网络出版时间:2022-04-19 13:29 网络出版地址:https://kns.cnki.net/kcms/detail/34.1065.R.20220415.1514.023.html

基于 MALDI-TOF MS 平台结合机器学习算法鉴别三唑耐药热带念珠菌

王金宇,张 可,夏翠萍,王中新

摘要 目的 利用基质辅助激光解吸电离飞行时间质谱(MALDI-TOF MS)平台数据分析和机器学习算法快速鉴别三唑(氟康唑、伏立康唑、伊曲康唑)耐药和敏感的热带念珠菌。**方法** 从临床各类标本中收集191株热带念珠菌,其中71株为三唑耐药热带念珠菌,120株为三唑敏感热带念珠菌。使用MALDI-TOF MS平台进行数据采集,并根据Mann-Whitney *U*-test及随机森林(RF)算法获得的重要性评分对耐药株及敏感株的质荷比特征进行分类和选择。利用RF算法及径向基函数核非线性支持向量机(RBF-SVM)构建分类模型,计算相同实验数据下RBF-SVM模型和RF模型的准确度、敏感度、特异度、F1值及受试工作者曲线下面积(AUC)以评估模型鉴别性能。**结果** 所有菌株经过MALDI-TOF MS平台分析后共得到76个独特的质谱峰。其中,通过特征降维处理后选择6个峰3 481、7 549、6 500、3 048、6 892、2 596 m/z作为模型建立的特征峰。RBF-SVM模型和RF模型的准确度均为0.84,AUC分数分别为0.930 5、0.927 3。**结论** 机器学习算法结合MALDI-TOF MS平台进行数据分析可作为一种快速区分三唑耐药热带念珠菌和三唑敏感菌株的方法。

关键词 基质辅助激光解吸电离飞行时间质谱技术;机器学习算法;热带念珠菌;支持向量机;随机森林算法

中图分类号 Q 939.9

文献标志码 A **文章编号** 1000-1492(2022)05-0801-04

2022-02-25 接收

基金项目:安徽高校自然科学研究项目(编号:KJ2015A337)

作者单位:安徽医科大学第一附属医院检验科,合肥 230022

作者简介:王金宇,男,硕士研究生;

王中新,男,副教授,硕士生导师,责任作者,E-mail:aywzhx87@163.com

doi:10.19405/j.cnki.issn1000-1492.2022.05.024

近年来,随着免疫抑制剂和广谱抗生素不合理地应用,以及各种侵入性诊疗的进行,临床真菌感染率、耐药率及病死率大幅升高[1]。为了指导临床抗生素的使用,需要及早检测出真菌对抗生素的耐药性。

基质辅助激光解吸电离飞行时间质谱(matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry,MALDI-TOF MS)以其快速、可靠的菌种鉴定能力,已广泛用于实验室菌种的快速鉴定[2-4]。同时,MALDI-TOF MS主要通过分析指纹图谱特征峰的差异来区分耐药株及敏感株[5],但是仅通过有限的特征很难准确区分。机器学习算法能够计算出数据的统计相关性和非线性特征之间的影响。为了充分利用MALDI-TOF MS数据中包含的信息来简化耐药性的测定[6-9],该研究引入机器学习算法来探讨一种快速区分三唑(氟康唑、伏立康唑、伊曲康唑)耐药和敏感热带念珠菌的方法。

1 材料与方法

1.1 菌株来源 收集2018年1月—2021年3月自安徽医科大学第一附属医院临床各类标本中191株热带念珠菌,其中120株为三唑敏感的热带念珠菌,71株为三唑耐药的热带念珠菌。所有分离株均通过MALDI-TOF MS平台进行鉴定。耐药性依据美国临床和实验室标准化协会(CLSI)指南,使用微量

by CCK-8 method, colony formation assays and flow cytometry respectively. Western blot was used to detect the expression level of CDK4 and cyclinD1. ***Results*** The expression levels of *TRIM59* mRNA and protein in colorectal cancer cells were significantly higher than those in normal colorectal mucosa cells ($P<0.05$). After knockdown of *TRIM59* expression, the proliferation activity and colony forming ability of HCT116 cells were significantly inhibited ($P<0.05$), and the cell cycle was arrested in G0/G1 phase. At the same time, the expression levels of CDK4 and CyclinD1 significantly decreased ($P<0.05$). ***Conclusion*** *TRIM59* was highly expressed in colorectal cancer cells. Down regulation of *TRIM59* expression can inhibit the proliferation and clone formation of colorectal cancer cells, suggesting that *TRIM59* may become a new target for gene therapy of colorectal cancer.

