Title	Development of a characterization method for food-related bacteria using Raman spectroscopy and machine learning
Author(s)	山本, 貴志
Degree Grantor	北海道大学
Degree Name	博士(農学)
Dissertation Number	甲第14810号
Issue Date	2022-03-24
DOI	https://doi.org/10.14943/doctoral.k14810
Doc URL	https://hdl.handle.net/2115/88852
Type	doctoral thesis
File Information	Yamamoto_Takashi.pdf

Development of a characterization method for food-related bacteria using Raman spectroscopy and machine learning

ラマン分光法と機械学習による
食品関連細菌の特性評価手法の開発

Laboratory of Agricultural & Food Process Engineering
Graduate School of Agricultural Science
Hokkaido University

Takashi Yamamoto

Acknowledgements

The present thesis is based on the studies carried out from 2019 to 2021 at Laboratory of Agricultural & Food Process Engineering, Graduate School of Agricultural Science, Hokkaido University; Food Development Laboratories, Nippon Shinyaku Co. Ltd.

I would like to express my sincere gratitude to my supervisor, Dr. Shigenobu Koseki for providing me this precious study opportunity as a Ph.D student in his laboratory. I especially would like to express my deepest appreciation, Dr. Kento Koyama for his elaborated guidance, considerable encouragement and invaluable discussion that make my study of great achievement and my study life unforgettable.

A special gratitude I give to Dr. James Nicholas TAYLOR for his continuous support and thoughtful guidance throughout Raman spectral analysis without whom this study would not have been possible. I am also grateful to all members of Laboratory of Agricultural & Food Process Engineering, Graduate School of Agricultural Science, Hokkaido University; Food Development Laboratories, Nippon Shinyaku Co. Ltd.

Last but not least, I was encouraged by the support of my family throughout my study. Words cannot express the gratitude I feel for the kindness of my wife and son.

Table of Contents

Table of Contents

List of Tables

Page

Table

List of Figures

Page

Figure

Chapter 1

General Introduction

1.1. Foodborne illness and food spoilage caused by food related microorganisms

The problem of microorganisms is inevitable when considering food safety and food quality preservation. While most of the foodborne illness are caused by viruses, hospitalizations and deaths were associated with food poisoning are due to bacteria. The bacteria causing foodborne illness are the leading cause of serious and fatal foodborne illnesses. More than 90 percent of foodborne illnesses in the world are caused by species of *Staphylococcus, Salmonella*, *Clostridium, Campylobacter*, *Listeria*, *Vibrio*, *Bacillus,* and *Escherichia coli* (Fung et al., 2018). When pathogens are transferred and attached on fresh food, they might cause food poisoning in humans, if the environmental conditions are such that the bacteria can multiply and produce toxins in the food (Black et al., 2018). In contrast, food wastage is mainly associated with spoilage, and microbial growth is one of the most common causes (Gram et al., 2002). The issue of food loss is gaining international attention and efforts to reduce it are being stepped up, as the SDGs include the reduction of food loss and waste. The Food and Agriculture Organization of the United Nations's Food Loss Index estimates that globally about 14% of all food excluding retail and post-harvest (Fao, 2019). Therefore, microbial control in food is an important matter to prevent foodborne illness and food spoilage.

1.2. Microbial control for food preservation

There are various methods for microbial control technology to maintain food safety and quality. For example, inhibition of microbial growth is achieved by controlling storage temperature and using additives such as acids and preservatives. Microbial inactivation includes heat treatment, irradiation, and pressurizing (Gould, 1996). In practice, a combination of several control methods may be used, which is commonly referred to as hurdle technology (Leistner, 2000).

In general, different types of bacteria have different resistance to stresses such as heat and low temperature, and as a result, their growth behavior in food is also different. In addition, it is difficult to estimate the risk of food spoilage and food poisoning caused by bacteria based on experience alone, because foods are made from a wide variety of raw materials and ingredients, and their manufacturing processes and storage conditions vary. Therefore, numerous studies have been conducted to predict the behavior of microorganisms in food under complex environmental conditions with many factors. Numerous studies are used to establish statistical models and develop databases such as ComBase (www.combase.cc). ComBase is a database that can retrieve the growth and inactivation of bacteria, which mainly cause foodborne illness (Baranyi, 2006; Jozsef Baranyi and Tamplin, 2004).

1.3. Identification of microorganisms

In order to utilize databases and predictive models that are useful for risk assessment and determination of effective control methods against food poisoning and spoilage bacteria, it is necessary to understand what kinds of bacteria exist in food. Most of the microbial identification is based on biochemical or genetic characteristics. Identification based on biochemical properties is a common method that has been used for about 100 years, using reactions to various substrates such as fermentation tests, assimilation tests, enzyme reaction tests, and inhibition tests. Identification based on genetic characteristics classifies microorganisms by comparing sequences obtained by sequencing with known sequences of each strain (**幸恵**, 2015). Particularly in bacteria, 16S rRNA is a region that is conserved across bacterial species and is stable across bacterial species, making 16S rRNA analysis the gold standard for bacterial identification (Schleifer, 2009).

On the other hand, identification of microorganisms by biochemical properties takes 5 to 7 days for determination. Identification by genetic analysis provides results in one to two days, but requires complicated work such as extraction of DNA and preparation of reagents for identification such as PCR primers (Roda et al., 2012). That is why rapid and simple microbial identification method compared with conventional biochemical and genetic methods, which requires only a few minutes of measurement time and does not require any pretreatment or reagents, is highly attractive.

1.4. Microbial identification by Raman spectroscopy

In addition to biochemical and genetic methods of microbial identification, there are other methods such as Immunoassays, MALDI/TOF mass spectrometry and vibrational spectroscopy. Among them, vibrational spectroscopy (Raman and infrared) are expected to be high discriminatory power (at the species, subspecies and strain level) methods (Roda et al., 2012). Raman spectroscopy is a means of analyzing chemical and biological components at the molecular level without reagents and resorting to complex sample preparation methods (Moreira et al., 2008). Therefore, Raman spectra have been reported to be useful in identifying microorganisms (de Biasio et al., 2013; Strola et al., 2014; Yan et al., 2021; Yilmaz et al., 2015). Not only that, but it has been reported that Raman spectra could be used to identify antimicrobial resistance in *E. coli* and *Staphylococcus* (Germond et al., 2018; Ho et al., 2019). However, there are no reports on the use of Raman spectroscopy for the characterization of microorganisms useful for the control of microorganisms in food, such as resistance to food additives and growth behavior in food.

1.5. Objective of this study

The purpose of this study is to develop a method for rapid evaluation of bacterial characterization of properties related to bacterial growth behavior such as stress tolerance by using Raman spectroscopy and chemometric analysis. In Chapter 2, I investigated a rapid evaluation method by using Raman spactra for stress tolerance to additives such as sodium acetate and glycine, which affect the growth of bacteria. In Chapter 3, I studied the development of a machine learning model to predict the growth behavior of unknown bacteria using Raman spectral information obtained from bacterial cells. The developed method will enable to rapidly obtain bacterial growth/no growth characteristic under certain environmental condition, which is useful for efficiently determining the production and storage conditions of foods.

Chapter 2

Classification of Food Spoilage Bacterial Species and their Food Additives Tolerance Using Chemometrics Analysis and Raman Spectroscopy

2.1. Introduction

As an alternative to genotypic methods, Raman spectroscopy has attracted attention in the classification of bacterial species. Raman spectroscopy is rapid, non-destructive, and relatively inexpensive compared to genotypic methods (Jaafreh et al., 2019). It has been reported that it is possible to classify bacterial species from Raman spectra (Huayhongthong et al., 2019) and that Raman spectroscopy enables classification not only by the species of microorganisms but also by the characteristics of the strains of microorganisms such as antimicrobial resistance (Germond et al., 2018). However, there are no reports on whether Raman spectroscopy techniques are useful not only for the identification of spoilage microorganisms in foods but also for the identification of stress tolerance to food additives. In reducing the risk of food spoilage from many types of microorganisms, it will be practical if microorganisms with different genus levels can be classified according to minimum inhibitory concentrations of food additives.

The purpose of this study was to investigate the possible use of Raman spectroscopy as a tool for classifying not only bacterial species contributing to food spoilage, but also their tolerances to food additives for extension of food shelf life. First, Raman spectra of 6 types of bacteria *Bacillus subtilis, Escherichia coli, Leuconostoc mesenteroides, Staphylococcus epidermidis, Staphylococcus saprophyticus*, and *Pseudomonas fluorescens* were investigated. Second, growth inhibition by adding NaCl and food additives (sodium acetate and glycine) was evaluated by the optical density method. Sodium acetate and glycine are widely used in Japan as materials to extend the shelf life of food and have different mechanisms for suppressing the growth of microorganisms (Inatsu et al., 2017). Third, I developed a machine learning model for classifying bacterial species and stress resistance from the Raman spectra of the 6 bacterial species. Bacterial species and stress tolerance classification with Raman spectra is easier and faster than current genotypic methods, and provides useful information for examining the risk-reducing effects associated with food spoilage.

2.2. Materials and Methods

2.2.1. Bacterial strains and sample preparation

As representative food spoilage and food related bacteria (Møretrø and Langsrud, 2017), I investigated six bacterial strains (*Bacillus subtilis* NBRC 13719, *Escherichia coli* NBRC 3301, *Leuconostoc mesenteroides* NBRC 100496, *Pseudomonas fluorescens* NBRC 14160, *Staphylococcus epidermidis* NBRC 100911, and *Staphylococcus saprophyticus* NBRC 102446), which were obtained from the National Institute of Technology and Evaluation Biological Resource Center (Tokyo, Japan). The strains were stored at -80℃ in tryptic soy broth (TSB, Difco) supplemented with 25% (w/w) glycerol. A platinum loop was used to transfer the frozen bacterial cultures, which were independently inoculated into 10 mL of TSB. The bacteria in TSB were incubated at 30℃ for 20 h with shaking at 200 rpm.

2.2.2. Raman spectra measurements

Raman spectra of bacterial cultures were collected using a laser Raman microscope (RAMAN touch, Nanophoton, Osaka, Japan). The excitation source was a 532 nm laser operated at 5 mW. A 50x/0.80 objective lens (Nikon TU Plan Fluor) with a laser spot size of approximately 400 nm was used to focus the excitation light onto the sample. Raman spectra were acquired with a 300 lines/mm grating for 10 s. The ten spectra were averaged to obtain the mean spectrum. The Raman shift was calibrated using silicon (520 cm^{-1}) before acquiring the spectra. The cultured cells were collected by centrifugation (12,000 g, 5 min.) at 4°C. The resulting pellet was washed twice with pure water and re-suspended in 1.0 mL of pure water. 1 µL aliquots of suspension were dropped onto a stainless-steel piece (SUS430, HIKARI, Osaka, Japan), which was air dried before starting the Raman measurement. Twenty single-cell Raman spectra were obtained for each species using a range of Raman shifts from 125 to 4690 cm^{-1} with increments of 5.0 cm^{-1}.

2.2.3. Bacterial growth experiment

I used NaCl, sodium acetate, and glycine as food additives. These food additives are known to be effective in suppressing the growth of a wide variety of bacteria (Hishinuma et al., 1969; Houtsma et al., 1993; Nanasombat and Chooprang, 2009). Among several concentrations of additives, the lowest concentration where no bacterial growth is observed is defined as the MIC, which indicates the resistance tolerance of the microorganism to concentration of the additive. MIC was determined for each microorganism by optical density measurements (Knight and McKellar, 2007). Four concentrations of sodium chloride {428 (2.5% w/v), 856 (5.0% w/v) ,1283 (7.5% w/v) and 1711 (10 % w/v) mM)}, sodium acetate (62.5, 125, 250 and 500 mM) and glycine (62.5, 125, 250 and 500 mM) and were added to the TSB to check the MICs of six food spoilage bacteria. The pH of TSB was adjusted to pH 6.5 using 1M HCl. 200 µL aliquots of each different concentration were dispensed into three wells in 96-well flat bottom culture plates (Corning 3595 96-well cell culture plate, Corning, NY, USA).

The cultured cells were collected by centrifugation (12,000 × *g*, 5 min.) at 4°C. The resulting pellet was washed twice with TSB and re-suspended in 1.0 mL of TSB. The number of acquired cells was counted using a hemocytometer (C-Chip DHC-N01, NanoEnTek Inc.). The cell cultures were diluted to 10^{7} cell/mL in TSB. 5 µL aliquots (10^{7} cell/mL) were transferred to a microplate well with 200 µL adjusted solution with additive and then incubated for 20 h at 30°C. After incubation, the optical density (absorbance at 600 nm, OD_{600}) of each well was measured by a microplate reader SH-9000 Lab (Corona Electric, Ibaraki, Japan). In this study, the smallest concentration of additive having OD_{600} less than 0.1 was used as the MIC threshold for that stress condition (Knight and McKellar, 2007).

2.2.4. Data preprocessing

The procedure from data preprocessing to acquisition of classification results was performed according to the schematic flow shown in Fig.2-1. The preprocessing procedure, which reduces effects of cosmic rays, baseline contamination, noise, and dimensionality, consists of eight parts. First, all spectra with cosmic spike(s) were searched by visual inspection and cosmic spikes removed using software for Raman spectroscopy (Raman viewer, Nanophoton). The maximum intensity within the selected ±120 cm^{-1} was replaced with the median intensity in that range. Second, concave baseline correction was performed on all spectra with a seventh-degree recursive polynomial fitting algorithm (Lieber and Mahadevan-Jansen, 2003; Taylor et al., 2019). Third, the spectra were smoothed with Savitzky-Golay polynomial filters (polyorder 2, 25 cm^{-1} window length) (Savitzky and Golay, 1964). Fourth, each Raman spectrum was standardized such that the mean intensity was 0 and the variance was 1. Fifth, spectra were truncated to the fingerprint region of 600 to 1800 cm$^{−1}$ (Yu et al., 2020). Sixth, since the measured spectra clearly contained outlier data, the outliers were detected and removed by Local Outlier Factor (Breunig et al., 2000) using the values of PC1 to PC10 obtained from principal component analysis (PCA) for the 20 spectra measured for each strain. The numbers of samples after outlier removal were 15 spectra for *B. subtilis,* 20 spectra for *E. coli*, 16 spectra for *L. mesenteroides,* 20 spectra for *P. fluorecens*, 20 spectra for *S. epidermidis* and 20 spectra for *S. saprophyticus*. Seventh, after outlier removal, the spectral data were divided into training and test data at a ratio of 70: 30 for each sample. Finally, principal component analysis (PCA) was used to transform the high-dimensional data into low-dimensional data. The training data was transformed from 241 dimensions to 10 dimensions by PCA. For test data, PCA was performed using the PCA formula obtained from the training data. Analysis proceeded using the dimensionally reduced PCA data.

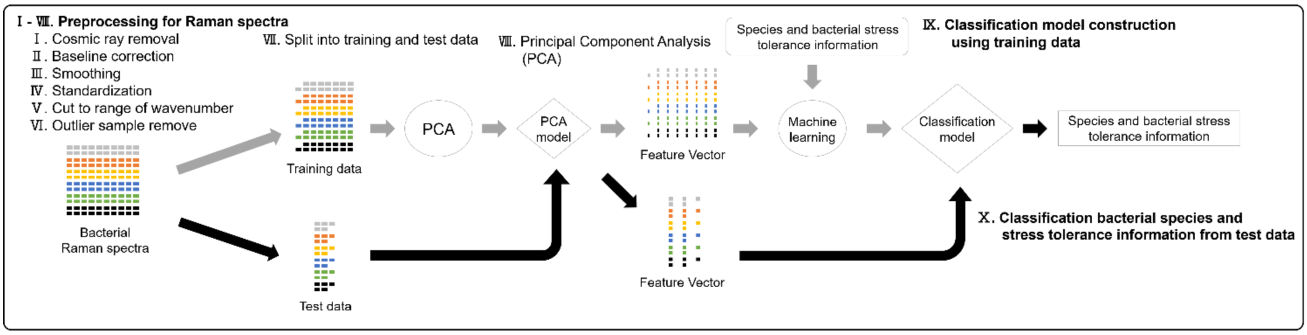

Ⅰ - Ⅶ. Preprocessing for Raman spectra
Ⅰ. Cosmic ray removal
Ⅱ. Baseline correction
Ⅲ. Smoothing
Ⅳ. Standardization
Ⅴ. Cut to range of wavenumber
Ⅵ. Outlier sample remove
Bacterial Raman spectra
Ⅶ. Split into training and test data
Training data
Test data
PCA
Ⅷ. Principal Component Analysis (PCA)
PCA model
Feature Vector
Feature Vector
Species and bacterial stress tolerance information
Machine learning
Ⅸ. Classification model construction using training data
Classification model
Species and bacterial stress tolerance information
Ⅹ. Classification bacterial species and stress tolerance information from test data

Fig. 2-1. The flow of model development. After pre-processing of data, the classification model is established. The numbers indicate the order of processing.

