Discrimination of Clinical Microorganism using Raman Spectroscopy and Pattern Recognition Techniques

Chi-Chang Lin

***Abstract*—The raised of prime interest on the bacteria detections are important because these biological systems normally have their essential functions in nature, however, some harmful micro-organisms may have negative effects associated with animals, human and also environments. An accurate, high sensitivity and rapid identification assay of cells is extremely important in areas such as medical diagnosis, biological research, and environmental monitoring. A direct, sensitive, and rapid method for the detection of clinical microorganism, including bacteria and phages, is demonstrated using Raman spectroscopy combined with pattern recognition techniques (PR for short). A preliminary study of the data set structure was performed using following steps procedures: data pre-treatment by exploratory data analysis (EDA) and standard normal variant (SNV), variable selection and reduction, and pattern recognition technique including Genetic Algorithm (GA), k-nearest neighbour (k-NN) and Partial least squares discriminant analysis (PLS-DA). The new processing algorithm is applied to three bacteria and their related phages (*Salmonella, Acinetobacter baumannii, Klebsiella pneumoniae*) with excellent results. The uniqueness of the selected variable an iterative random subsample, leave-one-out cross-validation (LOOCV) procedure is further performed. The Raman fingerprint combination with pattern recognition techniques showed prefect classification of microorganism with respect to clinical rapid diagnostics (LOOCV for bacteria, phages, and bacteria mixed with phages are 89.22%, 64.44% and 99.99%, respectively). This study represents a basic research tool that may allow researchers and medical doctors to rapidly detect, quantify, and classify bacterial type at subspecies and strain levels.**

***Index Terms*—Raman Spectroscopy, Clinical microorganism, Pattern recognition, Genetic algorithm, Partial least squares**

I. Introduction

A direct, sensitive, and rapid method for clinical detection and identification is an important basis of providing patients with medical diagnosis. Laboratory examinations of clinical isolates such as *S. aureus* from bodily fluids and *H. pylori* from duodenal ulcer specimens currently require highly time-consuming and complex processes; typically 10^6–10^8 cells are needed for the characterization of strains. In deed, the traditional way is nothing more than high specificity of species identification or biochemical tests such as Polymerase Chain Reaction (PCR), antigen-antibody reaction, etc. In recent years, a quantum jump progress has been made in rapid detection combined with nanotechnology and biomedical engineering. Detection tools including ultraviolet–visible spectroscopy, Fourier transform infrared spectroscopy, fluorescence spectroscopy, and nuclear magnetic resonance spectroscopy, are very common and these forward-looking technologies mentioned above plays an extremely important role not just in the fields of medical diagnosis, biological research but also pharmaceutical, food science, and environmental monitoring. However, they still exists some shortcomings, for example, *in vivo* experiments using fluorescence spectroscopy needs the target molecular reagents, FTIR is not easy to apply to the active species analysis due to interference of solution, PCR is fast, but the defects are complex steps and non-target amplification. Specimens for microbial culture from lesions are also a difficult procedure and costs time-consuming that cannot obtain information immediately. Therefore, an accurate, highly sensitivity and rapid identification assay for bacteria identification is extremely important [1, 2].

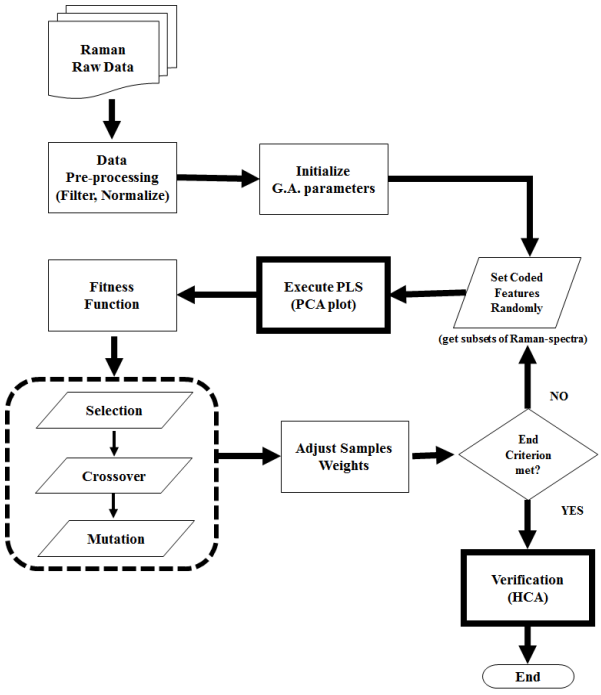

Fig. 1. The block diagram of pattern recognition.

