Aprendizado de máquinas e Espectroscopia Raman para identificação de fungos

***Resumo*—A identificação da espécie de um fungo causador de doenças é determinante na escolha do tratamento mais adequado. Contudo tal tarefa pode ser desafiadora dadas as semelhanças entre as espécies de mesmo gênero. Para tanto, pode ser utilizado o sequenciamento genético, entretanto, é uma técnica de elevado custo e requer equipamentos que não estão facilmente disponíveis. Este trabalho apresenta uma proposta de projeto para analisar dados do espectro Raman de alguns fungos através de técnicas de aprendizado de máquinas, de forma que, consiga-se identificar quais dados são relevantes para diferenciação e identificação de uma espécie específica dentre as analisadas, além de criar uma rede neural artificial treinada com esse propósito. A relevância desta proposição está na elaboração de um método mais barato e simples para identificar espécies de fungos com gêneros iguais.**

I. Introdução

O Reino Fungi é composto pelas mais variadas formas de seres, desde micro-organismos até cogumelos. Contudo, quando deseja-se diferenciar espécies de mesmo gênero (táxon da classificação biológica) isto torna-se uma tarefa árdua, pois tais diferenças só são perceptíveis nas estruturas celulares destes seres, como pode ser verificado na figura 1.

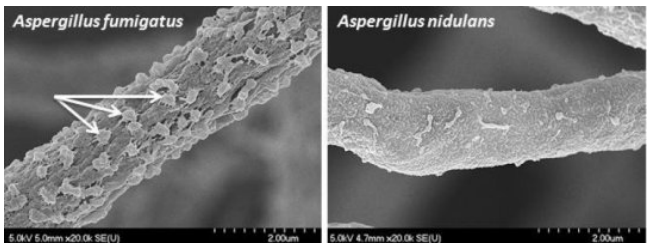

Figura 1. Diferença nas estruturas celulares das espécies fungos fumigatus e nidulans ambos do gênero Aspergillus. Fonte: [1]

Realizar o sequênciamento genético de uma amostra de células de um fungo é a forma mais assertiva de determinar a qual espécie o mesmo pertence. Mas, é um método com alto custo e requer um sequenciador genético a disposição. Por isto este trabalho propõe a utilização de dados de Espectroscopia Raman das células dos fungos para criar uma modelo de aprendizagem de máquina capaz de identificar a espécie de um fungo.

A Espectroscopia Raman é uma técnica que necessita de equipamentos simples para sua execução: microscópio ótico comum, laser de excitação, monocromador e um detector sensível (Figura 2). E, apesar de simples, ela pode gerar informação de qualidade sobre a composição da amostra que se analisa. Mas interpretar seus resultados é uma tarefa complexa, por isso, utilizar técnicas de aprendizado de máquina podem ajudar a identificar os padrões nos dados do espectro Raman.

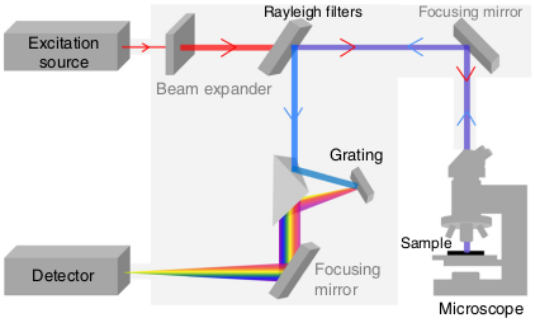

Figura 2. Esquema técnico das Espectroscopia Raman. Fonte: [2]

Nas próximas seções serão detalhadas a significância do sinal da Espectroscopia Raman para o caso de estudo, PCA(Principal Component Analysis) e Redes Neurais Artificiais e será realizada uma discussão do porquê estas técnicas podem trazer um valor significante à matéria de estudo, além de apresentar os resultados esperados ao fim desta pesquisa.

II. Aspectos teóricos

Espectroscopia Raman é uma técnica que foi inspirada em um fenômeno observado experimentalmente por Chandrasekhara Venkata Raman. Tal fenômeno consiste no espalhamento da luz em diferentes frequências ao incidir um laser sobre um material. Essas frequências luminosas se espalham devido a vibração entre as moléculas que foram excitadas pela luz incidida. A Figura 3 apresenta o esquemático do fenômeno descrito.

