

基于单细胞拉曼技术鉴定非结核分枝杆菌的方法研究

阮真¹, 朱鹏飞³, 张磊³, 陈荣泽³, 李洵融³, 付晓婷³,
黄正谷⁴, 周刚⁴, 籍月彤⁵, 廖璞^{1,2*}

1. 重庆医科大学检验医学院, 临床检验诊断学教育部重点实验室, 重庆 400016
2. 重庆市人民医院检验科, 重庆 400013
3. 中国科学院青岛生物能源与过程研究所单细胞中心, 山东 青岛 266101
4. 重庆市公共卫生医疗救治中心, 中心实验室, 重庆 400036
5. 青岛星赛生物科技有限公司, 山东 青岛 266101

摘要 非结核分枝杆菌(NTM)是除结核分枝杆菌复合群(MTC)和麻风分支杆菌以外的分枝杆菌总称。近年来 NTM 导致人类感染的发病率不断上升, 其感染的临床症状与 MTC 感染极为相似, 但两者治疗方案却存在差异, 临床亟须快速、准确的鉴定方法用于诊断 NTM 感染。单细胞拉曼光谱技术(SCRS)具有非标记、免培养、快速、准确、低成本等优势。据此, 我们提出了一种基于显微共聚焦单细胞拉曼光谱技术鉴定 NTM 的方法。通过对临床常见的六种 NTM(脓肿分枝杆菌、戈登分枝杆菌、偶发分枝杆菌、土分枝杆菌、鸟分枝杆菌以及堪萨斯分枝杆菌)的拉曼光谱进行处理比较, 并结合峰位注释进行分析。采用无监督低维可视化的 t -分布式随机邻域嵌入方法展示六种 NTM 的拉曼数据结构, 证明其数据在低维空间上的可分性后, 比较分类中常用的六种分类器[支持向量机分析(SVM)、K 最近邻分类算法(KNN)、偏最小二乘判别分析(PLS-DA)、随机森林(RF)、线性判别分析(LDA)、XG Boost]的效果。SVM 和 LDA 在 NTM 分类中效果最好, 分别达到了 99.4% 和 98.8% 的测试准确率; SVM 仅对于堪萨斯分枝杆菌(97.96%, 48/49)的分类准确性略低, 其余均为 100%; LDA 对于脓肿分枝杆菌(95.65%, 22/23)和戈登分枝杆菌(96.30%, 26/27), 其余也均为 100%。因此, 单细胞拉曼检测结合 SVM 分类器为 NTM 快速准确鉴定提供了富有潜力的新工具。

关键词 单细胞拉曼技术; 非结核分枝杆菌; 病原微生物鉴定; 支持向量机分析

中图分类号: O657.37 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2021)11-3468-06

引言

非结核分枝杆菌(non-tuberculosis mycobacteria, NTM)作为分枝杆菌属中除结核分枝杆菌复合群(*Mycobacterium tuberculosis complex*, MTC)和麻风分支杆菌以外的分枝杆菌, 截至目前共发现近 200 种, 致病能力各不相同^[1]。分枝杆菌具有相似的生物学性状: 多生长周期长(1~8 周)、有抗酸性。NTM 导致人类感染无论是临床症状还是其培养后的菌落常与 MTC 难以区分, 导致其易被误认为 MTC 进行治疗, 但常规的 MTC 治疗方案对 NTM 可能没有效果, 还会诱导耐药菌的产生^[2-3]。近年来 NTM 的分离率不断增加, 许多国家和地区在结核病的发病率不断下降的同时 NTM 的感

染率和耐药率却在不断上升^[4-5]。NTM 感染的诊断和治疗依赖于病原学依据^[6-7]。由于存在生物安全风险, 常规实验室多未对分枝杆菌进行鉴定。目前, 基因测序是 NTM 鉴定的金标准, 但操作相对繁琐, 需在专业实验室进行检测。商品化试剂盒用于鉴别 NTM 其敏感性受标本质量影响较大, 仅能用于常见 NTM 鉴定, 无法对新物种检测^[8-9]。当下亟须一种能快速、准确区分和鉴定 NTM 的新方法来帮助临床诊断 NTM 感染。

