUÇÃO 14
2. REVISÃO BIBLIOGRÁFICA 16
2.1. Fungos 16
2.2. Espectroscopia 18
2.2.1. O Infravermelho na Espectroscopia 19
2.2.2. Espectroscopia Infravermelha por Transformada de Fourier 21
2.3. Processamentos matemáticos 24
2.3.1. Filtro de Savitsky-Golay (S-G) 24
2.4. Machine Learning 26
2.4.1. Análise Discriminante Linear (LDA) **28**
2.4.2. Análise de Componentes Principais (PCA) **31**
3. OBJETIVOS 35
3.1. Objetivos específicos 35
4. METODOLOGIA 36
4.1. Objeto de estudo 36
4.2. Coleta dos fungos 37
4.3. Coleta dos espectros 38
4.4. Pré-processamento 39
4.5. Classificação 41
5. RESULTADOS 42
5.1. Discussão **51**
6. CONCLUSÃO 54
7. REFERÊNCIAS BIBLIOGRÁFICAS 55
8. APÊNDICE 61

1. INTRODUÇÃO

Anualmente, existem aproximadamente 4 milhões de brasileiros afetados por infecções de natureza fúngica. No mundo, considerando todas as infecções fúngicas graves e micoses superficiais, são mais de 11,5 milhões de casos anuais, com aproximadamente 1,5 milhões de mortes, o que ultrapassa o total de mortes por malária e tuberculose [2].

Para o correto tratamento das infecções fúngicas, em muitos casos, se faz necessário o conhecimento da espécie infectante. Dentre as técnicas para classificação da espécie, está a visualização dos fungos em material biológico coletado do paciente por meio de microscópio ótico [3]. As amostras podem ter diversas origens (pelos, pele, sangue, mucosas, fluidos e entre outros) e cada uma possui um meio de cultura e/ou armazenamento recomendados, de forma que o processamento da amostra também seja personalizado [4]. É muito comum, por exemplo, para estudos de Fungos como o *Aspergillus nidulans* e *Metarhizium anisopliae,* que sejam aplicadas técnicas de espectroscopia para estudo da conformação bioquímica das amostras, de forma a identificar a espécie por meio das diferenças nestas estruturas.

A técnica de espectroscopia no infravermelho por transformada de Fourier (FTIR) obtém informações bioquímicas de forma semiquantitativa por meio da análise da interação da radiação eletromagnética com a amostra biológica. Ela tem como vantagem a alta reprodutibilidade com o mínimo de preparo da amostra [5] e, quando aliada a técnicas de reconhecimento de padrões, consegue classificar e determinar espécies/cepas de microrganismos com sensibilidade e especificidade comparadas à técnica de PCR (Proteína-C Reativa) [6]. A. Naumann et. Al. [7] mostrou que esta metodologia consegue classificar 26 cepas pertencentes a 24 espécies do fungo Micélio.

Para o sucesso da utilização das técnicas de reconhecimento de padrões para classificação de espécies de fungo em dados de espectroscopia FTIR, se faz necessário um pré-processamento adequado no qual são removidos grande parte das contribuições não relacionadas à bioquímica da amostra em análise, tal como ruído e contaminantes (parafina e vapor d´água). Dentre as diversas técnicas de pré-

processamento para atenuar essas contribuições espúrias, podemos citar o filtro de suavização Savitzky-Golay (S-G). Este filtro é amplamente utilizado na área de espectroscopia vibracional aplicada ao diagnóstico médico.

Diante do exposto, o objetivo do presente trabalho é verificar os impactos na acurácia do classificador sobre os parâmetros de janelamento e derivação do filtro de suavização S-G. Para isto foi utilizado um conjunto de dados de espectroscopia FTIR de 6 cepas de fungos de gênero *Metarhizium*, pertencente ao filo *Ascomycota*, processados em MATLAB®.

2. REVISÃO BIBLIOGRÁFICA

2.1. Fungos

Assim como os animais são organismos que pertencem ao reino *animalia* e as plantas ao reino *plantae*, os fungos são organismos vivos que possuem um reino próprio: o reino *fungi*. São aproximadamente 1,5 milhões de espécies possivelmente coexistindo na Terra, tendo apenas cerca de 7% delas catalogadas principalmente de acordo com os métodos reprodutivos [8].

Os fungos se diferenciam dos organismos dos demais reinos devido a presença de ambos quitina e glucano ao mesmo tempo na composição de suas paredes celulares. São invisíveis a olho nu, mas algumas espécies possuem a capacidade de frutificar-se, gerando um grande acúmulo de microestruturas para a formação de uma cogumelos, trufas, bolores e mofos visíveis a olho nu. Sua presença é fundamental para os ciclos de vida da natureza, estando muito presentes (junto das bactérias) no papel de decomposição de matéria morta [8].

Os fungos estão muito presentes no dia a dia do ser humano, seja aplicado a alimentação, processos industriais na geração de enzimas (como as epóxido hidrolases do *Trichoderma reesei*, biorremediadoras aplicadas descontaminação de solo com petróleo) [9] e produção de medicamentos como a Penicilina (fungo *penicillium*) e a Ciclosporina A (*Tolypocladium niveum*) [10]. Somado a isso, os fungos podem também ser vistos como inimigos no caso de serem agentes causadores de doenças graves que afetam o organismo humano.

Apesar dos fungos serem considerados inofensivos até as últimas décadas, a diminuição das defesas naturais do organismo por enfermidades e/ou medicação antibiótica tornou o homem suscetível a infecções fúngicas que podem ser fatais. Grande parte dos casos é causado depois de danos ao sistema de defesa do organismo após medicação para transplantados, tratamentos de câncer, além de períodos de baixa imunidade após procedimentos invasivos como sondas cateteres em ambiente hospitalar [2]

Estima-se que um ser humano inspire de 200 a 2.000 poros por dia [2], sendo poros as estruturas reprodutivas de tamanhos microscópicos resistentes a

permanecer no ar por longos períodos, empregados pela maioria das espécies de fungos. Os conídios, por exemplo, são um tipo especializado de esporo de baixa atividade metabólica, produzido de forma assexuada por fungos filamentosos [11]. Tais organismos estão presentes no corpo humano, como é o caso do *Candida albicans* que estabelece colônias na boca, intestinos e pele, mas que só podem se tornar danosos caso a imunidade do organismo seja reduzida em níveis consideráveis por decorrência de outras enfermidades ou medicações.

É mostrado, portanto, que mesmo uma espécie presentemente reconhecida como inofensiva ou até mesmo utilizada em meios produtivos de forma ampla e em escala na indústria, pode se tornar agente causador de graves infecções fúngicas em determinados cenários. A grande variabilidade de gêneros e espécies abre margem para inúmeras possíveis ameaças que convivem com o ser humano. É preciso acompanhar de perto tais variações deste organismo a fim de manter controle e prevenção sobre cenários de disseminação e pandemia de doenças fúngicas.

A exemplo da amplitude da ação dos organismos do reino *fungi*, está o gênero *Aspergillus*, pertencente ao Filo *Ascomycota*. Os fungos deste gênero geram interesses médicos, agrícolas e industriais em função de diversas espécies como *A. nidulans*, *A. fumigatus*, *A. flavus* e *A. niger*. Dentre as citadas, a primeira espécie se destaca como organismo modelo para estudos de genética e biologia molecular, contribuindo amplamente para compreensão de muitos processos celulares de forma geral, mas também auxiliando especificamente no entendimento dos metabolismos secundários em fungos [12]. Por ser encontrada em solo ou em forma de bolor em alimentos [13] com capacidade de reprodução sexuada e assexuada conhecidas, tal espécie se tornou muito utilizada nas últimas décadas.

Apesar de o fungo *A. nidulan*s representar uma versão inofensiva e muito engenhosa para o meio acadêmico, a espécie *A. fumigatus* já é considerada um grande perigo por causar as aspergiloses. Tal família de infecção fúngica tem mortalidade que varia de 30 até 90% e é disseminada por meio de conídios pelo ar e depositado em superfícies. O estabelecimento de uma colônia nos pulmões causa a aspergilose bronco-pulmonar alérgica e a aspergilose, muito comum em pacientes imunodeprimidos [14].

Um segundo gênero do mesmo filo *Ascomycota*, é o *Metarhizium*. Tal gênero é mundialmente conhecido, estudado e aplicado, pois diversas espécies são utilizadas como biopesticidas para controle de pragas agrícolas e vetores de doenças. É um gênero de fungo facilmente isolado em amostras de solo [15], também sendo muito aplicado em estudos e pesquisas acadêmicas e industriais como organismo modelo, tendo algumas de suas espécies mais conhecidas descobertas, como *M. anisopliae*, *M. acridum* e *M. brunmeum.*

Os fungos *Metarhizium anisopliae* são utilizados para controle de insetos, pois agem tendo seus conídios penetrando a cutícula dos animais hospedeiros, onde aderem e germinam em estruturas produtoras de enzimas degradantes durante a penetração do tegumento [16]. Entre 3 e 10 dias após o contato, é possível observar a morte dos alvos do fungo biopesticidas em diversas espécies de insetos, explicando a ampla utilização do espécime pelo mercado agroindustrial [17].

Especialmente nos casos dos fungos reproduzidos por meio do espalhamento de conídios (alta disseminação) [14], é de grande interesse avaliar a possibilidade de se estudar métodos de identificação, diferencial e classificação dos fungos de acordo com suas linhagens, espécies, gêneros e filos. Os esporos especializados e hifas têm ação metabólica baixa [11] e, portanto, o estudo de sua estrutura bioquímica torna-se relevante para a identificação e diferenciação de amostras, seja para diagnóstico médico de uma infecção, seja para controle de pragas no meio agrícola.

2.2. Espectroscopia

Parte fundamental da análise computacional de amostras biológicas é a identificação das características mais adequadas a serem processadas a fim de se obter uma conclusão quanto a amostra em questão. No caso de fungos, as diferenças de composição de suas estruturas são o foco para o estudo, sendo que elas são consequência de variações bioquímicas de seus tecidos. Tais variações podem ser observadas pelo uso de técnicas espectroscópicas sobre a amostra biológica, como é o caso da microespectroscopia infravermelho, que permite a identificação de padrões espectrais em determinadas frequências de acordo com a presença de complexos bioquímicos específicos. A determinação desta região espectral se

relaciona com os diferentes tipos de transição energética que são tidos como foco do estudo. As transições entre os níveis eletrônicos podem ser estudadas com aplicação da luz ultravioleta até a visível, enquanto transições vibracionais de moléculas orgânicas ocorrem por meio de interações de dipolo na região do infravermelho próximo e médio e, por último, as transições rotacionais, que prevalecem na região de micro-ondas [18, 19], conforme **Figura 1**.

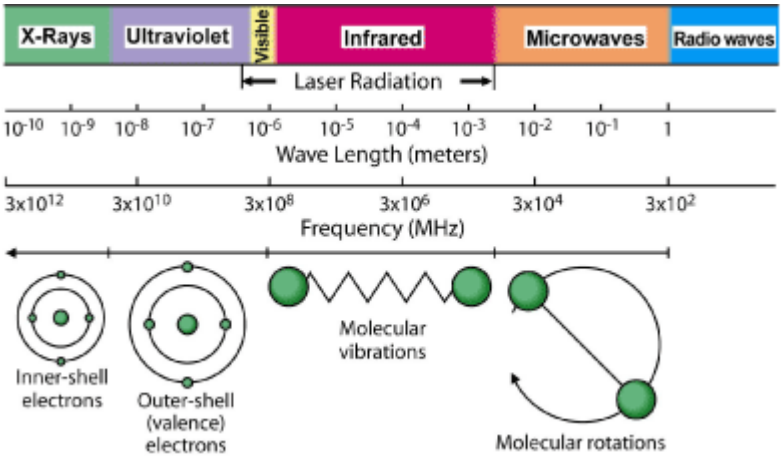


Figura 1 - *Espectro Eletromagnético. Observa-se que o infravermelho induz a vibração, enquanto micro-ondas levam a rotação [19].*

Por meio da análise e comparação desses padrões bioquímicos de diferentes estruturas de um mesmo fungo ou entre gêneros, espécies e/ou linhagens distintas, pode ser possível averiguar relações entre estas amostras e fornecer uma análise quantitativa complementar para diferenciá-las.

2.2.1. O Infravermelho na Espectroscopia

A luz infravermelha é o nome dado a uma onda eletromagnética de radiação não ionizante com comprimento de onda na faixa de 1 mm a 700 nm, que surge a partir da incidência de luz em um comprimento de onda específico, o qual pode ser absorvido por uma molécula. Ao se discutir a região espectral no infravermelho, tange-se o escopo do comportamento vibracional das moléculas, o qual pode ser descrito pela lei de Hooke, que teoriza sobre a força restauradora de um sistema oscilatório [20]. Especificamente no caso das moléculas, elas são formadas por átomos, que reagem as interações vibratórias do meio e formam o perfil vibracional molecular.

Esta ferramenta baseia-se no fato de que todo corpo é formado por uma estrutura físico-química que está em constante movimento, cuja frequência depende de quais ligações interatômicas existem no sistema, o que leva a um certo nível de energia discretizado vinculado a matéria em questão. Dado esse perfil vibracional e somando ao conceito de conservação da energia, tem-se que um corpo irá absorver a radiação de um feixe em frequências específicas que existam em sua estrutura e deixar passar àquelas que não são características dele, de forma que, conhecendo-se todas as frequências que foram emitidas sobre o corpo e mensurando as que passaram por ele (não absorvidas), é possível identificar quais foram absorvidas na matéria. Isto quer dizer que há uma maneira de identificar a composição química de compostos de acordo com a vibração característica de cada elemento e ligação molecular presente de acordo com a absorbância relacionada a ele [6], conforme **Figura 2**.

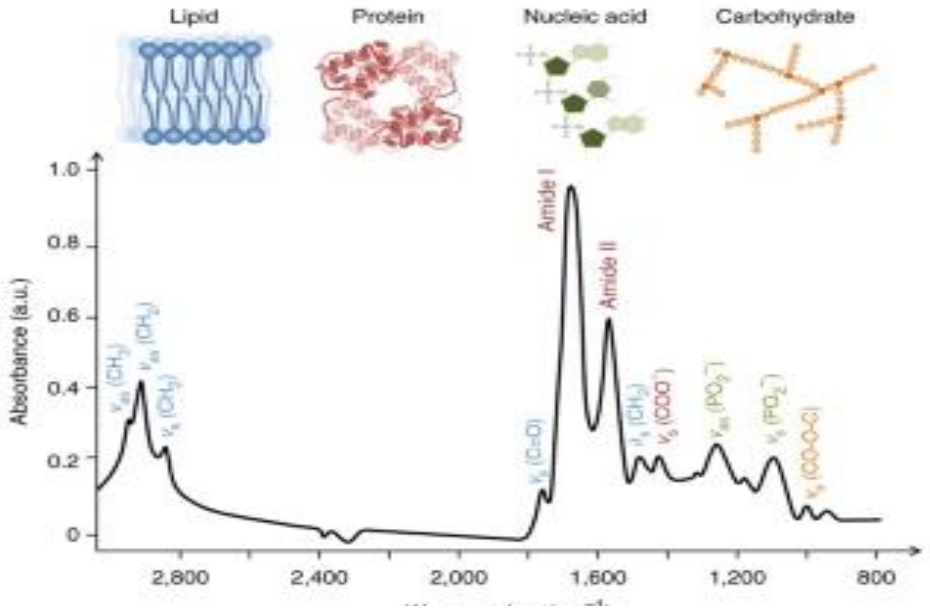


Figura 2 - *Espectro biológico típico, mostrando a correspondência biomolecular dos picos [21].*

A espectroscopia vibracional é uma técnica de aquisição de sinais e imagens muito aplicada em estudos relacionados às mudanças na estrutura da matéria e comportamento de partículas, o que tange o escopo de áreas como a engenharia de materiais, química, física e, recentemente, a medicina diagnóstica em conjunto com a engenharia biomédica. A aplicação no domínio biológico das ciências se deu graças ao avanço da tecnologia de fontes de fluxo de fótons no infravermelho (IR) e sensores de alto desempenho utilizados na detecção dos feixes de infravermelho, elevando a resolução até um nível suficiente para aplicação em amostras biológicas [22].

