Data analysis of microorganism with Raman

Li Y[1], Wang Y[1,2], Cui DY[1,2], Zhu L[3*] & Zhang CL[1,2*]

ABSTRACT:

For the people who study the components of microbial, especially biology researchers, it is essential of examining the possible contamination and investigating the relationship between microorganisms and the environment. However, the problem of how to identify unknown microbial in community is still puzzled some biology scientists.

In this study, based on R program language, a new analysis work-flow with machine learning, which applied to the classification of archaeas with Raman spectral images, was designed for identifying the unknown organisms, may provide a novel insight in microbial community.

[1] Department of Ocean Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China.

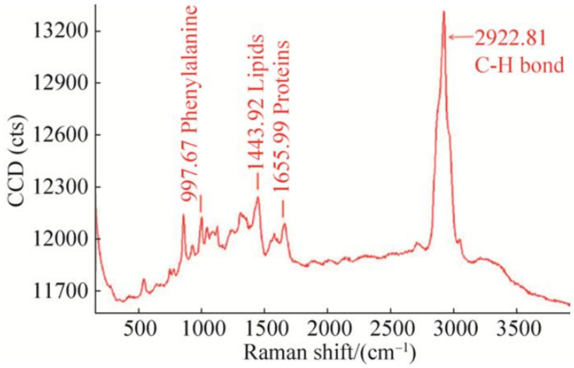
[2] Shenzhen Key Laboratory of Marine Archaea Geo-Omics, Department of Ocean Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China.

[3] School of Environmental Science & Engineering, Southern University of Science and Technology, Shenzhen, China.

CONTENTS

1. Introduction

Raman spectrum, which can provide information regarding distribution of chemical compounds of the considered biological entity, plays a crucial role in specials analysis. Since from the spectral images, each spectral wavelength can be used to reconstruct the distribution of a given compound, and every organisms have different groups of compounds, which representing a unique pattern of this organism. Usually, we can match the spectral wavelength on the opening database and identify what's the unknown organism is. However, owing to the complexity of spectral wavelength and the corresponding bond, it is still difficult for researchers to identify the detailed composition of this organisms[1-4]. Thus, it is necessary to simplify the matching steps, such as using a important or special wave peak to replace the full wave peaks.



Using a dataset of three types of 91 living archaeas, we demonstrate that the typically peaks can used to accurately classify cell types. This method is compared with the conventional Raman spectral analysis. We also propose to combine the information from whole spectral analyses and selected spatial features and show that this yields higher classification accuracy, which more than 90%. This method provides the basis for a novel and systematic analysis of cell-type investigation using Raman spectral imaging, which may benefit several studies and biomedical applications.

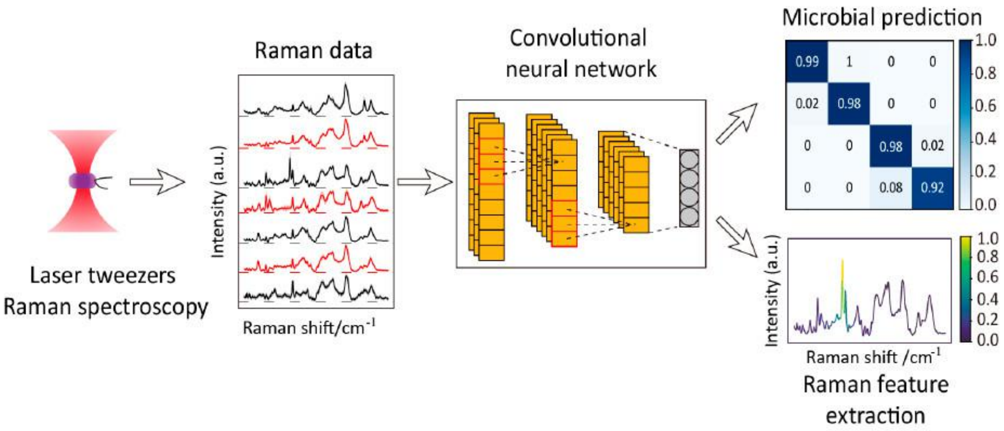


Fig.1 Workflow of data analysis for archaeas using Raman spectroscopy[5].

2. Material and methods

2.1. Microbial samples and Raman Spectral Acquisition

The microbial samples were directly collected from the thermal spring of YunNan, TengChong, which are provided by WangYi and Cui DongYu. The Raman Spectral data were provided by WangYi, including H1 x 32; H7 x 30; SCM1 x 29.