Manuscript received March 01, 2017; revised March 30, 2017. This work was supported in part by the Ministry of Science and Technology, R.O.C. under Grant 104-2632-M-029-001.

Dr. C. C. Lin is with the Department of Chemical and Materials Engineering, Tunghai University, Taiwan (corresponding author: 886-4-23590262 ext210; e-mail: chichang31@ thu.edu.tw).

ISBN: 978-988-14048-3-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Since Raman spectroscopy has the advantages of high sensitivity, non-destructive, non-invasive detection, and highly specific like a chemical fingerprint of a material and do not interfered by water for extracting useful information, make it become a remarkable novel instrument [3]. Previous studies have reported that using Raman spectroscopy to detect blood glucose concentration [4, 5]. Liu, R has reported *in-vivo* analysis of cells and microorganisms [6]. Briefly, Raman spectroscopy can provide a qualitative and quantitative comprehensive chemical signals on the molecular level of strains, but high complexity of the clinical microorganisms and their different hosting environment makes more difficult to discriminate the Raman fingerprints. In this study, the use of mathematical and statistical procedures in pattern recognition, which tries to construct a classifier that can classify implicit or unseen information based on the different strain spectral fingerprint and can be analyzed in relatively short time [7, 8].

II. Experimental Procedure

A. Microorganism samples

Strain selection for clinical infection including three different types of bacteria (*Salmonella, Acinetobacter baumannii, Klebsiella pneumoniae*) in order to facilitate identification, bacteria were named ED2, KM18 and NK5 respectively.

B. Spectra Acquisition

Bacteria were taken from the culture and centrifuged at speed 18,000 rpm for 30 minutes. Then taken out the bacteria from bottom solution and 0.002 mg/mL of sample was added for SERS. Raman spectra were measured on i-Raman system produced by BWTEK. The radiation of excitation was 785 nm, laser power was 12 mW and the spectra were measured at a resolution of 3 cm^{-1}. Each spectrum was an average of 2 accumulation for 60s exposure time and was stored from Raman shift 400 to 1800 cm^{-1}.

C. Data Processing and Pattern recognition

In general, data pre-treatment is needed prior to the application of multivariate data analysis techniques. Each raw data of fingerprint was pretreated by Savitzky-Golay filter for reducing the noise in a 5-point smoothing function [12]. To eliminate systematic differences among measurement, the same spectral region (Raman shift) chosen for the SNV transformation is performed on an individual sample basis, and the same calculation is performed on every single spectrum. Finally, because principal components maximize variance, the main EDA technique used here is PCA, which is mostly used as a tool in order to detect patterns in the measured data. After the above steps, a few outliers were found and removed prior to classify.

Variable selection is a preliminary step used in multivariate data analysis, particularly if the number of samples is relatively small, the number of variables is large and many of these variables contain redundant or noisy information. In this paper, a global optimization strategy can be applied for extracting features or variable selection (intensities at specific Raman-shift) purposes: Genetic-Algorithms (GA). The GA generates N populations (or called *chromosome*) of the same length with 0s and 1s binary sequences randomly, each of which represents a possible solution. Every fingerprint of Raman shift will be included in the subset if the corresponding bit (*gene*) is set to 1, or the Raman shift will not be included. Then, do PLS for all subsets, i.e., N chromosomes will generate a total of N PLS-principal components plots [10, 11].

