微生物学报
Acta Microbiologica Sinica
2024, 64(2): 432–442
http://journals.im.ac.cn/actamicrocn
DOI: 10.13343/j.cnki.wsxb.20230523

Review 综 述

基于基因组数据分析的细菌耐药基因识别与表型预测

周秀娟[1,2]，何逸尘[1]，张利达[1]，崔妍[1]，史贤明[1*]

1 上海交通大学农业与生物学院 中美食品安全联合研究中心 微生物代谢国家重点实验室，上海 200240
2 上海健康医学院健康与公共卫生学院，上海 201318

周秀娟, 何逸尘, 张利达, 崔妍, 史贤明. 基于基因组数据分析的细菌耐药基因识别与表型预测[J]. 微生物学报, 2024, 64(2): 432-442.
ZHOU Xiujuan, HE Yichen, ZHANG Lida, CUI Yan, SHI Xianming. Identification and prediction of bacterial antibiotic resistance *via* genomic data analysis[J]. Acta Microbiologica Sinica, 2024, 64(2): 432-442.

摘　要：利用基因组数据和生物信息学分析方法，快速鉴定耐药基因并预测耐药表型，为细菌耐药状况监测提供了有力辅助手段。目前，已有的数十个耐药数据库及其相关分析工具这些资源为细菌耐药基因的识别以及耐药表型的预测提供了数据信息和技术手段。随着细菌基因组数据的持续增加以及耐药表型数据的不断积累，大数据和机器学习能够更好地建立耐药表型与基因组信息之间的相关性，因此，构建高效的耐药表型预测模型成为研究热点。本文围绕细菌耐药基因的识别和耐药表型的预测，针对耐药相关数据库、耐药特征识别理论与方法、耐药数据的机器学习与表型预测等方面展开讨论，以期为细菌耐药的相关研究提供手段和思路。

关键词：细菌耐药；耐药数据库；生物信息学；耐药基因识别；耐药表型预测

资助项目：国家重点研发计划(2019YFE0119700)
This work was supported by the National Key Research and Development Program of China (2019YFE0119700).
*Corresponding author. Tel: +86-21-34206616, E-mail: xmshi@sjtu.edu.cn
Received: 2023-08-11; Accepted: 2023-10-08; Published online: 2023-10-19

Identification and prediction of bacterial antibiotic resistance *via* genomic data analysis

ZHOU Xiujuan[1,2], HE Yichen[1], ZHANG Lida[1], CUI Yan[1], SHI Xianming[1*]

1 State Key Laboratory of Microbial Metabolism, MOST-USDA Joint Research Center for Food Safety, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai 200240, China
2 College of Public Health, Shanghai University of Medicine & Health Sciences, Shanghai 201318, China

Abstract: Using genomic data and bioinformatics methods has become an important approach to rapidly identify the genes and predict the phenotypes of bacterial antibiotic resistance. Dozens of antibiotic resistance databases have been established, providing information and auxiliary tools for the identification and prediction of bacterial antibiotic resistance. As the bacterial genome data and antibiotic resistance phenotype data are increasing, the correlation between them can be established *via* big data and machine learning. Therefore, establishing efficient models predicting antibiotic resistance phenotypes has become a research hot topic. Focusing on the gene identification and phenotype prediction of bacterial antibiotic resistance, this review discusses the related databases, the theories and methods, the machine learning algorithms, and the prediction models. In addition, we made an outlook on the future prospects in this field, aiming to provide new ideas for the related studies.

Keywords: bacterial antibiotic resistance; antibiotic resistance database; bioinformatics; identification of antibiotic resistance genes; prediction of antibiotic resistance phenotypes

抗生素是 20 世纪伟大的生物学发现之一，在细菌感染的临床治疗和畜牧业的养殖过程中都发挥着极其重要的作用。然而，由于抗生素的大量使用和滥用及其在环境中长时间的大量残留[1-2]，使得细菌产生耐药性，并且耐药性不断增强，甚至在世界范围内广泛传播。耐药细菌和耐药基因成为一种新的污染物，给全球的公共卫生、人类健康和环境生态都带来了巨大威胁。

细菌耐药特征的识别主要通过传统培养法进行耐药表型的鉴定，并利用 PCR 等分子生物学手段进行耐药基因的筛查，虽然有很高准确率，但工作量较大。近年来，细菌基因组学技术高速发展，基因组数据的大量积累为细菌耐药特征的识别提供了丰富的信息储备。在生物信息学技术的辅助下，大数据可适用于发现新的耐药基因和耐药机制，也可适用于监测和追踪耐药基因及耐药菌的分布特征和传播规律。本文针对目前主要的耐药基因数据库、耐药特征识别理论和耐药表型预测方法等方面进行分析和介绍，以期为细菌耐药的相关研究提供新思路。

1 细菌耐药基因的生物信息学分析

采用传统分子生物学方法检测细菌耐药基因时，常常会花费较多的时间，且不能达到耐药基因的全检测，更不能实现耐药基因亚型的直接判读。目前测序的成本越来越低，与细菌耐药相关的数据库也越来越多[3]。国内外大量的研究表