单细胞拉曼光谱技术(single-cell Raman spectroscopy, SCRS)是基于拉曼光谱分析原理实现检测物质结构的新方法, 通过收集检测物的生物学信息(包括蛋白质、核酸、脂质等)组成生物体特有的“指纹图谱”, 并结合不同的分析方法实现对检测物的快速、准确区分^[10-11]。由于 SCRS 检测样本

收稿日期: 2021-03-25, 修订日期: 2021-07-04

基金项目: 国家重大科研仪器研制项目(31827801)资助

作者简介: 阮真, 1989 年生, 重庆医科大学检验医学院主管技师

* 通讯作者 e-mail: liaopu2015@163.com

e-mail: 224751928@qq.com

无需特殊处理、操作简便,20 世纪初 SCRS 开始应用于病原微生物研究在微生物鉴定方面展现了强大的应用前景^[12-13]。

本次实验选择了 NTM 菌群中六种最常导致临床感染的 NTM 菌株,通过直接单个细胞拉曼检测、原始拉曼光谱比较、峰位注释分析、比较常用的六种分类器效果,来寻找最适合 NTM 区分鉴定的分类方法,为临床快速、准确鉴定 NTM 提供新思路。

1 实验部分

1.1 样品

实验涉及 NTM 菌株共六种(表 1),由重庆市公共卫生医疗救治中心实验室提供。经灭活后送至北京睿博兴科生物技术有限公司(RuiBio BioTech)进行 16S rRNA 测序鉴定,确定细菌种类。

在 II 级生物安全柜内将 NTM 菌株接种在 Loewenstein-Jensen(L-J)培养基上,置于细菌培养箱中,37 °C 培养 14 d。培养后的 NTM 菌落于生理盐水中制备菌悬液。放入恒温金属浴中 100 °C 15 min,对 NTM 菌株灭活。已有研究证实高温灭活微生物细胞对后续拉曼检测结果影响不大^[14]。

表 1 菌株列表及获得单细胞拉曼光谱数

Table 1 List of species and the number of Raman spectra from single cells

细菌名称	光谱数
脓肿分枝杆菌 <i>Mycobacterium abscessus</i>	92
戈登分枝杆菌 <i>Mycobacterium gordonae</i>	99
偶发分枝杆菌 <i>Mycobacterium fortuitum</i>	67
土分枝杆菌 <i>Mycobacterium terrae</i>	73
鸟分枝杆菌 <i>Mycobacterium avium</i>	68
堪萨斯分枝杆菌 <i>Mycobacterium kansasii</i>	154

1.2 拉曼光谱检测

检测单细胞的拉曼光谱由临床单细胞拉曼耐药性快检仪(CAST-R, 青岛星赛)收集。具体配备含有 532 nm Nd:YAG 激光器(cobolt, Sweden),CCD 图像传感器(Andor, UK),100 倍长焦镜头(NA=0.95, Olympus, Japan)等。

取 1 mL 菌液,10 000 g 离心 2 min 收集菌体后用超纯水洗涤 3 次,重悬于 1 mL 超纯水中。取重悬后的菌液 1 μ L 于干净的氟化钙玻片上,将样品置于拉曼光谱仪的显微镜下,进行拉曼测量。每个样品随机选取 100 个单细胞和 4 个无细胞区域,逐一收集拉曼光谱。单细胞拉曼光谱测量参数为:输出激光为 532 nm,输出强度约为 100 mW,50% filter 下收集时间为 5 s。

1.3 光谱数据处理

采用 R 软件对 6 种单细胞拉曼光谱进行减背景、一阶导数变换和 Savitzky-Golay 卷积平滑处理,将处理后的数据按照 7:3 的比例划分成训练集和测试集,采用训练集数据构建各类预测模型,应用该模型对测试集进行预测,测试模型性能。处理后的光谱进行 *t*-分布式随机邻域嵌入,测试支持向量机分析(support vector machine, SVM)、K 最近邻分类

算法(K-nearest neighbor method, KNN)、偏最小二乘判别分析(partial least square-discriminate analysis, PLS-DA)、随机森林(random forests, RF)、线性判别分析(linear discriminant analysis, LDA)、XG Boost 等的分类效果。