2.2. Raman Spectral Data Procressing and Plotting

In this study, we use R 3.6.2 in windows system with some packages in the processing procedure(Table 1). Before formal data processing of Raman

Spectral Data, normalization was performed by machine of Raman Spectral. Thus, we can use this data in the data analysis directly.

During the data procedure: At first, the necessary packages were loaded, and the work dictionary was set as follows, which stores the 91 samples recorded in text file. Then, we get all the files name by functions “list.files()”, and we read those files in a “for cycle”. Once the loading of 91 samples files was complete, “NA” of wave-number was removed and the all samples were transformed in type of “tibble”, in which we calculate “mean”, “standard deviation”, “error” as new columns for following data visualization. After that, we draw spectra plot with R library “ggplot2”, and we find some peaks between different kinds of organism were different. Next, we draw “box-plot” and “PCA-plot” to see the difference between the group.

To verify the guess that the chemical bonds between different organisms were different, a statistical test was done by one-way Annova. We conduct three one-way annova analysis in H1~H7, H7~SCM1, H1 ~ SCM1 respectively, and all of comparable groups show significance of statistical.

2.3. Construction of predict model with machine learning

After loading library “caret”, we split 91 samples into train and test data sets, and the percentage of train and test was 4:1. The “SVM” model was set 10-fold cross-validation and repeats 5 times, the output was probability of each test samples. To describe the model's ability of predictions, we set that probability more than 0.5 were right classified in the analysis of results.

After building and validation of model with "SVM', we compare the predict and true label of samples, and R library "Pheatmap" was used to show the results of prediction.

Table 1 The libraries were used in the workflow of Raman Spectral data analysis in R

Packages	Main Functions
dplyr	Data clean and filter
tidyr	Data rearrange
ggplot2	Data visualization
ggbiplot	PCA and Plot
caret	Machine Learning
pheatmap	Cluster and Plot

3. Results and discussion

3.1. Archaea has characteristic chemical bonds

In the distribution of three types of archaeas, the intensity of wave numbers between 1200-1600, 2800-3000, and 3200-3600 have a clear difference(Fig.2).

There’s almost no difference in chemical bonds of “H1”(Fig.3), and with principal component analysis, “H7”, “SCM1”, “H1” can be clearly separated(Fig.4).

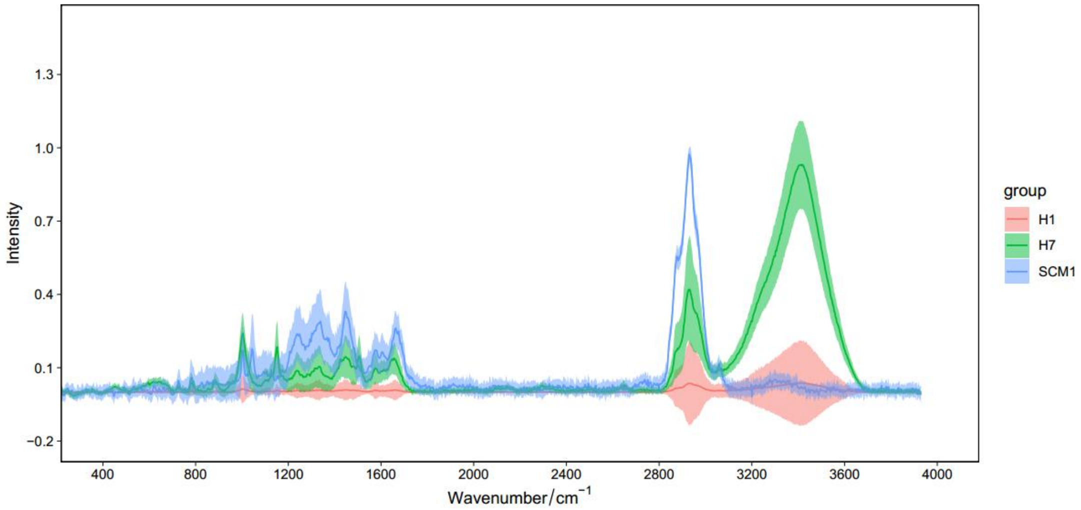


Fig.2 The distribution of three types of archaeas.