明[4]，利用生物信息学算法对基因组进行分析能快速预测耐药基因，可以作为现有常用方法的辅助和补充，减少人力与时间成本以及资源的耗费。

1.1 用于耐药基因筛查与预测的公共数据库

目前收录细菌耐药基因及其相关信息的公共数据库已有数十个，这些数据库提供了大量细菌基因组数据及耐药表型等相关信息，支持进一步的细菌耐药基因筛查与预测。这些数据库大致可分为 4 类：(1) 大型生物学数据库，包括 National Center for Biotechnology Information (NCBI)[5]、universal protein (Uniprot)[6]和 the Pathosystems Resource Integration Center (PATRIC)[7]等，这些数据库不仅收录了与耐药基因相关的信息，还包含大量其他的相关特征信息与数据，因此在实际使用时需要事先进行数据的筛选、整理和去冗余；(2) 通用型耐药数据库，即专门针对耐药基因建立的数据库，既包含较为全面的耐药基因序列，还归纳和注释了耐药基因的表达产物及对应表型等；(3) 特异型耐药数据库，即专注于某一特定抗生素、特定细菌种属或者特殊耐药元件的数据库；(4) 综合型抗性数据库，即同时收集了耐药基因、抗金属离子或抗杀菌剂等抗性基因的综合数据库，如包含抗生素、金属离子和杀菌剂的抗性基因的 MEGARes 数据库[8]，对研究多抗性间的协同与交叉作用及其调控具有重要意义。在生物信息学分析中，综合使用这些数据库有助于全面识别细菌的耐药基因。下面重点介绍通用型耐药数据库和特异型耐药数据库的主要特点与应用(表 1)。

1.1.1 通用型耐药数据库

马里兰大学生物信息学与计算生物中心于 2009 年发布了抗生素耐药基因数据库(antibiotic resistance genes database, ARDB)[9]，该数据库全面整理了当时的耐药基因信息，共筛选收录了 4 545 条耐药基因序列以及有文献支撑的序列描述信息。在 ARDB 的基础之上，McArthur 等[10]对序列信息进行了更新，建立了一个基于耐药数据共享的抗生素耐药性综合数据库(comprehensive antibiotic resistance database, CARD)。CARD 以耐药本体论(antibiotic resistance ontology, ARO)[10]为依据，关联抗生素、靶信息、耐药机制和基因位点变异等信息，不仅可以检索目标菌株中已确定的耐药基因，还可以通过抗性基因识别器(resistance gene identifier, RGI)预测潜在的耐药基因。

Dantas 实验室于 2014 年搭建了 Resfams 数据库[11]，该数据库的核心是基于蛋白家族和隐马尔可夫模型(hidden Markov model, HMMs)进行耐药基因的预测。类似的还有功能性抗生素耐药性宏基因组元件(functional antibiotic resistance metagenomic element, FARME)数据库[12]，后者包含 24 530 个不重叠的 HMMs。与 ARDB、CARD 数据库相比，Resfams 和 FARME 能够预测更多的未知耐药基因，能够更好地分析不可培养细菌中的耐药基因情况，以及耐药基因在环境菌株与临床培养株间的关联[3]。

此外，还有一些通用型耐药数据库是对其他数据库的数据进行了整合与去冗余处理。如 ARG-ANNOT 数据库[13]，该数据库对前文提到的大型生物数据库进行整理后得到 1 689 个耐药基因，基于此来预测细菌基因组中已知与潜在的耐药基因，并识别由基因突变产生的耐药；DeepARG 数据库[14]整合了 CARD、ARDB 和 Uniprot 数据库中的耐药基因信息，包含了基于机器学习算法的耐药基因分类鉴定工具，能够从宏基因组中鉴定耐药基因；SARG 数据库[15]整合了 ARDB、CARD 和 NCBI 数据库，主要用于鉴定并定量分析宏基因组中的耐药基因。这些通用型耐药数据库给耐药基因的快速识别提供了很好的平台，其中数据更新的程度与规范性成为制约数据库使用效力的重要因素之一。CARD