2 结果与讨论

2.1 六种 NTM 细胞平均拉曼光谱的外形比较及峰位归属分析

本研究共获得 6 种 NTM 菌株细胞拉曼特征光谱 553 例,表 1 记录了具体光谱分布。经基线扣除及归一化处理后,各样品的绝对峰值差异已经消除,各样品已基本处于分析的统一水平线。从不同类别细胞光谱的平均光谱可以看出,不同分枝杆菌的单细胞拉曼光谱的大致形态相似,但根据平均光谱的差异峰谱可以看出,不同分枝杆菌的单细胞拉曼光谱在局部波数范围内存在不同程度的差异(图 1)。

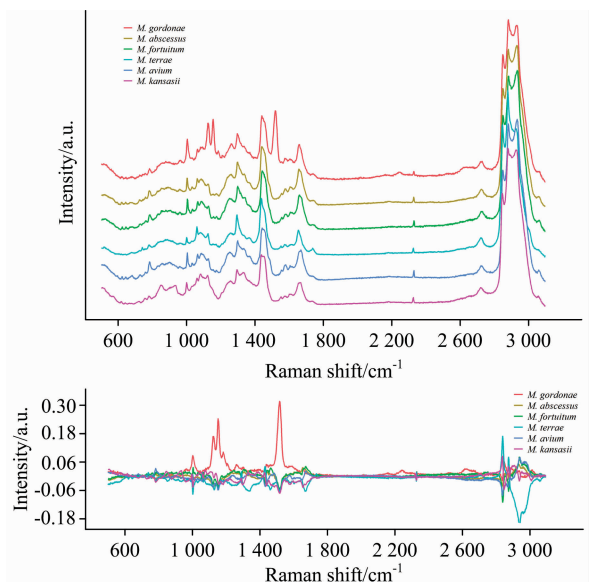


图 1 六种 NTM 细胞平均拉曼光谱及差异峰谱

Fig. 1 Original average Raman spectra and difference peaks of six NTMs

其中,戈登分枝杆菌与其他分枝杆菌的差异较大,主要表现在类胡萝卜素(1 153.82 和 1 518.22 cm^{-1})含量高于其他细胞,代表戈登分枝杆菌单细胞相较于其他分枝杆菌而言存在细胞色素的累积(表 2)。土分枝杆菌的单细胞拉曼光谱显示其苯丙氨酸(1 004.49 cm^{-1} 等)位置上与其他分枝杆菌的光谱相比强度较低,代表土分枝杆菌中苯丙氨酸的含量相对较少。分枝杆菌细胞壁中复合脂质含量较高,其主要成分为分枝菌酸(碳原子含量 70~90 个),分枝菌酸被认为是分枝杆菌不同种属间特异性的化合物,不同种的分枝杆菌菌酸含量存在差异^[15]。从 NTM 细胞的拉曼光谱图中也可观察到这种差异性,在 1 449.47 cm^{-1} 位置上,六种分枝杆菌的拉曼信号强度呈现阶梯式差异,代表六种 NTM 单细胞内菌酸含量呈现不同程度的差异。

表 2 细胞拉曼光谱的峰位归属

Table 2 Molecular structure and composition of Raman peaks

Peak position / cm^{-1}	Proteins	Lipids	Nucleic acids	Carbohydrates
644.126	C—S stretching & C—C twisting of proteins			
726.371		Characteristic for phospholipids		
783.39			U, T, C	
829.318			tyrosine (proteins), O—P—O stretch DNA	
854.267				C—O—C skeletal mode of α -anomers
1 004.49	Phenylalanine ring breath			
1 030.24				$\nu(\text{CO})$, $\nu(\text{CC})$, $\nu(\text{CCO})$
1 084.69		Typical phospholipids		
1 128.03				C—O stretching
1 153.82	C—C/C—N stretching			
1 172.2	C—H in-plane bending mode of tyrosine, (CH) phenylalanine			
1 211.05	$\nu(\text{C—C}_6\text{H}_5)$, tryptophan, phenylalanine			
1 248.78	Amide III			
1 340.43			CH stretching of adenine	
1 449.47		CH_2 bending and scissoring modes of phospholipids		
1 518.22				$\nu(\text{C}=\text{C})$ diagnostic for the presence of a carotenoid structure, most likely a cellular pigment $\nu(\text{C}=\text{C})$ carotenoid
1 574.87	Phenylalanine			
1 604.68	Phenylalanine, tyrosine, C=C			
1 669.75		Allyl C=C stretches		