Key words *TRIM59*; colorectal cancer; cell proliferation; cell cycle

肉汤稀释法对上述菌种进行药物敏感性实验。

1.2 仪器与试剂 Autof ms1000 全自动微生物质谱检测系统及配套试剂(郑州安图实验仪器有限公司)、生物安全柜(上海瑞仰净化装备有限公司)、UP700 恒温培养箱(英国 GreenPrima 公司)、微量移液器(德国 Eppendorf 公司)、科马嘉显色培养基(合肥天达诊断试剂有限公司)。

1.3 方法

1.3.1 MALDI - TOF MS 数据采集 数据采集流程:① 菌株接种,35 ℃温箱过夜培养 16 ~ 18 h;② 挑选生长良好的单个菌落均匀涂抹于靶板上;③ 加入 1 μl 甲酸;④ 加入 1 μl 基质溶液;⑤ 使用质谱仪 Autof ms1000 进行峰值采集。

1.3.2 特征峰选择 采用随机森林(random forest,RF)算法[10]对特征峰重要性进行评分,10 倍交叉验证保证结果的稳定,挑选出重要性排名前 10 的特征峰,使用 Mann-Whitney U-test 对特征峰进行相关性分析(表 1),检验均为双侧检验,$P<0.01$ 具有统计学意义。符合条件的峰值作为特征峰用于 RF 模型及径向基函数核非线性支持向量机(the radial basis function support vector machine,RBF-SVM)模型的开发。

1.3.3 RF 模型及 RBF-SVM 模型构建与性能评估 本实验引入 RF 模型及 RBF-SVM 模型对热带念珠菌敏感株和耐药株进行识别分类。RF 模型及 RBF-SVM 模型均基于 Python 环境开发的机器学习模块 scikit-learn[11]提供预封装的工具包进行构建。RF 模型调优:使用随机搜索交叉验证对参数决策树数量 、最大深度进行调优,然后用网格搜索在一定浮动范围内微调选择参数最优解。SVM 模型调优:再通过相同的方法,确定 RBF-SVM 的最佳核参数(γ)和最佳代价参数(C)。对 RF 模型及 SVM 模型进行 10 倍交叉验证以确保参数的稳定性。在模型的性能评估中,计算每种模型的准确性、AUC、F1 值、特异性和敏感性作为评价指标。此外绘制非线性分类器 RF 模型与 SVM 模型受试工作者特征(receiver operating characteristic curve,ROC)曲线,对模型进行更直观的比较。模型构建流程见图 1。

1.4 统计学处理 采用 Mann-Whitney *U*-test 对 MALDI-TOF 质谱峰特征进行分析,所有统计检验均为双侧检验,$P<0.01$ 为差异有统计学意义。

2 结果

2.1 数据采集结果 所有热带念珠菌应用 MALDI-TOF MS 进行光谱采集的结果均处于得分区间[9.0, 10.0],达到种水平置信度。耐药株及敏感株在质荷比2 000 ~ 20 000范围内的所得到的总光谱峰数分别为5 746、9 620个,特征峰得到 76 个。