表 1　常用的细菌耐药相关数据库

Table 1　Commonly used databases related to drug resistant bacteria

Name	Features	Applications	Name	Features	Applications
ARDB	(1) Contained 4 545 drug-resistant genes and their descriptive information (2) The data are no longer updated	Used for identifying drug resistance genes	ARGO	Only contained β-lactamases and vancomycins related resistance genes	Used for the prediction of resistance related genes for specific types of antibiotics
CARD	(1) Contained all resistance genes in the ARDB database and resistance gene prediction tools (2) Data updated every month	Used for predicting drug resistance genes; The most popular tool for identifying and predicting drug resistance genes	Lahey/BLDB	Only contained β-lactamases related resistance genes	
Resfams	(1) Based on protein families and hidden Markov models (HMMs) (2) Contained 166 HMMs	Used for identifying resistance genes and predicting unknown resistance genes; Analyzing the situation of drug-resistant genes in non-culturable bacteria and unknown environments	TBDReaMDB	Contained 946 specific mutation sites related to seven different antibiotics of *Mycobacterium tuberculosis*	Used for the prediction of drug resistance related genes in *Mycobacterium tuberculosis*
FARME	(1) Based on HMMs (2) Contained 24 530 non-overlapping HMMs		MUBII-TB-DB	Contained the mutation information of seven genes with important therapeutic value for *Mycobacterium tuberculosis*	
ARG-ANNOT	(1) Integrated and removed redundancy of large-scale biological databases (2) Contains 1 689 drug resistance genes	Used for predicting known and potential drug resistance genes; Identifying drug resistance caused by genetic mutations	u-CARE	Contained 107 drug resistance genes and corresponding 52 antibiotics information for *Escherichia coli*	Used for the prediction of resistance related genes in *Escherichia coli*
DeepARG	(1) Integrated CARD, ARDB, and Uniprot databases (2) Included a tool for classifying and identifying drug resistance genes based on machine learning algorithms	Used for identifying drug resistance genes from metagenomes	ResFinder	(1) Contained mutation information of 11 strains/species (2) Suitable for various high-throughput sequencing raw data	Used for identifying acquired drug resistance genes and their subtypes; Used for research on a large number of environmental samples, the most citations database
SARG	Integrated ARDB, CARD, and NCBI databases	Used to identifying and quantitatively analyzing drug resistance genes in metagenomes	PointFinder	Contained mutation information of inherent resistance genes on chromosomes	Used for identifying mutations in inherent resistance genes on chromosomes

数据库是目前耐药基因研究中最常用的工具，其数据更新较为频繁，每个月更新一次，保证了数据的时效性。然而，一些数据库(如 ARDB、ARG-ANNOT 等)存在数据不更新或者更新信息不明确等问题，直接限制了这些数据库的使用。

1.1.2 特异型耐药数据库

特异型耐药数据库具有针对性强的特点，在研究特定抗生素、特定细菌以及特殊耐药元件(或突变信息)的耐药特征方面发挥着重要作用。

抗生素耐药基因在线(antibiotic resistance genes online, ARGO)数据库[16]是典型的针对特定类型抗生素的数据库，也是最早建立的关于耐药基因的数据库，它只包含 β-内酰胺酶和万古霉素这两类抗生素的耐药基因。针对 β-内酰胺类抗生素的数据库还有 Lahey β-内酰胺酶列表(Lahey list of β-lactamases, Lahey)数据库[17]和 β-内酰胺酶数据库(β-lactamase database, BLDB)[18]。

针对特定菌株也有相应的耐药数据库，如针对结核分枝杆菌的 TBDReaMDB[19] 和 MUBII-TB-DB 数据库[20]。前者收录了该菌中与耐药相关的基因突变信息，包含针对 7 种不同抗生素的共 946 个特异突变位点；后者收录了对结核病具有重要治疗价值的 7 个基因的突变信息。针对大肠埃希菌的耐药数据库 u-CARE[21]，包括 107 个耐药基因及其对应的 52 种抗生素。

由 Zankari 等[22]建立的 ResFinder 数据库关注获得性耐药基因及其亚型之间的区别，是针对特殊耐药元件的数据库。该科研团队后续还建立了 PointFinder 数据库[23]，侧重于染色体上固有基因突变导致耐药变化的信息收集。目前 ResFinder 已收录了 11 个菌株/物种的突变信息，这些突变信息的识别，有助于揭示耐药突变的发生与演变过程，对抗生素的开发也具有重要意义。ResFinder 数据库是目前引用次数最多的数据库，无论是基因组数据还是未经处理的原始测序数据都能作为该数据库的输入数据，这样更适合大量环境样本的研究。

由于特异型数据库能够满足一线工作者有针对性的需求，这种类型的耐药数据库有明显的增多趋势。在畜牧业和农业生产领域广泛使用的抗生素(如四环素、磺胺类、氟喹诺酮、大环内酯类和氨基糖苷类)，重要的人畜共患细菌(如肺炎链球菌和金黄色葡萄球菌)以及重要的工程生产用菌(如乳酸杆菌)都可能成为特异型耐药数据库开发的热点。除了获得性耐药元件和染色体固有耐药突变外，有些耐药基因的诱导表达还受到调控蛋白控制(如调控 AmpC 的 AmpR 等)，这些调控基因的存在与否同样对细菌耐药性的形成至关重要，目前尚无数据库针对性地收录这些耐药基因的调控蛋白，这也是建立特异型耐药数据库的一个方向。

1.2 耐药基因识别的理论与方法

耐药基因识别的理论主要分为耐药本体论和泛耐药基因组两大类。(1) 耐药本体论[10]是在本体论(ontology)基础上发展起来的。本体论是探究世界本源的哲学理论，将其应用于生物信息学中形成了序列本体论(sequence ontology)，为了防止语义上的冗余或混淆，需要通过规范的、统一的词汇表来详细地描述对象。ARO 就是通过建立相似的词汇表来描述耐药信息，如可移动遗传元件获取外源基因等。耐药本体论的建立能够更全面地从基因组信息中识别耐药基因，并对其相应的抗生素、耐药机理等信息进行后续研究。目前使用最广泛的 CARD 数据库就是通过这一理论来实现耐药基因的快速识别。(2) 泛耐药基因组[24-25] (pan-resistome)是在泛基因组(pan-genome)基础上延伸出来的。提出泛基因组概念的目的是探究完全描述一个物种所需要的基因集合的大小[26]，分为核心基因组(core genome)和附属基因组(accessory genome)[27]。对于同一物