2.2.2. Espectroscopia Infravermelha por Transformada de Fourier

Uma técnica específica deste ramo é fundamentada pela aplicação da Transformada de Fourier no processo de aquisição deste espectro, o que caracteriza a *Transform Infrared Spectroscopy – FTIR.* Na técnica, a radiação é emitida por uma fonte térmica com emissão predominante na região do IR e passa por um aparato conhecido como Interferômetro de Michelson antes de interagir com a amostra.

Este aparato, de acordo com VELOSO, M. N. [18], consiste em um conjunto de espelhos, sendo um fixo e um móvel posicionados de forma que fiquem a 90º entre si. Além disso, um terceiro espelho semireflexivo (conhecido como *beamsplitter*) é posicionado exatamente na bissetriz dos dois primeiros espelhos. Quando o emitido passa pelo divisor, o feixe se reparte de forma que metade dos fótons seja refletida e a outra metade seja transmitida, sendo que o feixe é então refletido de volta para o divisor somado de uma variação de deslocamento (diferença de caminho ótico) devido ao movimento oscilatório constante presente no espelho móvel. Esta oscilação permite uma variação na fase dos feixes parciais (com metade dos fótons), o que gera uma variação de amplitude decorrente da interferência promovida [18].

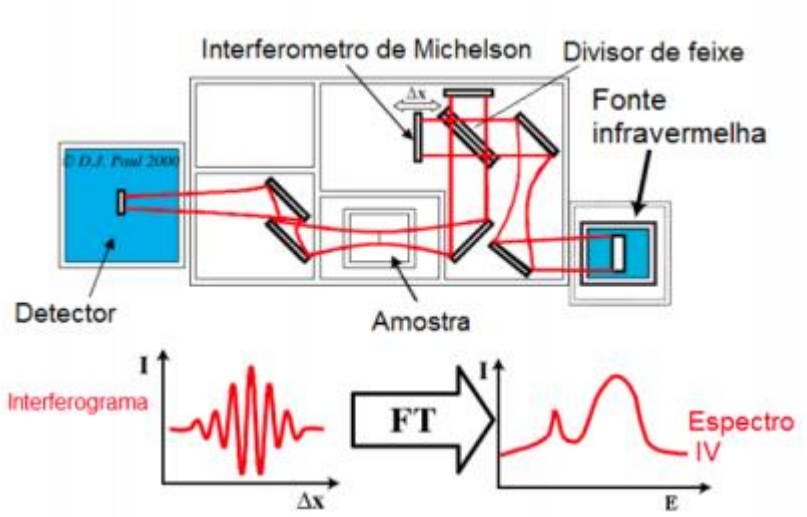


Figura 3 - *Espectrômetro por Infravermelho [18].*

O sinal proveniente do detector é chamado de interferograma, que é transformado por meio da aplicação da Transformada de Fourier, gerando, então, o espectro infravermelho. Esta robusta ferramenta matemática transforma sinais periódicos e aperiódicos para o espaço da frequência, onde os sinais aperiódicos (como o interferograma) podem ser representados como uma combinação linear de

exponenciais. O espectro de coeficientes resultantes gera a representação da transformada de Fourier, conforme **Figura 3**.

Já que a análise, então, baseia-se num estudo da frequência, basta que uma técnica seja aplicada para transformação das variáveis de intensidade de um feixe (sobre as amostras coletadas) para o domínio da frequência, o que leva a utilização de ferramentas como a Transformada de Fourier sobre o sinal de infravermelho, ou seja, *Fourier Transform Infrared Spectroscopy – FTIR*. O resultado é apresentado como um espectro de frequências, com a informação geralmente demonstrada como frequência em números de onda com a intensidade da absorbância em valores unitários arbitrários.

Para algumas amostras, quando possuem interferência de polímeros opacos como colas, parafina, tecidos e entre outros, o espectro de transmissão é prejudicado por haver baixo índice de refração e é necessário empregar técnicas especiais para utilização do feixe infravermelho para FTIR. Uma destas técnicas é conhecida como Refletância Total Atenuada (da sigla em inglês, ATR).

No ATR, é aplicada uma superfície com alto índice de refração (como um cristal de diamante, por exemplo) em contato com a amostra. Aplicando um feixe de onda eletromagnética no ângulo Θ no cristal, quando o feixe encontra a interface dos dois materiais (ATR e amostra), ocorre a reflexão interna total do feixe [24]. Tal ângulo deve ser superior a um ângulo crítico dado por θ_c, obtido pela Equação 1.

$$\theta_c = \sin^{-1}\frac{n_1}{n_2} \qquad \text{Equação 1}$$

Sendo n_1 o índice de refração do ATR e n_2 igual ao índice de refração da amostra.

O feixe sofrerá inúmeras reflexões e absorções na interface cristal-amostra, penetrando distâncias mínimas na amostra, sendo a radiação penetrante denominada onda evanescente. Quando a amostra absorve tais ondas, ocorre atenuação do sinal e isso é mensurado pelo detector na outra ponta do experimento, conforme **Figura 4**.

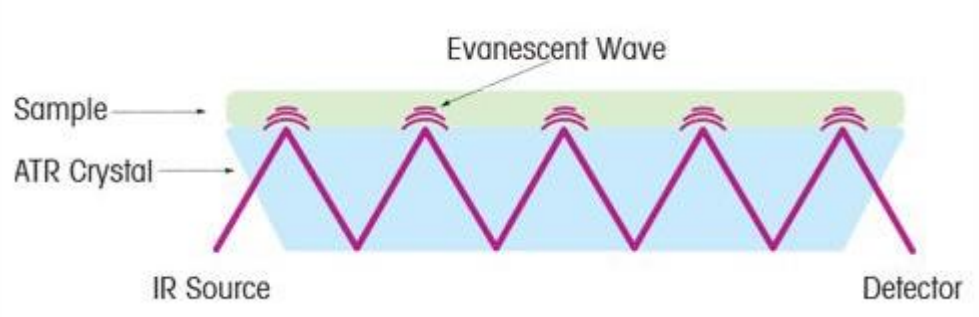


Figura 4 - *Técnica de ATR, onde o feixe de IR interage com a interface ATR-amostral e tem comprimentos de onda absorvidos pela amostra, o que é mensurado pelo detector. [25].*

A possibilidade de uma análise espectroscópica de alterações bioquímicas na estrutura de macromoléculas abriu portas para ramos dos mais diversos na ciência, como em estudos farmacêuticos, aplicações com engenharia de alimentos, estudos de biocompatibilidade de materiais e utilização para diagnóstico médico por meio da aplicação em material celular. Esta última tem uma importância fundamental na nova forma de se estudar medicina, uma aproximação mais preventiva e proativa do que remediativa, e definir parâmetros para identificar doenças e seus níveis de avanço contra o organismo enfermo, uma vez que torna possível evidenciar o início de uma patologia em níveis bioquímicos que são anteriores às consequências celulares e ao organismo como um todo. Esse fato colabora para a visão geral de que um tratamento contra uma infecção fúngica, por exemplo, é mais eficaz se iniciado com antecedência já identificando qual o fungo está afetando o organismo.

Como trata-se de uma análise com dados que são intensamente selecionados, filtrados e amplificados para só então construir um banco de informação que possa ser estudado e comparado com a literatura e a experiência clínica de profissionais da saúde, é preciso garantir que a informação contida em cada amostra seja confiável no quesito coleta e armazenagem. Isto quer dizer que o método para construção de uma amostra de tecido deve ser devidamente selecionado e seus passos bem detalhados, pois cada técnica aplicada na coleta influenciará na informação de espectro que será avaliada no futuro e, conhecendo-se bem o método, pode-se minimizar ou eliminar ruídos e interferências conhecidas.

Sistemas biológicos são caracterizados por processos bioquímicos complexos simultâneos e a análise estatística multivariada permite não só a análise da

informação em si para agrupamento de padrões, como também abstrair informação da relação entre as diferentes classes. Esse fato é de suma importância neste estudo devido ao foco na separação em classes entre diferentes linhagens, espécies e gêneros fúngicos. A interpretação e tratamento de tal volume de dados se torna cada vez mais dependente da aplicação de técnicas de computação automatizadas, baseadas em *Big Data* e inteligência artificial, mais especificamente em ferramentas associadas ao *machine learning* (aprendizado de máquina).

2.3. Processamentos matemáticos

Tratando-se de espectros de absorção, muitas técnicas matemático-computacionais podem ser aplicadas dependendo do tipo e quantidade de amostras estudadas, além de depender de qual é o objetivo de experimento. No caso dos espectros de absorbância, por exemplo, é interessante a utilização de filtros de suavização e separação de picos de absorbância nas bandas do espectro para reconhecimento de padrões das amostras. Além disso, a etapa de processamento dos dados coletados é de suma importância para minimização da interferência de ruído proveniente de contaminantes e espectros indesejados, baseline, distorções etc [26].

Além disso, técnicas de análise multivariada pautadas em *machine learning*, como a Análise de Componentes Principais (*PCA – Principal Component Analysis*) e Análise Discriminante Linear (*LDA – Linear Discriminant Analysis*), podem ser aplicadas em cenários de múltiplas variáveis, de forma a separar amostras em grupos de acordo com um número finito de características.

2.3.1. Filtro de Savitsky-Golay (S-G)

No caso de espectros de absorção, o sinal é medido em uma curva de absorbância pelo número de onda de radiação infravermelha, sendo que as bandas de absorção são na realidade sobreposições de vários picos [27]. A separação desses picos se torna essencial para a análise de amostras biológicas com bandas complexas de absorção, o que é praticado com aumento da resolução do sinal.

Uma das técnicas conhecidas para a separação dos picos sobrepostos para melhor definição das bandas de absorção é a derivação do sinal. Tal processo

matemático, quando aplicado na segunda ordem de derivada em sinais de FTIR, caracteriza-se por gerar uma curva com pontos mínimos nos locais onde havia picos de absorbância no espectro, conforme **Figura 5**.

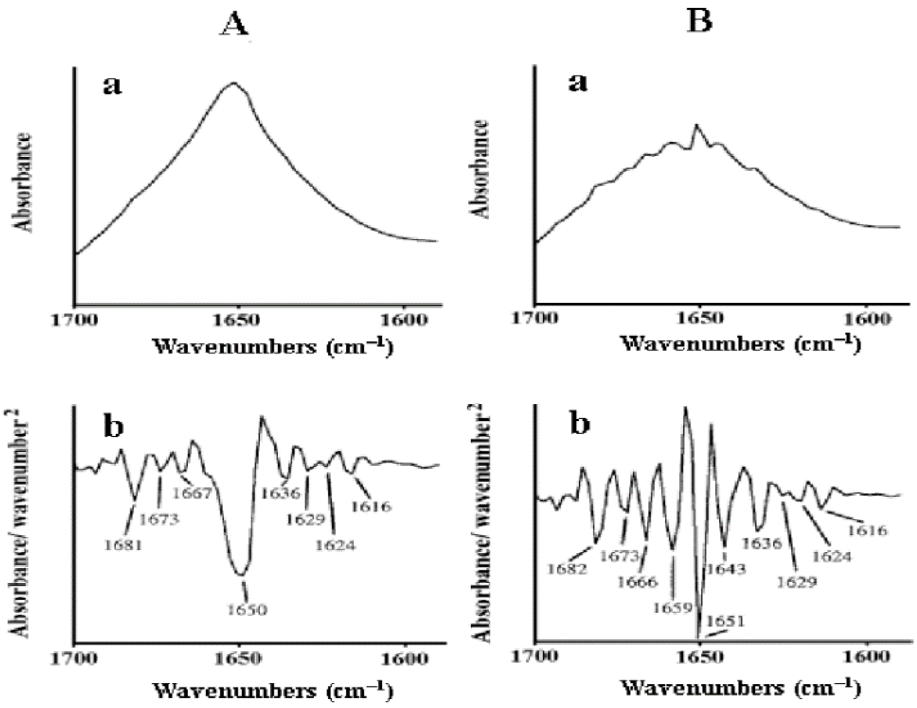


Figura 5 *- Espectros A e B de absorção simulado (a) e suas respectivas derivadas de segundo grau (b). É possível identificar os pontos de mínimo paralelos aos picos do espectro, caracterizando a separação dos picos de absorção [28].*

Aliado a derivada do sinal com intuito de uma melhor separação de bandas sobrepostas, é possível que o sinal tenha traços de ruídos provenientes de ruídos provenientes de baseline e contaminantes como vapor d'água. O baseline é comumente acarretado por pequenas instabilidades ou erros durante a coleta dos espectros, enquanto vapor d'água e contaminantes podem estar presentes no meio de cultura da amostra utilizada na coleta, o que gera interferência nas bandas de absorção desejadas. Tais fatores consistentes e interferentes podem ter seus efeitos minimizados por filtros suavizadores, ou seja, ferramentas que tem por objetivo a correção de tendências indesejadas do sinal.

O filtro de Savitzky-Golay (S-G) é uma das ferramentas mais usuais para a derivação combinada com a suavização do sinal de FTIR [29]. Esta técnica resulta num sinal suavizado em função da interpolação de um polinômio diferenciado pelo

grau *p* de derivada, de forma que o resultado da interpolação corrija o valor central de uma janela de tamanho *2m* + 1 definido no filtro [30], conforme **Figura 6**.