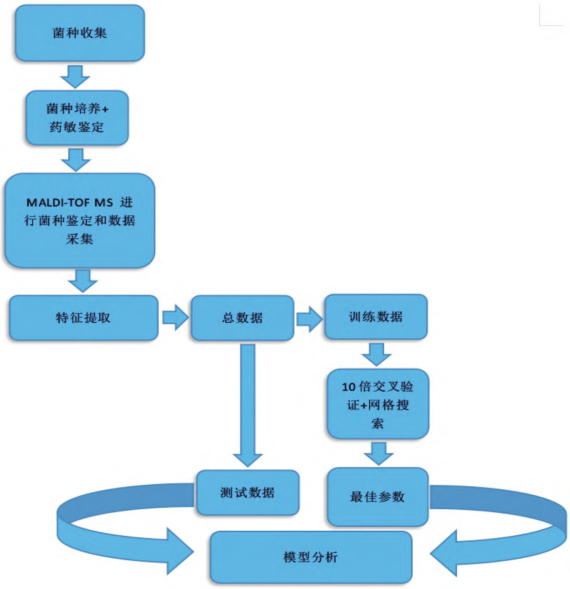

图 1 模型构建流程图

2.2 特征峰选择结果 特征峰3 481、7 549、6 500、3 048、6 892 m/z 经过双侧检验后 $P<0.01$,据有统计学意义。为了尽量减少数据内部信息的损失,根据 RF 算法的结果,基于 10 倍交叉验证,60% 以上的模型筛选出2 596 m/z 也纳入后续模型的构建。

2.3 RF 模型及 RBF-SVM 模型性能评估的结果 图 2 显示对于模型区分热带念珠菌中敏感株和耐药株的性能,最佳预测模型为 RBF-SVM 模型(AUC = 0.930 5,95% *CI*:0.868 1 ~ 0.955 3)。其中 RF 模型(AUC = 0.927 3,95% :0.830 1 ~ 0.949 9)具有相似的性能。表 1 列出了 RBF-SVM 模型和 RF 模型预测性能评估结果,与 RF 模型相比,RBF-SVM 模型敏感度为 0.91 低于 RF 模型,特异度为 0.73 高于 RF 模型。所有结果均进行 10 倍交叉验证确定。两种预测模型性能很接近且整体预测性能都能达到 0.8 以上。

表 1 RBF-SVM 模型和 RF 模型性能评估结果

分类模型	准确度	敏感度	特异度	F1 值	AUC
RBF-SVM	0.84	0.91	0.73	0.95	0.9305
RF	0.84	0.94	0.68	0.95	0.9273

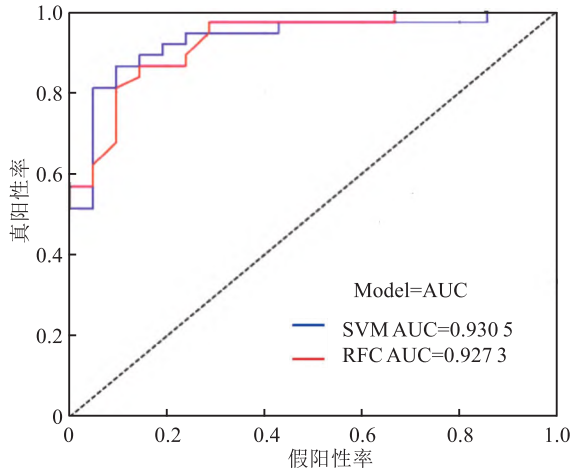

图 2 RBF-SVM 模型和 RF 模型性能评估结果

3 讨论

机器学习算法[11]主要包括支持向量机、RF、遗传算法、K 近邻算法等。目前,最佳的机器学习算法尚未明确,国内外研究[5,8,12-13]通过应用多种机器学习算法建立不同的分类模型,最终选择结果最优的模型作为最优分类模型,并且这些研究结果证明了支持向量机算法和 RF 算法在分类模型中的优越的表现,因此,本研究采用这两种算法对 MALDI-TOF MS 平台收集的光谱进行分析。

模型构建的重点在于模型的稳定性和可靠性。多数研究[8,12-13]通过交叉验证(5 或 10 倍)来避免模型的过拟合。模型构建图显示,在本研究实验流程中,通过 10 倍交叉验证来实现模型的稳定性和可靠性。RBF-SVM 模型和 RF 模型性能评估结果显示两种模型效能非常接近,这与 Wang et al[13]研究结果类似。但是本研究中两种模型的准确度仅为 0.84,这很可能与数据采集过程中多种因素有关,包括菌种反复冻融、靶点上菌落涂抹厚薄不均、基质液裂解不充分、MALDI-TOF MS 参数调优不佳等。