种来说，核心基因组中的基因差异体现在单核苷酸多态性(single nucleotide polymorphisms, SNPs)上，表明了菌株间的亲缘关系，SNPs 一致性高表明亲缘关系较近。附属基因组中基因的获得方式可通过进化过程中某些基因的获得或缺失后形成在部分群体内的稳定遗传，还可以通过可移动遗传元件获取外源基因[4]。类似地，泛耐药基因组(pan-resistome)将所有菌株共有的耐药基因划分为核心耐药基因组(core resistome)，其余的则划分为附属耐药基因组(accessory resistome)[4]，两者的耐药机理存在本质区别，需要分别进行研究。He 等[24]正是基于泛耐药基因组的概念，建立了细菌泛耐药基因组分析软件(pan-resistome analysis pipeline, PRAP)，可用于耐药基因的快速筛查和批量基因组的特征分析。此外，接合质粒和整合性接合元件等可移动元件通常携带多个获得性耐药基因，是附属耐药基因组的重要组成部分，也是分析耐药区结构变异的关键。Wang 等最新开发了细菌耐药移动元件一键式分析新软件 VRprofile2[28]，可用于细菌病原体中与抗生素耐药性相关的可移动元件并阐明依托可移动元件转移的多重耐药区的形成机制，为揭示“菌型-质粒-耐药基因”的关联性提供新思路。

目前，耐药基因识别的模型和方法主要有三大类：(1) 基于序列相似度，如局部比对算法的搜索工具 BLAST，在 BacMet、ResFinder 和 ARG-ANNOT 等多个数据库中都使用了基于序列相似度比对的基因功能预测工具，然而，当待检基因与参考基因间的进化距离较远时，两者的序列相似度往往较低，BLAST 方法就会失效。(2) 基于保守序列或保守结构域，如 HMMs，在 MEGARes、CARD、Resfams 和 ARGs-OAP 2.0 等数据库中增加了 HMMs 预测模型，由于充分考虑了不同保守度的氨基酸在相应位置的权重，实现了对进化距离较远蛋白质相关性的敏感检测，因此 HMMs 可用于预测未知的新耐药基因和与已知耐药基因亲缘关系较远的基因[11-12]，这一特征可应用于未知环境样本中耐药基因的筛选。(3) 基于机器学习，DeepARG 是一种基于深度学习的耐药基因预测方法[12]，它不依赖于预定方程模型，直接从数据中“学习”信息的机器学习过程，其中 DeepARG-SS 和 DeepARG-LS 分别是针对短序列和全基因长度的深度学习模型。这种深度学习模型可预测很多传统方法未获得的耐药基因，更重要的是除了单个基因组样本还可以用于宏基因组的分析，更适合于在各种未知环境样本和宏基因组样本中预测出更多的耐药基因。

2 细菌耐药表型的生物信息学预测

虽然上述耐药数据库及其相关软件可以用于细菌耐药基因的快速识别，但是菌株的耐药表型与已知的耐药基因并不是完全对应的关系，因此仍需对耐药表型进行测定，传统培养方法是目前测定耐药表型或最低抑菌浓度(minimum inhibitory concentration, MIC)的常规方法。随着各种类型数据库中细菌基因组数目的不断增加，以及耐药表型数据的大量积累，利用大数据和机器学习建立耐药表型或 MIC 值与基因组中耐药基因或元件序列信息之间的相关性已成为可能。近年来，基于全基因组测序信息预测细菌的耐药表型或 MIC 值，已经在金黄色葡萄球菌(*Staphylococcus aureus*)[29-30]、大肠埃希菌(*Escherichia coli*)[31-32]、肺炎克雷伯菌(*Klebsiella pneumoniae*)[33]、淋病奈瑟菌(*Neisseria gonorrhoeae*)[34]和沙门氏菌(*Salmonella*)[35]等致病菌中得到了证实。下文将对已报道的预测模型进行分类剖析，从而推测细菌耐药表型预测的可行性和新方向。

2.1 预测的目标与方法

根据预测目标的不同可分为两类：(1) 建立的模型只区分耐药(resistant)和敏感(susceptible)两种表型，即二分类训练。如 Her 等[31]在对 *Escherichia coli* 进行耐药和敏感二分类预测时，分别采用朴素贝叶斯(naive Bayes model)、随机森林(random forest)、支持向量机(support vector machine)等算法建立模型，结合遗传算法优化模型，用模型接受者操作特征曲线下面积(area under the receiver operating characteristic curve, AUROC)衡量预测准确性，此研究的 AUROC 能达到 0.97 (即预测准确度达到 97%)。(2) 建立的模型能精确到具体的 MIC 值，通常是通过回归(regression)算法建立基因组信息与 MIC 值(或 MIC 值的对数值)之间的具体模型。如 Nguyen 等[33]利用 1 668 株 *Klebsiella pneumoniae* 的基因组序列及其 20 种抗生素药敏实验结果，建立了从全基因组测序数据直接预测 MIC 值的模型，在 2 倍稀释浓度±1 范围评估方法下，15 种抗生素的预测准确率能达到 90%以上，平均准确率能达到 92%。