2.2.5. Classification model and model evaluation

There are two classification objectives in this study: discrimination by bacterial species and discrimination by level of stress tolerance. For example, there are 4 NaCl MIC classes, the 856 mM class contains *P. fluorescens*, the 1283 mM class contains *E. coli* and *L. mesenteroides*, the 1711 mM class contains *B. subtilis*, and the > 1711 mM class contains *S. epidermidis* and *S. saprophyticus*. The goal is to classify Raman spectra of the target bacteria into these classes to identify stress tolerance through machine learning. 3 types of stress conditions were assayed by adding sodium chloride, sodium acetate and glycine. Thus, I developed one discrimination model for species classification and three models classifying stress tolerance in target bacteria.

In this study, I used a support vector machine (SVM) with a radial basis kernel function as the classification model in the PC space. SVM is a popular machine learning model used for linear or nonlinear classification, regression, and outlier detection, and has been used as a discriminant model for bacterial species (Jaafreh et al., 2019; Meisel et al., 2012). SVM models were developed to classify microbial species and stress tolerance information based on the processed Raman spectral data. SVM requires adjustment of two parameters. The regularization parameter controls the penalty of misclassification, with small values assigning small penalties for misclassifications and large values assigning large penalties. The localization parameter, gamma, controls the size of the radial basis functions representing the data in the PC space. Small values of gamma lead to larger representations in the space, while larger values lead to smaller representations. The SVM parameters were tested using 10 parameters for the regularization parameter (10 values logarithmically spaced within the range 10^0 to 10^9) and 10 parameters for the gamma parameter controlling localization (10 values logarithmically spaced within the range 10^{-9} to 10^0) for a total of 100 parameters. Hyperparameters of SVM were determined for each classification model based on the results of stratified 5-fold cross validation. The regularization parameter and gamma were 10^2, 10^6, 10^1, 10^1 and 10^{-3}, 10^{-7}, 10^{-1}, 10^{-2}, in classification of species and classification of stress tolerance for NaCl, sodium acetate and glycine, respectively. To evaluate model performance, the results obtained from the test data were used as the model evaluation by calculated accuracy,

$$\text{accuracy (\%)} = \frac{Number\ of\ correctly\ classified\ evaluations}{Number\ of\ evaluations} \times 100. \qquad (1)$$

All computations were performed using software written in open-source python 3.7.1 (www.python.org).

2.3. Results

2.3.1. Raman spectra measurements

After preprocessing, the training set included 77 spectra (10 spectra for *B. subtilis*, 14 spectra for *E. coli*, 11 spectra for *L. mesenteroides*, 14 spectra for *P. fluorecens*, 14 spectra for *S. epidermidis*, 14 spectra for *S. saprophyticus*), and the test set included 34 spectra (5 spectra for *B. subtilis*, 6 spectra for *E. coli*, 5 spectra for *L. mesenteroides*, 6 spectra for *P. fluorecens*, 6 spectra for *S. epidermidis*, 6 spectra for *S. saprophyticus*). Figure 2-2 shows the average intensity of Raman spectra of the substrate, SUS430, and all bacteria from all preprocessed spectra. The spectrum of SUS430 had peaks near 900 cm^{-1} and 1550 cm^{-1}. Each of the bacteria had some peaks between 600 cm^{-1} and 1800 cm^{-1}, which were absent in SUS430.

The PCA plot is shown in Figure 2-3. Although there was overlap in the center of the plot, it was found that the bacterial species tended to disperse in relative directions. After the training and test data were mapped to the PC space and reduced to 10 dimensions, approximately 90% of the Raman spectral information was retained within the 10 dimensions. The first to tenth principal component explained 52.4%, 18.0%, 9.6%, 4.1%, 2.0%, 1.2%, 0.9%, 0.7%, 0.7% and 0.6% of the variance, respectively. The 10-dimensional PC representations of the spectra were then used as a feature set in the classification of bacterial species and bacterial stress tolerances with machine learning.

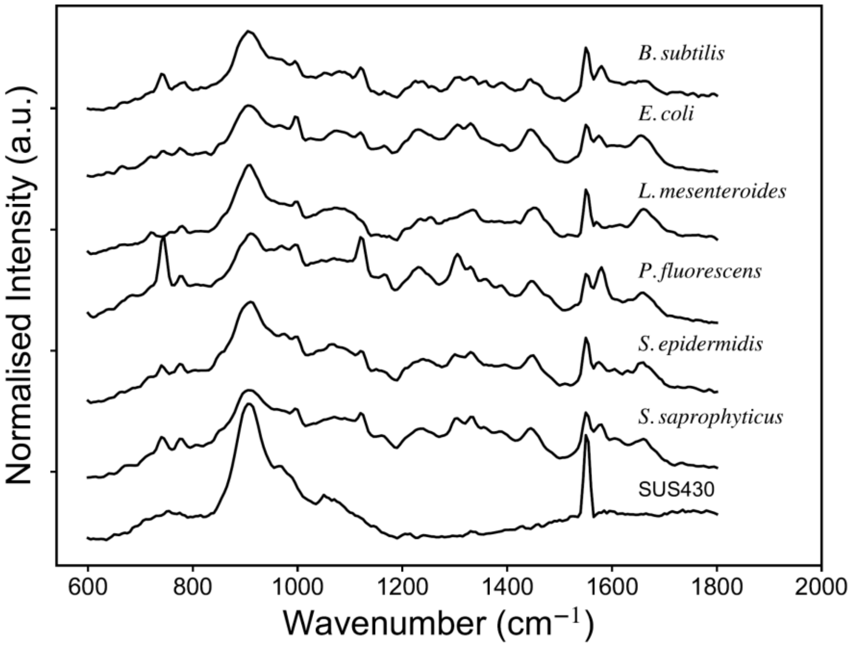


Fig. 2-2. Average, baseline corrected, smoothed and intensity-normalized spectra after preprocessing of six bacterial species and SUS430. Numbers of spectra are 15 for *B. subtilis*, 20 for *E. coli*, 16 for *L. mesenteroides*, 20 for *P. fluorecens*, 20 for *S. epidermidis*, 20 for *S. saprophyticus* and 20 for SUS430.

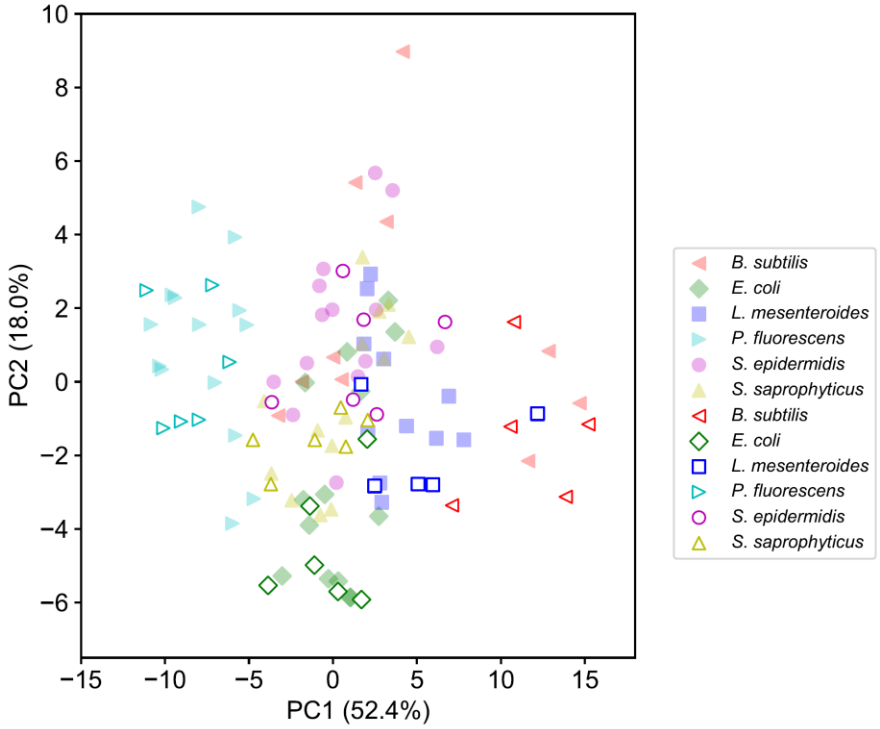


Fig. 2-3. Principal component analysis plot based on the first and second canonical variables of principal components derived from the training data set (closed symbol) and test data set (opened symbol).

2.3.2. Bacterial stress tolerance

TSB was adjusted according to each stress condition to obtain information on bacterial stress tolerances, and the optical density after 1 day was measured. OD_{600} values were collected from bacterial culture after 20 h at 30 °C under different TSB conditions are shown in Figure 2-4. In this test, the mean OD_{600} value of the three wells was defined as MIC when it was less than 0.1, which is shown as a dashed line (Fig. 2-4).

Based on the average OD_{600} value, the MIC for all bacteria are shown in Table 2-1. I considered each MIC value as a stress tolerance group, and each category was divided into 3 or 4 classes (NaCl was 4 classes, sodium acetate was 3 classes, and glycine 3 classes). Classes having NaCl MIC values of 856 mM, 1283 mM, 1711 mM, and >1711 mM contain *P. fluorescens, E. coli* and *L. mesenteroides, B. subtilis*, and *S. saprophyticus* and *S. epidermidis*, respectively. Classes having sodium acetate MIC values of <62.5 mM, 500 mM, and >500 mM contain *P. fluorescens, B. subtilis* and *E. coli*, and *L. mesenteroides, S.epidermidis* and *S. saprophyticus*, respectively. Last, classes having glycine MIC values of 250 mM, 500 mM, and > 500 mM contain *B. subtilis, E. coli* and *S. epidermidis*, and *L. mesenteroides, P. fluorescens, S. saprophyticus*, respectively. MIC values used as indices of stress tolerance differed depending on the bacterial species, and sodium acetate appears to be less inhibitory than glycine at similar concentrations.

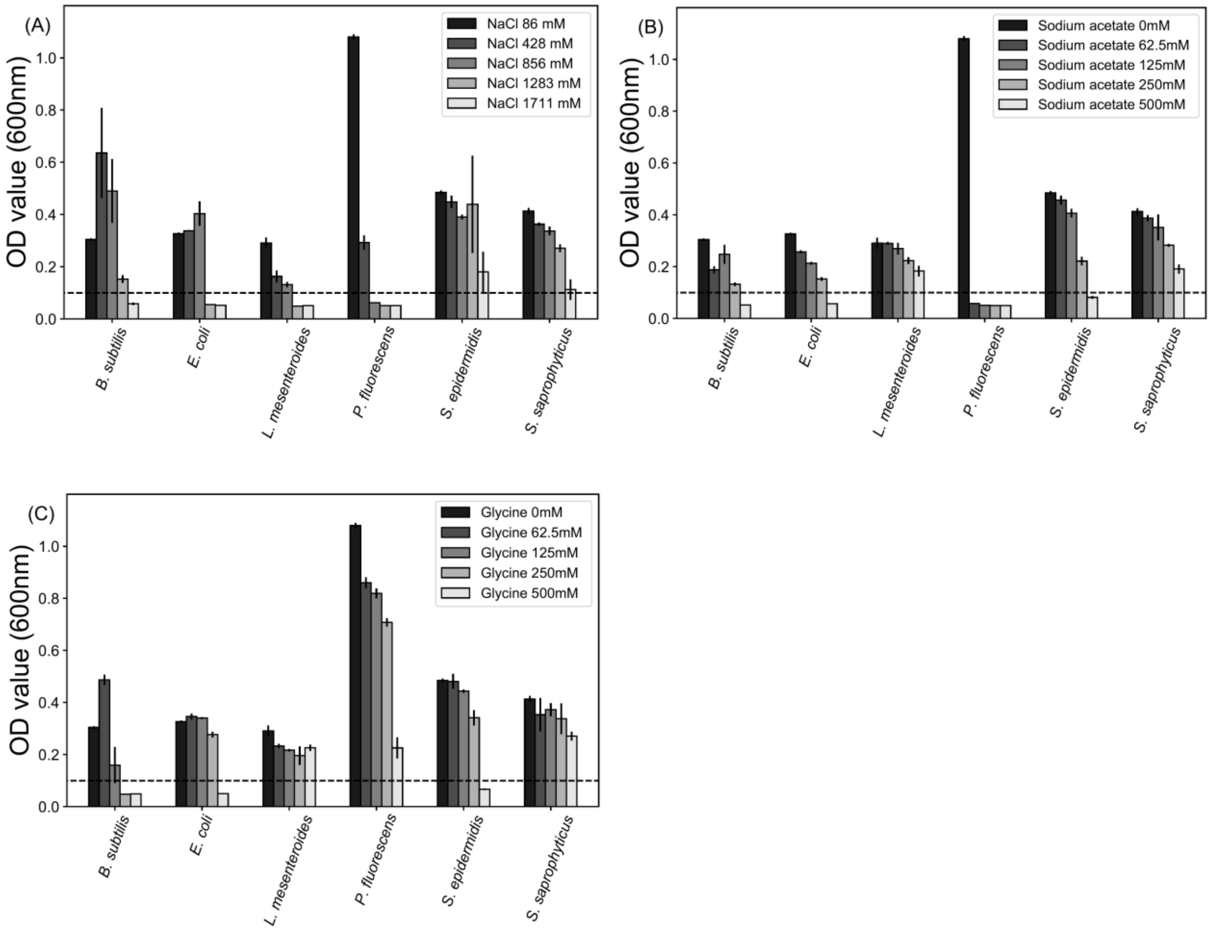


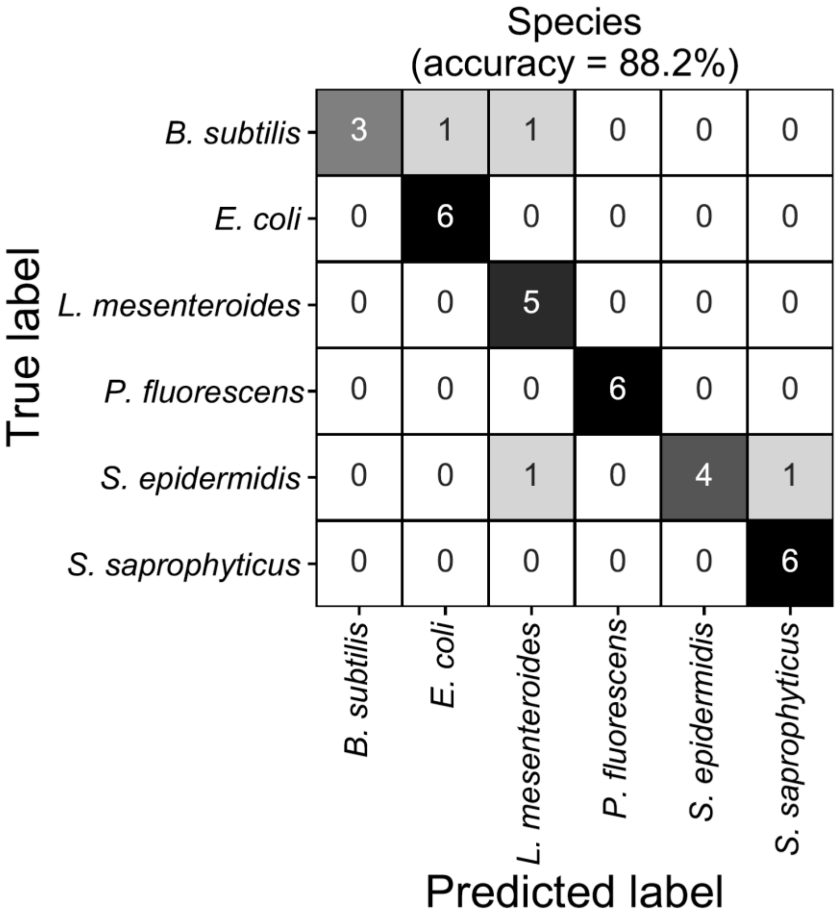
Fig. 2-4. The OD$_{600}$ values for each bacterial culture under different, adjusted TSB conditions. Difference in NaCl (A), sodium acetate (B) and glycine (C). Error bar is standard deviation for three replicate experiments. Dashed line shows indicate the OD$_{600}$ value of 0.1.