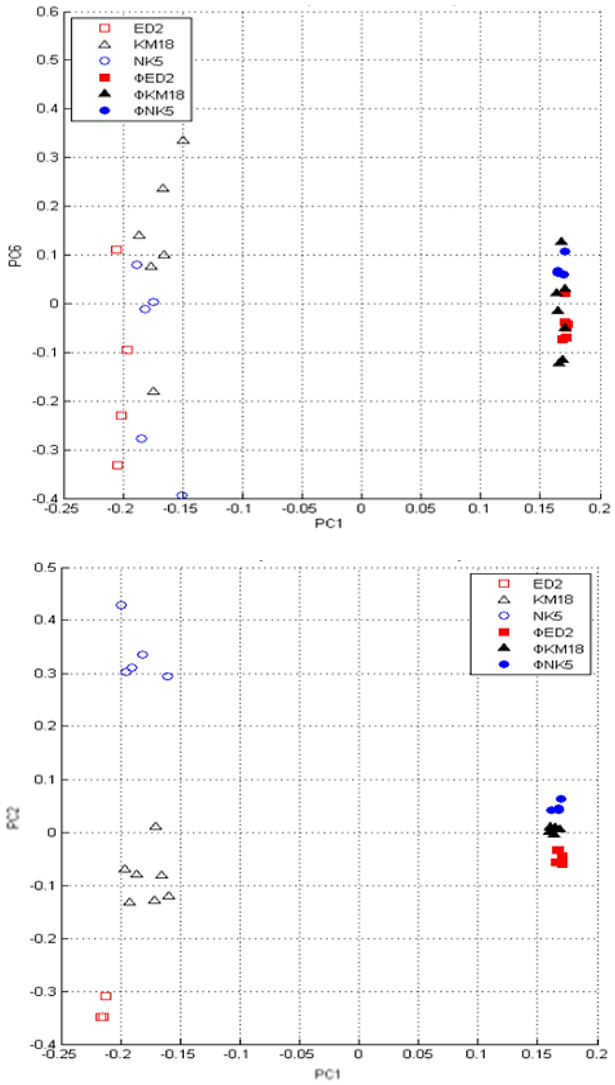

Fig. 2. The left figure shows the PCs plot before the calculation of our approach. The right figure shows the PCs plot after GA calculation. For each pair of chromosomes selected for crossover, two new offspring are generated using a three-point crossover and mutation probability set at 1%.

A PLS model will try to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space. In other words, the principle of PLS is to find the components in the independent variables (X) that describe as much as possible the relevant variations in the input variables and at the same time have maximal correlation with the dependent variables (Y). To facilitate the tracking and scoring of the principal component plots, the fitness function (here defined as f(d) cost) of GA was assessed the quality of the chromosome [12]. Each principal component plot generated for each feature subset is scored

ISBN: 978-988-14048-3-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

using the *k*-nearest neighbor (*k*-NN) classification algorithm. The whole evaluation is an iterative process, and the population in each iteration called a *generation*.

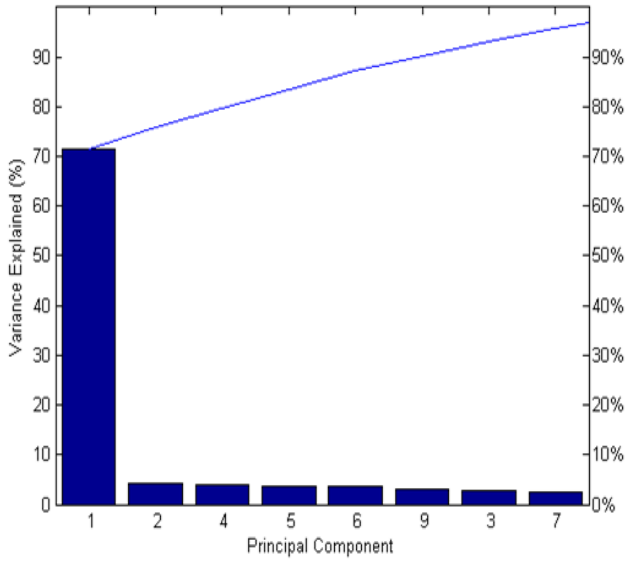

Fig. 3. The principal components of the data explain of the total cumulative variance.