本研究得到的 76 个特征峰中,并不是所有的特征峰都有助于敏感株和耐药株的区分,通过 Mann-Whitney *U*-test 得到的峰中,只有 3 481、7 549、6 500、3 048、6 892 m/z 具有统计学意义。Fangous et al[6]和 Rhoads et al[9]的研究通过单个或者几个特征峰来判断菌株的耐药性,也证明了在判断敏感株和耐药株时,并不是所有的峰都有意义。

在临床应用中,常规抗生素敏感性试验结果通常在真菌分离后至少需要 24 h 才能得到,成本也比较昂贵。抗生素治疗的不及时会导致住院时间延长、治疗费用增加,以及因不恰当的抗生素治疗增加住院死亡率。然而本实验所研究的 RF 模型及 RBF-SVM 模型的优点是速度快、成本低,可以快速获得热带念珠菌药敏结果,从而指导临床医生进行准确且快速的抗真菌感染治疗,这对于规范临床抗生素的使用以及因抗生素的滥用导致细菌耐药率逐年增高方面有着重要意义。

RF 和 RBF-SVM 模型虽然平均准确度都能够达到 0.8 以上,具有较好的分类识别能力,但是模型的普适性仍有待研究。如提取方法不同,Lu et al[14]研究使用了试管提取法,而本实验使用直接涂板法,这增加了数据采集时的不稳定;Liu et al[12]研究中特征峰是基于统计或多元回归进行选择的,相比之下,本实验直接使用 RF 算法来选择特征峰值;到目前为止,最佳的数据的降维处理方式还不明确,不同的降维分析方式对于结果的影响有待后续的研究。同时,MALDI-TOF MS 光谱质量范围通常仅为 2 ~ 20 ku,然而,与真菌耐药密切相关的一系列高分子量酶往往不在这一领域,如热带念珠菌中的羊毛甾醇 14-α 去甲基化酶分子量远大于 2 ku,细菌中的青霉素结合蛋白分子量约为 76 ku[15],这将导致菌株中一些重要信息无法在光谱中反应出来,使得分类模型无法发挥到最佳性能。

综上所述,该研究表明机器学习算法结合 MALDI-TOF MS 平台的方法可以一定程度上快速区分热带念珠菌的敏感株和耐药株。这种方法有助于指导临床医师更快速、精确地使用抗生素,从而减少患者住院时间和费用。但是机器学习算法结合 MALDI-TOF MS 平台方法仍处于起步阶段,在后续的研究中有必要解决样本量小、缺乏外部验证、重现性差等相关问题。

参考文献

[1] Castanheira M, Messer S A, Rhomberg P R, et al. Antifungal susceptibility patterns of a global collection of fungal isolates: results of the SENTRY antifungal surveillance program[J]. Diagn Microbiol Infect Dis,2016,85(2):200-4.

[2] Foster A G. Rapid identification of microbes in positive blood cultures by use of the vitek MS matrix-assisted laser desorption ionization-time of flight mass spectrometry system[J]. J Clin Microbiol, 2013,51(11):3717-9.

[3] Haigh J D, Green I M, Ball D, et al. Rapid identification of bacteria from bioMérieux BacT/ALERT blood culture bottles by MALDI-TOF MS[J]. Br J Biomed Sci,2013,70(4):149-55.

[4] Weis C V, Jutzeler C R, Borgwardt K. Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review[J]. Clin Microbiol

Infect, 2020,26(10):1310 - 7.

[5] de Bruyne K, Slabbinck B, Waegeman W, et al. Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning[J]. Syst Appl Microbiol,2011,34(1):20 - 9.

[6] Fangous M S, Mougari F, Gouriou S, et al. Classification algorithm for subspecies identification within the *Mycobacterium abscessus* species, based on matrix-assisted laser desorption ionization-time of flight mass spectrometry[J]. J Clin Microbiol, 2014,52(9):3362 - 9.

[7] Sogawa K, Watanabe M, Ishige T, et al. Rapid Discrimination between methicillin-sensitive and methicillin-resistant *Staphylococcus aureus* using MALDI-TOF mass spectrometry[J]. Biocontrol Sci,2017,22(3):163 - 9.