Table 2-1. The MIC information of six species bacteria classified based on the OD$_{600}$ value after growth in TSB adjusted to different concentrations of additives.

Species	MIC (OD$_{600}$ value≦0.1)		
	NaCl (mM)	Sodium acetate (mM)	Glycine (mM)
B. subtilis	1711	500	250
E. coli	1283	500	500
L. mesenteroides	1283	> 500	> 500
P. fluorescens	856	< 62.5	> 500
S. epidermidis	> 1711	500	500
S. saprophyticus	> 1711	> 500	> 500

2.3.3. Machine learning for bacterial species classification from Raman spectra

I constructed an SVM model that classifies bacterial species using the Raman spectra, and confirmed the classification accuracy of the model by 5-fold cross-validation using training data and test data. Confusion matrices showing classification results for 6 species of bacteria using 5-fold cross-validation of 77 training data and 34 test data are shown in Figure 2-5. Each data point was classified into one of 6 classes, each of which corresponded to a bacterial species. As a result of 5-fold cross validation, it was possible to classify validation data with an average accuracy of 87.0% (std. 4.2%) and test data with an accuracy of 88.2%. Validation and test accuracy were almost the same, and only *B. subtilis* and *S. epidermidis* were misclassified in the test data.



True label \ Predicted label	*B. subtilis*	*E. coli*	*L. mesenteroides*	*P. fluorescens*	*S. epidermidis*	*S. saprophyticus*
B. subtilis	3	1	1	0	0	0
E. coli	0	6	0	0	0	0
L. mesenteroides	0	0	5	0	0	0
P. fluorescens	0	0	0	6	0	0
S. epidermidis	0	0	1	0	4	1
S. saprophyticus	0	0	0	0	0	6

Fig. 2-5. Test set confusion matrix of bacterial species classification with Raman spectra and 5-fold cross validated SVM.

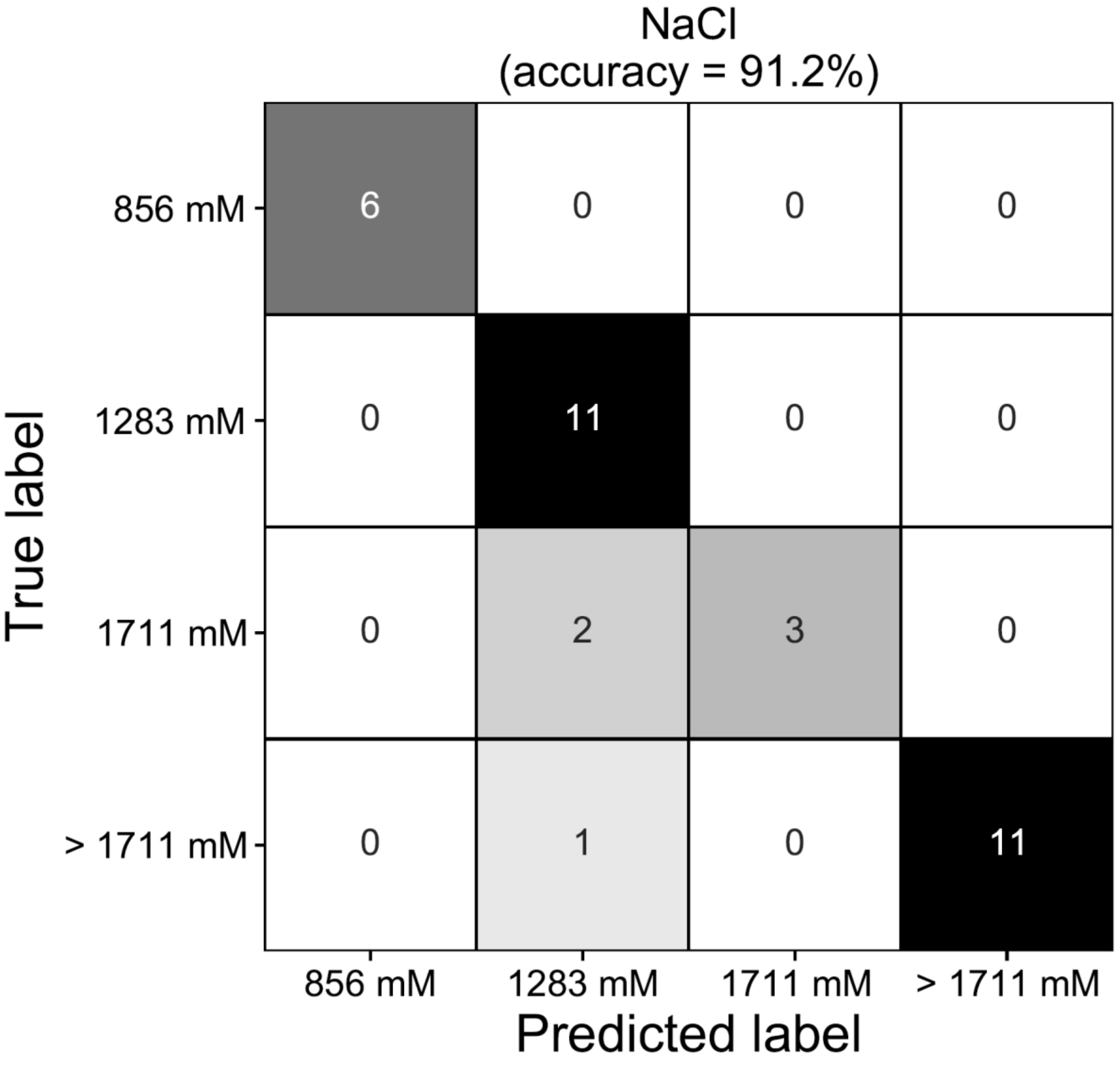
2.3.4. Machine learning for classification of bacterial stress tolerance from Raman spectra

I also constructed SVM models classifying the bacteria by their stress tolerances, and confirmed the classification accuracies of the models with 5-fold cross validation using training data and test data. The stress tolerance labels for bacteria were assigned using the MIC values obtained from OD_{600} measurements. For example, NaCl stress tolerance is divided using MIC values into 4 classes by growth at the various NaCl concentrations. Bacteria growth was not affected by NaCl (> 1711 mM) in one group, and the inhibition of microbial growth with 856 mM, 1283 mM or 1711 mM NaCl composed the other three groups, respectively (Table. 2-1). Other stress tolerances were also grouped similarly.

Results of the SVM model classification of the 4 NaCl classes are shown in Figure 2-6. The class having NaCl MIC of 856 mM class contains *P. fluorescens,* the 1283 mM class contains *E. coli* and *L. mesenteroides*, the 1711 mM class contains *B. subtilis*, and the > 1711 mM class contains *S. saprophyticus* and *S. epidermidis*. As a result of 5-fold cross-validation using training data, I were able to classify with an average accuracy of 93.6% (std. 6.8%), and the test data was classified with an accuracy of 91.2%, similar to that obtained with cross validation. As a result of the classification by the constructed SVM model, the bacteria with NaCl MICs of 856, 1283 mM and > 1711 mM could be classified with high accuracy, with most misclassifications occurring in the 1711 mM class.

Results for the 3-class SVM model classifying sodium acetate stress tolerance are shown in Figure 2-7. The classes having sodium acetate MIC values of < 62.5 mM, 500 mM, and > 500 mM contain *P. fluorescens, B. subtilis*, *E. coli* and *S.epidermidis,* and *L. mesenteroides* and *S. saprophyticus*, respectively. The classifications of sodium acetate stress tolerance were classified with an average accuracy of 94.8% (std. 2.6%) in a 5-fold cross-validation using the training data, and test data were be classified with an accuracy of 91.2%.

Finally, the 3-class classification of glycine stress tolerance is shown in Figure 2-8. The 250 mM glycine MIC class contains *B. subtilis,* the 500 mM class contains *E. coli* and *S. epidermidis*, and the > 500 mM class contains *L. mesenteroides, P. fluorescens, S. saprophyticus*. Classification of glycine stress tolerance yielded an average accuracy of 85.8% (std. 4.5%) in a 5-fold cross validation using training data, and the test data yielded 91.2%, which was as accurate as cross-validation.



	856 mM	1283 mM	1711 mM	> 1711 mM
856 mM	6	0	0	0
1283 mM	0	11	0	0
1711 mM	0	2	3	0
> 1711 mM	0	1	0	11

Fig. 2-6. Test set confusion matrix of NaCl stress tolerance classification with Raman spectra and 5-fold cross validated SVM. Groups of MIC are as follows: 856 mM: *P. fluorescens,* 1283 mM: *E. coli,* and *L. mesenteroides,* 1711 mM: *B. subtilis,* and >1711 mM: *S. saprophyticus* and *S. epidermidis*.

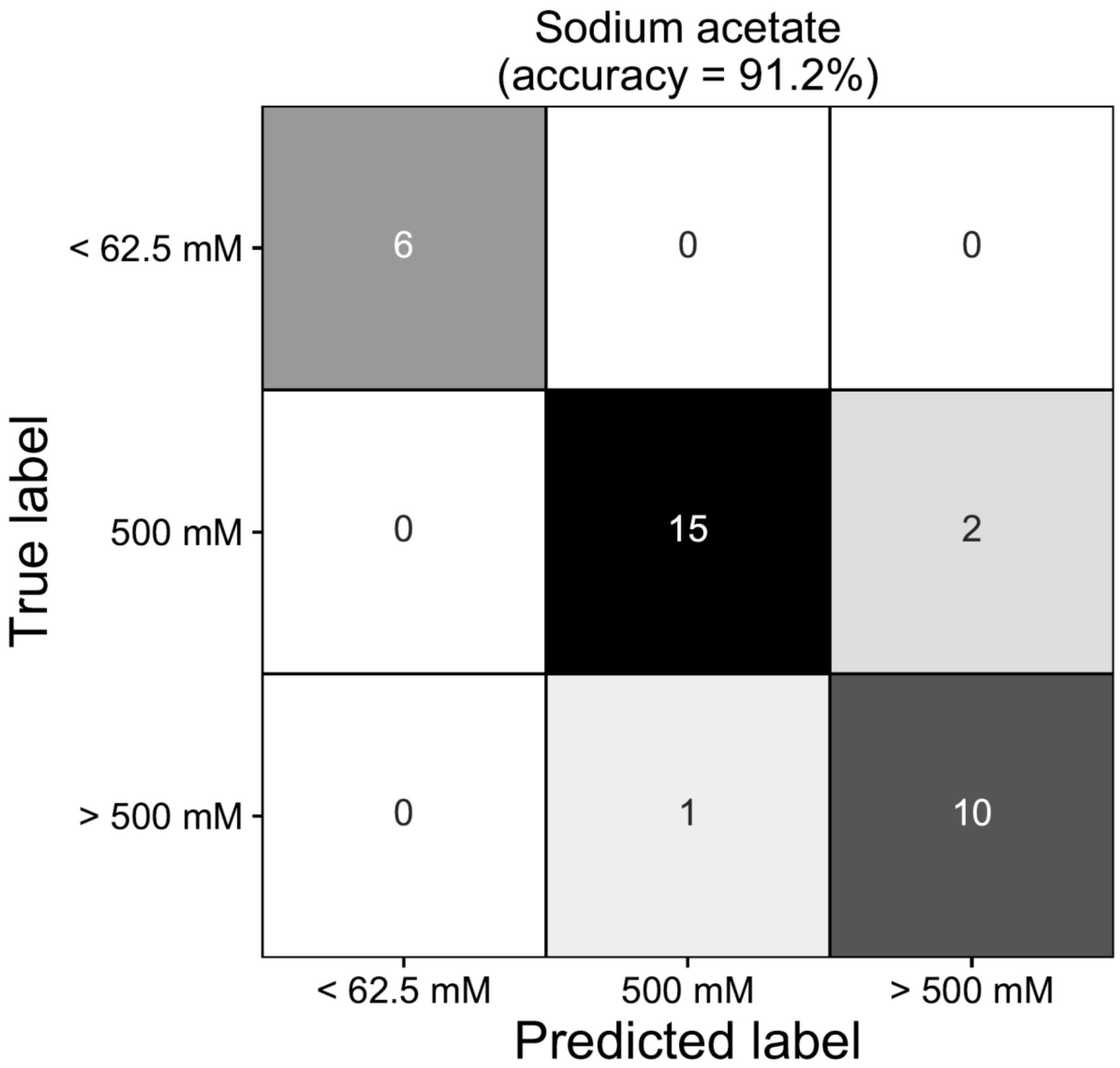


Fig. 2-7. Test set confusion matrix of sodium acetate stress tolerance classification with Raman spectra and 5-fold cross validated SVM. Groups of MIC are as follows: < 62.5 mM: *P. fluorescens*, 500 mM: *B. subtilis E. coli and S. epidermidis* and > 500 mM: *L. mesenteroides* and *S. saprophyticus*.