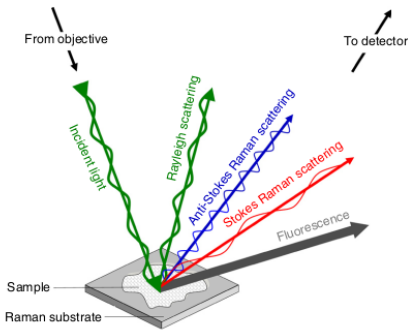

Figura 3. Esquema do efeito Raman. Fonte: [2]

O conjunto das frequências luminosas espalhadas compõe uma assinatura que reflete a composição do material analisado, cada tipo de estrutura molecular vibrará em uma intensidade, por isso, espalhará a luz em uma frequência diferente. Alguns

exemplos de assinaturas geradas podem ser verificados na figura 4.

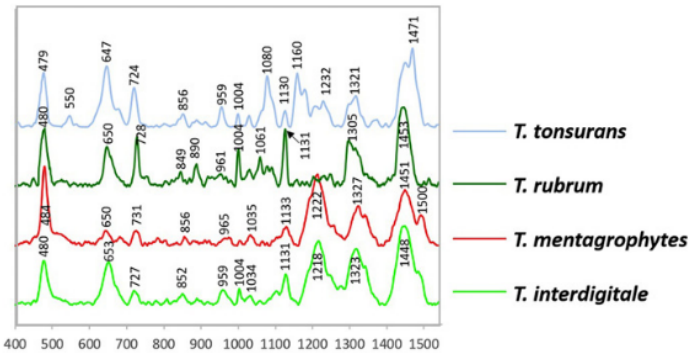

Figura 4. Exemplo de assinaturas obtidas através da Espectroscopia Raman. Cada cor reflete uma medição para espécies diferentes de fungos do gênero Trichophyton. Fonte: [3]

Quando se realiza uma medição do espectro Raman, é preparada uma solução aquosa adicionando uma pequena quantidade do material que se deseja analisar e se adiciona um substrato à amostra (geralmente metais nobres) que por sua vez servirão como amplificadores da intensidade luminosa espalhada. A escolha do substrato é parte crucial na obtenção dos dados do espectro Raman conforme [4], pois ruídos podem ser gerados pelo substrato escolhido distorcendo os resultados.

Uma vez obtidos os dados referentes ao espectro Raman é comum utilizar ferramental matemático para extrair da assinatura obtida ruídos advindos da fluorescência da composição material da amostra. Tal aspecto pode, por sua vez, mascarar a informação relevante do sinal que em geral é caracterizada por picos.

O PCA é uma das ferramentas utilizadas para eliminar os ruídos do espectro Raman. Este método matemático é capaz de identificar em um espaço amostral de n dimensões quais destas carregam informação realmente relevantes para análise. Contudo, a eficiência deste método está no agrupamento de informação em poucas dimensões. Caso exista grande variação entre várias das componentes das amostras, ao considerar apenas as de maior variância, muita informação relevante pode ser perdida comprometendo a análise que se deseja realizar [5]. Por isso plotar as relações entre as componentes de maior variância pode ter grande valor analítico, pois se a significância não estiver sendo perdida padrões de agrupamento poderão ser visualizados. A Figura 5 demonstra o surgimento destes padrões de agrupamento na análise de amostras de pólen realizadas em [4].

Entretanto, para materias biológicos, distinguir tais agrupamentos pode ser desafiador. Segundo [4] a utilizações de redes neurais artificias pode trazer grandes vantagens para este tipo de análise, tendo resultados melhores do que com outras estratégias como SVM(Support Vector Machines).

As Redes Neurais Artificias são uma técnica que procura simular o funcionamento do cérebro humano. São formadas por conjuntos de neurônios que são estruturas que avaliam através de uma função de ativação as entradas resultando uma saída. Estas saídas podem ser submetidas a outros

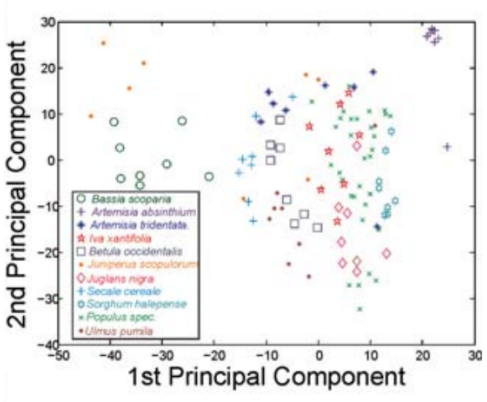

Figura 5. Análise das principais componentes do espectro Raman de amostras de pólen. Fonte: [4]

neurônios sucessivamente de forma que a mensagem possa ser trafegada por várias camadas. A disposições dos neurônios define a arquitetura da rede neural, de forma que, diferentes configurações podem trazer diversos benefícios. Uma vez que a mensagem tenha trafegado por todas as camadas o saída final deve corresponder a resposta para a questão analisada.