2.2 无监督学习对六种 NTM 的光谱数据结构分析

为验证光谱数据结构是否具有可分性,在训练分类器之前,采用低维投影可视化的方法测试数据在低维空间上是否可分。将原始光谱进行减背景,一阶导数变换及 Savitzky-Golay 卷积平滑处理(图 2),然后采用 t -SNE 的方法展示数据在二维空间上的区分程度。

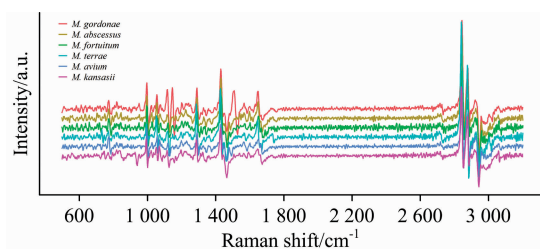


图 2 预处理后的六种 NTM 细胞平均拉曼光谱

Fig. 2 Pretreatment average Raman spectra of six NTMs

将高维的光谱数据在二维空间中进行展示(图 3)。根据结果可以看出,六种 NTM 可以较好的相互区分开,代表虽然属于同一菌属的分支杆菌在生物学性状上不易区分,但它们的拉曼光谱数据结构存在很好的可分性。其中土分枝杆菌和堪萨斯分枝杆菌相对其他种类的分支杆菌的区分程度更显著。

2.3 对六种 NTM 采用不同分类器训练后的测试效果比较

为获得更好的分类效果,本研究测试了六种常用的机器学习的方法构建分类器,分别为 SVM, KNN, PLS-DA, RF, LDA 和 XG Boost。六种分析方法对分枝杆菌单细胞拉曼光谱进行建模分析,采用相同的训练及测试过程,通过比较测试集的分类准确率选择最合适的分类方法。

首先按照 7:3 的比例将预处理后的数据随机划分成训练集和测试集,采用相同的训练集数据对不同方法的模型进行预测,应用模型对相同的测试集进行验证,测试模型性能。根据测试的结果显示 SVM 和 LDA 的预测准确率较高,

对六种分枝杆菌的分类效果较好, 预测准确率分别为 99.4% 和 98.8%, 而且几乎所有分类器的分类效果都能达到 80% 以上(表 3)。

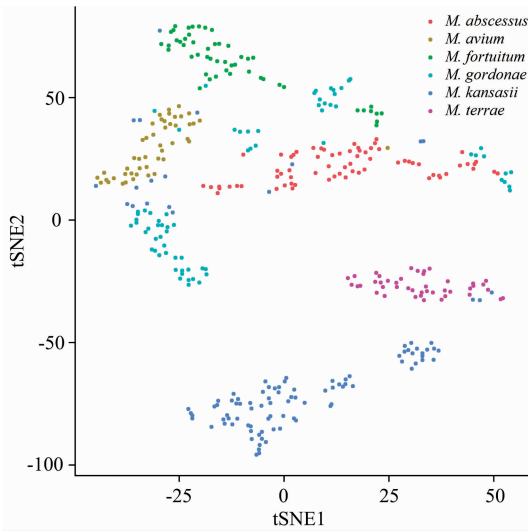


图 3 六种 NTM 单细胞拉曼光谱的 t-SNE 聚类结果
Fig. 3 t-SNE results of six NTMs SCRS

表 3 六种分类器的分类测试结果
Table 3 Results of six identification models

Model	Test dataset		
	Sample	Correct	Accuracy/%
SVM	164	163	99.4
K-NN	164	113	69.1
PLS-DA	164	139	85.0
RF	164	156	95.2
LDA	164	162	98.8
XGB	164	143	87.3

提取两个模型的混淆矩阵后可以发现, SVM 模型仅有一个单细胞预测错误, 将堪萨斯分枝杆菌分类成土分枝杆菌, 预测准确率为 97.96%(48/49); 其他种类的分类效果达到了 100%(表 4)。而 LDA 模型有两个单细胞预测错误: 将 1 个脓肿分枝杆菌细胞分类成偶发分枝杆菌, 预测准确率为 95.65%(22/23); 将 1 个戈登分枝杆菌细胞分类成土分枝杆菌, 预测准确率为 96.30%(26/27), 对其他分枝杆菌的分类效果可以达到 100%(表 5)。