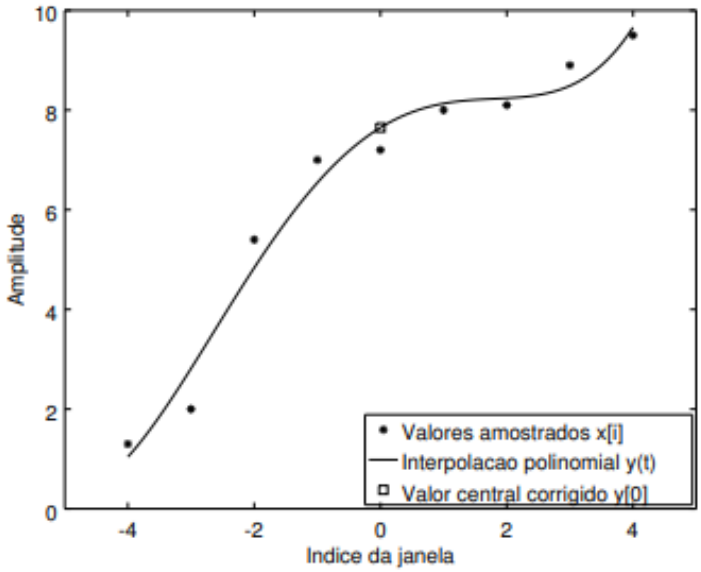


Figura 6 - *Demonstração hipotética da interpolação de pontos de um sinal amostrado [30].*

A aplicação do filtro de S-G permeia a escolha de um cenário que melhor suavize e derive o sinal, isto é, que promova o aumento da SNR e facilite a identificação dos picos de absorção para diferenciação de bandas. Tal cenário é determinado pelos parâmetros que definem como será realizada a filtragem, o que é definido pelo tamanho da janela *j(m)* (Equação 2) e pelo grau da derivada aplicada [30].

$$j(m) = 2m + 1 \qquad \text{Equação 2}$$

A variação desses dois parâmetros do filtro S-G gera uma infinidade de cenários que pode ser estudada, conforme será feito ao longo deste trabalho, onde a dimensão do janelamento estará diretamente ligado a suavização os espectros de absorção e a derivação terá por objetivo a melhor separação dos picos.

2.4. Machine Learning

As áreas da tecnologia e computação abrigam uma constelação de algoritmos, códigos, redes neurais e outras ferramentas que têm por objetivo aproximar o processamento de um software a inteligência humana e capacidade de aprendizado a partir de bancos de dados iniciais. Com a inteligência artificial (IA) e, mais especificamente na subárea de aprendizado de máquina (do inglês *machine* learning), é possível utilizar a computação para aprendizado com dados prévios de forma a prever comportamentos e reconhecer padrões de outros dados [31].

O conceito de aprendizado de máquina se torna cada vez mais presente e indispensável para a resolução problemas encontrados na área da medicina e engenharia biomédica devido ao alto volume e variabilidade de dados [32]. Tais áreas passam por uma revolução em termos de metodologia e ferramental, recebendo altos investimentos para desenvolvimentos de tecnologias voltadas a classificação de dados, principalmente para diagnóstico, tendo sucesso reconhecido em exemplos como o reconhecimento e classificação com 88% de acurácia de sinais de eletrocardiogramas com anormalidades rítmicas [33 – 34].

Técnicas de mesmo embasamento em *machine learning* e vêm sendo empregadas em classificação (análise discriminante) de imagens médicas [35] e espectros de absorção provenientes de dados transformados para o domínio da frequência, onde cada região de espectro pode ser classificada quanto a formação bioquímica da estrutura analisada. Algoritmos pautados em inteligência artificial (IA) buscam treinar o reconhecimento de determinadas estruturas e organismos a partir de espectros conhecidos (bancos de dados), de forma a se tornar robustos e automatizados para reconhecer dados de novos pacientes e fornecer apoio a diagnósticos, por exemplo.

Apesar de robustas e revolucionárias, tais técnicas são sensíveis a problemas relacionados a qualidade e extensão dos dados utilizados como base no treinamento do algoritmo ou sistema, mas também durante o uso com dados reais. O padrão reconhecido por IA pode se tornar corrompido se dados de treinamento possuem muitas informações não relacionadas à bioquímica da amostra tal como ruido e contribuições de contaminantes como o vapor d’água. Portanto, a classificação, assim, se torna suscetível a erros de interpretação [21].

É essencial que a base de informação seja coleta de forma padronizada e abrangente durante o treinamento e desenvolvimento de tais técnicas, de forma a garantir que o algoritmo responda bem quando utilizado com dados reais fora do estudo. No caso de estudo baseado em espectros de absorbância, por exemplo, é essencial que a coleta de imagens espectrais seja realizada de forma cuidadosa e leve em consideração o meio no qual o objeto de estudo foi estudado (meio de cultura de microrganismos, como fungos por exemplo), além da pureza e padronização das amostras a serem aplicadas na espectroscopia.

2.4.1. Análise Discriminante Linear (LDA)

Dentre as técnicas estatísticas pautadas em *machine learning* para classificação de dados estão as análises discriminantes. Tais técnicas são baseadas em criar bancos de dados que podem ser separados em classes de acordo com funções matemáticas, sendo o conjunto dessas funções os chamados modelos de classificação.

Parte-se do pressuposto de grupos que possuem, cada um, *n* elementos descritos individualmente por *p* características mensuradas na coleta de um banco de dados. Tais características dos elementos dentro de um grupo são analisadas estatisticamente de forma a criar o perfil que caracteriza tal grupo, o que o diferencia do perfil de outros grupos ou classes. A partir de um banco de dados conhecido, os chamados dados de treinamento, o modelo é desenvolvido buscando as relações estatísticas entre os elementos dentro de um mesmo grupo de forma a determinar parâmetros para a separação entre as classes.

Considerando-se, então, uma coleta de dados de um cenário desconhecido, é possível comparar estatisticamente as *p* características de *n* elementos para determinar em qual classe ou grupo a informação tem maior probabilidade de pertencimento. Tal determinação é pautada nas funções matemáticas criadas durante o treinamento do classificador, que criam os limites numéricos usados para decisão entre classificar um novo elemento em um grupo ou outro.

Uma das análises mais conhecida é a Análise Discriminante Linear de Fisher (LDA), que busca determinar limites em forma de linhas que separem as classes a

partir de projeções da base multidimensional inicial. A ideia desta técnica é encontrar o limite discriminante de forma a maximizar a distância entre elementos de classes distintas, ao mesmo tempo que minimiza a distância (ou variância) de elementos de uma mesma classe [36].

Analisando o esquema presente na **Figura 7**, é possível avaliar um cenário hipotético de *dataset* com elementos diferenciados em dois grupos identificados pelas cores azul e vermelha. Tais elementos são caracterizados individualmente pelas variáveis experimentais x_1 e x_2 conforme demonstrado nos eixos do gráfico de dispersão.

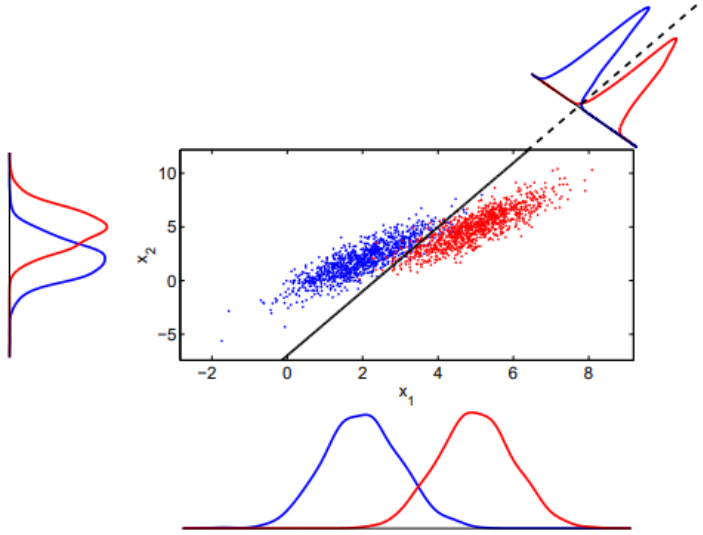


Figura 7 - *Visão esquemática da Análise Discriminante Linear de Fisher (LDA) [36].*

De acordo com os conceitos de LDA, as classes podem ser separadas por um limite definido por funções matemáticas, o que é simbolizado por uma linha preta que corta os dados.

Avaliando o cenário em que apenas a variável x_2 é utilizada para caracterizar os elementos da amostra, é possível ver a projeção à esquerda do gráfico de dispersão. Tal projeção mostra que os elementos das duas classes se sobrepõem em demasia a partir da análise individual desta variável, ou seja, teria uma classificação que colocaria muitos elementos como pertencentes a uma mesma classe, sendo que na realidade eles seriam de grupos diferentes. Isso demonstra que a variável x_2 não é uma boa opção para ser usada como classificadora.

Da mesma maneira, analisando o cenário em que a varável x_1 é utilizada para caracterização dos elementos da amostra, é possível avaliar a projeção abaixo do gráfico de dispersão. Esta projeção demonstra que os elementos das duas classes são melhor caracterizados de acordo com seus grupos se comparada com a outra variável individualmente, mas ainda com uma considerável sobreposição de informações. Isso significa que a variável x_1 teria um melhor desempenho como classificador, mas ainda longe do ideal.

Com isto, nota-se que nenhuma das variáveis presente nos eixos da **Figura 7** são suficientes individualmente para classificar os elementos de forma eficiente, ou seja, é necessário avaliar de forma combinada as características do *dataset* de forma a encontrar uma função que melhor separe as classes (representada pela linha preta na diagonal do gráfico).

O melhor cenário de separação é encontrando pela maximização do Critério de Fisher *J(W)* [36], definido pela razão entre S_W (matriz de variância dos elementos de cada classe) e S_B (matriz de variância entre as classes), equilibrado por uma matriz de projeção *W*, conforme Equação 3.

$$J(W) = \frac{WS_BW'}{WS_WW'} \qquad \text{Equação 3}$$

Tal maximização é atingida encontrando o valor *W* que otimiza a Equação 3 para os dados utilizados no treinamento do classificador. Ao aplicar um novo elemento desconhecido e sem classe definida, o método de Fisher pede que as variáveis que o caracterizam (vetor *x)* sejam combinadas linearmente com o vetor *W* para determinar a projeção *y* conforme Equação 4.

$$y = W'x \qquad \text{Equação 4}$$

O centroide do vetor *y* é então comparado com os vetores característicos de cada classe por meio de alguma técnica de mensuração de distâncias (como a

distância euclidiana, por exemplo), de forma que o cenário com a menor distância seja definido para determinar o melhor grupo para a classificação desse novo elemento [36].

É importante ressaltar que a performance do classificador é diretamente dependente dos critérios que definiram a etapa de treinamento da análise discriminante. Isto é, o total de variáveis originais e a relevância que possuem na classificação são fundamentais para que o classificador performe de maneira satisfatória com novos elementos fora do treinamento. Neste sentido, se um classificador for treinado com um número muito baixo de variáveis de caracterização, ele terá um desempenho ruim na classificação por não conseguir encontrar diferenças relevantes entre novos elementos (é o cenário de *underfitting*). Da mesma forma, se for utilizado um número muito grande de variáveis de um *dataset* em um treinamento de classificador, ele se tornará específico demais para os dados do treinamento e terá erros na classificação quando novos elementos forem aplicados (é o cenário de *overfitting*).

Técnicas estatísticas de transformação de variáveis podem ser aplicadas as características originais das amostras, com o intuito de diminuir o número de variáveis (diminuição das dimensões) originais e remover relações indesejadas de parcialidade entre elas. Tal transformação de domínio pode ser aplicada por meio da Análise de Componentes Principais (PCA) anteriormente a aplicação da análise discriminante, de forma a fazer ajustes de otimização no cenário original antes de desenvolver e utilizar o classificador LDA.

2.4.2. Análise de Componentes Principais (PCA)

A Análise de Componentes Principais (Principal Component Analysis – PCA) é uma ferramenta matemática aplicada sobre vetores aleatórios compostos de *n* variáveis vinculadas a informação estudada, de forma a reestruturar a análise dos dados sobre a variância e covariância [37]. A ideia é reconfigurar as variáveis originais em combinações lineares delas mesmas, formando as chamadas componentes principais (PCs, da sigla *Principal Components*), cujo número total *p* deve ser menor do que o número inicial de variáveis *n*, mas sem perder muita informação.

O cálculo das componentes é realizado a partir da decomposição das matrizes de covariância do vetor aleatório de variáveis originais. De tal matriz são extraídos os autovalores e autovetores, cuja combinação com as variáveis originais resulta nas componentes principais [38].

$$PC_p = \sum_{n=1}^{n} (v_{p,n} . A_n) \qquad \text{Equação 5}$$

A Equação 5 demonstra a combinação linear necessária para o cálculo das *p* novas variáveis (componentes principais, PC_p). A combinação é feita entre pesos $v_{p,n}$ (que ordenam as novas variáveis de acordo com sua importância de forma decrescente), o que é feito por meio da variância, combinados com as *n* variáveis originais A_n. Neste sentido, as novas componentes são ordenadas e descrevem o cenário original em função das variâncias, de forma que as primeiras PC's possuem os maiores pesos, ou seja, as maiores variâncias. Com isto, é possível selecionar um número de *p=1* até *p=m* de PC's com pesos (ou variâncias) que, quando somados, expliquem uma parcela relevante da variância total do cenário original. A medida de exemplo, se a primeira componente (PC_1) de um sistema representa 75% da variância original, a segunda componente (PC_2) representa 15% da variância original e a terceira componente (PC_3) represente 5% da variância original, então a utilização destas 3 componentes principais juntas para formar o novo sistema de variáveis terá 95% da variância original explicada, o que mede o quão fiel o novo sistema é com relação ao original [38, 39].

No novo sistema de componentes, no caso dos espectros de absorbância, é possível realizar, por exemplo, a análise de *Loading Plots*. Tal gráfico traz a informação dos pesos $v_{p,n}$ em função do número de onda dos espectros, de forma que se torna possível analisar ondas que demonstram o comportamento combinado de bandas de espectros [39].

A **Figura 8**, extraída da base teórica da tese de mestrado de PEREIRA, T. M. [39], representa um exemplo de *dataset* gerado por soma de gaussianas para criar dois grupos de 100 espectros, cujas médias aparecem como as ondas A e B e a diferença entre estas médias espectrais como A - B. Nesta representação gráfica é

possível notar um *shift* das bandas na região de 1020cm^{-1} à 1080cm^{-1} e uma diferença de intensidade de absorbância entre 1220cm^{-1} e 1370cm^{-1}.