Para cada entrada de um neurônio é atribuído um peso, que deve ser calibrado de forma a obter uma saida esperada. Na aprendizagem supervisionada, esse processo de calibração dos pesos é comumente chamado de treinamento da rede neural. Em que é informado a rede para determinadas entradas qual o resultado esperado, deste modo, a rede consegue determinar quais os melhores pesos a utilizar em seus neurônios. Assim, em uma execução posterior para um conjunto de entradas a rede será capaz de inferir um resultado. A determinação dos pesos dá-se por um processo iterativo de propagação de erros denominado back propagation, em que, cada iteração define um delta que deve ser acrescido ao peso para minimizar o erro em relação ao valor esperado na saída. Por isso, é comum que se utilize das amostras disponíveis 80% para realizar o treinamento e 20% para validar se o treinamento realizado gerou resultavo efetivo.

Dentre as principais aplicações de uma Rede Neural Artificial pode-se citar: aproximação de funções, previsão de séries temporais, classificações e reconhecimento de padrões.

III. Materiais e métodos

Para a realização deste trabalho, dados do espectro Raman serão fornecido em arquivos de formato txt. Tais dados correspodem as amostras de variadas espécies de fungos que serão obtidas e preparadas adequadamente para a Espectropia Raman pelos pesquisadores que fazem parte desse grupo de pesquisa. No arquivo txt estarão contidas duas colunas, a primeira com a médida em cm^{-1} referente a frequência luminosa e a segunda com a medida em unidade arbitrária da intensidade luminosa. A quantidade de linhas geradas reflete a calibragem definida no aparelho de microscopia que realiza a coleta do dados do espectro.

Para cada espécie serão fornecidos 10 arquivos com informações de diferentes medições do espectro Raman para um mesmo fungo. Espera-se analisar 2 gêneros de fungos

compostos por duas espécies diferentes cada um. O primeiro gênero a ser analisado é o Candida, de forma a validar o conceito formulado para identificação de fungos já que, neste caso, existe uma base de conhecimento mais difundido na literatura sobre os valores e formas de interpretação do espectro Raman. O segundo gênero a ser analisado é o Fonsecaea, em que, deseja-se diferenciar duas as espécies pedrosoi e pugnacius (Figura 6).

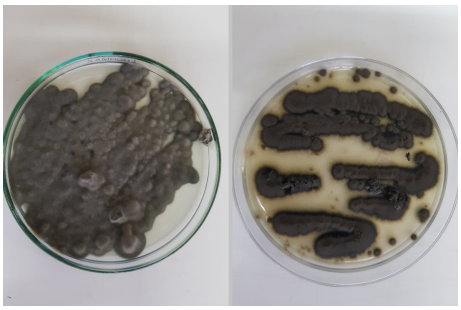

Figura 6. Fungo Fonsecaea pedrosoi e Fonsecaea pugnacius respectivamente. Fonte: Pesquisadora Juliana Thaler

Para cada gênero de fungo as amostras serão submetidas a uma correção de linha base e a uma normalização e então submetidas ao PCA para que apenas as informações mais relevantes do espectro sirvam como dados de entrada para a rede neural artificial a ser criada e treinada. A quantidade de componentes escolhidas será feita mediante análise da variância das mesmas, bem como, das plotagens de comparação entre as componentes. Para o treinamento da rede neural artificial serão utilizados 80% dos dados disponíveis e os outros 20% serão utilizados na validação da efetividade da mesma.

IV. Discussão e resultados esperados

A escolha de aplicar a rede neural artificial apenas as componentes de maior relevância do espectro Raman é feita com o propósito de eliminar ruídos que possam estar contidos nas amostras avaliadas. [2] cita o PCA como uma poderosa ferramenta na eliminação de ruídos e reconstrução do espectro Raman apenas com as principais componentes. [6] conclui em seu experimento que quando não é possível visualizar agrupamentos de forma clara apenas utilizando-se de PCA, treinar as redes neurais artificiais com os resultados do PCA pode aprimorar a classificação das amostras.

Para que o PCA exerça sua função da forma esperada, são necessários aplicar dois procedimentos as amostras: correção de linha base e normalização. A correção de linha base é necessária, por conta do mascaramento do sinal relevante que o espectro Raman sofre em virtude da fluorescência [6], fenômeno o qual a Espectroscopia Raman está sujeita. A normalização, por sua vez, deve ocorrer porque as amostras podem ter sido obtidas sobre diferentes calibrações dos aparelhos de microscopia, gerando diferentes intervalos nas frequências amostradas. Na Figura 7 está demonstrado um exemplo gráfico da correção de linha base e normalização.