表 4 SVM 的混淆矩阵

Table 4 Prediction results of SVM confusion matrix

SVM	Prediction					
	<i>M. abscessus</i>	<i>M. gordonae</i>	<i>M. fortuitum</i>	<i>M. terrae</i>	<i>M. avium</i>	<i>M. kansasii</i>
<i>M. abscessus</i>	23	0	0	0	0	0
<i>M. gordonae</i>	0	27	0	0	0	0
<i>M. fortuitum</i>	0	0	16	0	0	0
<i>M. terrae</i>	0	0	0	29	0	0
<i>M. avium</i>	0	0	0	0	21	0
<i>M. kansasii</i>	0	0	0	1	0	48

表 5 LDA 的混淆矩阵

Table 5 Prediction results of LDA confusion matrix

LDA	Prediction					
	<i>M. abscessus</i>	<i>M. gordonae</i>	<i>M. fortuitum</i>	<i>M. terrae</i>	<i>M. avium</i>	<i>M. kansasii</i>
<i>M. abscessus</i>	22	0	1	0	0	0
<i>M. gordonae</i>	0	26	0	1	0	0
<i>M. fortuitum</i>	0	0	16	0	0	0
<i>M. terrae</i>	0	0	0	29	0	0
<i>M. avium</i>	0	0	0	0	21	0
<i>M. kansasii</i>	0	0	0	0	0	49

3 结 论

通过对六种不同的分类器的效果比较, 提出一种基于特征拉曼光谱训练 SVM 模型的方法来鉴定六种非结核分枝杆菌。在利用已知种类的单细胞拉曼光谱样本构建数据库的基础上, 可以实现对未知非结核分枝杆菌样本的鉴定。单细胞

拉曼技术有快速、简易、成本低等优势, 未来通过大量临床样本的检验及数据库的扩充, 具备为临床医生提供更快速、便捷、低成本的临床辅助诊断方案的巨大潜力。

此外, 还可以结合拉曼分选与单细胞测序(RACS-Seq), 验证拉曼光谱分类和鉴定的结果, 并揭示全基因组水平、单碱基精度的遗传信息^[16]。针对大肠杆菌, 我们前期已经证明, 利用 RAGE-Seq 能从临床尿液样本中利用单细胞拉曼光

谱直接识别和分选出耐受特定抗生素的临床大肠杆菌, 并进行了精确到一个细胞的全基因组测序, 覆盖度可达 99.5%^[17]。这一高覆盖保证了基因组上所有耐药基因突变均得以全面、精确地揭示。因此, 本研究基于拉曼光谱的 NTM 分类和鉴别, 为一个针对 NTM 的 RACS-Seq 奠定了基础。这种在单个病原微生物细胞精度的临床分析, 代表着

生命单元和进化单位水平的病原识别、基因组追踪与药敏机制研究, 因此对于新一代的疾病防控平台具有重要的意义。

综上, 显微共聚焦单细胞拉曼光谱技术与各种机器学习的方法可以对生物样本进行种类鉴定, 从生物大分子的结构、组成及相对含量等方面给出依据, 可为临床诊断及微生物学研究提供有力的技术支撑。