根据预测方法的不同也可分基于序列相似度和基于机器学习两类。(1) 基于序列相似度的表型预测，即利用序列比对软件(如 BLAST)从全基因组序列中筛查与已知耐药基因同源的序列，再根据这些基因的已知功能推测对应的耐药性。如 Gordon 等[29]依托 NCBI 等数据库中与金黄色葡萄球菌耐药相关的文献和序列，建立了耐药基因及其突变与耐药表型间的关系目录；再以 501 株金黄色葡萄球菌的全基因组序列为测试对象，利用 BLAST 从基因组中查找与耐药目录中相符合的位点，它们被定义为耐药决定性位点，通过与耐药表型的比较发现，这些耐药决定位点对 12 种不同抗生素预测的灵敏度和特异度分别达到 97%和 99%。(2) 基于机器学习的表型预测，即不参考已知功能基因，基于统计分析建立从基因组序列预测耐药表型或 MIC 值的模型。如 Moradigaravand 等[32]利用 1 936 株大肠埃希菌，建立了针对 11 种抗生素的耐药表型预测模型，效果最佳的梯度提升树(gradient boosting decision tree)模型的平均准确率达到 91%。同样地，Nguyen 等[35]基于 5 278 株非伤寒沙门氏菌，利用极端梯度提升算法，建立了针对 15 种抗生素 MIC 值的预测模型，平均准确率达到 95%。

2.2 预测的机器学习与算法

机器学习是一种拟人化的定义，指让计算机像人一样进行模拟学习，在现有的实例(或者用于训练的数据)中学到规律与经验，并在完成新任务上表现出优秀的性能[36]。在耐药表型预测中，通常所说的学习任务指的是“根据基因组数据得出耐药表型和/或 MIC 值”，其中“大量的细菌基因组”是用于学习的原始数据，“对应的耐药表型和/或 MIC 值”是数据标签，如何“将基因组转化为预测因子矩阵”是对原始数据(或元数据)的处理与高维特征的提取，最后用核心算法实现“将高维特征与标签联系起来并记忆”的过程[4]。

对基因组序列等元数据进行前处理是运用机器学习的方法进行表型预测的重要步骤。根据基因组及耐药特征的提取方法不同，元数据的前处理分为 k-mer 和泛基因组两类。(1) k-mer 方法，即对输入的每一条序列进行片段化处理和统计。具体步骤是以 1 个碱基为步长，依次截取 k 个碱基长度的序列片段放入训练集中，在训练集中统计所有菌株中出现的 k-mer 片段，基于一定的顺序存放于数据库中；再单独统计每个基因组中每个片段出现的次数，将此数值按与 k-mer 片段一致的顺序储存在一维向量中，这就形成了每个基因组的序列特征信息。由于原始 reads 或组装好的基因组序列都可以将作为 k-mer 方法

的输入文件,并且这些输入数据不区分基因组中的编码区和非编码区,大大减少对原始序列的预处理过程,这 k-mer 方法用于数据前处理的一个优势。此外,k-mer 方法不需要提前验证菌株包含的耐药基因和表型特征,可以很大程度上避免因为先验知识有错误造成的预测不准。然而,k-mer 间存在大量的重叠序列,这些冗余信息导致整体矩阵增大,当处理较多基因组时需要消耗更多的计算资源,这是此方法一个很大的不足。如 Nguyen 等[33,35]使用 k-mer 方法分别构建非伤寒沙门氏菌和肺炎克雷伯菌的耐药预测模型,当菌株数目增加到 4 500 株时,占用的内存就已经超过了其服务器 1.5 TB 的上限。因此,随着逐渐增多的菌株全基因组测序数据及其耐药信息,这种方法可能将变得不适用。(2) 泛基因组方法,即对基因组信息的压缩与分类过程,具体来说是将基因组序列分为核心耐药基因的 SNP 位点和附属耐药基因的有无,再对其分别进行编码,构成预测因子矩阵,从而减少了基因组信息占用的内存[4],这种预处理方法已经在多种细菌中得到应用,包括结核分枝杆菌、肠杆菌(即肺炎克雷伯菌、大肠埃希菌和沙门氏菌)、非发酵菌、金黄色葡萄球菌、淋病奈瑟菌和肺炎链球菌。此外,Her 等[31]发现采用 CARD 注释的附属耐药基因能达到较好的预测效果。尽管 Moradigaravand 等[32]将种群结构、分离时间等其他信息也加入了泛基因组预测的矩阵中,但附属耐药基因组是对最终的预测结果起关键作用的因素。

