基于计算方法的抗菌肽预测

曹隽喆 顾 宏
(大连理工大学控制科学与工程学院 辽宁 大连 116024)

摘 要 抗菌肽是由生物体免疫系统所产生的能抵抗微生物感染的一种小分子多肽,因其具有高效低毒的广谱抗菌活性且几乎无耐药性问题,被看做是抗生素的最佳替代品,对解决抗生素滥用问题具有重要的意义.抗菌肽预测是生物信息学的一个重要研究内容,对抗菌肽及其抗菌功能进行预测能有效帮助了解抗菌肽的作用机理,为抗菌肽药物的设计和改造提供理论依据.基于计算方法的抗菌肽预测是采用数学理论、计算机技术和生物信息学方法,通过对抗菌肽数据的分析来挖掘出抗菌肽的生物特征和抗菌活性之间的关联,从而自动地对抗菌肽的类别做出推断.由于不依赖于生物实验,而是依靠有效的算法和计算机的高速计算能力来完成预测工作,计算方法具有高效快捷、成本低廉等特点,且具有良好的可操作性和批量处理能力,非常适合大规模预测任务,因此已经引起了国内外学者越来越多的关注.文中对国内外的相关研究成果进行了阐述和总结,包括抗菌肽生物信息数据库、主流的预测方法和预测方法的性能检验等.抗菌肽数据库是专门针对抗菌肽建立的数据库,收录了大量的抗菌肽数据,使用者不仅可以从中提取所需要的信息,还可以使用数据库所提供的各类在线工具对数据进行处理.文中对常见的一些抗菌肽数据库进行了介绍,给出相关数据库的数据收录情况、功能特点和网址链接等,以方便读者查询使用.接着文中介绍了目前主要使用的抗菌肽预测方法,包括基于经验分析的预测方法和基于机器学习的预测方法,前者是根据已知的经验规则或者模式对某类抗菌肽的一些生化属性和抗菌活性之间的关联进行统计或建模来对该类抗菌肽进行识别,而后者则是利用机器学习技术,通过对抗菌肽的已知数据信息进行学习,建立合理的预测算法从中找出抗菌肽的特点和规律,并将其推广到未知多肽数据来进行预测.随后文中又给出了预测方法的评估方法和评价指标,这些性能检验结果既是评估一个方法预测性能好坏的标准,又是与其他方法进行比较的依据.最后,文中对抗菌肽预测的发展进行了思考和讨论,并展望了未来的研究方向.

关键词 抗菌肽预测;计算方法;特征提取;机器学习;算法设计
中图法分类号 TP399 **DOI 号** 10.11897/SP.J.1016.2017.02777

A Review on Prediction of Antimicrobial Peptides Based on Computational Methods

CAO Jun-Zhe GU Hong
(*School of Control Science and Engineering, Dalian University of Technology, Dalian, Liaoning* 116024)

Abstract Antimicrobial peptides represent a diverse class of natural small peptides derived from innate immune system of organisms to combat microorganism infection, and are considered as the best potential candidate substitution of antibiotics because antimicrobial peptides have properties of high efficiency, low toxicity, broad spectrum antimicrobial activity without drug resistance. Prediction of antimicrobial peptides is an important part of bioinformatics. Predicting antimicrobial peptides and their functional information can assist to comprehend their mechanism and provide theoretical supports for designing and improving antimicrobial peptide medicines. By using mathematical theory, computer technology and bioinformatics method, prediction of antimicrobial

收稿日期:2016-05-27;在线出版日期:2017-05-04.本课题得到国家自然科学基金(61502074)、中国博士后科学基金资助项目(2016M591430)、大连理工大学基本科研业务费科研项目(DUT17RC(4)09)资助.**曹隽喆**,男,1984 年生,博士,讲师,主要研究方向为机器学习、数据挖掘、生物信息学.E-mail:caojunzhe@dlut.edu.cn.**顾 宏**(通信作者),男,1961 年生,博士,教授,主要研究领域为机器学习、生物信息学、大数据技术.E-mail:guhong@dlut.edu.cn.