Quanto a arquitetura da rede neural artificial, diferentes configurações podem ser experimentadas de forma a identificar

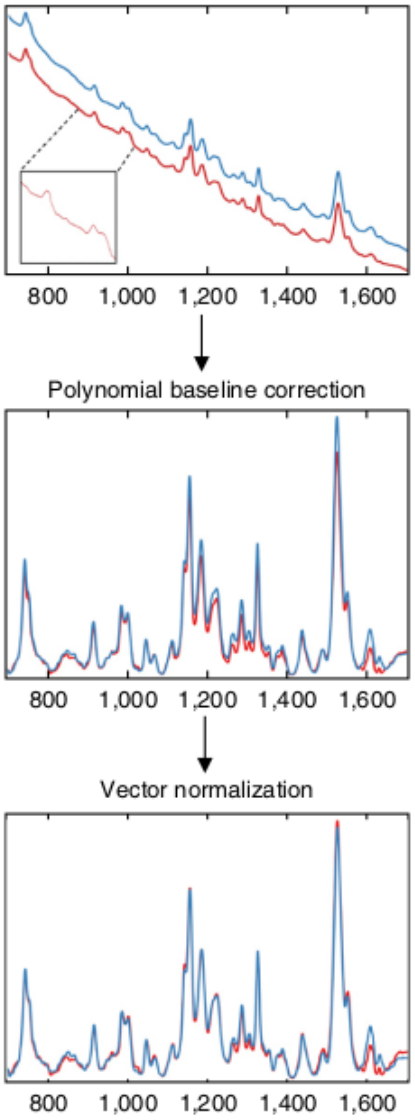

Figura 7. Gráfico da transformação da assinatura do espectro Raman após as operações de correção de linha base e normalização. Fonte: [2]

qual atinge o melhor resultado consumindo o menor quantidade de recursos computacionais.

Ao fim dessa pesquisa espera-se obter duas redes neurais artificiais treinadas capazes de a partir do dados do espectro Raman de um fungo identificar a qual espécie ele pertence. Contudo, elas não serão genéricas, elas serão aptas a distinguir espécies entre gêneros iguais. Então a primeira rede neural artificial será capaz de distiguir espécies de fungos do gênero Candida e a segunda distinguir espécies do gênero Fonsecaea.

V. Conclusão

A relevância deste trabalho dar-se-á pela possibilidade criar um método de identificação de espécies de fungos mais simples do que o sequenciamento genético. Impactando a velocidade e diminuindo os custos na obtenção de diagnósticos por infecções/alergias fúngicas.

RECONHECIMENTO

Ao Professor Hugo Vieria Neto, Ph.D. pelas aulas ministradas na displina de Metodologia Científica do programa de pós graduação da Universidade Tecnológica Federal do Paraná.

REFERÊNCIAS

[1] M. J. Lee, H. Liu, B. M. Barker, B. D. Snarr, F. N. Gravelat, Q. Al Abdallah, C. Gavino, S. R. Baistrocchi, H. Ostapska, T. Xiao, B. Ralph, N. V. Solis, M. Lehoux, S. D. Baptista, A. Thammahong, R. P. Cerone, S. G. Kaminskyj, M. C. Guiot, J. P. Latgé, T. Fontaine, D. C. Vinh, S. G. Filler, and D. C. Sheppard, "The Fungal Exopolysaccharide Galactosaminogalactan Mediates Virulence by Enhancing Resistance to Neutrophil Extracellular Traps," *PLoS Pathogens*, vol. 11, no. 10, pp. 1–22, 2015.

[2] H. J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, K. Esmonde-White, N. J. Fullwood, B. Gardner, P. L. Martin-Hirsch, M. J. Walsh, M. R. McAinsh, N. Stone, and F. L. Martin, "Using Raman spectroscopy to characterize biological materials," *Nature Protocols*, 2016.

[3] D. Pankin, I. Kolesnikov, A. Vasileva, A. Pilip, V. Zigel, and A. Manshina, "Spectrochimica Acta Part A : Molecular and Biomolecular Spectroscopy Raman fi ngerprints for unambiguous identi fi cation of organotin compounds," vol. 204, pp. 158–163, 2018.

[4] S. Seifert, V. Merk, and J. Kneipp, "Identification of aqueous pollen extracts using surface enhanced Raman scattering (SERS) and pattern recognition methods," *Journal of Biophotonics*, vol. 9, no. 1-2, pp. 181–189, 2016.

[5] R. G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, J. M. Roger, B. Walczak, and R. Tauler, "Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools," *Analytical and Bioanalytical Chemistry*, vol. 409, no. 25, pp. 5891–5899, 2017.

[6] R. E. De Góes, L. V. M. Fabris, M. Muller, and J. L. Fabris, "Light-assisted detection of methanol in contaminated spirits," *Journal of Lightwave Technology*, vol. 34, no. 19, pp. 4499–4505, 2016.