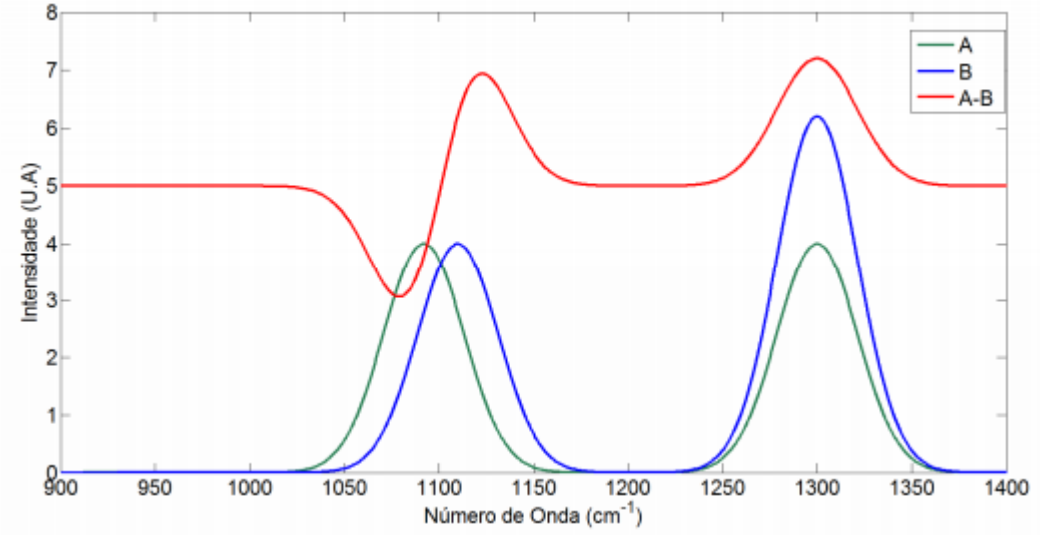


Figura 8 - *Espectros aleatórios médios gerados por soma de gaussianas [39].*

Aplicando-se a transformada de variáveis PCA para obtenção das componentes principais deste conjunto de dados, é possível avaliar o *Loading Plot* da **Figura 8**, onde é possível notar as 3 PC's que juntas ultrapassam 99% da variância original dos dados, conforme **Figura 9**. Neste gráfico, a PC_1 apresenta um comportamento muito semelhante aos espectros médios originais e, em análise, representam a média dos espectros de A e B. Olhando para a PC_3 é possível notar um comportamento de função assimétrica circulado em vermelho, o que é devido ao *shift* de bandas apresentado na **Figura 8**, enquanto a função simétrica em formato de vale circulado em amarelo tem como origem a diferença de magnitude das bandas A e B presentes na **Figura 8**.

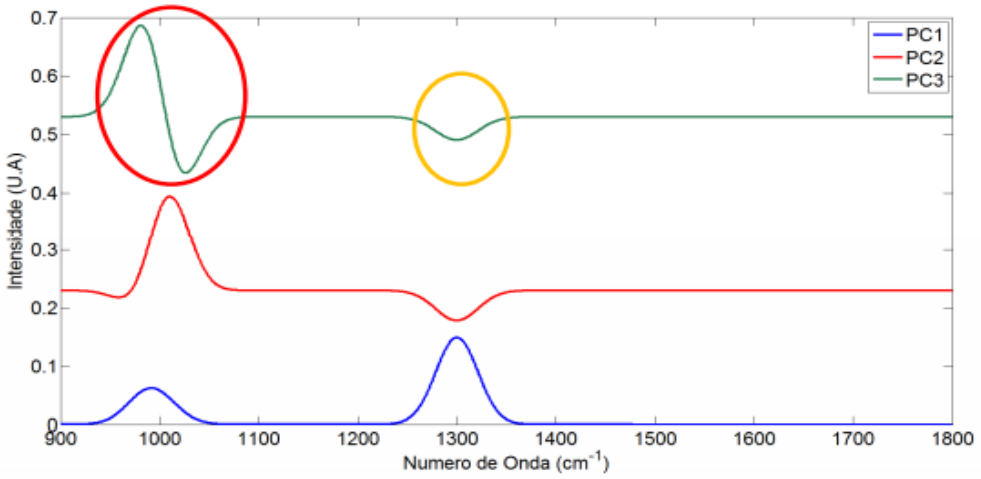


Figura 9 - *Loading Plot dos espectros médicos demonstrados na Figura 2 [39].*

O exemplo hipotético acima é bastante simplista, mas exemplifica bem como a redução de variáveis por PCA pode auxiliar na diferenciação de classes. Entretanto, quando se tem amostras biológicas como as cepas de fungos, os espectros são bem mais complexos do que a soma de gaussianas da **Figura 8** e são formados por sobreposição de várias bandas de absorção. Além disso, a presença de contaminantes como parafinas e vapor d'água em determinadas regiões do espectro tem grande interferência sobre a variância dos dados [39], o que é chave para cálculo das componentes principais, ou seja, tais ruídos causam distorção dos *loading plots* e dificultam a análise e classificação.

Entende-se, então, que a maneira mais otimizada de se reduzir o espaço original de variáveis a um cenário computacionalmente mais simples, é por meio de uma PCA, entretanto não necessariamente a direção dos autovetores das novas PCs terão, individualmente, a melhor indicação para separação de classes [40]. Para este fim, segue-se a aplicação do algoritmo de classificação supervisionada LDA, sendo que esta combinação de técnicas matemáticas PCA-LDA traz resultados interessantes na separação de classes de espectros de absorção, podendo ser otimizada de acordo com a variação dos parâmetros de pré-processamento dos dados.

3. OBJETIVOS

O objetivo do presente trabalho é determinar os melhores parâmetros no filtro de suavização de Savitzky-Golay que otimize a classificação, por LDA-PCA, de espectros FTIR de seis linhagens do fungo do gênero *Metarhizium*.

3.1. Objetivos específicos

Os objetivos específicos do trabalho são:

- Avaliar a melhor região espectral (restrição espectral) para classificação dos fungos *Metarhizium.*
- Avaliar os melhores parâmetros de janelamento e derivação na filtragem de sinal para construção de classificadores LDA-PCA.
- Avaliar a técnica de PCA como uma ferramenta para redução de dimensionalidade e melhoria da razão sinal ruído (SNR).

4. METODOLOGIA

Dividindo-se em 4 etapas principais, a metodologia aqui descrita inicia-se pela determinação e separação do objeto de estudo (quais linhagens de qual gênero de fungo seriam analisadas).

Em seguida, foi realizada a coleta dos espectros dos objetos de estudo em laboratório, enquanto a terceira etapa correspondeu ao pré-processamento de tais dados coletados. Por fim, a informação pré-processada foi analisada por meio de técnicas estatísticas multivariadas tais como Análise de Componentes Principais (PCA) e Análise Discriminante Linear (LDA).

Com as quatro etapas concluídas, resultados gráficos e quantitativos foram extraídos para análise e conclusão sobre o objetivo do trabalho em classificar as linhagens de fungo colocadas como foco do estudo, comparando diferentes graus de derivação e dimensão de janela de filtragem na técnica de Savitzky-Golay.

Toda a parte relacionada à separação das espécies e meios de cultura, coleta dos espectros e digitalização dos dados foi realizada como parte do desenvolvimento do projeto de mestrado [24] de Marina Ribeiro Batistuti, mestranda orientada pelo Prof. Dr. Luciano Bachmann, pela faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo (2012).

4.1. Objeto de estudo

O foco deste estudo foi dado na diferenciação de linhagens de um mesmo gênero específico do Reino *Fungi* e do Filo *Ascomycota*, que é o gênero Metarhizium. Tal gênero foi empregado ao estudo com análise de 6 de suas linhagens conhecidas, conforme **Figura 10**, frequentemente encontrados na formulação de bioinseticidas.

As linhagens utilizadas no estudo foram três da espécie *Metarhizium acridum* (ARSEF 324, ARSEF 3391 e ARSEF 7486), uma da espécie *Metarhizium anisopliae* (ARSEF 5749) e duas da espécie *Metarhizium brunmeum* (ARSEF 1095 e ARSEF 5626). Todas as linhagens foram obtidas da USDA – ARSEF collection of Entomopathogenic Fungal Cultures (U.S. Plant, Soil and Nutrition Laboratory, Ithaca, NY).

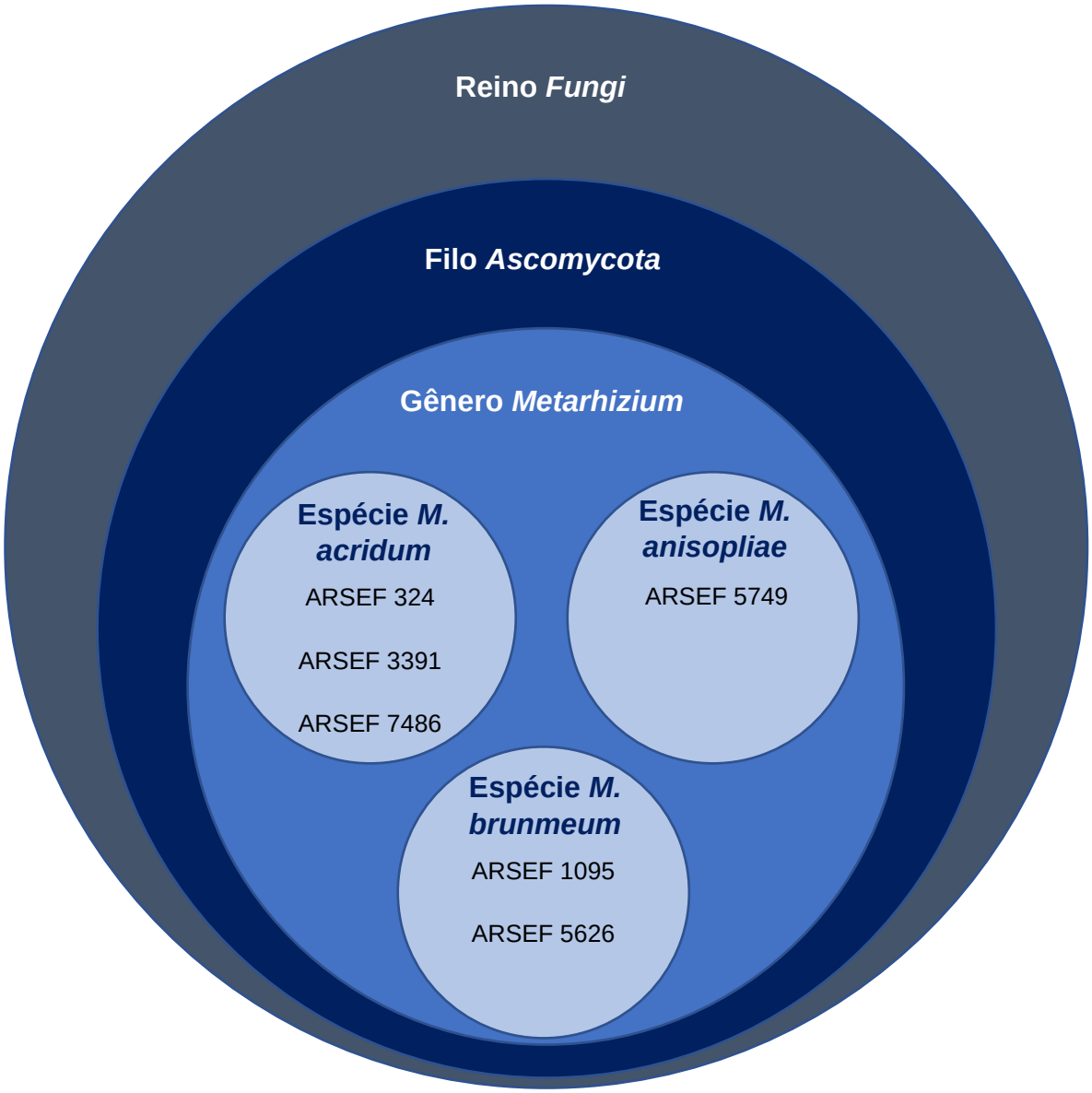


Figura 10 - *Classificação Taxonômica de seis linhagens distribuídas em 3 espécies do fungo de gênero Metarhizium (3 linhagens da espécie M. acridum, 1 linhagem da espécie M. anisopliae e 2 linhagens da espécie M. brunmeum). Tal gênero pertencente ao filo Ascomyota.* ***Fonte:*** *Próprio autor.*

4.2. Coleta dos fungos

O cultivo dos fungos foi inspecionado pelo Prof. Dr. Gilberto Úbida Leite Braga, que deu apoio a mestranda Marina Ribeiro Batistuti e orientador Prof. Dr Luciano Bachmann na coleta das espécies de fungos para a dissertação de mestrado [24] pela faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo (2012).

Placas de Petri 90 x 15mm foram empregadas para receber o meio de cultura, dotado somente de 23mL de meio PDA (Potato Dextrose Agar ou Ágar de batata) (Difco, EUA). Tais placas foram preparadas no dia anterior a inserção dos fungos para que houvesse tempo de analisar as mesmas, minimizando assim as chances de contaminação.

O fungo foi preparado para inoculação com 2mL de suspensão de conídios em uma solução de 0,01% (v/v) de *tween-80* (empresa Sigma – Aldrich, EUA) alocados em tubos de vidro (empresa Schott GL 18, EUA). Foi empregada uma câmara de contagem de Neubauer (empresa Boeco, Alemanha) para fixar a concentração de conídios das suspensões em $1x10^8$ células mL^{-1}. Cada placa de Petri recebeu 100µL da suspensão preparada, sendo três placas utilizadas para cada linhagem, de forma a garantir a reprodutibilidade do experimento. As 18 placas foram incubadas por 12 dias na ausência de luz em uma Incubadora B.O.D.411D (empresa Nova Ética, Brasil) a 28°C. Após o período de incubação, as placas foram levadas diretamente para coleta de espectros de infravermelho.

4.3. Coleta dos espectros

Foi utilizado um espectrômetro Nicolet 380 (empresa Thermo Nicolet, EUA) acoplado a um acessório de Reflexão Totalmente Atenuada (ATR) Durascope (empresa Detection, EUA), equipado de um cristal de diamante que auxilia na detecção de espectros entre 700 e 4000cm^{-1}, enquanto espectros de 400 a 700cm^{-1} possuem uma alta absorbância por si só. A resolução dos espectros coletados foi de 4cm^{-1} na região espectral completa entre 400 e 4000cm^{-1}. A área útil do equipamento (onde a amostra foi posicionada) possuía 250µm de diâmetro.

Com os conídios coletados no laboratório onde os meios de cultura foram cultivados, eles foram levados para o laboratório onde primeiramente a absorbância apenas do *background* foi coletada, ou seja, apenas do cristal do ATR. Depois os conídios foram depositados e comprimidos sobre o cristal pelo aparelho de espectroscopia antes da coleta.

A um passo de 1cm^{-1} e 32 repetições por espectro, pôde ser realizada a aquisição dos espectros de absorção, sendo dez espectros coletados de cada uma

das placas, sendo que cada linhagem possuía 3 placas de Petri para cultivo, isto é, totalizando 30 espectros de absorção coletados por amostra.

4.4. Pré-processamento

Com os dados provenientes da linha de pesquisa de mestrado de Marina Ribeiro Batistuti, pôde-se desvencilhar da metodologia de coleta e estudo para então focar na nova análise desenvolvida no presente projeto.

Os espectros adquiridos foram colocados em um arquivo denominado *Metarhizium.mat* para MATLAB®, estando divididos em estruturas de dados por linhagem do fungo *Metarhizium*.

- 32 Espectros de ARSEF 324 – Estrutura M324
- 34 Espectros de ARSEF 3391 – Estrutura M3391
- 32 Espectros de ARSEF 7486 – Estrutura M7486
- 30 Espectros de ARSEF 5749 – Estrutura M5749
- 30 Espectros de ARSEF 1095 – Estrutura M1095
- 31 Espectros de ARSEF 5626 – Estrutura M5626

Os dados foram carregados e processados em ambiente MATLAB®, utilizando rotinas desenvolvidas *in house* pelo Prof. Dr. Thiago Martini Pereira (Universidade Federal de São Paulo, ICT - São José dos Campos) para filtragem de Savitzky-Golay.