peptides based on computational methods analyzes antimicrobial peptide data to explore the connection between the biological feature and antibacterial function of antimicrobial peptides, to make decisions automatically for the samples' attribution. Being independent of biology experiments, the computational method relies on the effective algorithms as well as the computing power of computers to perform the prediction missions, therefore, this kind of approach is low-cost, efficient, fast, and has excellent operability and processing batch ability to be quite proper for dealing with predicting tasks under large scale data, and then it has already attracted more and more attentions of both domestic and foreign scholars. This paper summarizes related researches at home and abroad, including antimicrobial peptide databases, current predicting methods for antimicrobial peptides, and the performance validation of prediction methods. The antimicrobial peptide databases are a class of databases specially created for researching antimicrobial peptides, which collect a mass of antimicrobial peptide data including information on antimicrobial peptides' amino acid residue sequences, sources of the recorded data, activities as well as functions and beyond. In addition, the users of these databases can not only download and extract the information they need but also process data by using the various online analysis tools provided by the databases. This article introduces some main open-access antimicrobial peptide databases, presents their current inclusion of collected data, sample categories, functions, characters, Web sites with links, and so on, to provide convenience and guidance for readers when they try to use these databases. And then some mainstream methods for predicting antimicrobial peptides are proposed, including the approaches based on empirical analysis and the ones based on machine learning. The empirical analysis method gathers statistics of data and establishes a mathematical model for the connection between some antimicrobial peptide's biochemistry properties and antimicrobial activities, according to known experiences, rules or pattern, to recognize this kind of antimicrobial peptides. And the machine learning method aims to mine and learn existing data information in the databases to design a proper algorithm, and finds out the antimicrobial peptide's features and laws, and then extends the relevance to unseen peptide samples to deduce their functions for prediction. After that, this paper also introduces the model evaluation methods and validation criterions, which can both evaluate the performance of a prediction approach and provide a reference for comparing the effects of different algorithms. Finally, we discuss the development of antimicrobial peptides prediction, and propose some meaningful research directions in future.

Keywords antimicrobial peptide prediction; computational method; feature extraction; machine learning; algorithm design

1 引 言

抗菌肽(Antimicrobial Peptides, AMPs)[1]是一类具有天然抗菌活性的小分子多肽,具有广谱高效的抗菌活性,且不会使病菌对其产生耐药性[2-6]. 因此,抗菌肽被认为是抗生素的最佳替代品,对解决日益严重的抗生素滥用问题具有十分重要的意义,在制药、食品、基因工程、农业和养殖业等多个领域具有远大的应用前景和发展价值[7-8].

然而,目前关于抗菌肽作用机制的理论依据较为缺乏,抗菌肽以何种机制杀死病菌、具体的作用过程如何、哪些特征对抗菌活性具有重大影响等关键问题至今依然没有完全弄清楚[9],这对抗菌肽的人工制备造成了很大的困难. 尤其是近年来一些抗菌肽被发现具有多效抗菌活性,能够同时对多种不同类型的微生物都具有杀灭效果,比如银屑素[10]能同时杀灭大肠杆菌和丝状真菌,而乳铁素[11]则对细菌、真菌、病毒、癌细胞都有抑制和抵抗作用. 这类多效抗菌肽具有更加广泛强效的抗菌能力,在临床应用

上具有更强的实用性，但其作用机理更加复杂，人工改造和设计更为困难，特别需要深入地探索和研究.

