2024 CMBEC46
Toronto, ON
CMBES/SCGB May 28-30, 2024

Raman Microspectrometry and Machine Learning: Accelerated Bacterial Identification

J. Levine[1,2], N.K. Kunchur[1,2], K. Allison[2,3], J. Overhage[2,3], E. Cassol[2,3] and L.B. Mostaco-Guidolin[1,2]

[1]Carleton University/Systems and Computer Engineering, Ottawa, Canada
[2]Tissue Engineering and Applied Materials (TEAM) Hub, Ottawa. Canada
[3]Carleton University/Health Sciences, Ottawa, Canada

Abstract— **The combination of Raman microspectrometry (RM) and Machine Learning (ML) enables the identification and categorization of bacterial species. In this work, supervised ML models are used to differentiate amongst complex biochemical Raman signatures unique to distinct bacterial species.**

Keywords— **Machine Learning, Raman spectroscopy, Bacterial Colonies**

I. INTRODUCTION

Bacterial infections are among the leading causes of death globally. Diagnoses are slow, and infection is costly to treat [1]. When awaiting diagnosis, general, broad antibiotics are prescribed, with the CDC indicating nearly 30% of patients being treated inappropriately [2]. Bacterial species differ in their unique molecular composition, which can rapidly be identified non-invasively using RM. Through the combination of RM with ML, we can provide an efficient label-free mode to support bacterial identification. Supervised ML models can recognize minute molecular differences amongst the RM spectra of bacterial species. Using Principal Component Analysis (PCA), decision trees (DT) and Random Forests (RF) we can accurately classify the biomolecular chemistry of six common bacterial species.

II. METHODOLOGY

A: Sample Preparation

The spectral signatures of six bacteria prevalent in clinical infections are acquired: *Staphylococcus aureus*, *Corynebacterium striatum*, *Pseudomonas aeruginosa*, *Escherichia coli*, *Enterococcus faecalis*, *Stenotrophomonas maltophilia*. The bacterial sample preparation follows the protocol described in [3] except for Lennox-Broth agar used as a substitute for Mueller-Hinton agar. A total of four replicates are to be done per species, totaling 400 spectra per species.

B: Raman Signal Acquisition

On average, 100-200 spectra per bacterial species have been imaged using the WiTec Alpha 300R Raman Microscope. Large area scans of 10µm x 10µm are defined within the sample, automatically collecting a total 100 spectra per scan. Each pixel within the defined area is representative of a spectrum unique to that location. With RM spectra capturing the biochemical signatures unique to each bacterial species. PCA, DTs and RFs can objectively classify the variation in minute spectral differences.

C: Data preparation for Machine Learning

Spectra are first normalized and cleaned, removing outliers within the dataset defined using IQR method. A train, test and validation split of 0.7, 0.15, 0.15 is applied. The models' performances are evaluated based on their performance against the test set.

III. RESULTS

Preliminary analysis shows clear distinctions between the species, as seen in Figure 1 (A). DTs and RFs demonstrate an accuracy of 94.4% and 98.5%, respectively in bacterial species classification.

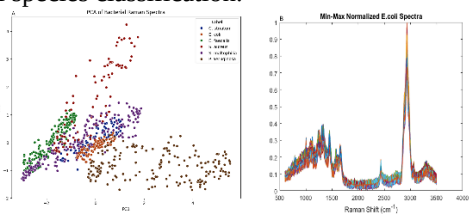

Figure 1: (A) PCA of all the spectra for each species; (B) Illustrative Raman spectra collection of *E. coli*

CONCLUSIONS

This work demonstrates promise in combining novel RM imaging systems with supervised ML techniques (DTs, RFs) in the classification of common bacterial species. ML algorithms can, with high levels of accuracy, identify individual species based on their Raman signature. This approach offers a faster and more precise diagnosis of bacterial infection, without the requirement of extensive sample preparation.

ACKNOWLEDGMENT

This work is supported by NSERC-DG (RGPIN-2021-04185).

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1. C.-S. Ho et al., Nature Communications, 10.1 (2019): 4927
2. Tamma P.D., et al., Jama, *315*(17), 1839-1841.
3. Wang. K. Chen, et al., *Applied and environmental microbiology*, *86*(20), e00924-20.