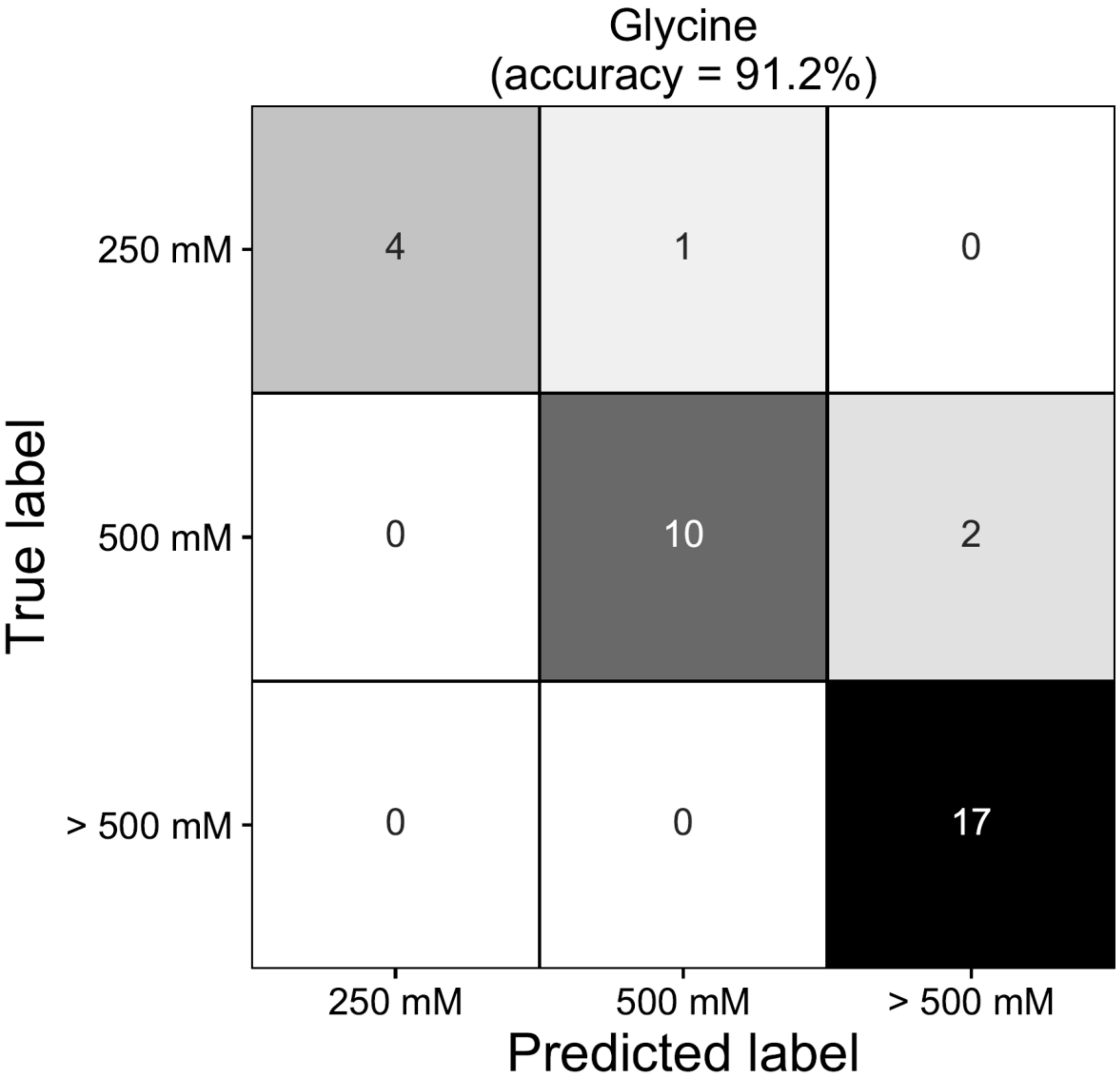


Fig. 2-8. Test set confusion matrix of glycine stress tolerance classification with Raman spectra and 5-fold cross validated SVM. Groups of MIC are as follows: 250 mM: *B. subtilis*, 500 mM: *E. coli* and *S. epidermidis*, and > 500 mM: *L. mesenteroides, P. fluorescens* and *S. saprophyticus*.

2.4 Discussion

In this work, I investigated Raman spectra obtained from bacterial cells cultured in TSB (with no other additives) for identifying food spoilage bacteria and classifying their stress tolerances with machine learning techniques (Fig. 2-1). The 0.1 OD_{600} value for six strains was estimated in solutions of TSB adjusted by NaCl, sodium acetate, and glycine to annotate stress tolerances of six bacteria (Fig.2-4 and Table 2-1). As a result, our model enabled classification of species with 88.2% accuracy (Fig. 2-5) and classification of stress tolerances with 91.2% accuracy (Figs. 2-6, 2-7, and 2-8). Therefore, the combination of Raman spectroscopy and machine learning may provide a model that classifies bacterial species and the stress tolerance involved in microbial growth.

Pre-processing of Raman spectra and the construction of the machine learning classification model for bacterial species were reported previously. Jaafreh et al. (Jaafreh et al., 2019) reported that Raman spectra in the fingerprint region of bacteria were subjected to principal component analysis and used to the first 10 PCs (about 90% of cumulative variance explained) as feature data for classifying bacterial species of spoilage and pathogenic microorganisms commonly found in poultry meat, resulting in accurate classification of 100%. In our study, the equipment, measurement conditions, and target bacteria were different from those reported in the past, but the SVM model classified food spoilage bacterial species with 90% accuracy from the bacterial Raman spectrum (Fig. 2-5). Our result also showed that the bacterial species associated with food spoilage can be classified with Raman spectroscopy and chemometrics methods.

The main purpose of this study was to classify not only bacterial species, but also stress resistance using Raman spectra. The identification of stress resistance would help control the growth of microorganisms to prevent food spoilage. The combination of Raman microscopy and chemometrics can group *E. coli* antimicrobial resistance to almost 100%

accuracy (Germond et al., 2018), and Raman spectroscopy and deep learning were able to classify methicillin-resistant and -susceptible *Staphylococcus aureus* with about 89% accuracy (Ho et al., 2019). Our study investigated the resistance of each microorganism to stress by NaCl, sodium acetate, and glycine. As a result, different bacterial species may exhibit the same stress resistance, such as the same glycine and sodium acetate resistance of *L. mesenteroides* and *S. saprophyticus* (Table 2-1). All stress tolerance classification models constructed using the same preprocessing steps (Fig. 2-1) were able to classify microorganisms by stress tolerance with 91.2% accuracy, which is equivalent to the classification of bacterial species (Fig. 2-6, 2-7 and 2-8). Gene classification techniques using 16S rRNA cannot classify stress tolerance as in this study and may require more complex and time-consuming methods such as whole genome analysis. Conversely, Raman measurement takes less than an hour, even if the spectrum is measured multiple times, which is quicker and easier than culture and genotypic methods. Unlike current morphological observation and gene extraction methods Raman spectra can be quickly measured, and measurement does not burden the person in charge of bacterial identification. In the microbial classification study using Raman spectroscopy, there was no example of classifying multiple food spoilage-related microorganisms by multiple stress tolerances. Considering that I have classified stress tolerances with a similar same chemometrics method, the Raman spectrum may contain a variety of information about the stress tolerance properties of microorganisms.

Although this study was limited to only a few bacterial species and types of stress resistance the possibility of using Raman spectroscopy in classifying various stress tolerance in bacteria was demonstrated. To ensure food quality, there are many stresses applied to microorganisms such as heat stress and cold stress (Gould, 1996), and additives as in this study. Target bacteria and their control points differ depending on the product. Therefore, it is necessary to investigate various bacterial properties depending on the food material and processing method. Stress-resistant proteins can be produced depending on the prior history of cells, which induce a variety of resistance even in the same species (Begley and Hill, 2015). I continue to accumulate and analyze data on the relationship between stress tolerance of various microorganisms and Raman spectra. If the characteristics of spoilage bacteria isolated from the actual food manufacturing can be obtained from Raman spectrum, a problem in the product related to bacterial contamination can be immediately fed back to the food manufacturing site and developers. It is useful for efficient risk analysis and changing in manufacturing processes and formulations.

Chapter 3

Combining Raman Spectroscopy and Machine Learning to Predict the Growth/No growth of Unknown Strains

3.1. Introduction

Predicting the growth behavior of microorganisms is useful as a means to quickly and easily set deadlines and conditions for food safety and quality assurance (Stavropoulou and Bezirtzoglou, 2019). The statistical models in food microbiology can be used to quantitatively assess the impact of processing and storage conditions on the microbial behavior of products (Walls and Scott, 1997). The ComBase (http://www.combase.cc) was developed as a database to provide easy access to records of bacterial behavior reported in research facilities and publications (József Baranyi and Tamplin, 2004). In recent years, new efforts have been made to predict microbial behavior in foods from the accumulated data by utilizing advanced artificial intelligence technology (Hiura et al., 2021)

On the other hand, the identification of the detected microorganisms is necessary for the utilization of databases such as ComBase and predictive models of microbial growth behavior. Genetic methods are used to identify microorganisms, while reliable, are time-consuming and require complex technical operations such as DNA extraction. There are many types of microorganisms associated with food (Møretrø and Langsrud, 2017). It is inefficient to perform the identification process every time to predict the behavior of microorganisms. In addition, even if the bacteria are of the same species, each strain may have different sensitivity to factors related to bacterial growth behavior, such as pH and NaCl (Dengremont and Membré, 1995). In other words, more accurately characterize unknown bacteria in food, a method that can be evaluated at the strain level is needed.

Chapter 3 Combining Raman Spectroscopy and Machine Learning to Predict the Growth/No growth of Unknown Strains

In Chapter 2, Raman spectroscopy and chemometrics can be useful for identifying stress tolerance of spoilage bacteria to food additives. Based on the results of Chapter 2, I hypothesized that by combining Raman spectroscopy with predictive microbiology, which uses modeling to predict the growth characteristics of bacteria. It would be possible to evaluate the growth characteristics of unknown bacteria detected in food using only Raman spectral measurements. This research provides an opportunity to overcome the problem of predictive microbiology, which requires highly accurate identification of bacterial species, such as genetic analysis, to use a model to predict bacterial growth characteristics.

The purpose of this study is to develop a model to predict the growth behavior of microorganisms isolated from food without identifying the microorganisms using Raman spectroscopy and machine learning. First, 21 strains of bacteria from cut vegetables were isolated, which then were measured for obtaining the Raman spectra. Second, the growth of the isolated bacteria was evaluated in liquid medium with different sodium acetate concentrations at different incubation temperatures and times by optical density method using. Third, I developed a machine learning model to predict the growth / no growth of isolated bacteria based on their Raman spectra and the conditions of sodium acetate concentration, incubation temperature, and incubation time and compared the results of identification by 16S rRNA analysis with the model used for prediction. Finally, I evaluated whether the machine learning model could predict the behavior of unknown strains that were not included in the training data. If a prediction model combining Raman spectroscopy and machine learning can be used to predict the behavior of unknown strains of bacteria, this technology can contribute to solving challenges in the field of food predictive microbiology.

3.2. Material and Methods

3.2.1. Isolation of bacteria strains from fresh-cut vegetables

I collected seven kinds of cut vegetable with various ingredients at supermarkets and convenience stores in Kyoto, Japan. The details of the fresh-cut vegetable were described in Appendix 1. To isolate the bacteria from the fresh-cut vegetable samples, the following methods were used. A 25 g portion of fresh-cut vegetable was homogenized in 225 g of phosphate buffered saline (PBS) for 3 min using a homogenizer (Pro-media SH-IIM, ELMEX, Japan) to obtain a suspension. Dilute the suspension appropriately with PBS and each 1 mL of the suspension was transferred to sterile Petri dishes, mixed with Standard Plate Count agar (SPC agar, NISSUI, Japan). Once the medium had solidified, incubated for 48 h at 35 °C under aerobic conditions. The major colonies obtained from the incubated SPC agar were randomly isolated and put on SPC agar by using loop. In this study, 3 strains were isolated from each sample of cut vegetable. 21 strains were obtained in total. The isolated colonies were stored at 5 ° C.

3.2.2. Identification of the isolated bacteria strains

The isolated bacteria were identified according to the method described in the 17th edition of the Japanese Pharmacopoeia (http://www.mhlw.go.jp/topics/bukyoku/iyaku/yakkyoku/english.html). Briefly, genomic DNA of the isolate bacteria was extracted from the culture with a PrepManTM Ultra Sample Preparation Reagent (Life Technologies, USA) and directly used as a PCR template to amplify the divergent region of 16S rRNA gene using 10 F primer (5′-GTTTGATCCTGGCTCA-3′) and 800R primer (5′-TACCAGGGTATCTAATCC-3′). The PCR products (approximately 740bp) were purified using a ExoSAP-IT (Affymetrix part of Thermo Fisher Scientific, USA) for sequencing. Sequencing reactions were performed in a

BioRad DNA Engine Dyad PTC-220 Peltier Thermal Cycler using the BigDye™ Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific), according to manufacturer's instructions. Single-pass sequencing was performed on each template using 10 F primer. The fluorescent-labeled fragments were purified from the unincorporated terminators either by ethanol precipitation method or using the BigDye XTerminator™ Purification Kit (Thermo Fisher Scientific). The samples were analysed by a 3730xl DNA Analyzer (Thermo Fisher Scientific). The 16S rRNA gene sequences obtained were compared with those of the type strains available in the EzBioCloud (https://www.ezbiocloud.net/)(Yoon et al., 2017) to determine an approximate phylogenetic affiliation of each strain. The corresponding sequences of closely related type species were retrieved from GenBank database using the EzBioCloud server. The evolutionary history was inferred by the Maximum Likelihood method and Kimura 2-parameter model (Kimura, 1980) using the MEGA X (Kumar et al., 2018).

3.2.3. Sample preparation and Raman spectra measurements

In subsequent experiments, the isolated bacteria in the stock were transferred to SPC agar by platinum loop and then used after culturing for 48h at 35 °C under aerobic conditions. Raman spectra of single cell were collected using a laser Raman microscope (RAMAN touch, Nanophoton, Osaka, Japan). The excitation source was a 532 nm laser operated at 10 mW. A 20x/0.45 objective lens (Nikon TU Plan Fluor) with a laser spot size of approximately 720 nm was used to focus the excitation light onto the sample. Raman spectra were acquired with a 300 lines/mm grating for 30 s. The 2 spectra from a single spot were averaged to obtain the mean spectrum. The Raman shift was calibrated using silicon (520 cm^{-1}) before acquiring the spectra. The cultured cells were suspended into 50 µL of pure water using 1 µL loop. 1 µL aliquots of suspension were dropped onto a stainless-steel piece

(SUS430, HIKARI, Osaka, Japan), which was air dried before starting the Raman measurement. 20 Raman spectra were obtained for each strains using a range of Raman shifts from 125 to 4690 cm^{-1} with increments of 5.0 cm^{-1}.

3.2.4. Bacterial growth experiment

In this study, I test whether we can predict the growth/non-growth of unknown strain under certain culture conditions by using Raman spectra and machine learning. Therefore, I decided to predict the growth/no growth of unknown strain under different conditions of sodium acetate, a food additive involved in bacterial growth, and low temperature stress. In order to construct a prediction model, I first confirmed the growth/no growth of bacteria under the following conditions using the turbidity method. The condition of this experiment was two incubation temperature (5 and 10 °C), three sodium acetate concentration (0, 0.25, and 0.5% w/v), and eight incubation time (0, 1, 2, 3, 4, 5, 6, and 7 day). The experimental condition was in total 48.

The pH of TSB was adjusted to pH 6.5 using 1M HCl. The cultured cells were suspended in TSB using a 1 µL loop and diluted with TSB so that the bacterial concentration was 10^4 cfu/mL in TSB at the start of culture. 100 µL aliquots of the suspension were dispensed into a well in 96-well flat bottom culture plates (Corning 3595 96-well cell culture plate, Corning, NY, USA) with three replicates for each condition. The optical density (absorbance at 600 nm, OD$_{600}$) of each well was measured by a microplate reader SH-9000 Lab (Corona Electric, Ibaraki, Japan) at the beginning of the experiment and every other day thereafter. The average of 3 wells OD$_{600}$ values was calculated for each condition. In this study, the OD$_{600}$ value was set to 0.1 as the threshold (Yamamoto et al., 2021). If the OD$_{600}$ value is less than 0.1, the population was regarded as no growth. If the OD$_{600}$ value is 0.1 or more, bacterial population was regarded as growth.

3.2.5. Feature extraction from single-cell Raman spectra

The procedure from data preprocessing to acquisition of classification results was performed according to the schematic flow shown in Figure 3-1. The preprocessing procedure, which reduces effects of cosmic rays, baseline contamination, smoothing, and standardization, consists of four parts. First, all spectra with cosmic spike(s) were searched by visual inspection and cosmic spikes removed using software for Raman spectroscopy (Raman viewer, Nanophoton). The maximum intensity within the selected ±120 cm^{-1} was replaced with the median intensity in that range. Second, concave baseline correction was performed on all spectra with a seventh-degree recursive polynomial fitting algorithm (Lieber and Mahadevan-Jansen, 2003; Taylor et al., 2019). Third, the spectra were smoothed with Savitzky-Golay polynomial filters (polyorder 2, 25 cm^{-1} window length) (Savitzky and Golay, 1964). Fourth, each Raman spectrum was standardized such that the mean intensity was 0 and the variance was 1.