Primeiramente todas as seis linhagens foram agrupadas em uma única matriz para que os dados pudessem ser pré-processados conjuntamente. Em seguida, foi aplicada uma redução da taxa de amostragem (*downsample*) a um passo de 4 pontos, ou seja, a partir do início do espectro, foi suprimida a amostra dos 4 pontos seguintes e mantida a posterior, sendo esse processo repetido até o final do dado original. Tal ação foi tomada para padronização com um passo de $4cm^{-1}$, que é usualmente aplicado para análises de classificação de espectros de absorção na área de bioespectroscopia. Isso porque os dados desta pesquisa foram coletados a um passo de $1cm^{-1}$. Este processo suaviza a curva do espectro de absorção.

Em seguida, foi iniciada a filtragem pelo método de Savitzky-Golay, em que o espectro foi suavizado em função de um grau de derivação e da dimensão de uma

janela. Neste ponto foi criada uma rotina para extrair diferentes cenários de um espectro filtrado, variando o grau de derivada e a dimensão da janela aplicada. Seis cenários foram estudados variando a derivada entre primeiro e segundo grau, com a janela variando de 11, 13 ou 15 pontos. Os parâmetros de P1 a P6 foram escolhidos baseando-se no que há de mais comum nas pesquisas da comunidade científica da área de bioespectroscopia [1, 44].

- P1 – Derivada de 1º grau e Janela de 11 pontos.
- P2 – Derivada de 1º grau e Janela de 13 pontos.
- P3 – Derivada de 1º grau e Janela de 15 pontos.
- P4 – Derivada de 2º grau e Janela de 11 pontos.
- P5 – Derivada de 2º grau e Janela de 13 pontos.
- P6 – Derivada de 2º grau e Janela de 15 pontos.

Neste ponto, para cada cenário pôde-se realizar a restrição espectral conforme necessidade, utilizando uma ferramenta simples de corte do espectro utilizando de um limite inferior e superior. Do espectro geral de 400 a 4000cm^{-1}, as principais regiões analisadas foram:

- 900 a 1800cm^{-1};
- 900 a 1350cm^{-1};

Com a secção selecionada do espectro, uma Variação Normal Padrão (Normalização por SNV) foi aplicada de forma a normalizar o sinal, subtraindo o espectro médio de cada espectro e depois dividindo pelo desvio padrão dos mesmos.

Com os dados normalizados, aplicou-se a técnica de análise por PCA (*Principal Component Analysis*) de forma a extrair o gráfico de s*catter plot com as duas primeiras componentes* e o *loading plot* com cinco componentes. Tais dados foram utilizados para avaliar a separação dos grupos de acordo com cada variação de parâmetros do filtro de Savitzky-Golay.

4.5. Classificação

Com espectros pré-processados foi possível analisar os dados por meio de técnica de análise discriminante em função das classes determinadas pelas linhagens fúngicas estudadas. Para este trabalho, foi definida a aplicação de Análise Discriminante Linear de Fisher (LDA) com validação cruzada do tipo K-fold de 10 camadas (*folds*). Nesta validação, o classificador é desenvolvido com *k*-1 particionamentos do conjunto de dados original e o único subgrupo não utilizado é então aplicado para validação. Tal processo é repetido *k* vezes e a partir de todas as validações, obtém-se a acurácia do classificador, minimizando efeitos de Overfitting dos dados. A técnica de Análise de Componentes Principais (PCA) foi empregada com objetivo de reduzir a dimensionalidade dos dados e assim otimizar a relação sinal ruído a partir da ordenação da variância das novas variáveis. Para o presente trabalho, selecionaram-se as primeiras componentes principais que representam 95% da variância acumulada, o que é comumente empregado em trabalhos acadêmicos da área de bioespectroscopia [24].

5. RESULTADOS

Este trabalho inicia-se com a análise do espectro médio das seis linhagens de fungo do gênero Metarhizium, de forma a analisar os dados completos e comparar as regiões espectrais. Com foco em uma região específica do espectro, pôde-se realizar o estudo de comparação de parâmetros de pré-processamento. Variando-se o grau de derivada e a dimensão das janelas aplicadas ao método de Savitzky-Golay, foi possível determinar qual cenário apresentaria os melhores resultados de classificação das linhagens do fungo aqui estudado.

O espectro de absorção típico de um fungo é formado por polissacarídeos, ligações do tipo CH2 e CH3 de lipídios, ligações P=O de fosfodiésteres, amida, sebo e vapor d'água [43]. Os espectros médios das seis linhagens do fungo foram plotados de forma completa, com número de onda variando de 500 a 4000cm^{-1}, conforme **Figura 11**, representando toda a estrutura bioquímica do fungo.

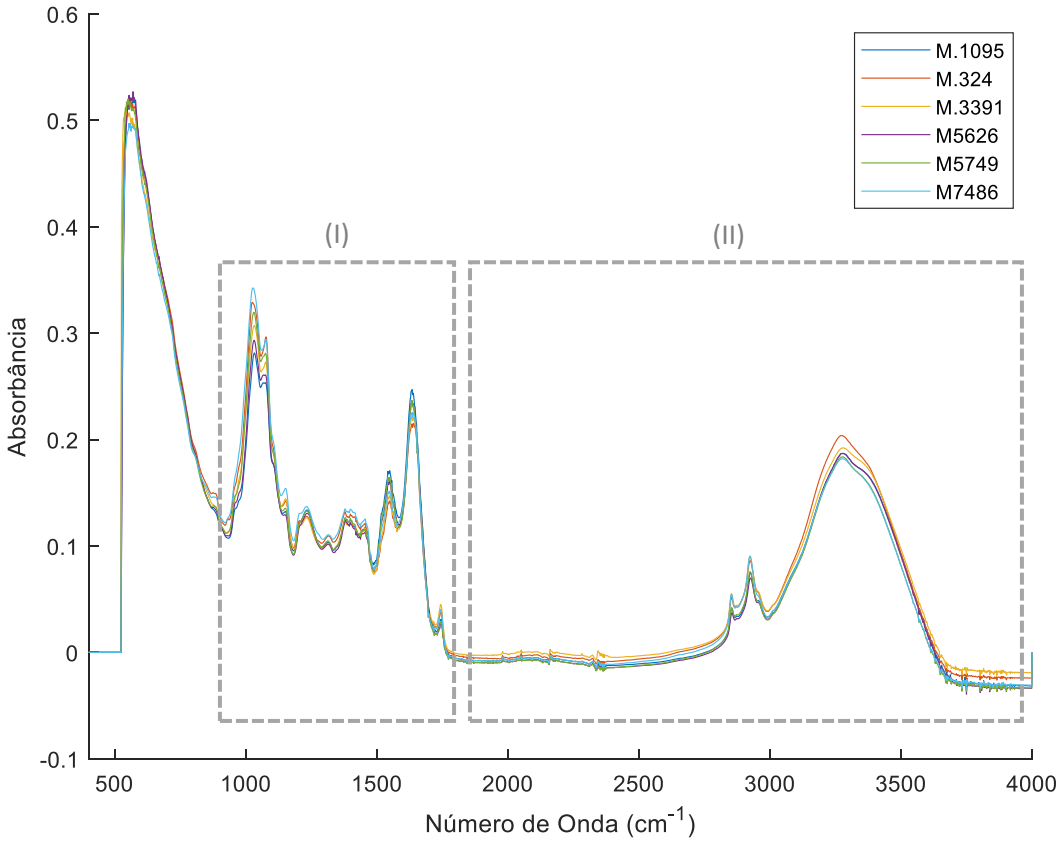


Figura 11 - *Espectro médio completo das seis linhagens de fungo do gênero Metarhizium utilizadas neste trabalho. O intervalo (I) define a região de interesse, onde concentra-se mais informação bioquímica relevantes para classificação dos fungos. O intervalo (II) representa uma região mais ruidosa.*

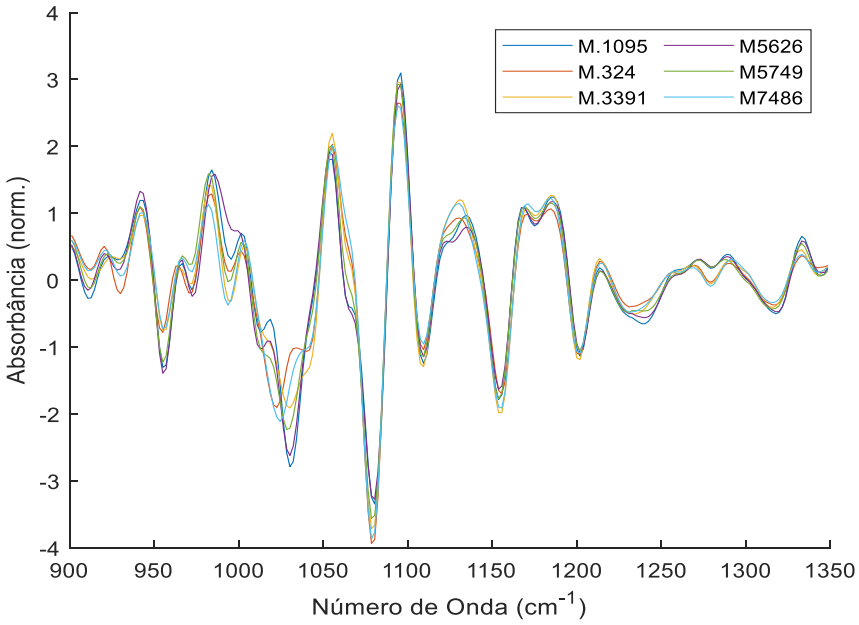
Existem regiões de maior importância no que se diz respeito a informação para classificação de linhagens fúngicas, o que leva a ideia de seccionar apenas parte do espectro. Os dados, portanto, não precisam ser levados de forma completa adiante no estudo, uma vez que a literatura reporta que a região mais importante para classificação de fungos fica entre 900 e 1800cm^{-1}**,** representado pelo intervalo (I) da **Figura 11**. O intervalo (II) da figura representa uma região de baixa absorbância (entre 1800 e 2600cm^{-1}), seguida de uma região de vapor d'água (entre 2600 e 3700cm^{-1}) [24]. Esta última pôde ser suprimida para foco apenas na região de maior importância em classificação.

Dentro do espectro já seccionado na **Figura 11** (I), observou-se ainda que é possível manter dois cenários: o intervalo de 900 a 1350cm^{-1} ou manter o intervalo completo de 900 a 1800cm^{-1}. Essa possibilidade se deu avaliando que a região acima de 1350cm^{-1} parecia ser mais ruidosa e então distorcer a classificação que seria feita em seguida, conforme **Figura 12**.

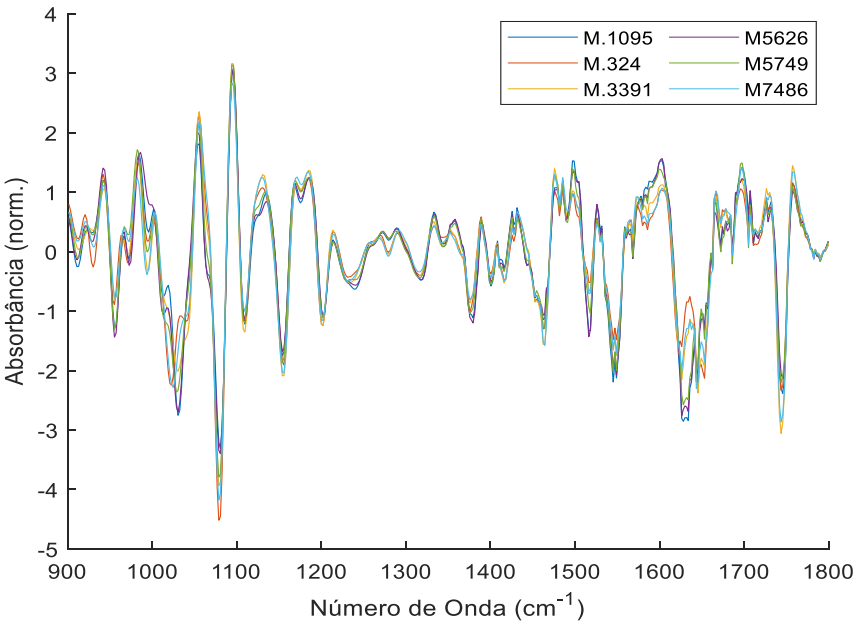
Analisando o espectro médio normalizado presente na **Figura 12** (b), nota-se que a região acima de 1350cm^{-1} apresenta uma distorção característica de alta frequência que ocorre de forma consistente devido à modos rotacionais de vapor d´água atmosférico. Apesar desta região ter diversos modos vibracionais relacionados a Proteínas, e Lipídios [14], a contribuição espectral do vapor d´água pode piorar a acurácia do classificador.

A região limitada de 900 a 1350cm^{-1} mostrada na **Figura 12** (a) tem uma característica mais suavizada e com regiões mais separadas (como acontece entre 1000 e 1060cm^{-1}). Esta parte do espectro é formada principalmente por Polissacarídeos, Riboses e Desoxirriboses, Glicogênio, Fosfato, Carotenoides, Colágeno e Amida III.

Além da análise do espectro médio normalizado, foi possível gerar e comparar o PCA - Scatter Plot dos dois intervalos, conforme a **Figura 13**. É visível que os grupos se separam melhor no espectro de 900 a 1350cm^{-1} (a) do que no espectro que se estende até 1800cm^{-1} (b).