为了能充分了解抗菌肽的相关知识，对抗菌肽进行预测是探索抗菌肽作用机制和规律的重要途径. 为了从多肽中发现抗菌肽，传统的方法是采用实验手段对多肽进行处理，通过观测其是否具有抗菌活性来得到识别结果，这类实验方法虽然识别的准确率较高，然而过程却比较复杂，需要耗费大量的人力、费用和时间，且无法对抗菌肽的活性进行预测. 随着高通量蛋白质组学的发展，蛋白质和多肽序列数量急剧增长，人们需要从海量多肽样本中鉴别出有效的抗菌肽，并对其潜在的抗菌活性进行科学预测，而实验方法因其固有的缺陷已经远远无法满足需求，因此迫切需要找到其他行之有效的方法对抗菌肽的功能信息加以识别和预测[12]. 而随着生物信息学近年来的迅速发展，基于计算方法的智能预测成为目前解决上述问题最为有效的手段[13].

计算方法是通过对数据库中数据信息的提取和挖掘，采用智能计算的方式，将实际的生物预测问题抽象成为数学问题，并建立相关的算法来处理. 计算方法不仅具有精度高、成本低、高效快捷等优点，而且相应的生物信息数据库和各类计算工具还具有良好的可操作性和批量处理能力，能够为相关研究者提供方便自由的服务. 更为重要的是这些方法能够挖掘到数据中隐含的信息，提炼出不易发觉的规律和关联. 机器学习、数据挖掘和模式识别等计算方法已经被广泛应用在蛋白质亚细胞定位预测、基因识别等诸多分子生物学问题中，并取得了良好的成果.

计算方法也十分适用于抗菌肽的预测问题. 很多研究表明，与其他多肽相比，抗菌肽不仅具有一些独特的结构特征和序列模式，抗菌肽之间还存在某些共性[14]. 例如文献[15]针对抗菌肽的一级结构，分析了多条抗菌肽序列 N 端和 C 端前 15 个氨基酸残基的构成(如图 1 和图 2 所示)，发现抗菌肽的 N 端通常富含亮氨酸、丙氨酸等非极性氨基酸，C 端则通常富含赖氨酸、甘氨酸等极性氨基酸. 而抗菌肽和非抗菌肽的序列组成则有着较为明显的区别，如图 3 所示，抗菌肽序列中的半胱氨酸、甘氨酸等非电离极性氨基酸的含量高于非抗菌肽，而天冬氨酸、谷氨酸等酸性氨基酸含量则低于非抗菌肽. 另外，各类抗菌肽在两亲性和电荷性等方面具有一定相似性[16]，而不同抗菌肽间的活性差异则与其氨基酸残基排列方式、肽链结构等关系密切[17]，某些特定的氨基酸组合或蛋白质二级结构也常出现在特定功能的抗菌肽中，而一些特殊位置上的氨基酸残基则具有很强的保守性[18].

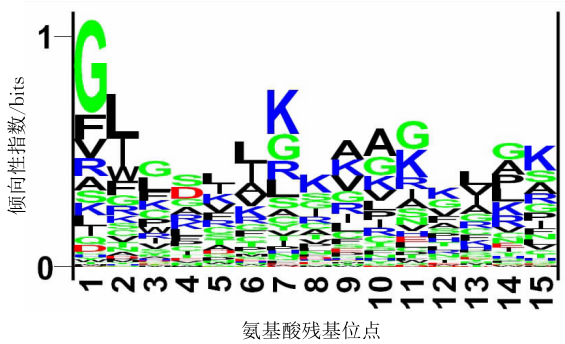


图 1　抗菌肽 N 端前 15 个位点上的氨基酸残基序列标识[15]，残基标识的大小为氨基酸在该位点出现的倾向性指数，该指数越大表示该氨基酸被分配到该位点的可能性就越大，图 2 同

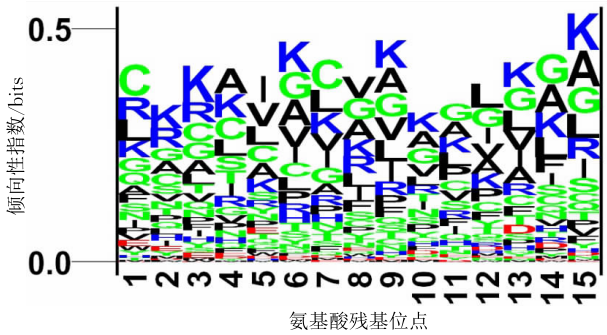


图 2　抗菌肽 C 端前 15 个位点上的氨基酸残基序列标识[15]，残基标识的大小为氨基酸在该位点出现的倾向性指数，该指数越大表示该氨基酸被分配到该位点的可能性就越大