Finally, linear discriminant analysis (LDA) was used to transform the high-dimensional data into low-dimensional data (Fisher, 1938). LDA is a method of maximizing the distance between pre-specified classes and finding the axis of dimensionality reduction (Schumacher et al., 2014). LDA is useful as a method of reducing the dimensionality of microbial Raman spectra while maintaining their unique characteristics (add reference). In this study, the classes were specified for each strain. The Raman spectral data was transformed from 914 dimensions to 10 dimensions by LDA.

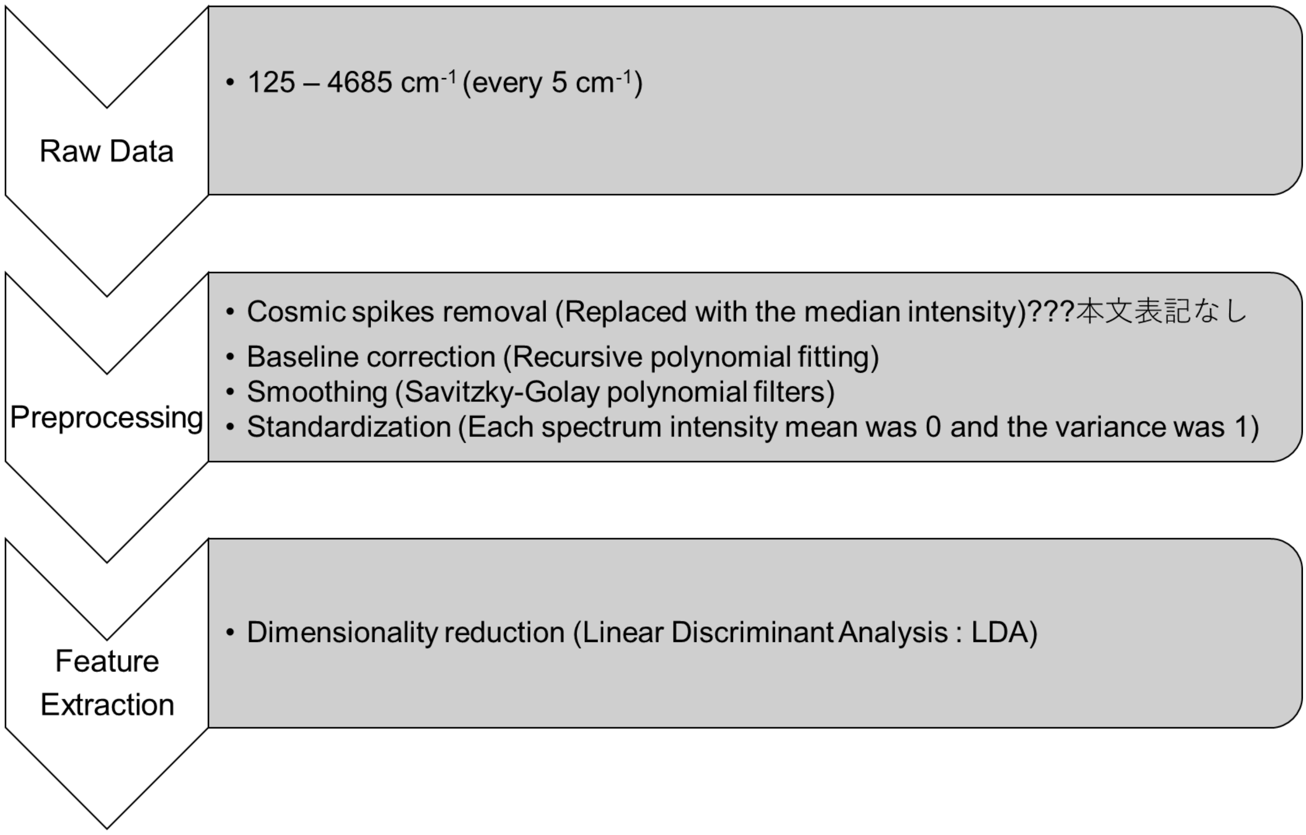

Raw Data
• 125 – 4685 cm-1 (every 5 cm-1)
Preprocessing
• Cosmic spikes removal (Replaced with the median intensity)???本文表記なし
• Baseline correction (Recursive polynomial fitting)
• Smoothing (Savitzky-Golay polynomial filters)
• Standardization (Each spectrum intensity mean was 0 and the variance was 1)
Feature Extraction
• Dimensionality reduction (Linear Discriminant Analysis : LDA)

Fig. 3-1. The flow of feature extraction from Raman spectra.

3.2.6. Aggregated hierarchical clustering

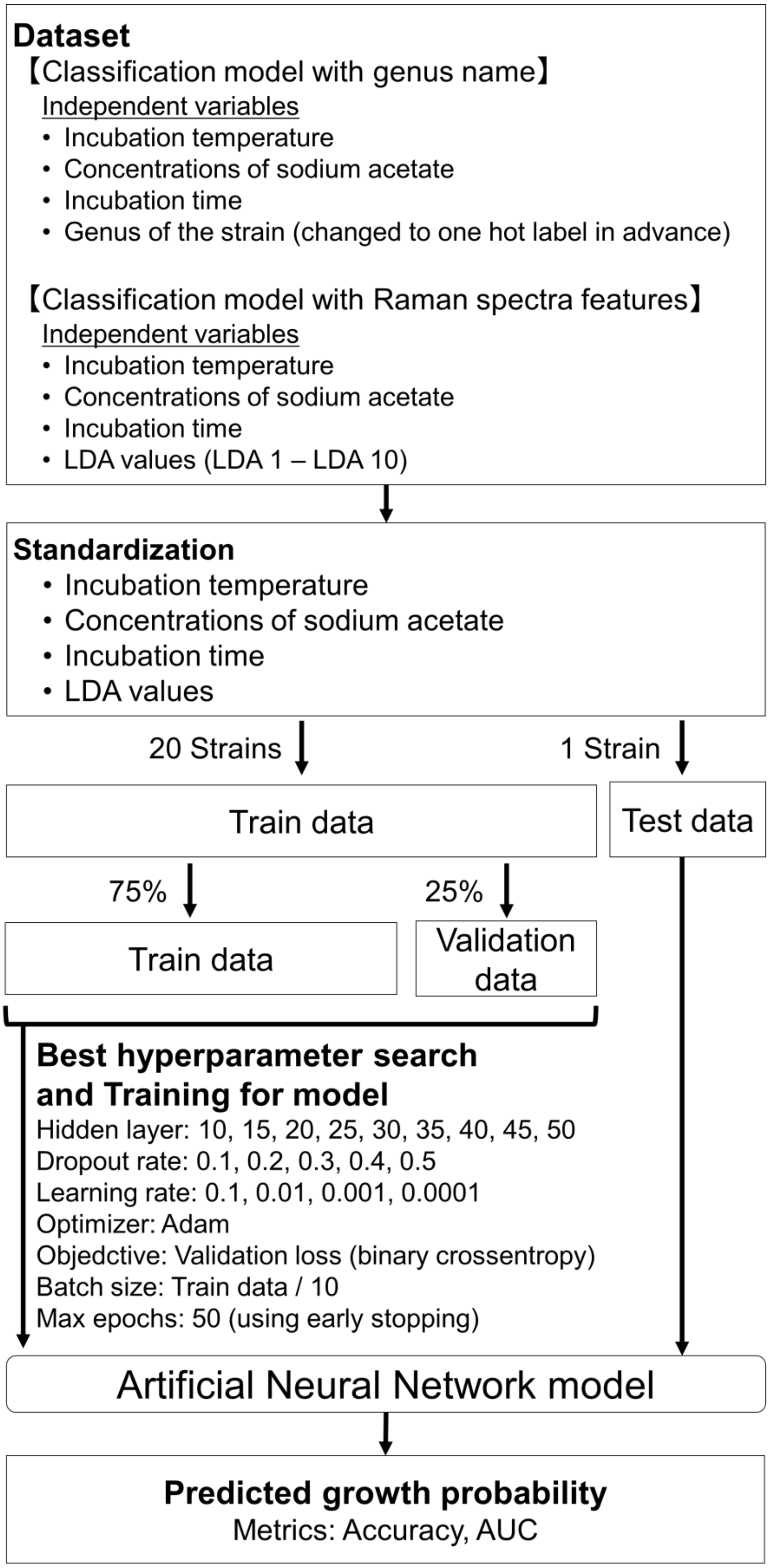
Aggregated hierarchical clustering (Ward, 1963) was performed using 10 dimensional LDA values for comparison with a phylogenetic tree based on 16S rRNA information. Using the SciPy (Jones et al., 2016) Python package, each strain was sorted by hierarchical clustering using mean linkage and Euclidean distance metrics based on the mean of each value from LDA1 to LDA10.

3.2.7. Model development and evaluation

In this study, we constructed 2 models to predict bacterial growth/no growth from incubation temperature, sodium acetate concentration, incubation time, and bacterial genus or Raman spectra features of the bacteria. One is a conventional model (Kuroda et al., 2019) that uses the genus name of each strain as the independent variable (classification model with genus name), and the other is a model that uses the LDA value of each strain obtained from the Raman spectrum (classification model with Raman spectra features). Independent variables of classification model with genus name were temperature, concentrations of sodium acetate, incubation time, and genus of the strains. Since the species is categorical data, it was changed to one hot label in advance. In contrast, independent variables of classification model with Raman spectra was used temperature, concentrations of sodium acetate, incubation time and LDA value (LDA 1 to LDA 10). The dependent variable was probability of growth in the range of 0-1. Based on the results obtained (growth: 1; no growth: 0), a artificial neural network (ANN) with a single hidden layer was built using TensorFlow (Martín Abadi et al., 2015) and Keras (Chollet and Others, 2015) package of python. ANN can approximate complex behaviors such as microbial growth without the need to assume the type of relationship or the degree of nonlinearity between the various independent and dependent variables (Hajmeer and Basheer, 2002).

The flow from model building to obtaining the prediction results is shown in Figure 3-2. The data set used in the classification model with genus name included a total of 1,008 data for 48 conditions which growth/no growth was confirmed for each of the 21 strains. The data set used in the classification model with Raman features included a total of 20,160 data for the Raman spectral features (LDA value, 20 measurements for each strain) multiplied by 48 conditions that were found to be growth/no growth for each of the 21 strains. First, incubation temperature, concentration of sodium acetate, incubation time, and LDA value were standardized in advance. Then, out of the 21 strains, 20 were used for training data and one was used for test data. In the classification model with genus name, only *Erwinia* sp., *Pseudomonas* sp., and *Rahnella* sp., were tested because if there was only one strain in a genus, there would be no data for training if it was used as test data. Since the classification model with Raman spectra features has no such problem, a total of 21 patterns were tested with all strains. The ANN model constructed consisted of an input layer, one hidden layer, and an output layer (Fig. 3-3). The hidden layer of activation function was rectified linear unit (Nair and Hinton, 2010). Dropout (Srivastava et al., 2014) used for overfitting prevention. The optimizer used Adam (Kingma and Ba, 2014). The sigmoid activation function was used the output layer. The batch size was set to one tenth of the training data. The number of epochs was 50. I also added the early stopping function (Raskutti et al., 2014) to prevent overfitting. Learning was set to stop if no loss reduction was observed during the 5 epochs. Binary cross entropy was used as the loss function (de Boer et al., 2005; Lee and Song, 2019). Hyperparameter determination and training of the model was performed using 25% of the training data as validation data. Hyperparameters related to the overall performance of the ANN, such as the number of hidden layer units, dropout rate, and learning rate, were selected using the keras tuner(O’Malley et al., 2019).

The performance of the constructed models was evaluated by accuracy and area under the curve (AUC) as metrics of predictive models. The data was classified into two classes (growth/no growth, threshold = 0.5 of growth probability). The AUC was calculated by plotting the true positive rate against the false positive rate at various thresholds to create a receiver operating characteristic (ROC) curve and taking a value between 0 and 1 in the region under the ROC curve. A value of 0.5 means that the model prediction is random, and a value of 1 means that the model predicts perfectly (Hanley and McNeil, 1982).


Dataset
【Classification model with genus name】
Independent variables
• Incubation temperature
• Concentrations of sodium acetate
• Incubation time
• Genus of the strain (changed to one hot label in advance)
【Classification model with Raman spectra features】
Independent variables
• Incubation temperature
• Concentrations of sodium acetate
• Incubation time
• LDA values (LDA 1 – LDA 10)
Standardization
• Incubation temperature
• Concentrations of sodium acetate
• Incubation time
• LDA values
20 Strains
1 Strain
Train data
Test data
75%
25%
Train data
Validation data
Best hyperparameter search and Training for model
Artificial Neural Network model
Predicted growth probability
Metrics: Accuracy, AUC

Hidden layer: 10, 15, 20, 25, 30, 35, 40, 45, 50
Dropout rate: 0.1, 0.2, 0.3, 0.4, 0.5
Learning rate: 0.1, 0.01, 0.001, 0.0001
Optimizer: Adam
Objedctive: Validation loss (binary crossentropy)
Batch size: Train data / 10
Max epochs: 50 (using early stopping)

Fig. 3-2. The flow of growth probability prediction model construction.

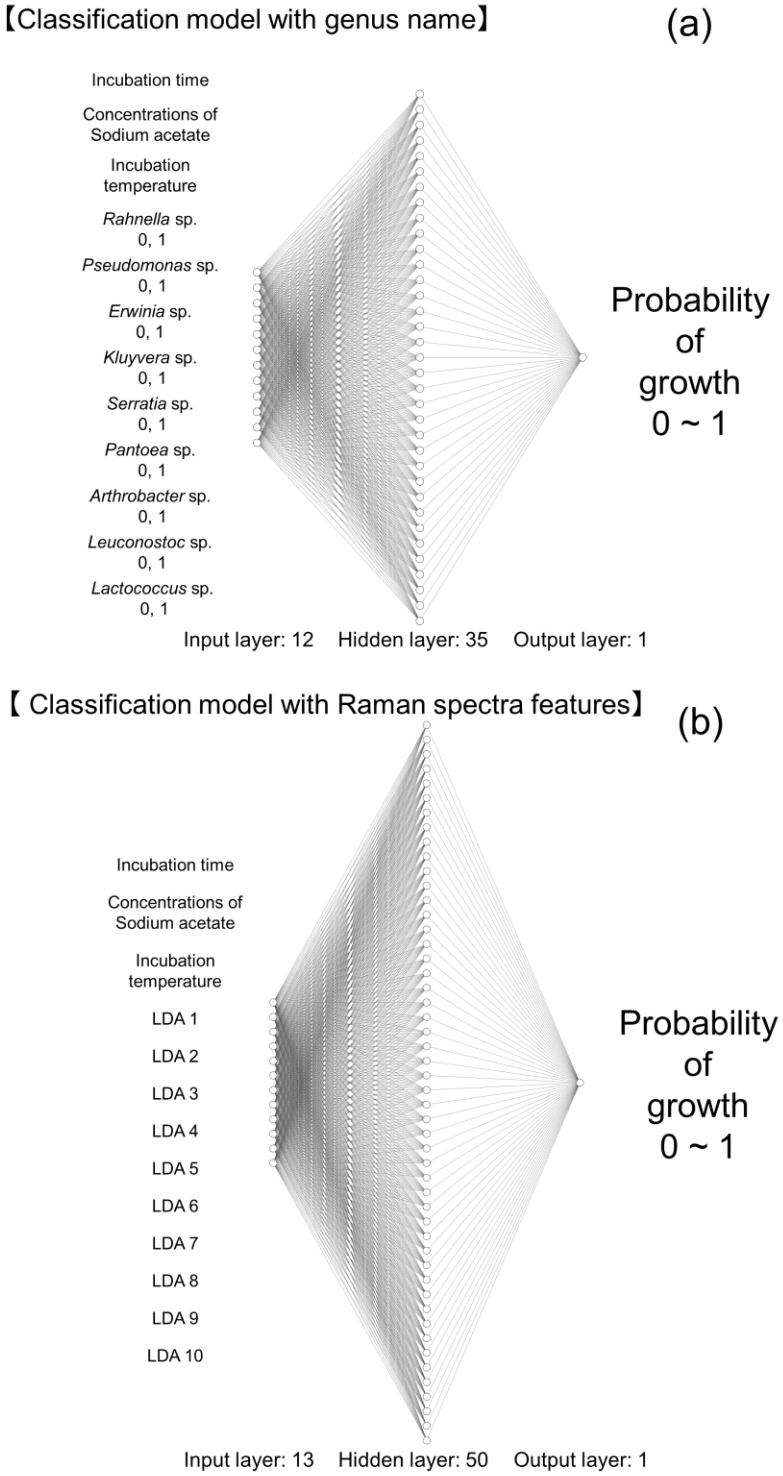


Fig. 3-3. Schematic diagram of multilayer perceptron neural network models for predicting probability of Ranella A1 strain growth. Classification model with genus name (Conventional model that requires microbial identification) (a) and classification model with Raman spectra features (Developed model that uses Raman spectroscopy instead of microbial identification) (b).