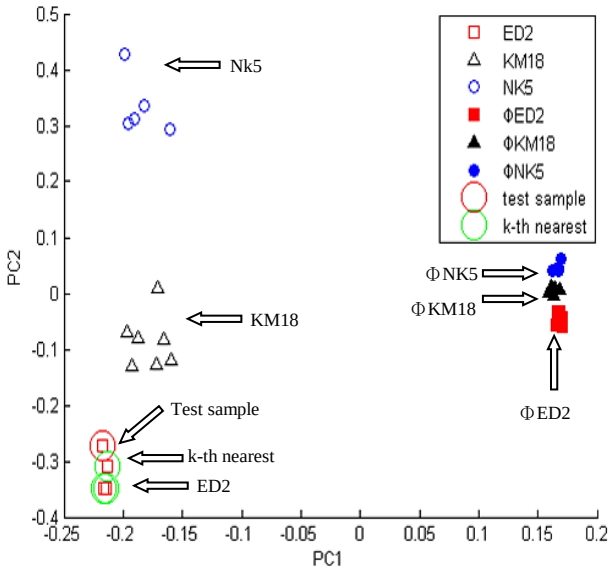

Fig. 4. An example of predict ability of classified model.

In each generation, the process including *selection*, *crossover*, and *mutation* is performed through repetitive steps until the fitness value equals 100 or a specified number of generation has been met. After that, the classified Raman fingerprint of microorganism will be shown in the form of PLS plot and hierarchical clustering analysis. Moreover, the leave-one-out cross-validation (LOOCV) procedure is further performed [13].

One of the important aspects for assessment of pattern recognition is how the results of a statistical analysis will generalize to an independent data set. A variant of k-fold cross-validation is the so-called leave-one-out cross-validation, which removes only one sample at a time from the training set and considers it as a test set. The LOOCV model allows illustrating that the model obtained by the pattern recognition technique is good enough to perform classification of unknown samples. Because GAs may have a tendency to converge towards local optima or even arbitrary points rather than the global optimum of the problem, i.e., the optimal solution of chromosome may not always be exactly the same, we try to repeat LOOCV 10 times in order to assess the overall reliability of training modules. In our approach, the algorithm flowchart is defined in Fig. 1.