[8] Mather C A, Werth B J, Sivagnanam S, et al. Rapid detection of vancomycin-intermediate *Staphylococcus aureus* by matrix-assisted laser desorption ionization-time of flight mass spectrometry[J]. J Clin Microbiol,2016,54(4):883 - 90.

[9] Rhoads D D, Wang H, Karichu J, et al. The presence of a single MALDI-TOF mass spectral peak predicts methicillin resistance in *staphylococci*[J]. Diagn Microbiol Infect Dis,2016,86(3):257 - 61.

[10] Rigatti S J. Random Forest[J]. J Insur Med,2017,47(1):31 - 9.

[11] Abraham A, Pedregosa F, Eickenberg M, et al. Machine learning for neuroimaging with scikit-learn[J]. Front Neuroinform, 2014, 8:14.

[12] Liu X, Su T, Hsu Y S, et al. Rapid identification and discrimination of methicillin-resistant *Staphylococcus aureus* strains *via* matrix-assisted laser desorption/ionization time-of-flight mass spectrometry[J]. Rapid Commun Mass Spectrom, 2021, 35(2): e8972.

[13] Wang H Y, Chen C H, Lee T Y, et al. Rapid detection of heterogeneous vancomycin-intermediate *Staphylococcus aureus* based on matrix-assisted laser desorption ionization time-of-flight: using a machine learning approach and unbiased validation[J]. Front Microbiol,2018,9:2393.

[14] Lu J J, Tsai F J, Ho C M, et al. Peptide biomarker discovery for identification of methicillin-resistant and vancomycin-intermediate *Staphylococcus aureus* strains by MALDI-TOF[J]. Anal Chem, 2012,84(13):5685 - 92.

[15] Varghese R, Neeravi A, Subramanian N, et al. Analysis of amino acid sequences of penicillin-binding proteins 1a, 2b, and 2x in invasive *Streptococcus pneumoniae* nonsusceptible to penicillin isolated from children in India[J]. Microb Drug Resist, 2021, 27(3):311 - 9.

Identification of triazole-resistant *Candida tropicalis* based on MALDI-TOF MS platform and machine learning algorithm

Wang Jinyu, Zhang Ke, Xia Cuiping, Wang Zhongxin

(*Dept of Clinical Laboratory, The First Affiliated Hospital of Anhui Medical University, Hefei* 230022)

Abstract ***Objective*** To rapidly identify triazole (fluconazole, voliconazole, iriconazole) drug resistance and sensitive *Candida tropical* using matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry (MALDI-TOF MS) platform data analysis and machine learning algorithms. ***Methods*** A total of 191 *Candida tropical* were collected from various clinical specimens, 71 of which were triazole-resistant *Candida tropical* and 120 were triazole-sensitive *Candida tropical* strains. Data acquisition was performed using the MALDI-TOF MS platform, and the mass and charge ratio features of resistant and susceptible strains were classified and selected based on the Mann-Whitney Rank-sum Test (Mann-Whitney *U*-test) and the importance score obtained by the Random Forest (RF) algorithm. The classification model was constructed using the RF algorithm and a nonlinear support vector machine with a radial basis function kernel(RBF-SVM), calculating the accuracy, sensitivity, specificity, F1 value and the area under the subject worker curve (AUC) of the RBF-SVM model under the same experimental data to evaluate the model discrimination performance. ***Results*** All strains obtained 76 unique mass spectrum peaks after analysis on the MALDI-TOF MS platform. Among them, six peaks 3 481, 7 549, 6 500, 3 048, 6 892, 2 596 m/z were selected as the model feature peaks established by the feature dimensionality reduction treatment. The accuracy of both the RBF-SVM and RF models was 0.84, and the AUC scores were 0.930 5 and 0.927 3, respectively. ***Conclusion*** Machine learning algorithms combined with the MALDI-TOF MS platform for data analysis can serve as a possible method to rapidly distinguish triazole-resistant *Candida tropical* and triazole-sensitive strains.

Key words matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry; machine learning algorithms; *Candida tropicalis*; support vector machine; random forest algorithm