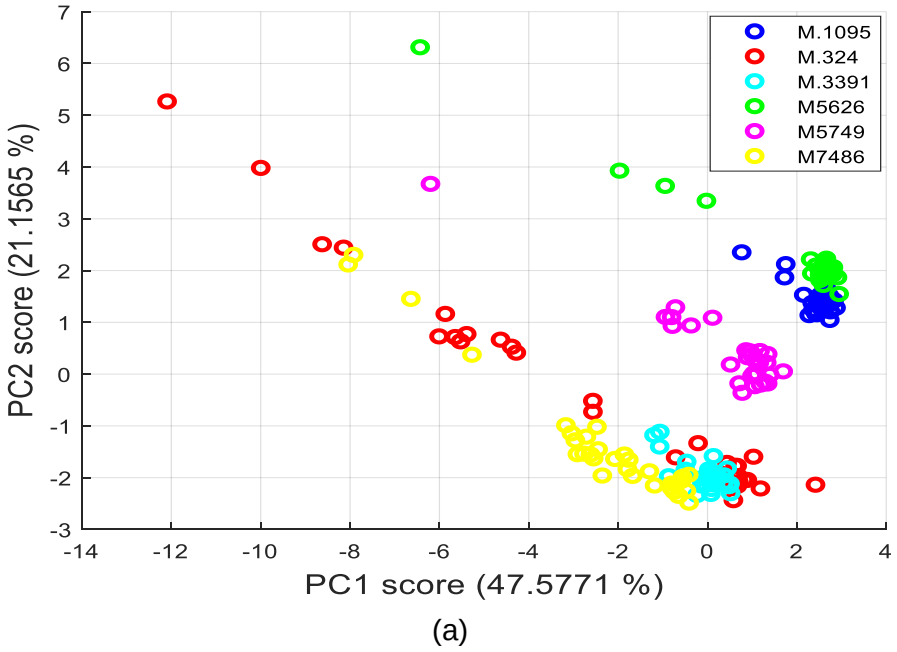


(a)

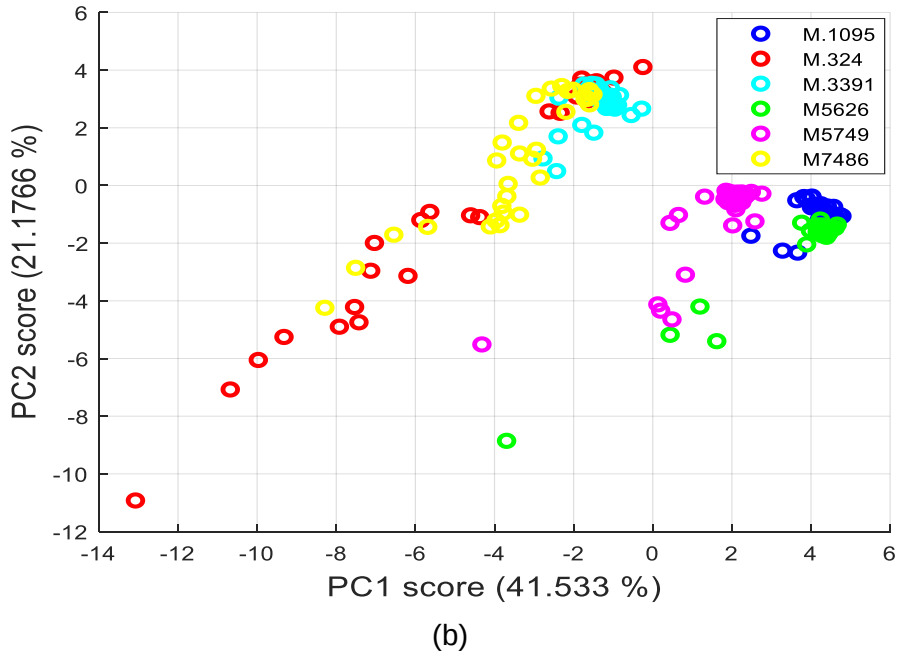


(b)

Figura 12 - *Espectro médio normalizado filtrado pelo método de Savitzky-Golay com derivação de segundo grau e janela de dimensão 11, sendo (a) o espectro de 900 a 1350cm^{-1} e (b) o espectro de 900 a 1800cm^{-1}.*



(a)



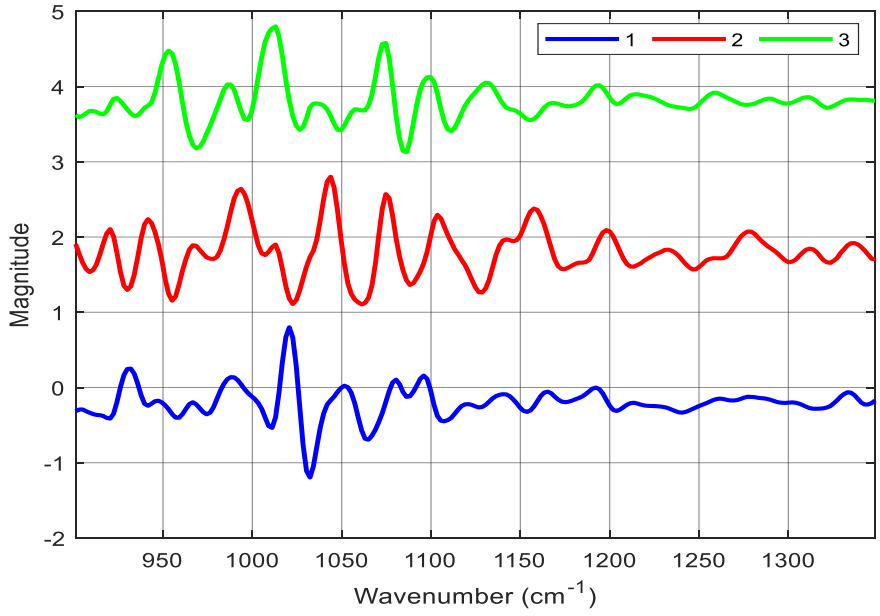
(b)

Figura 13 - *PCA - scatter plot do espectro médio normalizado filtrado pelo método de Savitzky-Golay com derivação de segundo grau e janela de dimensão 11, sendo (a) o espectro de 900 a 1350cm^{-1} e (b) o espectro de 900 a 1800cm^{-1} (que apresenta uma rotação de 90º.*

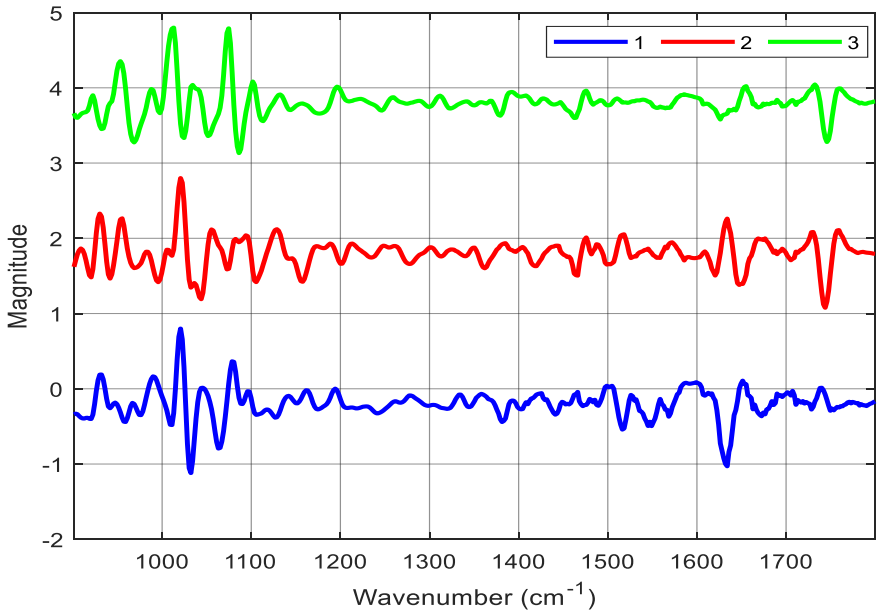
Ainda foi possível efetuar a plotagem do Loading plot para estes mesmos cenários de espectros, conforme **Figura 14**, com 3 variáveis analisadas.

Percebe-se nesta figura uma análise semelhante a anteriormente feita quanto ao scatter plot, onde as curvas da banda truncada em 1350cm^{-1} (a) são mais

suavizadas se comparadas às mesmas presentes em (b), o que se dá devido a presença de ruído oriundo do vapor d'água na região entre 1350 e 1800cm^{-1}.



(a)



(b)

Figura 14 - *PCA - Loading plot do espectro médio normalizado filtrado pelo método de Savitzky-Golay com derivação de segundo grau e janela de dimensão 11, sendo (a) o espectro de 900 a 1350cm^{-1} e (b) espectro de 900 a 1800cm^{-1}.*

Percebe-se nesta figura uma análise semelhante a anteriormente feita quanto ao scatter plot, onde as curvas da banda truncada em 1350cm^{-1} (a) são mais

suavizadas se comparadas às mesmas presentes em (b), o que se dá devido a presença de ruído oriundo do vapor d'água na região entre 1350 e 1800cm^{-1}.

Dadas estas análises de comparação entre regiões do espectro, conclui-se que a restrição espectral de 900-1350cm^{-1} (**Figura 12**-a) apresenta uma melhor separação entre grupos. Observamos que o grupo M5749 (pontos lilás) está mais bem separado na **Figura 13**-a e a dispersão dos pontos de cada grupo é um pouco menor do que na análise de PCA usando a região de 900-1800 cm^{-1}.

A melhor área do espectro para dar seguimento às análises de classificação seria a faixa de 900 a 1350cm^{-1}, a qual fornece informação suficiente, mas sem interferência de ruído.

Desta maneira, com o objetivo de identificar uma forma relevante de se diferenciar as linhagens do fungo de gênero Metarhizium cujos espectros foram coletados, estabeleceu-se já de início a análise comparativa dos espectros médios normalizados e filtrados via Savitzky-Golay. Esta primeira avaliação permitiu o estudo da forma das curvas de cada linhagem fúngica com a variação do grau da derivada aplicada (primeiro ou segundo grau) e de qual janela seria utilizada no pré-processamento (de tamanho 11, 13 ou 15). Isto garantiu um cenário onde foi possível avaliar algumas regiões de frequência interessantes, destacadas nos blocos conforme **Figura 15**.

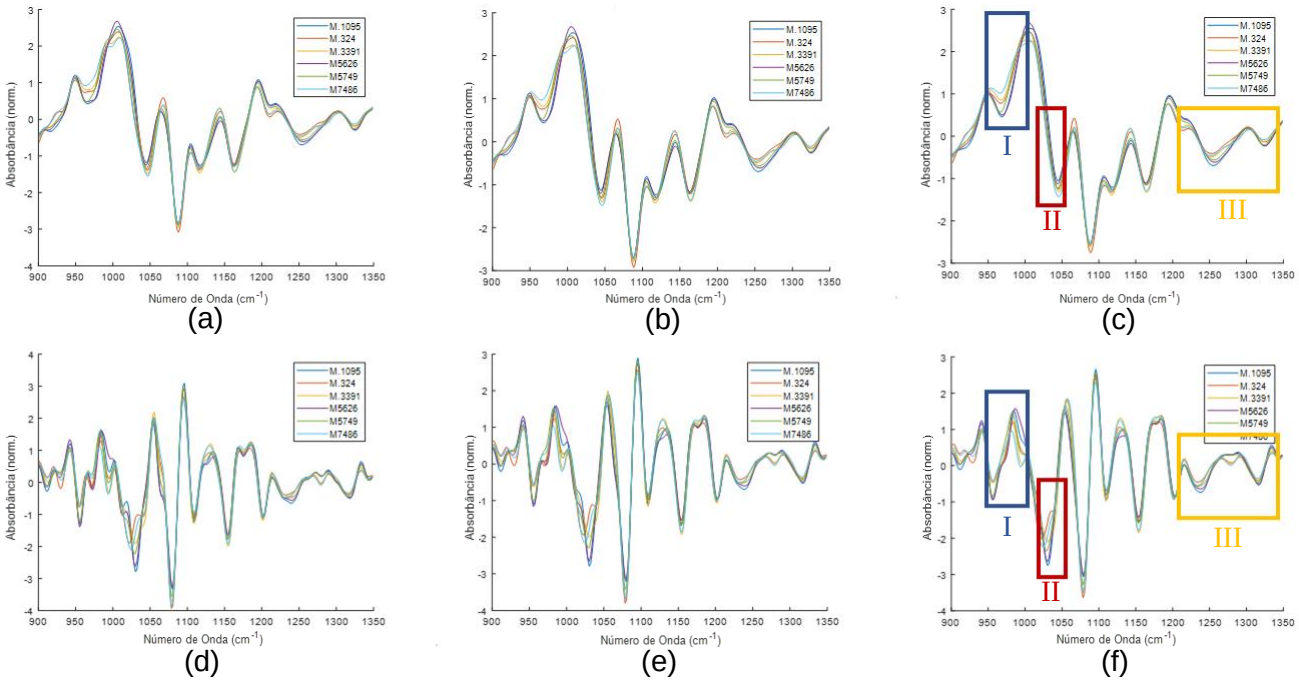
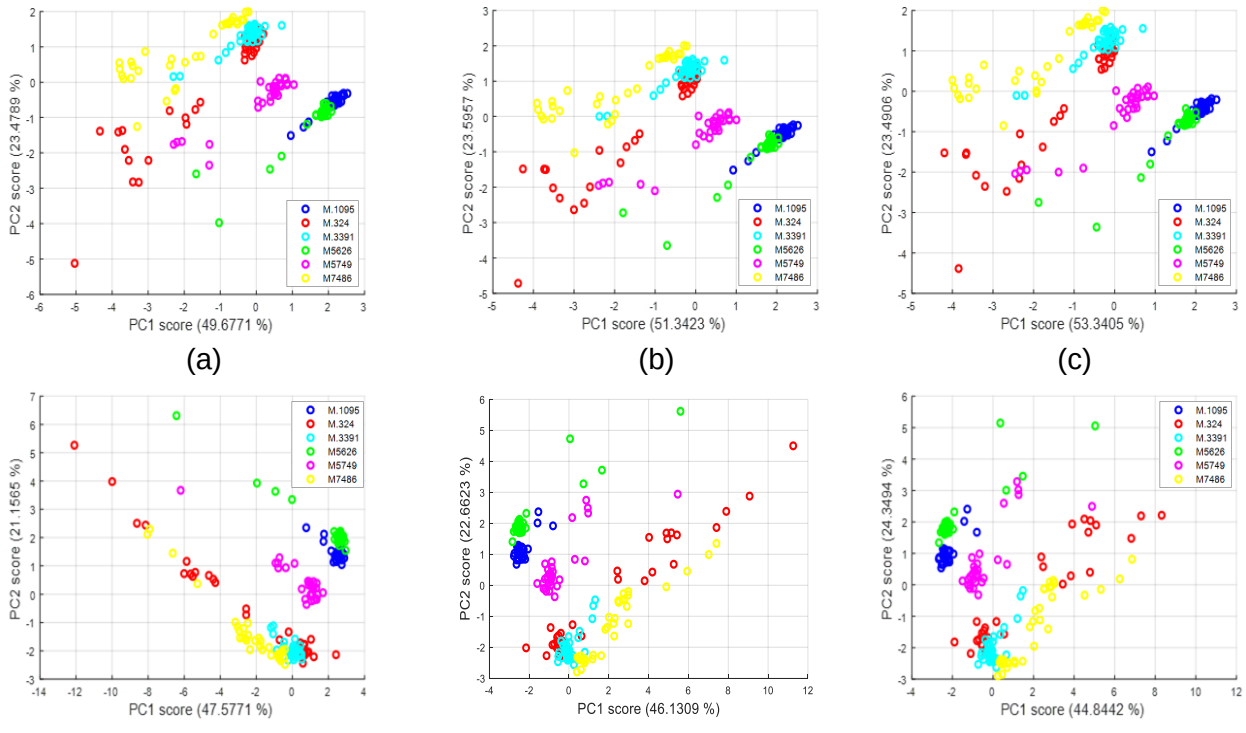


Figura 15 - *Espectro médio normalizado das seis linhagens do gênero Metarhizium, comparando a variação de janela e derivada. (a) – 1º Derivada com janela de 11 pontos; (b) - 1º Derivada com janela de 13 pontos; (c) – 1º Derivada com janela de 15 pontos; (d) – 2º Derivada com janela de 11 pontos; (e) - 2º Derivada com janela de 13 pontos; (f) – 2º Derivada com janela de 15 pontos.*

Nota-se que as janelas menores (de tamanho 11 e 13) apresentaram derivadas mais ruidosas (como esperado), mesmo que mantenham mais informações do que a janela de tamanho 15, uma vez que esta última apresente uma maior capacidade de suavização da curva. Além disso, ao utilizar a segunda derivada e a maior janela, é possível perceber uma melhor separação das bandas devido a uma menor resolução, conforme destacado nos quadros I, II e III da **Figura 15** – (c) e (f). Estes casos correspondem a três determinadas regiões específicas do espectro que tem uma relação direta com a bioquímica do fungo, sendo que a região I está amplamente relacionada a Desoxirriboses, a região II é composta principalmente de Glicogênio e Riboses, e a região III formada de Colágeno Fosfato e Amida III [24].