根据机器学习的理论不同,可以分为监督学习 (supervised learning)、无监督学习(unsupervised learning)、强化学习(reinforcement learning)和迁移学习(transfer learning)等。其中,目前最常用的方法是监督学习,即利用有标记的训练数据对模型的参数进行调整[37]。在生物信息学中,监督学习最常见的两类任务是“分类”和“回归”,其区别在于“分类”的标签是有限的离散变量,如在预测耐药表型时,标签是“耐药”“敏感”或“中介” 3 个离散变量;而“回归”的标签是连续变量,如预测 MIC 值时,标签是连续变化的浓度值,如 2.0、2.1、2.2 mg/L 等。根据核心算法的原理和适用条件不同可分为广义线性模型、决策树和神经网络三类。(1) 广义线性模型(generalized linear models),是常见的监督学习算法。以支持向量机算法为例,它既可以处理线性可分的二元分类,还能借助径向基函数(radial basis function)等非线性的核函数处理非线性划分边界问题。(2) 决策树(decision tree),通过在每个节点上设置判断条件将数据进行划分,并增加节点处信息纯度,通过多个节点和叶子构造一棵决策树,再通过多个决策树完成决策的集成方法,包括 Bagging[38] (如随机森林算法)和 Boosting[39] (如梯度提升树)这 2 种常用集成方法。现有的研究表明,梯度提升树算法无论在耐药表型还是 MIC 值的预测中均表现最佳[32,35]。(3) 神经网络,是逐渐兴起的一种监督学习方法,被称作深度学习[40],即通过模仿神经元结构及电信号传输的方法,来不断调整神经网络层中的参数。DeepARG[14]就是通过深度学习方法建立起来的预测软件,目前已经用于单个基因组样本和宏基因组的分析,可预测很多传统方法未获得的耐药基因,虽然这类方法还未正式进入耐药表型的预测,但是可以预见这类深度学习的方法将成为耐药表型预测的研究热点和发展趋势。

3 展望

近年来,基于基因组学数据的各种耐药数据库和耐药特征分析软件正在逐步开发,但不同数据库在基因序列的收集方式以及相关信息的注释等方面存在差异或者缺失,特别是获得性耐药

基因的亚型、固有耐药基因突变位点和耐药基因的调控蛋白的收录信息还不全面，完善数据质量和更新重要信息是数据库维护进而提高其使用效力的重要工作。大多数耐药数据库配套的软件仅限于对耐药基因的快速鉴定，并未涉及对鉴定结果的后续分析，当输入的基因组数量较多时仍需用户自己对输出的文本文件进行解读。因此，依托和整合这些耐药数据库，构建快速的细菌耐药特征识别和可视化分析软件也是未来的重要发展方向之一。为了更好地满足科研和临床的迫切需要，也亟需更多针对性较强的药物特异型和菌株特异型数据库的出现。目前数据库收录的基因组精细完成图越来越多，可以对耐药基因的上下游环境、可移动元件(质粒、转座子和基因组岛等)、调控蛋白进行统计分析和比对，挖掘出更多的耐药相关信息。可移动元件是耐药基因水平转移的重要载体，耐药基因并不总在可移动元件的内部，并且同一类耐药基因可能借助多种不同的可移动元件进行传播，因此，将耐药基因和可移动元件进行整合分析也将成为未来耐药数据库发展的方向之一。

目前许多研究已证实，基于机器学习可以从不同的细菌全基因组来预测耐药表型或 MICs，然而，这种方法得到的预测结果是否可以安全地用于治疗决策仍存在争议，一个重要的原因是机器学习的算法模型的理论水平或可靠性仍有尚未解决的问题。如基因组信息的预处理方法会直接影响计算资源的消耗；模型的参数可能会对细菌基因型和表型之间的联系产生干扰；训练集数据的平衡性和代表性也对最终结果产生影响等。其中，基因组和耐药数据的质量和代表性是保证耐药预测准确性的重要前提，目前公共数据库中含耐药数据的基因组信息大多来源于北美和欧洲，发展中国家的细菌基因组和耐药数据较少且不规范，这就导致了数据的不平衡和代表性不足。对现有的模型进行更新和完善，还需要考虑不同地区或国家抗生素的使用情况。当数据存在不平衡的情况时，仅使用准确率来评价模型也是不全面的。此外，前人常使用±1 或±2 稀释因子提高预测结果的可信度，但这些策略不能从根本上解决预测准确度的问题，只是对现阶段的药敏实验产生的数据误差的矫正。细菌的耐药特征不仅受基因的控制，还可能与调控蛋白、细菌的代谢过程相关，结合多组学数据建立模型能更加精确地反映基因组信息及基因表达水平与 MICs 的关系。综上所述，更精确的药敏实验测定结果、更加平衡的数据分布，可以提升模型预测 MICs 的准确性，多组学数据联合分析也会成为未来耐药特征识别与预测的焦点。

参考文献

[1] 张宁, 李淼, 刘翔. 土壤中抗生素抗性基因的分布及迁移转化[J]. 中国环境科学, 2018, 38(7): 2609-2617.
ZHANG N, LI M, LIU X. Distribution and transformation of antibiotic resistance genes in soil[J]. China Environment Science, 2018, 38(7): 2609-2617 (in Chinese).

[2] ZHU YG, ZHAO Y, LI B, HUANG CL, ZHANG SY, YU S, CHEN YS, ZHANG T, GILLING MR, SU JQ. Continental-scale pollution of estuaries with antibiotic resistance genes[J]. Nature Microbiology, 2017, 2: 16270.