3.3 Results

3.3.1. Isolated from fresh-cut vegetables and genotypic identification

In this study, from seven types of cut vegetables (Sample A to Sample G), a total of 21 strains were isolated. Based on the results of the 16S rRNA sequence in the first half (about 720-740 bp) of each strain, a closely related species estimation (Table 3-1) and a phylogenetic tree were prepared (Fig. 3-3). Several species were obtained from *Pseudomonas* sp. and *Rahnella* sp. as closely related species (Table 3-1). However, the phylogenetic tree revealed that the species are closely related (Fig. 3-3), so the genus was used as the identification result in this study. Nine genera were isolated from cut vegetables, each with seven strains of *Pseudomonas* sp. (Pseudomonas B1, Pseudomonas B3, Pseudomonas E1, Pseudomonas E2, Pseudomonas E3, Pseudomonas F1,and Pseudomonas F3), six strains (Rahnella A1, Rahnella A2, Rahnella A3, Rahnella G1, Rahnella G2 and Rahnella G3) of *Rarnella sp.*, two strains (Erwinia B2, Erwinia C2) of *Erwinia* sp., each one strain of *Arthrobacter* sp. (Arthrobacter D3), *Kluyvera* sp. (Kluyvera C3), *Lactococcus* sp. (Lactococcus F2), *Leuconostoc* sp. (Leuconostoc C1), *Serratia* sp. (Serratia D1) and *Pantoea* sp. (Pantoea D2).

Table 3-1. Similarity of 16S rRNA of each isolation retrieved by EzBioCloud.

Cut-Vegetable sample No.	Named Strain	Top-hit taxon	Top-hit strain	Similarity (%)
A	Rahnella A1	*Rahnella aceris*	SAP-19	100.0
	Rahnella A2	*Rahnella aquatilis*	CIP 78.65	99.2
	Rahnella A3	*Rahnella aquatilis*	CIP 78.65	99.3
B	Pseudomonas B1	*Pseudomonas viridiflava*	DSM 6694	100.0
	Erwinia B2	*Erwinia persicina*	NBRC 102418	99.6
	Pseudomonas B3	*Pseudomonas rhodesiae*	CIP 104664	99.7
C	Leuconostoc C1	*Leuconostoc holzapfelii*	BFE 7000	99.2
	Erwinia C2	*Erwinia persicina*	NBRC 102418	99.9
	Kluyvera C3	*Kluyvera intermedia*	NBRC 102594	100.0
D	Serratia D1	*Serratia myotis*	12	99.6
	Pantoea D2	*Pantoea eucalypti*	LMG 24198	99.9
	Arthrobacter D3	*Arthrobacter oryzae*	KV-651	99.6
E	Pseudomonas E1	*Pseudomonas grimontii*	CFML 97-514	100.0
	Pseudomonas E2	*Pseudomonas viridiflava*	DSM 6694	100.0
	Pseudomonas E3	*Pseudomonas kitaguniensis*	MAFF 301498	99.9
F	Pseudomonas F1	*Pseudomonas extremorientalis*	KMM 3447	99.9
	Lactococcus F2	*Lactococcus lactis* subsp. *cremoris*	NCDO 607	99.9
	Pseudomonas F3	*Pseudomonas extremorientalis*	KMM 3447	100.0
G	Rahnella G1	*Rahnella aquatilis*	CIP 78.65	100.0
	Rahnella G2	*Rahnella aquatilis*	CIP 78.65	99.7
	nam	*Rahnella aquatilis*	CIP 78.65	99.3

Coverage of 16S rRNA: 48.9 ~ 49.8 %

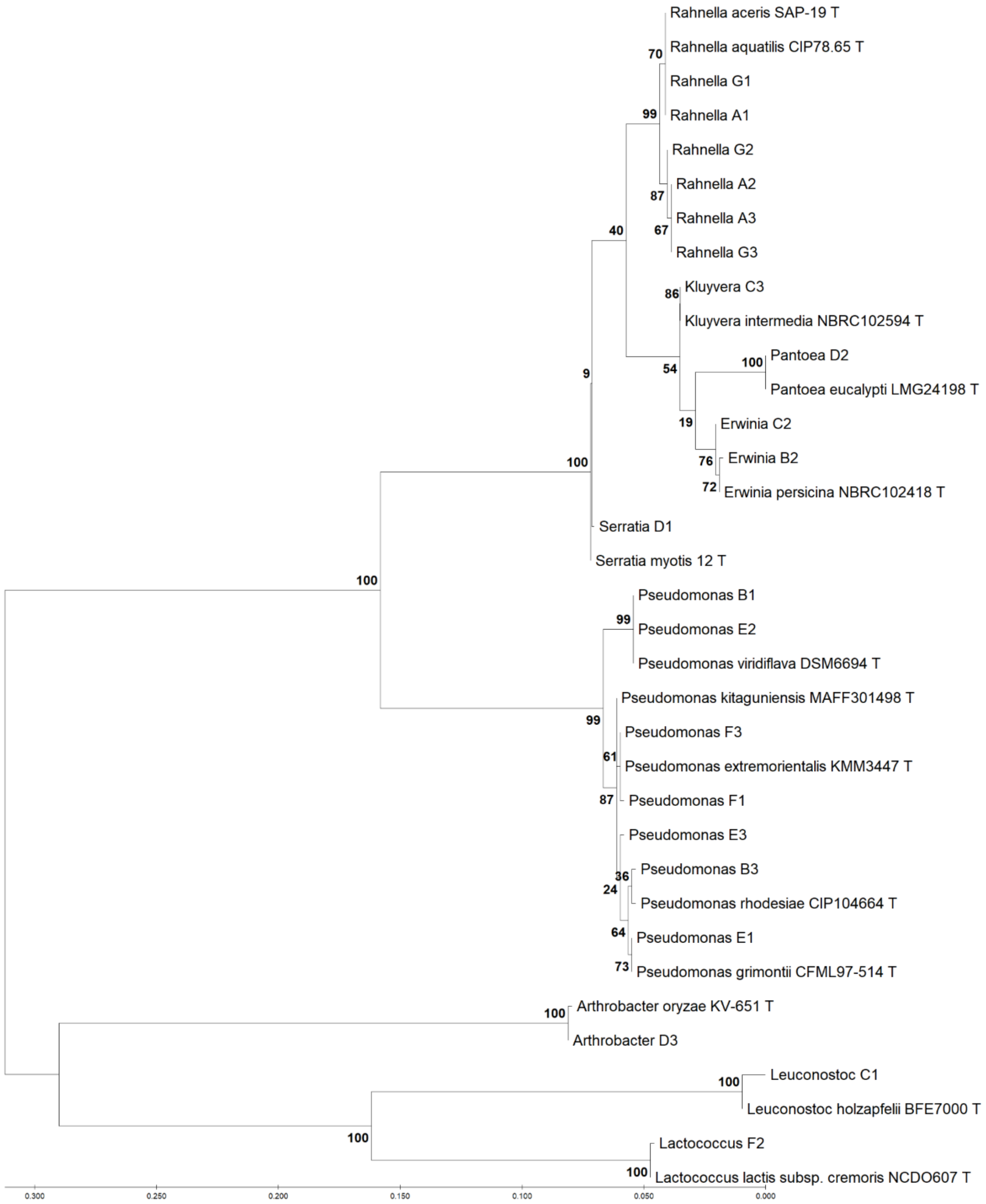

Rahnella aceris SAP-19 T
70
Rahnella aquatilis CIP78.65 T
Rahnella G1
99
Rahnella A1
Rahnella G2
Rahnella A2
87
Rahnella A3
40
67
Rahnella G3
86
Kluyvera C3
Kluyvera intermedia NBRC102594 T
100
Pantoea D2
9
54
Pantoea eucalypti LMG24198 T
19
Erwinia C2
100
76
Erwinia B2
72
Erwinia persicina NBRC102418 T
Serratia D1
Serratia myotis 12 T
100
Pseudomonas B1
99
Pseudomonas E2
Pseudomonas viridiflava DSM6694 T
Pseudomonas kitaguniensis MAFF301498 T
99
Pseudomonas F3
61
Pseudomonas extremorientalis KMM3447 T
87
Pseudomonas F1
Pseudomonas E3
Pseudomonas B3
36
24
Pseudomonas rhodesiae CIP104664 T
64
Pseudomonas E1
73
Pseudomonas grimontii CFML97-514 T
100
Arthrobacter oryzae KV-651 T
Arthrobacter D3
100
Leuconostoc C1
Leuconostoc holzapfelii BFE7000 T
100
Lactococcus F2
100
Lactococcus lactis subsp. cremoris NCDO607 T
0.300
0.250
0.200
0.150
0.100
0.050
0.000

Fig. 3-4. Phylogenetic tree of the isolates from the fresh-cut vegetable including the type strains, based on the partial 16S rRNA gene sequence.

3.3.2 Extracted features obtained from Raman spectra

20 spectra were obtained for each strain. Figure 3-5 showed the preprocessed and averaged Raman spectrua of each strain and the substrate, SUS430. Based on the Raman spectra alone, the only strain that showed a difference was Pantoea D2, while the other strains did not show a difference useful for identification (Fig. 3-5). Figure 3-6 showed a pair plot created based on the LDA value obtained from the Raman spectrum. the pair plot created using LDA1-LDA5 confirmed the cluster of each strain (Fig. 3-6). Figure 3-7 showed the dendrograms created based on LDA1 to LDA10, the close distances in the 16S rRNA classification such as Rahnella sp. (Rahnella A1, Rahnella A2, Rahnella A3, Rahnella G1, Rahnella G2 and Rahnella G3) were close distances in the dendrograms derived from Raman information.

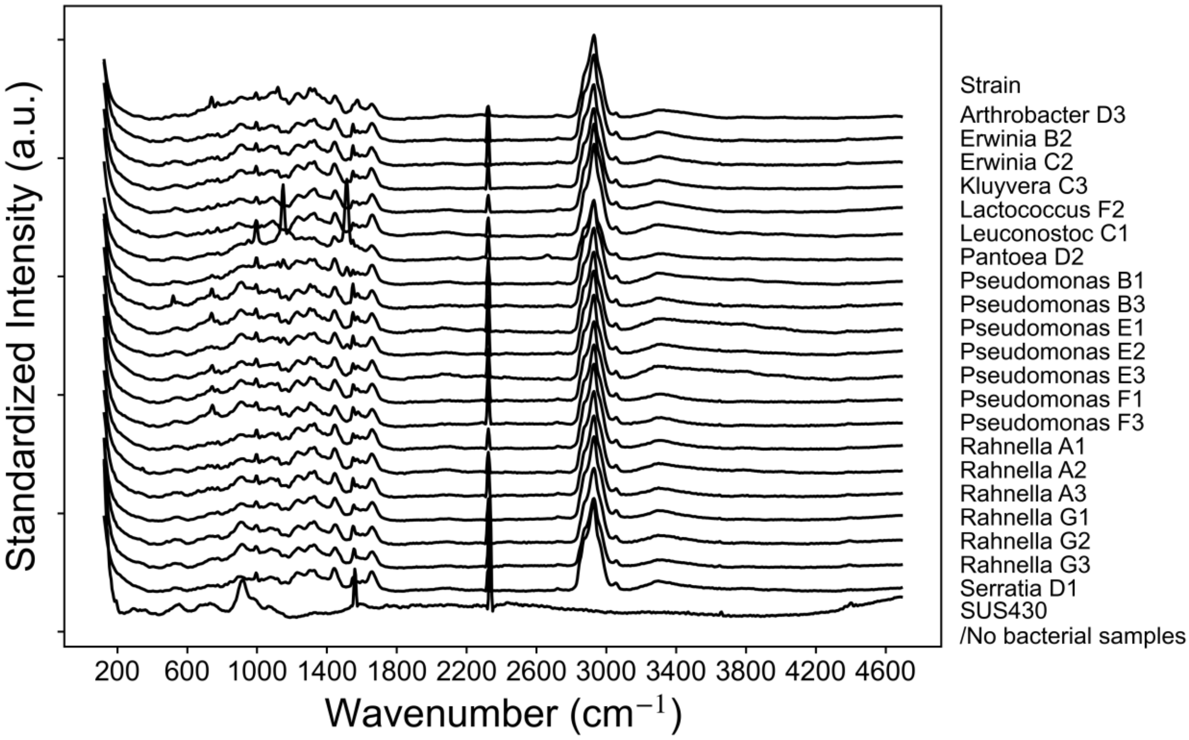


Fig. 3-5. Average spectra of strain after preprocessing that cosmic ray removed, baseline corrected, smoothed and intensity standardized. SUS430 was the result of measurement without bacterial samples. Numbers of spectra are 20 for each sample.

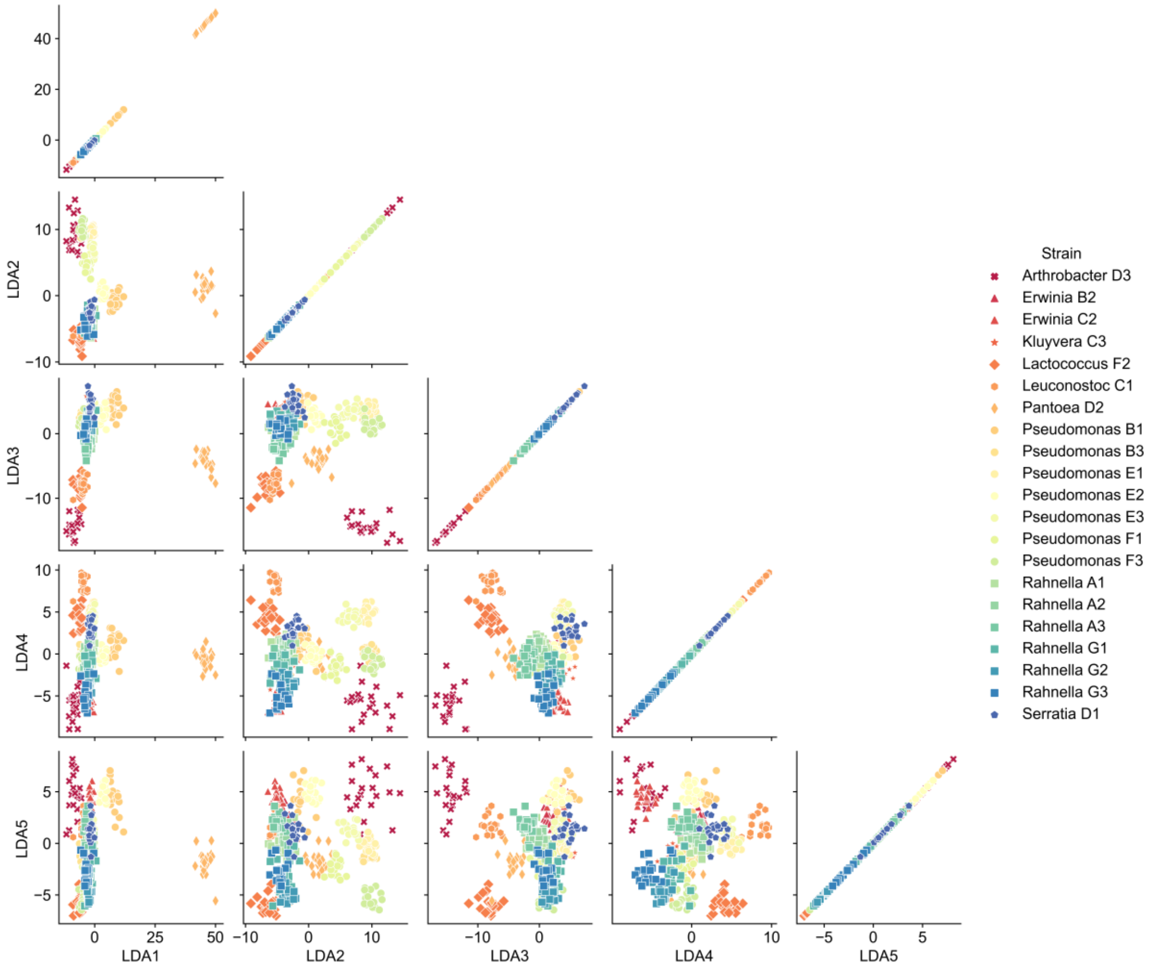


Fig. 3-6. A matrix of scatter plot of LDA 1 to LDA 5 derived from Raman spectra. Color-coded by strain and each symbol have a different type of genus by similarity of 16S rRNA of each isolation (Table 1), as follows, ×: *Arthrobacter* sp., ▲: *Erwinia* sp., ★: *Kluyvera* sp., ◆: *Leuconostoc* sp., ⬢: *Lactococcus* sp., ♦: *Pantoea* sp., ■: *Rahnella* sp., ●: *Pseudomonas* sp., ⬟: *Serratia* sp.