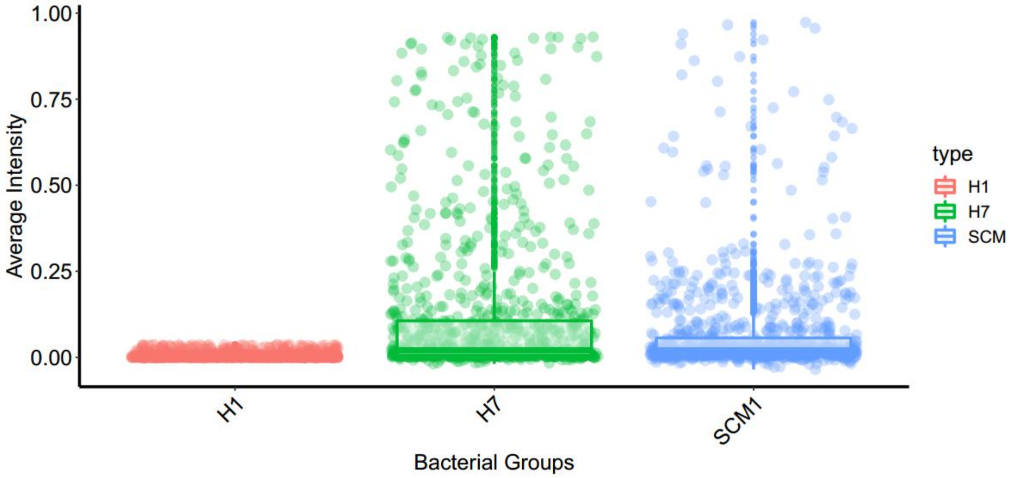


Fig.3 The distribution of three types of archaeas between groups.

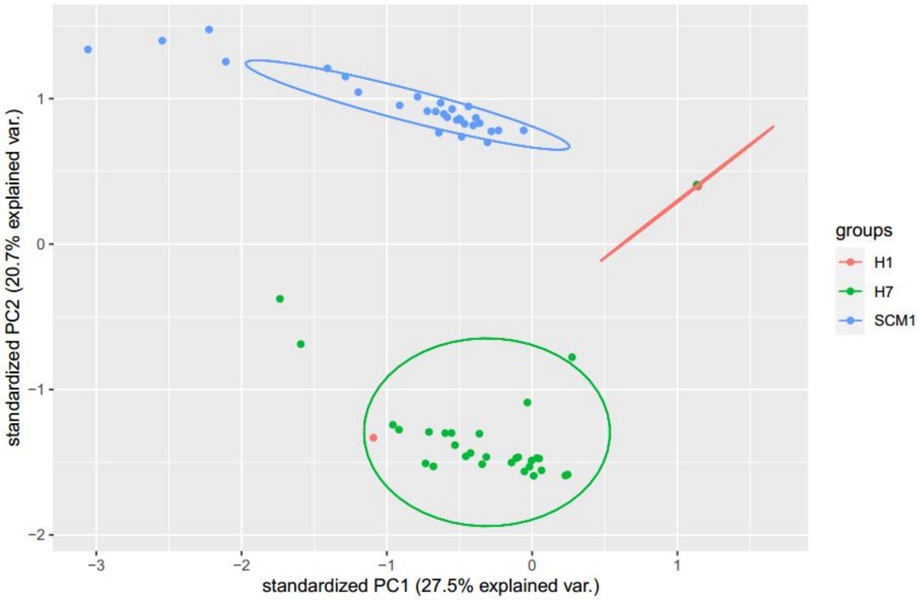


Fig.4 The PCA distribution of three types of archaeas between groups.

3.2. Chemical bonds can be used to distinguish types of archaea

In the model of machine learning using SVM, except one samples "H7" was classified to "H1" wrongly, mostly samples were gain high possibility of classification, which means some important chemical bonds may be used to classified archaea(Fig.5).