[3] 杨兵, 梁晶, 刘林梦, 李雪佩, 王荃, 任一. 耐药基因数据库概述[J]. 生物工程学报, 2020, 36(12): 2582-2597.
YANG B, LIANG J, LIU LM, LI XP, WANG Q, REN Y. Overview of antibiotic resistance genes database[J]. Chinese Journal of Biotechnology, 2020, 36(12): 2582-2597 (in Chinese).

[4] 何逸尘. 沙门氏菌泛耐药基因组分析及耐药表型预测[D]. 上海: 上海交通大学硕士学位论文, 2020.
HE YC. Pan-resistome analysis and prediction of antibiotic resistance phenotypes of *Salmonella*[D]. Shanghai: Master's Thesis of Shanghai Jiao Tong University, 2020 (in Chinese).

[5] BENSON D, BOGUSKI M, LIPMAN D, OSTELL J. The national center for biotechnology information[J].

Genomics, 1990, 6(2): 389-391.
[6] The Uniprot Consortium. UniProt: a worldwide hub of protein knowledge[J]. Nucleic Acids Research, 2018, 47(D1): D506-D515.
[7] WATTAM AR, DAVIS JJ, ASSAF R, BOISVERT S, BRETTIN T, BUN C, CONRAD N, DIETRICH EM, DISZ T, GABBARD JL, GERDES S, HENRY CS, KENYON RW, MACHI D, MAO C, NORDBERG EK, OLSEN GJ, MURPHY-OLSON DE, OLSON R, OVERBEEK R, et al. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center[J]. Nucleic Acids Research, 2016, 45(D1): D535-D542.
[8] LAKIN SM, DEAN C, NOYES NR, DETTENWANGER A, ROSS AS, DOSTER E, ROVIRA P, ABDO Z, JONES KL, RUIZ J, BELK KE, MORLEY PS, BOUCHER C. MEGARes: an antimicrobial resistance database for high throughput sequencing[J]. Nucleic Acids Research, 2017, 45(D1): D574-D580.
[9] LIU B, POP M. ARDB-antibiotic resistance genes database[J]. Nucleic Acids Research, 2009, 37(database issue): D443-D447.
[10] MCARTHUR AG, WAGLECHNER N, NIZAM F, YAN A, AZAD MA, BAYLAY AJ, BHULLAR K, CANOVA MJ, de PASCALE G, EJIM L, KALAN L, KING AM, KOTEVA K, MORAR M, MULVER MR, O'BRIEN JS, PAWLOWSKI AC, PIDDOCK LJ, SPANOGIANNOPOULOS P, SUTHERLAND AD, et al. The comprehensive antibiotic resistance database[J]. Antimicrobial Agents and Chemotherapy, 2013, 57(7): 3348-3357.
[11] GIBSON MK, FORSBERG KJ, DANTAS G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology[J]. The ISME Journal, 2015, 9(1): 207-216.
[12] WALLACE JC, PORT JA, SMITH MN, FAUSTMAN EM. FARME DB: a functional antibiotic resistance element database[J]. Database (Oxford), 2017, 2017: baw165.
[13] GUPTA SK, PADMANABHAN BR, DIENE SM, LOPEZ-ROJAS R, KEMPF M, LANDRAUD L, ROLAIN JM. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes[J]. Antimicrobial Agents and Chemotherapy, 2014, 58(1): 212-220.
[14] ARANGO-ARGOTY G, GAMER E, PRUDEN A, HEATH LS, VIKESLAND P, ZHANG L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data[J]. Microbiome, 2018, 6(1): 23.
[15] YIN X, JIANG XT, CHAI B, LI L, YANG Y, COLE JR, TIEDJE JM, ZHANG T. ARGs-OAP v2.0 with an expanded SARG database and hidden Markov models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes[J]. Bioinformatics, 2018, 34(13): 2263-2270.
[16] SCARIA J, CHANDRAMOULI U, VEMA SK. Antibiotic resistance genes online (ARGO): a database on vancomycin and β-lactam resistance genes[J]. Bioinformation, 2005, 1(1): 5-7.
[17] BUSH K, JACOBY GA. Updated functional classification of β-lactamases[J]. Antimicrobial Agents and Chemotherapy, 2010, 54(3): 969-976.
[18] NAAS T, OUESLATI S, BONNIN RA, DABOS ML, ZAVALA A, DORTET L, RETAILLEAU P, IORGA BI. Beta-lactamase database (BLDB)-structure and function[J]. Journal of Enzyme Inhibition and Medicinal Chemistry, 2017, 32(1): 917-919.
[19] SANDGREN A, STRONG M, MUTHUKRISHNAN P, WEINER BK, CHURCH GM, MURRAY MB. Tuberculosis drug resistance mutation database[J]. PLoS Medicine, 2009, 6(2): e1000002.
[20] FLANDROIS JP, LINA G, DUMITRESCU O. MUBII-TB-DB: a database of mutations associated with antibiotic resistance in *Mycobacterium tuberculosis*[J]. BMC Bioinformatics, 2014, 15: 107.
[21] SAHA SB, UTTAM V, VERMA V. u-CARE: user-friendly comprehensive antibiotic resistance repository of *Escherichia coli*[J]. Journal of Clinical Pathology, 2015, 68(8): 648-651.
[22] ZANKARI E, HASMAN H, COSENTINO S, VESTERGAARD M, RASMUSSEN S, LUND O, AARESTRUP FM, LARSEN MV. Identification of acquired antimicrobial resistance genes[J]. Journal of Antimicrobial Chemotherapy, 2012, 67(11): 2640-2644.
[23] ZANKARI E, ALLESØE R, JOENSEN KG, CAVACO LM, LUND O, AARESTRUP FM. PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens[J]. Journal of Antimicrobial Chemotherapy, 2017, 72(10): 2764-2768.
[24] HE YC, ZHOU XJ, CHEN ZY, DENG XY, GEHRING A, OU HY, ZHANG LD, SHI XM. PRAP: pan resistome analysis pipeline[J]. BMC Bioinformatics, 2020, 21(1): 20.