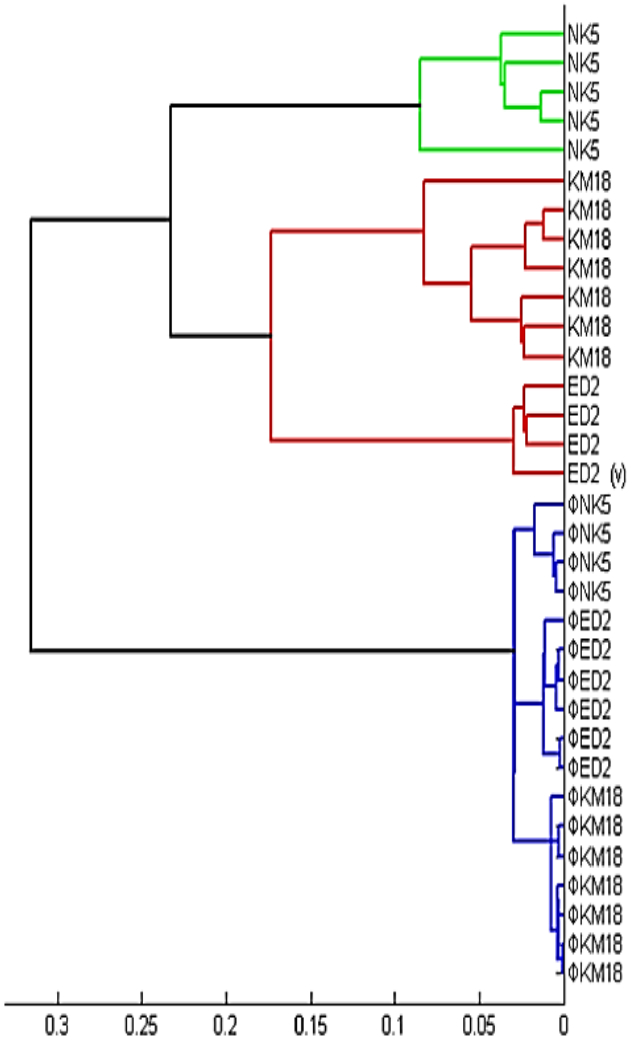

Fig. 5. The hierarchical clustering analysis result. The X-axis is Euclidean distance between each sample from PCs plot.

III. Result and Discussion

Six different variety of clinical microorganism were collected, and Raman fingerprint samples were measured 7 times for each strain. They entirely contained 42 samples. After procedures of pretreatment, two outliers were detected from SNV, one of the class ΦED2 and the other of the class ΦNK5, and deleted. Another seven outliers were removed by EDA, three for the class ED2, two for the class NK5 and two for the class ΦNK5. As can be seen clearly, after elimination of the outliers, the final data set consisted of a total of 33 samples: 4 sample of the class ED2, 7 of the class KM18, 5of

ISBN: 978-988-14048-3-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

the class NK5, 6 of the class ΦED2, 7 of the class ΦKM18, and 4 of the class ΦNK5. The following results are demonstrated using different bacteria mixed with its related bacteria phages.

The plot of the two largest principal components (defined by the first and second PCs of the normalized training set) of the 32 points Raman fingerprints of the microorganism that before and after GA calculation shows in figure 2. The Raman fingerprint combination with multiple pattern recognition techniques shows prefect classification results of six microorganisms.

Figure 3 illustrates that the two largest principal components of the data explain 74% of the total cumulative variance. Figure 4 shows an example of predict ability of our classified model. For LOOCV, this procedure can be treated as a blind testing, one of sample ED2 considered as a test set which is red circle in the figure. The green circles are kth nearest sample of the test sample ED2. Because there are three ED2 in the training set, the data set using the decision criteria k is equals 3. If test sample ED2 has its nearest neighbors 3 ED2 class, then defined prediction ability 3/3•100 = 100%.

The hierarchical clustering analysis result is next performed from the PCs plot (Fig. 5). It can be obviously seen that the microorganisms is divided into six partitions in HCA. The label ED2 affixed with (v) denotes test sample. HCA results are summarized in Table 1. The prediction abilities for the class ED2, NK5, ΦED2 and ΦNK5 are excellent (100%), but less accurate for the class KM18 (86%) and the class ΦKM18 (71%).

The classification error, which were divided into false negative, for example, a sample KM18 has been wrongly assigned to class ED2, and false positive, namely, one sample has been wrongly classified to class ED2, one to class ΦED2, and one to class ΦNK5. It means that 3.4% (1/29) of the samples were accepted as ED2 or ΦNK5, and 3.7% (1/27) of the samples were wrongly considered as ΦED2.

Table 1. Prediction ability results of HCA classification of different bacteria and bacteria phages.

	ED2	KM18	NK5	ΦED2	ΦKM18	ΦNK5	False neg.
ED2	4 (100%)						0/4
KM18	1	6 (86%)					1/7
NK5			5 (100%)				0/5
ΦED2				6 (100%)			0/6
ΦKM18				1	5 (71%)	1	2/7
ΦNK5						4 (100%)	0/4
False pos.	1/29	0/26	0/28	1/27	0/28	1/29	

IV. Conclusion

A direct, sensitive, and rapid method for the detection of clinical microorganism is demonstrated using their Raman spectroscopy combined with pattern recognition techniques. The present result shows that both bacteria and related bacteria phages can be discriminated by their Raman fingerprints.