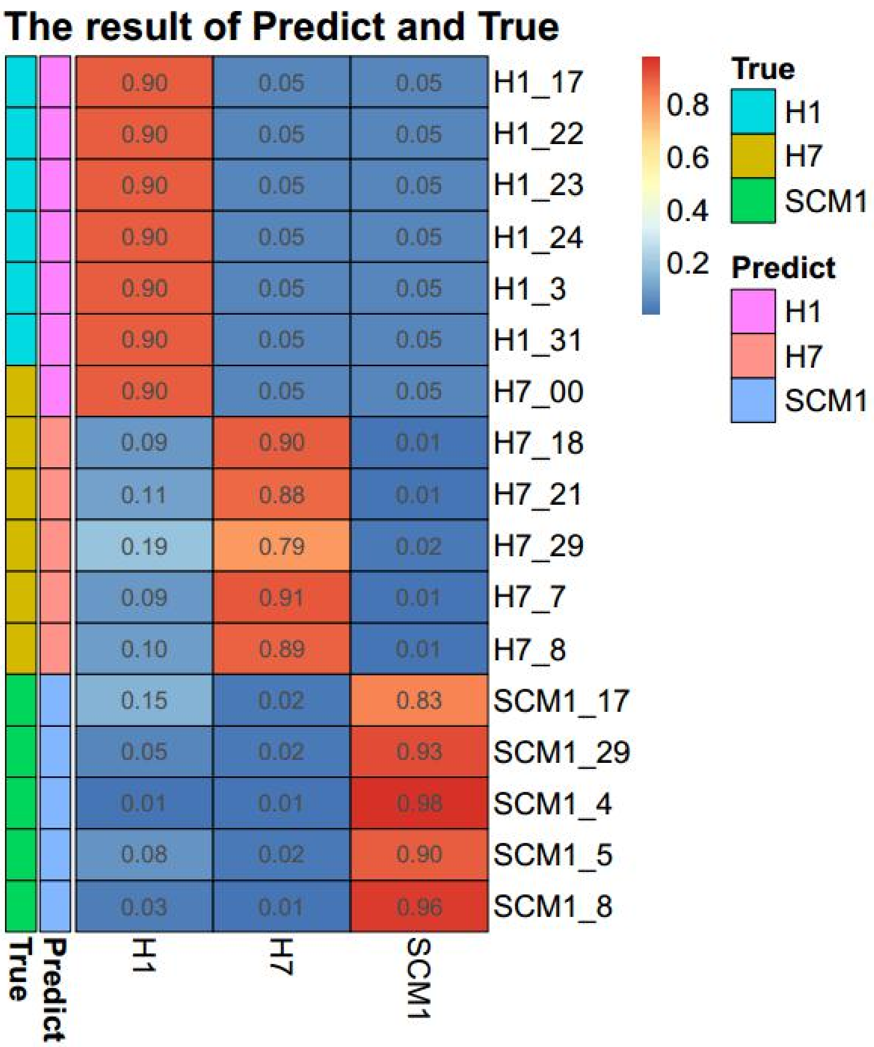


Fig.5 The Pheatmap of possibility of predict and true results, archaea H7_00 was classified to H1 wrongly.

4. Conclusion

In the research of 91 different types of archaeas samples collected from hot spring, we obtained 1024 chemical bonds of those archaeas with Raman Spectral. With R language, we conduct a series of analysis, including data rearrange, data clean, data filter, data visualization, statistical test, machine learning and so on with the spectral data.

We find that the difference spectral peaks of chemical bonds between different types of organisms has significance, which means chemical bonds may represent some difference of organisms, including environment, coding protein, genome and so on. What's more, we find that there was little difference between "H1" groups, but "H7" and "SCM1" have opposite phenomenon,

Then we build a machine learning model with "SVM", which shows high accuracy more than 90%, the only one samples "H7" wrong classified to "H1" may be attributed to the carelessness of the experimenter.

In the future planning of analysis, we want to make it clear that mechanism of production of the difference of peaks, and whether can we use important chemical bonds to the classification of all archaeas or not, which may need the access of large open databases of archaeas[6-7].

Reference

[1] Collins, C. H.; Lyne, P. M.; Grange, J. M.; Falkinham, J. O., III Collins and Lyne’ s *Microbiological Methods*, 8th ed.; Arnold: London, **2004**; Vol. 456

[2] Madigan MMartinko JParker J. *Brock biology of micro-organisms*. Fourteenth edition.

[3] Liu Cong, Xie Wei, He Lin, Zhang Chuanlun. Advances in the application of Raman microspectroscopy in microbe research. *Acta Microbiologica Sinica*, **2020**, 60(6): 1051-1062.

[4] Germond A, Ichimura T, Chiu LD, Fujita K, Watanabe TM, Fujita H. Cell type discrimination based on image features of molecular component distribution. *Sci Rep*. **2018 Aug 6**;8(1):11726.

[5] Lu W, Chen X, Wang L, Li H, Fu YV. Combination of an Artificial Intelligence Approach and Laser Tweezers Raman Spectroscopy for Microbial Identification. *Anal Chem*. **2020 May 5**;92(9):6288-6296.

[6] Germond A, Ichimura T, Horinouchi T, Fujita H, Furusawa C, Watanabe TM. Raman spectral signature reflects transcriptomic features of antibiotic resistance in Escherichia coli. *Commun Biol*. **2018 Jul 2**;1:85.

[7] Zhu Lei, etc. Computing and Programming for Environmental Research. *ESE5023*. **2020**; https://zhu-group.github.io/ese5023/Schedule.html