[25] 周秀娟, 崔妍, 何逸尘, 张利达, 史贤明. 沙门氏菌泛耐药基因组特征分析[J]. 微生物学报, 2021, 61(8): 2358-2369.
ZHOU XJ, CUI Y, HE YC, ZHANG LD, SHI XM. Characteristics analysis of pan-resistant genome of *Salmonella*[J]. Acta Microbiologica Sinica, 2021, 61(8): 2358-2369 (in Chinese).

[26] TETTELIN H, MASIGNANI V, CIESLEWICZ MJ, DONATI C, MEDINI D, WARD NL, ANGIUOLI SV, CRABTREE J, JONES AL, DURKIN AS, DEBOY RT, DAVIDSEN TM, MORA M, SCARSELLI M, MARGARIT Y ROS I, PETERSON JD, HAUSER CR, SUNDARAM JP, NELSON WC, MADUPU R, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome"[J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(39): 13950-13955.

[27] VERNIKOS G, MEDINI D, RILEY D, TETTELIN H. Ten years of pan-genome analyses[J]. Current Opinion in Microbiology, 2014, 23: 148-154.

[28] WANG M, GOH YX, TAI C, WANG H, DENG Z, OU HY, WANG M, GOH YX, TAI C, WANG H, DENG Z, OU HY. VRprofile2: detection of antibiotic resistance-associated mobilome in bacterial pathogens[J]. Nucleic Acids Research, 2022, 50(W1): W768-W773.

[29] GORDON NC, PRICE JR, COLE K, EVERITT R, MORGAN M, FINNEY J, KEARNS AM, PICHON B, YOUNG B, WILSON DJ, LLEWELYN MJ, PAUL J, PETO TE, CROOK DW, WALKER AS, GOLUBCHIK T. Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing[J]. Journal of Clinical Microbiology, 2014, 52(4): 1182-1191.

[30] BRADLEY P, GORDON NC, WALKER TM, DUNN L, HEYS S, HUANG B, EARLE S, PANKHURST LJ, ANSON L, de CESARE M, PIAZZA P, VOTINTSEVA AA, GOLUBCHIK T, WILSON DJ, WYLLIE DH, DIEL R, NIEMANN S, FEUERRIEGEL S, KOHL TA, ISMAIL N, et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*[J]. Nature Communications, 2015, 6: 10063.

[31] HER H, WU Y. A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the *Escherichia coli* strains[J]. Bioinformatics, 2018, 34(13): i89-i95.

[32] MORADIGARAVAND D, PALM M, FAREWELL A, MUSTONEN V, WARRINGER J, PARTS L. Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data[J]. PLoS Computational Biology, 2018, 14(12): e1006258.

[33] NGUYEN M, BRETTIN T, LONG SW, MUSSER JM, OLSEN RJ, OLSON R, SHUKLA M, STEVENS RL, XIA F, YOO H, DAVIS JJ. Developing an *in silico* minimum inhibitory concentration panel test for *Klebsiella pneumoniae*[J]. Scientific Reports, 2018, 8(1): 421.

[34] EYRE DW, de SILVA D, COLE K, PETERS J, COLE MJ, GRAD YH, DEMCZUK W, MARTIN I, MULVEY MR, CROOK DW, WALKER AS, PETO TEA, PAUL J. WGS to predict antibiotic MICs for *Neisseria gonorrhoeae*[J]. Journal of Antimicrobial Chemotherapy, 2017, 72(7): 1937-1947.

[35] NGUYEN M, LONG SW, MCDERMOTT PF, OLSEN RJ, OLSON R, STEVENS RL, TYSON GH, ZHAO S, DAVIS JJ. Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal *Salmonella*[J]. Journal of Clinical Microbiology, 2019, 57(2): e1218-e1260.

[36] MCKINNEY BA, REIF DM, RITCHIE MD, MOORE JH. Machine learning for detecting gene-gene interactions: a review[J]. Applied Bioinformatics, 2006, 5(2): 77-88.

[37] KÄLL L, CANTERBURY JD, WESTON J, NOBLE WS, MACCOSS MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets[J]. Nature Methods, 2007, 4(11): 923-925.

[38] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[39] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system: proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining[C]. San Francisco, 2016, 785-794.

[40] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.