Moreover, this study represents a high sensitivity Raman spectroscopy-based analysis tool that may allow researchers and medical doctors to identify and classify clinical bacterial Raman fingerprints at subspecies and strain levels more rapidly and easily.

Finally, about 100% of classification can be achieved between Raman fingerprints of the bacteria and its related bacteria phages.

Acknowledgment

We thank Dr. Ming-Tse Kuo (Department of Ophthalmology, Kaohsiung, Chang Gung Memorial Hospital) for his discussion on bacteria and statistical analysis methods. We also thank Mr. Cheng-Yao Lin and Mr. Chi–Jie Kao (Department of Chemical Engineering and Materials Engineering, Tunghai University) for their supports on both software processing and discussions of Raman spectroscopy techniques.

References

[1] D. I. Ellis, R. Goodacre, "Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy," *Analyst* vol. 131, no. 8, pp. 875-885, 2006.

[2] Y.S. Huang, T. Karashima, M. Yamamoto, H.O. Hamaguchi, "Molecular-level investigation of the structure, transformation, and bioactivity of single living fission yeast cells by time- and space-resolved Raman spectroscopy," *Biochemistry* vol. 44, no. 30, pp. 10009-10019, 2005.

[3] S. Efrima, L. Zeiri, "Understanding SERS of bacteria," *J. Raman Spectrosc.* vol. 40, no. 3, pp. 277-288, 2009.

[4] Z.Q. Niu, Y. Fang, "Surface-enhanced Raman scattering of single-walled carbon nanotubes on silver-coated and gold-coated filter paper," J. *Colloid Interf. Sci.* vol. 303, no. 1, pp. 224-228, 2006.

[5] M.T. Kuo, C.C. Lin, H.Y. Liu, M.Y. Yang, H.C. Chang, "Differentiation between infectious and noninfectious ulcerative keratitis by Raman spectra of human teardrops: a pilot study," *Invest. Ophthalmol. Vis. Sci.* vol. 53, no. 3, pp. 1436-1444, 2012.

[6] R. Liu, J. F. Liu, "Applications of Raman based techniques to on-site and in-vivo analysis," *Trends Anal. Chem.* vol. 30, 1462-1476, 2011.

[7] C.-C. Lin, C.-W. Chang, "AuNPs@mesoSiO$_2$ composites for SERS detection of DTNB molecule," *Bio. Bioelect.* vol. 51, pp. 297-303, 2014.

[8] C.-C. Lin, Y.-M. Yang, Pei-Han Liao, Duo-Wen Chen, H.-P. Lin, H.-C. Chang, "A filter-like AuNPs@mesoSiO2 SERS substrate for S. aureus detection" *Bios. Bioelect.* vol. 53, pp. 519-527, 2014.

[9] V. Mazet, "Background removal from spectra by designing and minimizing a non-quadratic cost function," *Chem. & Int. Lab Sys.* vol. 76, pp. 121-133, 2005.

[10] M.-T. Kuo, C.-C. Lin, H.-Y. Liu, H.-C. Chang, "Tear Analytical Model Based on Raman Microspectroscopy for Investigation of Infectious Diseases of Ocular Surface, " *Invest. Ophthalmol. Vis. Sci.* vol. 52, pp. 4942-4950, 2011.

[11] M.-T. Kuo, C.-C. Lin, H.-Y. Liu, M.-Y. Yang, H.-C. Chang, "Differentiation between Infectious and Non-infectious Ulcerative Keratitis by Raman Spectra of Human Teardrops: A Pilot Study," *Invest. Ophthalmol. Vis. Sci.* vol. 53, pp. 1436-1444, 2012

[12] B. K. Lavine, "Raman spectroscopy and genetic algorithms for the classification of wood types," *Spectroscopy*, vol. 55, no. 8, pp 960-966, 2001.

[13] C.-C. Lin, C.-Y. Lin, C.–J. Kao and C.-H. "Hung, High efficiency SERS detection of clinical microorganism by low cost silver-mirror reacted filter membrane and pattern recognition techniques," *Sens. Acuat. B* vol. 241, pp. 513-521, 2017.