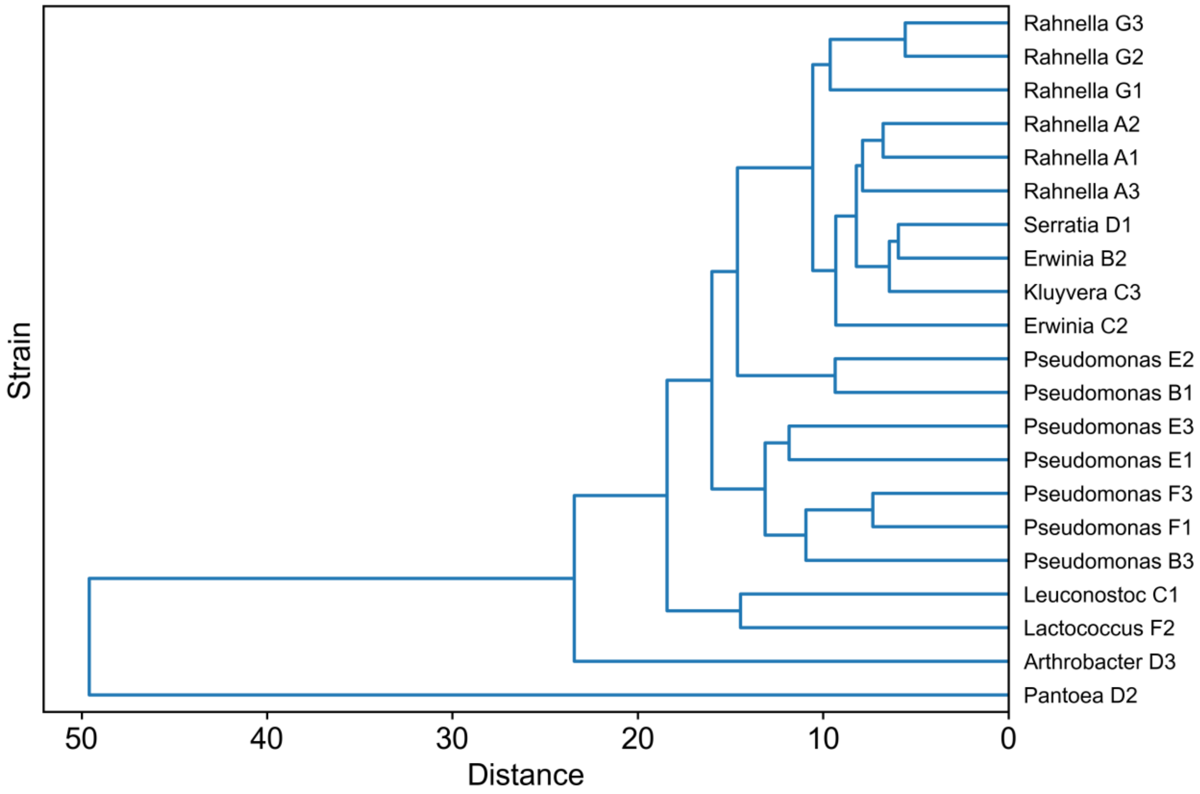


Fig. 3-7. Dendrogram of an agglomerative hierarchical cluster analysis performed on LDA values of Raman spectra of twenty-one strains. Ward's algorithm was used as cluster analysis. The preprocessed spectra were used.

3.3.3 Model performance

Using the results of genus identification by 16S rRNA analysis and growth/no growth data of microorganisms obtained by optical density method, I constructed a model that predicts growth/no growth based on genus, incubation time, sodium acetate concentration, and temperature of microorganisms, and a model in which genus is replaced by Raman spectral information and compared the prediction results when each strain is used as test data.

Table 3-2 summarized the results of calculating each metrics using the predicted values of growth/no growth by classification model with genus name and Raman spectra features as test data for each strain. In the results predicted by the classification model with Raman spectra features, the strains with the highest evaluation was Pseudomonas E2, with accuracy of 1.00 and AUC of 1.00. The lowest evaluation strain was Pantoea D2 with accuracy of 0.78 and AUC of 0.52. For *Erwinia* sp., *Pseudomonas* sp. and *Rahnella* sp., compared the results predicted by the classification model with genus name those predicted by the Raman spectra features model. The difference was less than 0.1 for all strains, and the classification model with Raman spectra features was able to predict the growth/no growth of unknown bacteria with high accuracy. Figure 3-8 and 3-9 shows the prediction of growth probability using classification model with Raman spectra features and the observation of growth/no growth in each condition. Pseudomonas E2 accurately represented the boundary between growth and no growth at both 5°C and 10°C (Fig. 3-8). On the other hand, Pantoea D2, the boundary between growth and no growth was not predicted at both 5°C and 10°C (Fig. 3-9).

Table 3-2. Accuracy and AUC when predicting growth/no growth by classification model with genus name and Raman spectra features. In all 21 strains, 1 strain was used as test data, and the remaining 20 strains were used as training data. The output of the models is probability of growth. The threshold of the probability is 0.5.

Test strains	Classification model with			
	Genus name		Raman spectra features	
	Accuracy	AUC	Accuracy	AUC
Arthrobacter D3	-	-	0.90	0.91
Erwinia B2	0.96	1.00	0.87	0.92
Erwinia C2	1.00	1.00	0.95	0.97
Kluyvera C3	-	-	0.94	0.95
Lactococcus F2	-	-	0.82	0.86
Leuconostoc C1	-	-	0.87	0.88
Pantoea D2	-	-	0.78	0.52
Pseudomonas B1	0.90	0.95	0.93	0.96
Pseudomonas B3	0.96	0.97	0.84	0.83
Pseudomonas E1	0.90	0.90	0.83	0.82
Pseudomonas E2	0.92	0.96	1.00	1.00
Pseudomonas E3	1.00	1.00	0.93	0.97
Pseudomonas F1	0.85	0.85	0.87	0.87
Pseudomonas F3	0.96	0.99	0.86	0.89
Rahnella A1	0.94	0.98	0.95	0.98
Rahnella A2	0.92	0.91	0.95	0.97
Rahnella A3	0.98	1.00	0.95	0.96
Rahnella G1	0.98	1.00	0.96	0.97
Rahnella G2	0.96	0.98	0.96	0.97
Rahnella G3	0.98	0.98	0.96	0.97
Serratia D1	-	-	0.92	0.94

-: Prediction could not be performed for strains that have only one strain in the genus, such as Arthrobacter D3 because they cannot train.

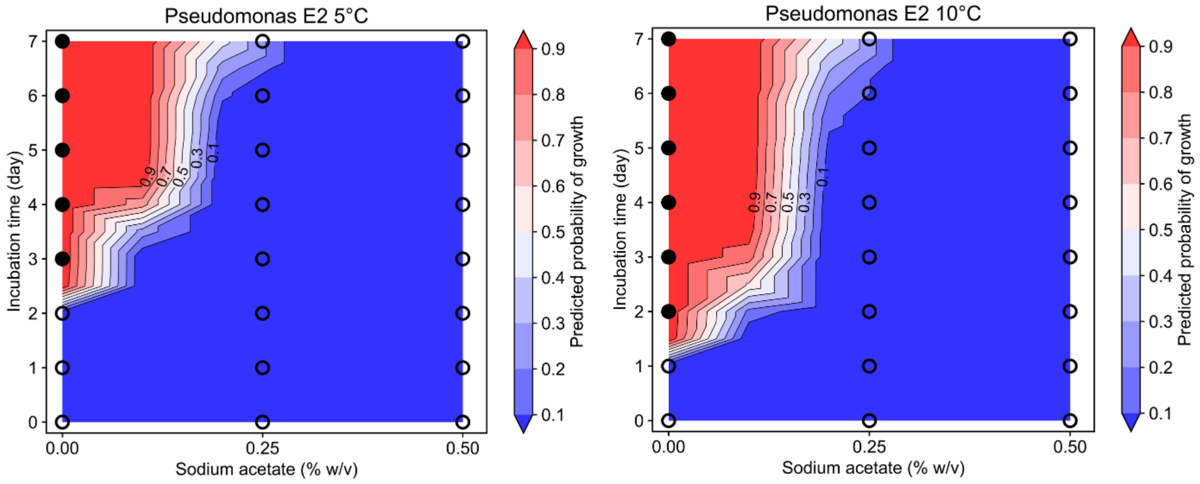


Fig. 3-8. The representative growth probability contour lines of Pseudomonas E2 strain predicted from total 180 conditions using classification model with Raman spectra features and the observation of growth (●)/no growth (○) in TSB various incubation time (0, 1, 2, 3, 4, 5, 6, and 7day), sodium acetate concentration (0, 0.25, and 0.5% w/v) and at 5°C and 10°C. Probability growth was calculated as the average value of the test results for 20 spectra.

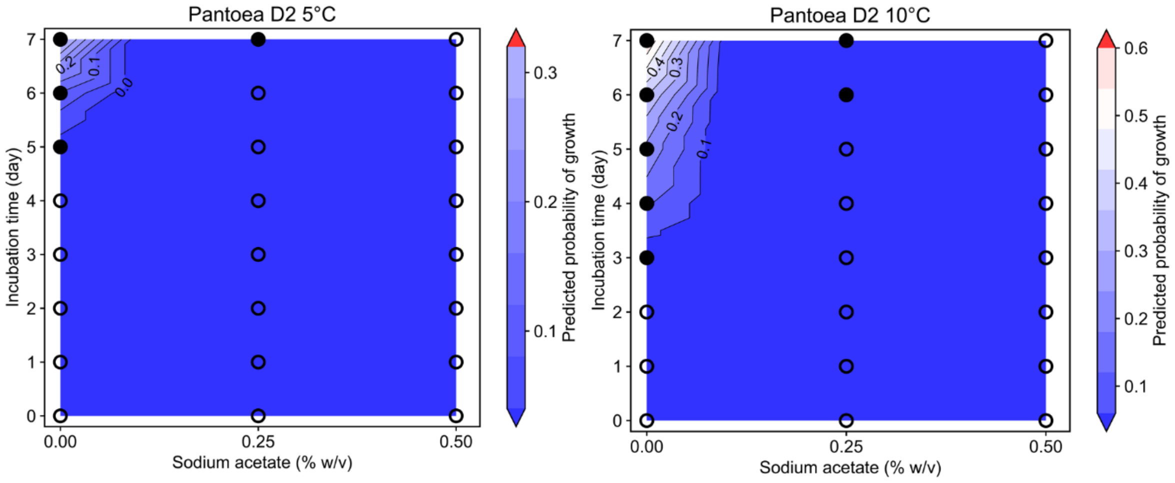


Fig. 3-9. The representative growth probability contour lines of Pantoea D2 strain predicted from total 180 conditions using classification model with Raman spectra features and the observation of growth (●)/no growth (○) in TSB various incubation time (0, 1, 2, 3, 4, 5, 6, and 7day), sodium acetate concentration (0, 0.25, and 0.5% w/v) and at 5°C and 10°C. Probability growth was calculated as the average value of the test results for 20 spectra.

3.4 Discussion

In this study, I developed and evaluated a model to predict the growth behavior of unknown bacterial strains isolated from food using Raman spectroscopy and machine learning. Raman spectra of 21 bacterial strains isolated from cut vegetables (Table 3-1) were measured and feature values were extracted (Figs. 3-5, 3-6, and 3-7). Bacterial growth behavior was evaluated by the optical density method, where the number of incubation time, temperature, and concentration of sodium acetate were used to determine growth/no growth in the liquid medium. Using the acquired data, I developed a machine learning model to predict bacterial growth/no growth. The model predicted the growth/no growth of unknown strains with about 90% accuracy (Table 3-2, Figs. 3-9 and 3-10). Therefore, the combination of Raman spectroscopy and machine learning had the potential to provide the growth behavior of unknown bacteria.

Bacteria were randomly isolated from cut vegetables, resulting in 21 strains in 9 genera (Table 3-1). Thus, since there are various kinds of bacteria in food, identification of the bacteria is necessary at first to consider bacterial control using databases and prediction models. The method developed in this study, which does not require the identification of microorganisms, is an innovative method in the food field where various microorganisms are detected.

Raman spectra can be used to classify microorganisms by species and phylogeny (Yan et al., 2021). As a result of confirming the cluster distance of each strain for the Raman spectra-derived features created in this study, it was similar to the closeness of the distance between closely related species in the results obtained by 16S rRNA analysis (Figs. 3-6 and 3-7). Therefore, it was considered that the features created in this study could be used as potential information to represent the characteristics of each strain.

The new model, in which the identification information required for the previous model was replaced by Raman spectral-derived features, was used to predict the probability of optimal growth according to incubation temperature, sodium acetate concentration, and incubation time, resulting in an AUC of about 0.9 for strains other than Pantoea D2 strain (Table 3-2). The AUC is useful for comparing multiple models (Saito and Rehmsmeier, 2015) and is 0.5 for classifiers whose prediction is random and 1.0 for perfect classifiers (Hanley and McNeil, 1982). If the AUC exceeds 0.9, the model is highly accurate (Fischer et al., 2003). The developed classification model using Raman spectral features has shown the potential to be used for evaluating the growth characteristics of unknown strain. Pantoea D2, had an AUC of 0.5, which was lower than the other strains (Table 3-2). The reason for the low prediction performance was probably due to the fact that the Raman features of Pantoea D2 differed significantly from those of other strains (Fig. 3-7). Therefore, when using this model, it is important to use it together with techniques to visualize Raman features, such as the clustering done in this study (Fig. 3-7), to estimate the reliability of the results.

A unique feature of the model developed in this study is the direct replacement of microbial identification information with Raman spectral information and prediction of unknown strain. There are a lot of reports that Raman spectroscopy can be used for the identification of food-related microorganisms (Huayhongthong et al., 2019; Meisel et al., 2014; Yilmaz et al., 2015). Even if identification is possible, it is difficult to predict growth behavior of unknown strains because there is no database or model. This is because conventional databases and models for predicting microbial behavior require the preparation of growth data for each species in advance in order to make predictions. The model developed in this study showed the possibility of predicting with high accuracy even bacterial species that cannot be predicted by the model using 16S rRNA analysis (Table 3-2). In other words, by simply measuring the Raman spectrum of an unknown strain, it is possible to predict its growth characteristics. This is thought to be because the LDA of the Raman spectrum can be used to summarize the molecular information of a strain that cannot be expressed by a 0 or 1 strain name.