References

- [1] Tortoli E, Fedrizzi T, Meehan C J, et al. *Infection, Genetics and Evolution*, 2017, 56: 19.
- [2] Baldwin S L, Larsen S E, Ordway D, et al. *PLOS Neglected Tropical Diseases*, 2019, 13(2): e0007083.
- [3] Huang W C, Yu M C, Huang Y W. *Journal of the Formosan Medical Association*, 2020, 119: 532.
- [4] WANG Li-xia, CHENG Shi-ming, CHEN Ming-ting, et al(王黎霞, 成诗明, 陈明亭, 等). *Chinese Journal of Antituberculosis(中国防痨杂志)*, 2012, (8): 485.
- [5] Mcgrath E E, Anderson P B. *The Lancet*, 2007, 370(9581): 28.
- [6] Griffith D E, Aksamit T, Brown-Elliott B A, et al. *American Journal of Respiratory and Critical Care Medicine*, 2007, 175(4): 367.
- [7] Brown-Elliott B A, Nash K A, Wallace R J Jr. *Clinical Microbiology Reviews*, 2012, 25(3): 545.
- [8] Seagar A L, Prendergast C, Emmanuel F X, et al. *Journal of Medical Microbiology*, 2008, 57: 605.
- [9] Kim S, Park E M, Kwon O J, et al. *The International Journal of Tuberculosis and Lung Disease*, 2008, 12(11): 1344.
- [10] LIANG Hao-yue, CHENG Xue-lian, YANG Wan-zhu, et al(梁昊岳, 程雪莲, 杨晚竹, 等). *Spectroscopy and Spectral Analysis(光谱学与光谱分析)*, 2020, 40(12): 3670.
- [11] RUAN Zhen, ZHU Peng-fei, FU Xiao-ting, et al(阮真, 朱鹏飞, 付晓婷, 等). *Microbiology China(微生物学通报)*. doi: 10.13344/j.microbiol.china.200703.
- [12] Petry R, Schmitt M, Popp J. *Chemphyschem*, 2003, 4(1): 14.
- [13] Montanari L B, Martins C H G, De Mello C A, et al. *Journal of Water and Health*, 2018, 16(2): 311.
- [14] Harz M, Kiehnopf M, Stockel S, et al. *Journal of Biophotonics*, 2009, 2(1-2): 70.
- [15] Jorgensen J H, Pfaller M A. *Manual of Clinical Microbiology*. American Society for Microbiology, 2015.
- [16] He Y, Wang X, Ma B, et al. *Biotechnology Advances*, 2019, 37(6): 107388.
- [17] Xu T, Gong Y, Su X, et al. *Small*, 2020, 16(30): e2001172.

Study on Identification of Non-Tuberculosis Mycobacteria Based on Single-Cell Raman Spectroscopy

RUAN Zhen¹, ZHU Peng-fei³, ZHANG Lei³, CHEN Rong-ze³, LI Xun-rong³, FU Xiao-ting³, HUANG Zheng-gu¹, ZHOU Gang⁴, JI Yue-tong⁵, LIAO Pu^{1, 2*}

1. Key Laboratory of Diagnostic Medicine, Ministry of Education; School of Laboratory Medicine, Chongqing Medical University, Chongqing 400016, China
2. Department of Laboratory Medicine, Chongqing General Hospital, Chongqing 400013, China
3. Single-Cell Center, Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao 266101, China
4. Center Lab, Chongqing Public Health Medical Center, Chongqing 400036, China
5. Qingdao Single-Cell Biotechnology Co., Ltd., Qingdao 266101, China

Abstract Non-tuberculosis mycobacteria (NTM) are the collection of mycobacteria other than *Mycobacterium tuberculosis* complex (MTC) and *Mycobacterium leprosy*. The clinical symptoms of NTM are very similar to MTC infection, yet their treatments are different, thus rapid and accurate identification methods of NTM are urgently needed. Single-cell Raman Spectroscopy (SCRS) is label-free, and independent of cultivation, thus it is deemed a rapid and efficient technology with low cost. Here we propose an SCRS based method to identify NTM based on confocal SCRS. We selected six common NTM species in the clinic, *Mycobacterium abscessus*, *Mycobacterium goodsonae*, *Mycobacterium fortuitum*, *Mycobacterium fortuitum*, *Mycobacterium avium* and *Mycobacterium kansasii*. The unsupervised low-dimensional visualization *t*-distribution random

neighborhood embedding method for the data structures proved the separability of data in the low-dimensional space. Performance of six commonly classifiers, including Support Vector Machine (SVM), K-Nearest Neighbor method (KNN), Partial Least Square-Discriminate Analysis (PLS-DA), Random Forests (RF), Linear Discriminant Analysis (LDA) and XG Boost was compared, with SVM and LDA achieving an accuracy of 99.4% and 98.8% respectively in NTMs classification. SVM offers 100% classification accuracy for every species, except *Mycobacterium kansasii* which is slightly lower (97.96%, 48/49), while LDA offers 100% accuracy for each species except *Mycobacterium abscessus* (95.65%; 22/23) and *Mycobacterium goodii* (96.30%, 26/27). Therefore, SCRS combined with SVM can accurately classify NTMs and thus provide a new tool for the rapid diagnosis of NTM.

Keywords Single-cell Raman Spectroscopy; *Non-tuberculosis mycobacteria*; Pathogenic microorganism identification; Support Vector Machine

(Received Mar. 25, 2021; accepted Jul. 4, 2021)

* Corresponding author