Após o pré-processamento e destaque de algumas regiões por meio da análise qualitativa, realizou-se Análise de Componentes Principais (*PCA – Principal Component Analysis*). Desta forma, foi possível promover uma comparação entre cada grupo, usando diferentes parâmetros do filtro de Savitzky-Golay. A **Figura 16** mostra os gráficos de *scatter plot* (diagramas de dispersão).



(a) (b) (c)

(d) (e) (f)

Figura 16 - *Espectro médio analisado após Análise de Componentes Principais das PCs 1 e 2, contendo as seis linhagens do fungo do gênero Metarhizium analisadas. (a) – 1º Derivada com janela de 11 pontos; (b) - 1º Derivada com janela de 13 pontos; (c) – 1º Derivada com janela de 15 pontos; (d) – 2º Derivada com janela de 11 pontos; (e) - 2º Derivada com janela de 13 pontos; (f) – 2º Derivada com janela de 15 pontos.*

Nota-se que ao utilizar a janela de 15 pontos, há uma maior remoção de informação e por isso resulta em uma maior suavização do sinal. Este cenário apresenta uma separação melhor dos grupos quando analisado visualmente, o que é fruto de uma maior remoção do ruído se comparado com as janelas de 11 e 13 pontos. Complementando-se a esse fato, as janelas menores promovem uma preservação maior de informação e por isso tendem a possuir um resultado de PCA melhor, contudo, este resultado pode ser comprometido devido a maior quantidade de ruído presente.

Para ter uma melhor compreensão dos valores das componentes principais plotadas em **Figura 16**, foram plotados os *loading plots*, conforme Figura 13. Para a análise destes gráficos é necessário frisar que não existe sentido em comparar cenários de diferentes graus de derivação e, portanto, é necessário fixar o grau da derivada e então comparar apenas a variação de dimensão de janelas. Na **Figura 17**, portanto, é necessária uma análise horizontal de comparação, mas não vertical.

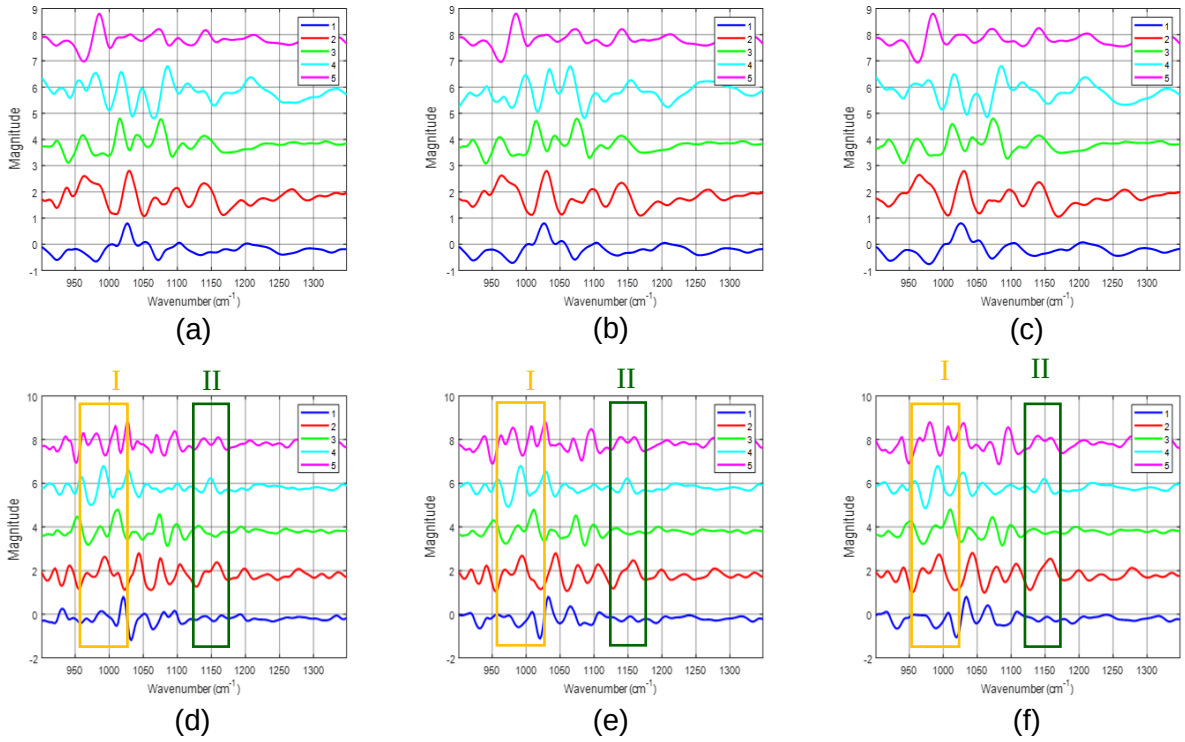


Figura 17 - *Loading Plots com a variação de dimensão das janelas e graus de derivada. (a) – 1º Derivada com janela de 11 pontos; (b) - 1º Derivada com janela de 13 pontos; (c) – 1º Derivada com janela de 15 pontos; (d) – 2º Derivada com janela de 11 pontos; (e) - 2º Derivada com janela de 13 pontos; (f) – 2º Derivada com janela de 15 pontos. Duas regiões relevantes para comparação estão destacadas nos loading de segundo grau de derivada, identificados pelos algarismos I e II.*

Nota-se que no gráfico de Janela de 11 pontos e segunda derivada (d) a primeira componente encontra-se rotacionada de 180º se comparada com a mesma PC em (e) e (f). Isso acontece durante o processo de pré-processamento, onde o eixo no qual a informação se dispõe é diferente, mas não afeta a análise por não conter consequências quantitativas.

Percebe-se que não há grande variação nos *Loadings* quando comparados, mas é possível notar algumas suavizações especialmente nas PCs 2 e 5 quando comparadas entre si na segunda derivada, especificamente na região de 950 a 1025cm^{-1} e entre 1125 e 1175cm^{-1} destacados respectivamente pelos quadros I e II. Nota-se um pico triplo com um *shoulder* a direita que é suavizado no quadro I à medida que se aumenta a dimensão das janelas de filtragem, assim como acontece com o pico duplo presente nos quadros identificados com o algarismo II.

Foi construído um classificador onde algumas análises numéricas puderam ser extraídas dos espectros processados de forma rápida e automática, o que foi o caso do nº de PCs necessárias para a variância mínima explicada de 95% e a acurácia de cada classificador de acordo com cada cenário derivação e janelamento, o que é visto na **Tabela 1**.

Grau da Derivada	Nº de pontos da Janela	Nº PCs (#)	Acurácia (%)	Sensibilidade (%)	Especificidade
1º	11	5	99,5	99,5	99,8
1º	13	5	98,9	99	99,7
1º	15	4	92,1	91,8	98,5
2º	11	6	98,4	98,5	99,7
2º	13	6	99,5	99,5	99,8
2º	15	5	100	100	100

Tabela 1 - *Apresentação dos resultados de Nº de PCs necessárias para alcançar 95% de variância explicada, Acurácia, Sensibilidade e Especificidade de cada cenário analisado do espectro, variando o grau da derivada e o número de pontos da janela de filtragem.*

Analisando-se a tabela percebe-se que a cenário com grau de derivação igual a 2 e janelamento em 15 oferece o melhor resultado quantitativo (5 PCs capazes de explicar 95% de variância com acurácia de 100%). No caso do cenário de grau 1 de derivação e janela de dimensão 15, é interessante analisar que se precisou de um baixo número de PCs para atingir a variância mínima, o que leva a conclusão de que é um cenário com menor chance de overfitting do dataset.

A técnica de PCA ordena as componentes de acordo com variância de forma decrescente. Considerando um cenário na qual o sinal coletado possui uma SNR maior do que 1, a ordenação de PCs acaba também ordenando o a SNR de forma decrescente, ou seja, as primeiras PCs carregam também menor ruído.

As segundas derivadas amplificam as diferenças no sinal entre os grupos analisados. Pode-se afirmar que o processo de “derivação” dos espectros funciona de forma similar ao processo de deconvolução de bandas. É de se esperar que a quantidade de informação no dataset das 2º’s derivadas é maior, levando a necessidade de um maior número de PCs para explicar a variância mínima de 95%. Ao passo que as primeiras derivadas possuem uma relação sinal/ruído menor e por isso uma acurácia com melhor desempenho no geral, com menor número de PCs necessário.

5.1. Discussão

Anualmente ocorrem mais de um milhão de mortes em decorrência de complicações dos mais de 11,5 milhões de casos de infecções fúngicas no mundo [2], uma tendência crescente muito ligada ao aumento de pacientes imunodeprimidos que oferecem ambiente propício para tais patógenos [1]. A crescente onda de casos coloca em pauta a necessidade de diagnóstico mais eficiente para um tratamento mais assertivo, o que depende da identificação de qual é exatamente o fungo que acomete cada paciente [44].

Em muitos hospitais, o diagnóstico destas infecções pode levar de 1 a 2 semanas quando técnicas de análise fisiológica do microrganismo são aplicadas [3]. Uma maior velocidade de diagnóstico é possível dependendo de análises do material genético, como no caso do PCR, uma técnica de sequenciamento do DNA do patógeno fúngico encontrado em amostras do paciente. Tal método é considerado um *gold standard* por ter alto grau de acuracidade na determinação da espécie infectante [45, 46], mas também um alto custo [47, 48], principalmente vinculado com a necessidade de reagentes específicos, chamado de primers, para cada tipo de cepa do fungo.

Diante disto, novas técnicas de identificação de fungos são necessárias, de forma a trazer uma maior rapidez na identificação do patógeno fúngico levando a um prognostico melhor e uma mais rápida recuperação do paciente. Neste cenário, análises espectroscópicas se colocam como uma alternativa rápida e de baixo custo para auxílio no diagnóstico dos patógenos do reino *Fungi*, oferecendo dados qualitativos e quantitativos para classificação de amostras em suas determinadas espécies e linhagens [46]. No caso dos espectros de absorção, informações muito importantes da estrutura bioquímica das amostras podem ser obtidas pela análise dos picos de absorbância na região de 900-1800 cm^{-1} [6], tendo a aplicação de FTIR já consolidada na identificação de câncer [29, 49], de células acometidas por infecção viral [50].

Com um conjunto de dados de espectro coletados, é possível realizar aplicação de uma das inúmeras técnicas de classificação baseadas em *machine learning* [32, 35]. O presente trabalho utilizou-se da combinação de LDA e PCA, que é colocada como uma das primeiras a serem aplicadas em estudos de classificação devido a sua simplicidade e baixo custo computacional, o que não impediu que fossem obtidos altos níveis de acurácia.

Aliado à classificação por análise discriminante linear, demonstrou-se que a filtragem de Savitzky-Golay tem grande impacto no processamento dos dados, muito vinculado a suavização do sinal e separação das bandas de espectro para uma melhor classificação. No presente trabalho foram aplicados os parâmetros mais comuns em pesquisas da área de bioespectroscopia [41, 42], mostrando que a variabilidade dos cenários de filtragem afeta diretamente a capacidade do classificador.

Os dados de espectroscopia no infravermelho, apesar de muito ricos em informação bioquímica para classificação, possuem grandes interferências e artefatos em sua base de dados, vinculados a flutuações de baseline e contaminantes como o vapor d'água [21]. Brunn et al. [51] demonstrou um método robusto para remoção das contribuições indesejadas do vapor d'água em espectros de FTIR na região de 1350 a 1800cm^{-1}. Contudo este método baseia-se em uma análise matemática de alto custo computacional e ainda depende de uma calibração prévia do equipamento de FTIR, ou seja, tem obstáculos que dificultam o emprego em larga escala. Neste sentido, aplicação da restrição espectral de regiões relevantes para identificação de fungos, mas sem a interferência de vapor d'água, se torna uma alternativa interessante para contornar tal problemática.

Neste contexto, os resultados encontrados pelo presente trabalho mostram que é possível realizar a classificação de fungos do gênero *Metarhizium* utilizando dados de FTIR somente da região 900-1350cm^{-1} com alta sensibilidade e especificidade.

Trabalhos futuros podem trazer a aplicação de outros modelos de classificadores e filtros de suavização e melhoramento da separação de bandas de espectros. As diversas técnicas disponíveis com seus mais distintos parâmetros de aplicação geram uma infinidade de cenários para possíveis análises comparativas com o intuito de encontrar o melhor método de classificação de cepas de fungos para diagnóstico médico.

6. CONCLUSÃO

Após o processamento dos dados e com os resultados obtidos neste trabalho, foi possível concluir que:

- A restrição espectral da região de 900 a 1350cm^{-1} possui grande potencial para classificação de fungos da do gênero *Metarhizium*.
- Os parâmetros de grau de derivada e tamanho de janela são importantes para construção de classificadores durante a etapa de processamento dos dados, contribuindo para suavização e otimização da separação de picos de absorção.

- A técnica de PCA é uma ferramenta importante para construção de classificadores, promovendo a redução do número de variáveis e ordenação das componentes principais em função da variância. Tal ordenação traz consigo a maximização da razão sinal-ruído (SNR).

7. REFERÊNCIAS BIBLIOGRÁFICAS

[1] GUARRO, J., GENE, J., STCHIGEL, A. M. Developments in fungal taxonomy. Clinical Microbiology Reviwes. v. 12, p. 454, 1999.

[2] FIORAVANTI, C. O ataque silencioso dos fungos. < https://revistapesquisa.fapesp .br/wp-content/uploads/2016/05/042_Fungos.pdf > Maio de 2016.

[3] THOMPSON, R. L., WRIGHT, A. J. General principles of antimicrobial therapy. Mayo Clinic Proceedings. v. 73, p. 995-1006, 1998

[4] Brasil. Agência Nacional de Vigilância Sanitária. Microbiologia Clínica para o Controle de Infecção Relacionada à Assistência à Saúde. < https://www20.anvisa.gov.br/segurancadopaciente/index.php/publicacoes/item/deteccao-e-identificacao-de-fungos-de-importancia-medica > Junho de 2014.

[5] Diem, M., et al., Molecular Pathology via Infrared and Raman Spectral Imaging1). Ex-vivo and In-vivo Optical Molecular Pathology, 2014: p. 45-102.

[6] Naumann, D., D. Helm, and H. Labischinski, Microbiological characterizations by FT-IR spectroscopy. Nature, 1991. 351(6321): p. 81-82.

[7] Naumann, A., A novel procedure for strain classification of fungal mycelium by cluster and artificial neural network analysis of Fourier transform infrared (FTIR) spectra.

[8] HARRINGTON T. C., RIZZO D. M. Defining Species in the Fungi. In: Worrall J.J. (eds) Structure and Dynamics of Fungal Populations. Population and Community Biology Series, v. 25. Springer, Dordrecht. Disponível em < https://doi.org/10.1007/978-94-011-4423-0_3 > Acesso em 02 mar. 2021.