A model for predicting the growth behavior of unknown strains using Raman spectra-derived features has not been reported. It has the potential to solve problems in the field of predictive microbiology, such as the need to identify microorganisms in order to predict their behavior using databases and models, and the difficulty in predicting the growth behavior of unknown strains. In the future, it is expected that the prediction performance and versatility will be improved by expanding the data. In this study, I were able to predict the growth behavior of an unknown strain of bacteria with high accuracy simply by measuring the Raman spectrum. This technology is expected to make it possible to quickly and efficiently determine storage conditions and the amount of additives to use, even in foods such as cut vegetables where many different strains of bacteria are likely to be detected.

Chapter 4

Summary

In this study, I investigated a method for evaluating food related bacterial characteristics by Raman spectra and chemometric analysis.

Chapter 2 presented an approach for evaluating the stress tolerance of bacteria to food additives. Laser Raman microscopy was used to obtain Raman spectra of six food-associated bacteria. Each bacterium was inoculated and cultured in liquid medium under the conditions of different concentrations of NaCl, sodium acetate, and glycine. Using the obtained Raman spectral and the grouping information of bacteria by stress tolerance, I constructed a classification model using a support vector machine. The classification model was able to classify the groups of bacterial stress tolerance with an accuracy of about 90%. The result indicated that Raman spectra and machine learning can evaluate the resistance of bacteria in food to stresses related to their growth. This technology is expected to be used for efficient determination of the amount of additives in food to prevent food spoilage.

In chapter 3, unknown bacterial species were investigated for evaluating the growth/no growth in some conditions by Raman spectra and chemometric analysis. Twenty-one strains were isolated from cut vegetables, and the features of each strain were summarized by the observed Raman spectra using the LDA. The growth behavior of each strain was evaluated by inoculating the bacterial cells into liquid medium with different concentrations of sodium acetate and evaluating the growth/no growth at various incubation temperatures and incubation periods using the optical density method. The Raman spectra and growth behaviors were trained into a neural network model to predict the growth/no growth of an unknown strain. As a result, the developed machine learning model was able to predict the growth behavior of an unknown strain with an accuracy of about 90%. The result demonstrated that the developed approach would be applied to evaluating characteristic of unknown bacteria. Although unknown bacteria are often detected from foods in the market, this technology enables to estimate the growth behavior of unknown bacterial strains simply by measuring their Raman spectra. Therefore, this technology is expected to immediately provide useful information for microbial control of foods.

The results of my investigation showed that Raman spectra and chemometric analysis can be used to predict bacterial characteristics useful for the control of food-related bacteria, such as their tolerance to food additives and growth behavior in various environments. My developed methods would be useful for quickly setting up manufacturing and storage conditions to prevent food quality deterioration.

.

References

Baranyi, J., 2006. Using the ComBase database and associated software tools to predict microbial responses to food environments. Food Manufacturing Efficiency 1, 13.

Baranyi, Jozsef, Tamplin, M., 2004. ComBase: A common database on microbial response to food environments. J. Food Prot. 67, 1834–1840.

Baranyi, József, Tamplin, M.L., 2004. ComBase: a common database on microbial responses to food environments. J. Food Prot. 67, 1967–1971. https://doi.org/10.4315/0362-028x-67.9.1967

Begley, M., Hill, C., 2015. Stress adaptation in foodborne pathogens. Annu. Rev. Food Sci. Technol. 6, 191–210. https://doi.org/10.1146/annurev-food-030713-092350

Black, E.P., Hinrichs, G.J., Barcay, S.J., Gardner, D.B., 2018. Fruit Flies as Potential Vectors of Foodborne Illness. J. Food Prot. 509–514. https://doi.org/10.4315/0362-028X.JFP-17-255

Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J., 2000. LOF: identifying density-based local outliers, in: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, SIGMOD '00. Association for Computing Machinery, New York, NY, USA, pp. 93–104. https://doi.org/10.1145/342009.335388

Chollet, F., Others, 2015. Keras [WWW Document]. URL https://keras.io

de Biasio, M., McGunnigle, G., Leitner, R., Popp, J., Rösch, P., Balthasar, D., 2013. Identification of single bacteria using micro Raman spectroscopy, in: 2013 Seventh International Conference on Sensing Technology (ICST). ieeexplore.ieee.org, pp. 34–39. https://doi.org/10.1109/ICSensT.2013.6727612

References

de Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y., 2005. A Tutorial on the Cross-Entropy Method. Ann. Oper. Res. 134, 19–67. https://doi.org/10.1007/s10479-005-5724-z

Dengremont, E., Membré, J.M., 1995. Statistical approach for comparison of the growth rates of five strains of Staphylococcus aureus. Appl. Environ. Microbiol. 61, 4389–4395. https://doi.org/10.1128/aem.61.12.4389-4395.1995

Fao, I., 2019. The state of food and agriculture 2019. Moving forward on food loss and waste reduction. FAO, Rome 2–13.

Fischer, J.E., Bachmann, L.M., Jaeschke, R., 2003. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. Intensive Care Med. 29, 1043–1051. https://doi.org/10.1007/s00134-003-1761-8

Fisher, R.A., 1938. The statistical utilization of multiple measurements. Ann. Eugen. 8, 376–386. https://doi.org/10.1111/j.1469-1809.1938.tb02189.x

Fung, F., Wang, H.-S., Menon, S., 2018. Food safety in the 21st century. Biomed. J. 41, 88–95. https://doi.org/10.1016/j.bj.2018.03.003

Germond, A., Ichimura, T., Horinouchi, T., Fujita, H., Furusawa, C., Watanabe, T.M., 2018. Raman spectral signature reflects transcriptomic features of antibiotic resistance in Escherichia coli. Commun Biol 1, 85. https://doi.org/10.1038/s42003-018-0093-8

Gould, G.W., 1996. Methods for preservation and extension of shelf life. Int. J. Food Microbiol. 33, 51–64. https://doi.org/10.1016/0168-1605(96)01133-6

Gram, L., Ravn, L., Rasch, M., Bruhn, J.B., Christensen, A.B., Givskov, M., 2002. Food spoilage—interactions between food spoilage bacteria. Int. J. Food Microbiol. 78, 79–97. https://doi.org/10.1016/S0168-1605(02)00233-7

References

Hajmeer, M., Basheer, I., 2002. A probabilistic neural network approach for modeling and classification of bacterial growth/no-growth data. J. Microbiol. Methods 51, 217–226. https://doi.org/10.1016/s0167-7012(02)00080-5

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Hishinuma, F., Izaki, K., Takahashi, H., 1969. Effects of Glycine and d-Amino Acids on Growth of Various Microorganisms. Agric. Biol. Chem. 33, 1577–1586. https://doi.org/10.1080/00021369.1969.10859511

Hiura, S., Koseki, S., Koyama, K., 2021. Prediction of population behavior of Listeria monocytogenes in food using machine learning and a microbial growth and survival database. Sci. Rep. 11, 10613. https://doi.org/10.1038/s41598-021-90164-z

Ho, C.-S., Jean, N., Hogan, C.A., Blackmon, L., Jeffrey, S.S., Holodniy, M., Banaei, N., Saleh, A.A.E., Ermon, S., Dionne, J., 2019. Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. Nat. Commun. 10, 4927. https://doi.org/10.1038/s41467-019-12898-9

Houtsma, P.C., de Wit, J.C., Rombouts, F.M., 1993. Minimum inhibitory concentration (MIC) of sodium lactate for pathogens and spoilage organisms occurring in meat products. Int. J. Food Microbiol. 20, 247–257. https://doi.org/10.1016/0168-1605(93)90169-h

Huayhongthong, S., Khuntayaporn, P., Thirapanmethee, K., Wanapaisan, P., Chomnawang, M.T., 2019. Raman spectroscopic analysis of food-borne microorganisms. LWT 114, 108419. https://doi.org/10.1016/j.lwt.2019.108419

Inatsu, Y., Weerakkody, K., Bari, M.L., Hosotani, Y., Nakamura, N., Kawasaki, S., 2017. The efficacy of combined (NaClO and organic acids) washing treatments in controlling Escherichia coli O157: H7, Listeria monocytogenes and spoilage bacteria on shredded cabbage and bean sprout. LWT-Food Science and Technology 85, 1–8.

Jaafreh, S., Valler, O., Kreyenschmidt, J., Günther, K., Kaul, P., 2019. In vitro discrimination and classification of Microbial Flora of Poultry using two dispersive Raman spectrometers (microscope and Portable Fiber-Optic systems) in tandem with chemometric analysis. Talanta 202, 411–425. https://doi.org/10.1016/j.talanta.2019.04.082

Jones, E., Oliphant, T., Peterson, P., Others, 2016. SciPy: Open source scientific tools for Python, 2001.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111–120. https://doi.org/10.1007/BF01731581

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. arXiv [cs.LG].

Knight, K.P., McKellar, R.C., 2007. Influence of cinnamon and clove essential oils on the D- and z-values of Escherichia coli O157:H7 in apple cider. J. Food Prot. 70, 2089–2094. https://doi.org/10.4315/0362-028x-70.9.2089

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol. Biol. Evol. 35, 1547–1549. https://doi.org/10.1093/molbev/msy096

Kuroda, S., Okuda, H., Ishida, W., Koseki, S., 2019. Modeling growth limits of Bacillus spp. spores by using deep-learning algorithm. Food Microbiol. 78, 38–45. https://doi.org/10.1016/j.fm.2018.09.013

References

Lee, H., Song, J., 2019. Introduction to convolutional neural network using Keras; an understanding from a statistician. Communications for Statistical Applications and Methods 26, 591–610. https://doi.org/10.29220/CSAM.2019.26.6.591

Leistner, L., 2000. Basic aspects of food preservation by hurdle technology. Int. J. Food Microbiol. 55, 181–186. https://doi.org/10.1016/s0168-1605(00)00161-6

Lieber, C.A., Mahadevan-Jansen, A., 2003. Automated method for subtraction of fluorescence from biological Raman spectra. Appl. Spectrosc. 57, 1363–1367. https://doi.org/10.1366/000370203322554518

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, Xiaoqiang Zheng, 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.

Meisel, S., Stöckel, S., Elschner, M., Melzer, F., Rösch, P., Popp, J., 2012. Raman spectroscopy as a potential tool for detection of Brucella spp. in milk. Appl. Environ. Microbiol. 78, 5575–5583. https://doi.org/10.1128/AEM.00637-12

Meisel, S., Stöckel, S., Rösch, P., Popp, J., 2014. Identification of meat-associated pathogens via Raman microspectroscopy. Food Microbiol. 38, 36–43. https://doi.org/10.1016/j.fm.2013.08.007

Moreira, L.M., Silveira, L., Jr, Santos, F.V., Lyon, J.P., Rocha, R., Zângaro, R.A., Villaverde, A.B., Pacheco, M.T.T., 2008. Raman spectroscopy: A powerful technique for biochemical analysis and diagnosis. Spectrosc. Int. J. 22, 1–19. https://doi.org/10.1155/2008/942758

Møretrø, T., Langsrud, S., 2017. Residential Bacteria on Surfaces in the Food Industry and Their Implications for Food Safety and Quality. Compr. Rev. Food Sci. Food Saf. 16, 1022–1041. https://doi.org/10.1111/1541-4337.12283

Nair, V., Hinton, G.E., 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. Icml.

Nanasombat, S., Chooprang, K., 2009. Control of Pathogenic Bacteria in Raw Pork using Organic Acid Salts in Combination with Freezing and Thawing. Agriculture and Natural Resources 43, 576–583.

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., Others, 2019. Keras Tuner. Github. [(accessed on 31 January 2021)].

Roda, A., Mirasoli, M., Roda, B., Bonvicini, F., Colliva, C., Reschiglian, P., 2012. Recent developments in rapid multiplexed bioanalytical methods for foodborne pathogenic bacteria detection. Microchim. Acta 178, 7–28. https://doi.org/10.1007/s00604-012-0824-3

Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10, e0118432. https://doi.org/10.1371/journal.pone.0118432

Savitzky, A., Golay, M.J.E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal. Chem. 36, 1627–1639. https://doi.org/10.1021/ac60214a047

Schleifer, K.H., 2009. Classification of Bacteria and Archaea: past, present and future. Syst. Appl. Microbiol. 32, 533–542. https://doi.org/10.1016/j.syapm.2009.09.002

References

Schumacher, W., Stöckel, S., Rösch, P., Popp, J., 2014. Improving chemometric results by optimizing the dimension reduction for Raman spectral data sets. J. Raman Spectrosc. 45, 930–940. https://doi.org/10.1002/jrs.4568

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.

Stavropoulou, E., Bezirtzoglou, E., 2019. Predictive Modeling of Microbial Behavior in Food. Foods 8. https://doi.org/10.3390/foods8120654

Strola, S.A., Baritaux, J.-C., Schultz, E., Simon, A.C., Allier, C., Espagnon, I., Jary, D., Dinten, J.-M., 2014. Single bacteria identification by Raman spectroscopy. J. Biomed. Opt. 19, 111610. https://doi.org/10.1117/1.JBO.19.11.111610

Taylor, J.N., Mochizuki, K., Hashimoto, K., Kumamoto, Y., Harada, Y., Fujita, K., Komatsuzaki, T., 2019. High-Resolution Raman Microscopic Detection of Follicular Thyroid Cancer Cells with Unsupervised Machine Learning. J. Phys. Chem. B 123, 4358–4372. https://doi.org/10.1021/acs.jpcb.9b01159

Walls, I., Scott, V.N., 1997. Use of predictive microbiology in microbial food safety risk assessment. Int. J. Food Microbiol. 36, 97–102. https://doi.org/10.1016/s0168-1605(97)01260-9

Ward, J.H., 1963. Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. 58, 236–244. https://doi.org/10.1080/01621459.1963.10500845

Yamamoto, T., Taylor, J.N., Koseki, S., Koyama, K., 2021. Classification of food spoilage bacterial species and their sodium chloride, sodium acetate and glycine tolerance using chemometrics analysis and Raman spectroscopy. J. Microbiol. Methods 190, 106326. https://doi.org/10.1016/j.mimet.2021.106326

Yan, S., Wang, S., Qiu, J., Li, M., Li, Dezhi, Xu, D., Li, Daixi, Liu, Q., 2021. Raman spectroscopy combined with machine learning for rapid detection of food-borne pathogens at the single-cell level. Talanta 226, 122195. https://doi.org/10.1016/j.talanta.2021.122195

Yilmaz, A.G., Temiz, H.T., Acar Soykut, E., Halkman, K., Boyaci, I.H., 2015. Rapid Identification of P seudomonas aeruginosa and P seudomonas fluorescens Using R aman Spectroscopy. J. Food Saf. 35, 501–508.

Yoon, S.-H., Ha, S.-M., Kwon, S., Lim, J., Kim, Y., Seo, H., Chun, J., 2017. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int. J. Syst. Evol. Microbiol. 67, 1613–1617. https://doi.org/10.1099/ijsem.0.001755

Yu, S., Li, H., Li, X., Fu, Y.V., Liu, F., 2020. Classification of pathogens by Raman spectroscopy combined with generative adversarial networks. Sci. Total Environ. 726, 138477. https://doi.org/10.1016/j.scitotenv.2020.138477

幸恵関口, 2015. MALDI-TOF MSによる微生物同定の現状と活用にあたっての留意点. 腸内細菌学雑誌 29, 169−176. https://doi.org/10.11209/jim.29.169

Development of a characterization method for food-related bacteria using Raman spectroscopy and machine learning

March, 2022

Takashi Yamamoto

Laboratory of Agricultural & Food Process Engineering
Graduate School of Agricultural Science
Hokkaido University
Kita-9, Nishi-9, Kita-ku, Sapporo, Hokkaido, Japan