[9] BUENO I. K., MALLER, A. Caracterização das linhagens mutantes do fungo Trichoderma reesei. Cascavel, 2018. 61 p. Dissertação (Mestrado em Ciências Farmacêuticas) – Universidade Estadual do Oeste do paraná. Disponível em < http://tede.unioeste.br/handle/tede/3702 > Acesso em 02 mar. 2021.

[10] ISAAC, C. E., JONES, A., & PICKARD, M. A. Production of cyclosporins by Tolypocladium niveum strains. Antimicrobial agents and chemotherapy, v. 34, p. 121–127, 1990. Disponível em < https://doi.org/10.1128/aac.34.1.121 > Aceso em 3 mar. 2021.

[11] BRAGA, G. U. L.; DESTEFANO, R. H. R.; MESSIAS, C. L. Oxygen consumption by Metarhizium anisopliae during germination and growth on different carbon sources. Journal of Invertebrate Pathology, v. 74, p. 112-119, 1999.

[12] BENNET, J. Aspergillus: a primer for the novice. Medical Mycology, v. 47, p.5-12, 2009.

[13] LATGE, J. Aspergillus fumigatus and aspergilosis. Clinical Microbiology Reviews, v.12, p. 310, 1999.

[14] MOWAT, E. et al. The characteristics of aspergillus fumigatus mycetoma development: is this a biofilm? Medical Mycology, v. 47, p. 120-126, 2009.

[15] BRAGA, G. U. et al. Effects of uvbirradiance on conidia and germinants of the entomopathogenic hyphomycete Metarhizium anisopliae: a study of reciprocity and recovery. Photochem Photobio, v. 73, p. 140-6, 2001

[16] ST LEGER, R. J.; GOETTEL, M.; ROBERTS, D. W.; STAPLES, R. C. Penetration events during infection of host cuticle by Metarhizium anisopliae. Journal of Invertebrate Pathology, v. 58, n. 1, p. 168-170, 1991.

[17] CHANDLER, D.; DAVIDSON, G.; PELL, J. K.; BALL, B. V.; SHAW, K.; SUNDERLAND, K. D. Fungal biocontrol of acari. Biocontrol Science Technology, v. 10, n. 3, p. 357-384, 2000.

[18] VELOSO, M. N. Avaliação in vitro dos efeitos da radiação ionizante em tecido ósseo bovino por espectroscopia ATR-FTIR e análise dinâmico-mecânica. Dissertação (Mestrado em Tecnologia Nuclear - Materiais) - Instituto de Pesquisas Energéticas e Nucleares, Universidade de São Paulo, São Paulo, 2013. Disponível em < https://teses.usp.br/teses/disponiveis/85/85134/tde-10012014-111938/publico/2013VelosoAvaliacao.pdf > . Acesso em 03 mar. 2021.

[19] Electromagnetic Energy Facts, and a Qualitative View. < https://www.ices-emfsafety.org/electromagnetic-energy/ > 2019.

[20] WILSON, Keith; WALKER, John (Ed. 7). Principles and techniques of biochemistry and molecular biology. Cambridge university press, 2010

[21] BAKER, M., TREVISAN, J., BASSAN, P. et al. Using Fourier transform IR spectroscopy to analyze biological materials. Nat Protoc 9, 1771–1791 (2014).

[22] Clinical application of FTIR imaging: new reasons for hope < https://www.ncbi.nlm.nih.gov/pubmed/20828847 > Outubro de 2010

[23] PIERŚCIŃSKI, K. et al. Analiza właściwości laserów kaskadowych pod kątem zastosowań w systemach łączności w otwartej przestrzeni. (Polish). Przeglad Elektrotechniczny, [s. l.], v. 94, n. 9, p. 1–9, 2018. Disponível em < http://pe.org.pl/articles/2018/9/1.pdf > Acesso em 2 mar. 2021.

[24] BATISTUTI, M. R. Classificação de fungos através da espectroscopia no infravermelho por transformada de Fourier. 2012. 174 f. Dissertação (Mestrado – Programa de Pós-graduação em Física aplicada à Medicina e Biologia) – Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto – SP, 2012.

[25] Reflectância Total Atenuada (ATR). Tecnologia de Amostragem ATR para Aplicações FTIR. < https://www.mt.com/br/pt/home/products/L1_AutochemProducts/ReactIR/attenuated-total-reflectance-atr.html > Fevereiro de 2021.

[26] DIEM, M., MAZUR, A., LENAU, K., SCHUBERT, J., BIRD, B., MILJKOVIĆ, M., KRAFFT, C., POPP, J. (2013), Molecular pathology via IR and Raman spectral imaging. Journal of Biophotonics. v. 6, N. 11–12, p. 855–886, Dezembro de 2013. Disponível em < https://doi.org/10.1002/jbio.201300131 > Acesso em 01 mar. 2021.

[27] HARIS, P. I.; CHAPMAN, D. Does fourie-transform infrared-spectroscopy provide useful information on protein structures. Trends in Biochemical Sciences, v. 17, p. 328-333, 1992.

[28] RASTOGI, Sumeet & Singh, Jagdish. (2003). Passive and iontophoretic transport enhancement of insulin through porcine epidermis by depilatories: Permeability and Fourier transform infrared spectroscopy studies. AAPS PharmSciTech. 4. E29. 10.1208/pt040329.

[29] FORATO, L. A.; BERNARDES, R.; COLNAGO, L. A. Study of resolution enhancement methods for analysis of secondary structure of proteins by ftir. Quimica Nova, v. 21, p. 146-150, 1998.

[30] FIGUEIREDO, N. S. Aplicação de Filtros de Savitzky Golay no processamento de Sinais de Eletrocardiografia. Itajubá, 2018. 61 p. Dissertação (Mestrado em Ciências em Engenharia Elétrica na área de concentração em microeletrônica) – Universidade Federal de Itajubá. Disponível em < https://repositorio.unifei.edu.br/xmlui/bitstream/handle/123456789/1562/dissertacao_2018114.pdf?isAllowed=y&sequence=1 > Acesso em 01 mar. 2021.

[31] BERTOZZO, R. J. Aplicação de Machine Learning em Dataset de Consultas Médicas do SUS. Florianópolis, 2019. 100 p. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) - Departamento de Informática e Estatística, Universidade Federal de Santa Catarina. Disponível em < https://repositorio.ufsc.br/handle/123456789/202663 > Acesso em 28 fev. 2021.

[32] LOBO, L. C. Inteligência Artificial e Medicina. Revista Brasileira de Educação Médica, Rio de Janeiro, v.41, n.2, p.185-193, Junho de 2017. Disponível em < http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-55022017000200185&lng=en&nrm=iso >. Acesso em 28 fev. 2021.

[33] LI Q., RAJAGOPALAN C, CLIFFORD G. A machine learning approach to multi-level ECG signal quality classification. Comput Methods Programs Biomed. 2014; 117 (3): 435-47.

[34] OBERMEYER Z., EMANUEL E. Predicting the future: big data, machine learning, and clinical medicine. N Engl J Med. 2016; 375 (13): 1216-1219.

[35] NARULA S., SHAMEER K., SALEM O., DUDLEY J., SENGUPTA P. Machine-Learning Algorithms to Automate Morphological and Functional Assessments in 2D Echocardiography. J Am Coll Cardiol. 2016; 68 (21): 2287-2295.

[36] LIMA, F. A. Microespectroscopia infravermelha de processos inflamatórios e tumores de cólon. Ribeirão Preto, 2016. 119 p. Tese (Doutorado em Física aplicada a Medicina e Biologia) – Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo. Disponível em < https://www.teses.usp.br/teses/disponiveis/59/59135/tde-03062016-151554/publico/TESE_Fabricio_2016.pdf > Acesso em 28 fev. 2021.

[37] MORAIS, Camilo L. M.; LIMA, Kássio M. G. Principal Component Analysis with Linear and Quadratic Discriminant Analysis for Identification of Cancer Samples Based on Mass Spectrometry. J. Braz. Chem. Soc., São Paulo, v.29, n. 3, p. 472-481, Mar. 2018.

[38] MINGOTI, S. ANÁLISE DE DADOS ATRAVÉS DE MÉTODOS DE ESTATÍSTICA MULTIVARIADA. [21]: UFMG, 2005.

[39] PEREIRA, T. M. Análise das diferenças bioquímicas nos tecidos e lesões tireoidianas por imageamento espectral obtidos por espectroscopia no infravermelho (FTIR). São Paulo, 2013. 119 p. Tese (Doutorado em Ciências na Área de Tecnologia Nuclear – Materiais) – Instituto de Pesquisas Energéticas e Nucleares, autarquia associada à Universidade de São Paulo. Disponível em < https://www.teses.usp.br/teses/disponiveis/85/85134/tde-09122013-141944/publico/2013PereiraAnalise.pdf > Acesso em 28 fev. 2021.

[40] KITANI, E.; THOMAZ, C. ANÁLISE DE DISCRIMINANTES LINEARES PARA MODELAGEM E RECONSTRUÇÃO DE IMAGENS DE FACE. Departamento de Engenharia Elétrica – Centro Universitário da FEI p. 2-3 (2005).

[41] HERAUD, P., CHATCHAWAL, P., WONGWATTANAKUL, M. et al. Infrared spectroscopy coupled to cloud-based data management as a tool to diagnose malaria: a pilot study in a malaria-endemic country. Malaria Journal. v. 18, Artigo. 348. Outubro de 2019. Disponível em < https://doi.org/10.1186/s12936-019-2945-1 > Acesso em 01 mar. 2021.

[42] BENETTI, C. Estudo da reparação óssea por espectroscopia ATR-FTIR após remoção de fragmento da região mandibular com laser de Er, Cr:YSGG ou broca multilaminada. São Paulo, 2014. 119 p. Tese (Doutorado em Ciências na Área de Tecnologia Nuclear – Materiais) – Instituto de Pesquisas Energéticas e Nucleares, autarquia associada à Universidade de São Paulo. Disponível em < http://pelicano.ipen.br/PosG30/TextoCompleto/Carolina%20Benetti_D.pdf > Acesso em 01 mar. 2021.

[43] WARTERWIG, S. IR and Raman Spectroscopy – Fundamental Processing. [S.1.]: Wiley-VCH, 2003.

[44] DOERN, G. et al. Clinical impact f rapid in-vitro susceptibility testing and bacterial identification. Journal of Clinical Microbiology, v. 32, p. 1757-1762, 1994.

[45] TENOVER, F et al. Development of pcr assays to detect ampicillin resistance genes in cerebrospinal-fluid samples containing haemophilus-influenzae. Jouenal of Clinical Microbiology. v. 32, p. 2729-2737, 1994.

[46] ERUKHIMOVITCH, V. et al. FTIR microscopy as a method for identification of bacterial and fungal infections. Journal of Pharmaceutical and Biomedical Analysis. v. 37, p. 1105-1108, 2005.

[47] VANEECHOUTTE, M., VANELDERE, J. The possibilities and limitations of nucleic acid amplification technology in diagnostic microbiology. Journal of Medical microbiology. v. 46, p. 188-194, 1997.

[48] FREDRIKS, D. N., RELMAN, D. A. Sequence-based identification of microbial pathogens: A reconsideration of Koch's postulates. Clinical Microbiology Review. v. 9, p. 18-33, 1996.

[49] SUREWICZ, W., MANTSCH, H. H., CHAPMAN, D. Determination of protein secondary structure by Fourier transform infrared spectroscopy: a critical assessment. Biochemistry, v. 32, p.389-394, 1993.

[50] HULEIHEL, M. et al. Spectroscopic characterization of normal primary and malignat cells transformed bt retroviruses. Applied Spectroscopy. v. 56, p. 640-645, 2002.

[51] BRUUN S. W., KOHLER A., ADT I., SOCKALINGUM G. D., MANFAIT M., MARTENS H. Correcting Attenuated Total Reflection - Fourier Transform Infrared Spectra for Water Vapor and Carbon Dioxide. Applied Spectroscopy. v. 60 p. 1029-1039, 2006.

8. APÊNDICE

```
%% Gerando Arquivos com variações de derivada e janela:

close all % Fechar janelas abertas.
clear % Limpar variáveis.
clc % Limpar a command window.

load('metarhizium.mat') % Carregar dados Metarhizium.

data = spcgroup(M1095, M324, M3391, M5626, M5749, M7486); % Agrupar cepas.
data = spcdsample(data,4); % Downsample.

spcplot(data) % Plotar o espectro médio.
title("Dados Metarhizium"); % Título do gráfico.
xlim([min(data.wnp) max(data.wnp)]); % Ajuste do eixo x.
xlabel('Número de Onda (cm^-^1)'); % Rótulo do eixo x.
ylabel('Absorbância'); % Rótulo do eixo y.

savefig('Dados Metarhizium'); % Salvar figura.

lim_inf_cut = input('Entre com o limite inferior do corte: '); % Definir
limite inferior da restrição espectral.
lim_sup_cut = input('Entre com o limite superior do corte: '); % Definir
limite superior da restrição espectral.

for dev = 1:2 % Loop para variar o grau de derivação do filtro de S-G.
    for jan =11:2:15 % Loop para variar a dimensão da janela do filtro de
S-G.
        data2 = data; % Nova variável com os dados.
        data2 = spcgolay(data2,dev,dev,jan); % Filtro de S-G
        data2 = spcut(data2,lim_inf_cut,1350); % Restrição espectral.
        data2 = spcnorm(data2);

        h(1) = figure;
        spcplot(data2) % Espectro Médio.
        title({['Dados Metarhizium ', num2str(lim_inf_cut), '-'
num2str(lim_sup_cut)]; ['Dev = ', num2str(dev), ' Jan = ', num2str(jan)]});

        h(2) = figure;
        spcscores(data2, 1, 2) % Gráfico de scatter plot.
        title({['Scatter Plot', num2str(lim_inf_cut), '-'
num2str(lim_sup_cut), '- 1vs2']; ['Dev = ', num2str(dev), ' Jan = ',
num2str(jan)]});

        h(3) = figure;
        spcload(data2, 1, 2, 3, 4, 5) % Gráfico de loading plot.
        title({['Loading Plot', num2str(lim_inf_cut), '-'
num2str(lim_sup_cut), '- 1vs2']; ['Dev = ', num2str(dev), ' Jan = ',
num2str(jan)]});

        savefig(h(1), ['SPCplot_dev_', num2str(dev), '_jan_',
num2str(jan)]); % Salvar figura.
        savefig(h(2), ['SPCscatter_dev_', num2str(dev), '_jan_',
num2str(jan)]); % Salvar figura.
        savefig(h(3), ['SPCload_dev_', num2str(dev), '_jan_',
num2str(jan)]);% Salvar figura.
```

```
        data_LDA = [data2.g data2.p]; %Consolidação de dados e índices para
LDA.
        file = string(['data_dev=', num2str(dev), '_jan=' num2str(jan),
'.mat']); % Nome do arquivo para salvar.
        save(file, "data2", "data_LDA"); % Geração do arquivo com dados.
    end
end
```