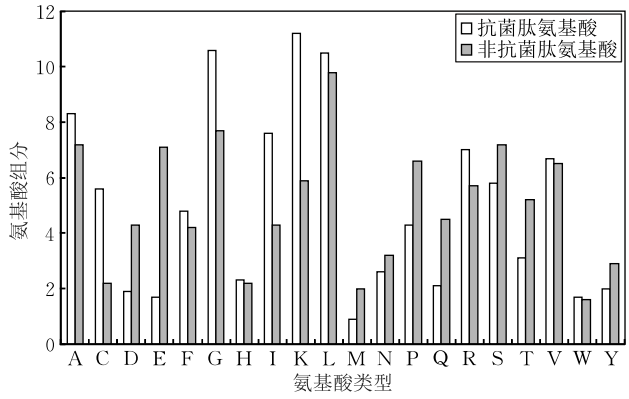


图 3　抗菌肽和非抗菌肽的氨基酸组成总体比较[15]

虽然实验方法会揭示抗菌肽的某些性质，但哪些性质是抗菌肽所独有的，而哪些是与其他多肽类所共有的往往很难直观地确定，很可能多种不同性质融合在一起才能产生抗菌肽独特的模式，一些潜在的关联也无法通过实验来获取. 而计算方法则通过对抗菌肽数据的挖掘来抽取出有效信息，并对此

进行学习、分析和预测，找出实验方法难以发现的内涵性规律，建立起多肽特征与抗菌活性之间的关联关系，深层次地探索、挖掘和理解抗菌肽的本质信息. 因此，预测的最终目的不是对实验结果的简单统计和总结，而是要从已知现象出发推断出未知的功能和构象，这才是计算方法的最大优势所在.

随着生物信息学的不断发展，基于计算的抗菌肽预测研究也取得了长足进步，国内外出现了各类卓有成效的成果. 由于实现预测主要的两点是数据和方法，因此现有的成果主要就集中在建立抗菌肽数据库和设计有针对性的预测算法这两方面，而预测算法则又可以分为基于经验分析的方法和基于机器学习的方法这两大类. 这些学术成果结合抗菌肽数据信息，将数学、统计学、计算机科学、信息技术等与分子生物学相结合，成为预测抗菌肽各类功能的有效工具. 总体来说，国际上与抗菌肽预测问题相关的研究成果相对多一些，而国内见诸报道的成果则相对较少，本文主要对抗菌肽数据库和预测算法方面的国内外研究进展进行介绍.

2 抗菌肽生物信息数据库

分析预测离不开数据，过去的十多年里国内外研究人员陆续建立了多个抗菌肽数据库，收录了大量的抗菌肽数据. 使用者不仅可以从中提取所需要的信息，还可以使用数据库所提供的各类工具对数据进行处理.

大型蛋白质数据库 UniProt[19] 和 PDB[20] 就收录了多条具有抗菌功能的蛋白质，绝大部分抗菌肽都可以在这些蛋白质中找到，并且其收录的数据通常都具有经人工检验过的信息来源和功能注释. 但是这两个数据库主要收录的是蛋白质，通常不能直接用于预测，而是需要合理地拆分蛋白质肽链才能提取出有效的短链抗菌肽序列. 因此有很多研究人员专门为抗菌肽研究建立了抗菌肽数据库，这些数据库中收录的都是能够直接用于预测的抗菌肽序列. 抗菌肽数据库按收录内容可分为综合数据库和专题数据库两类，综合数据库收录了各种来源各种类型的抗菌肽数据，而专题数据库则根据研究角度的不同只收录特定的抗菌肽数据.

主要的综合数据库包括 APD[21-23]、DBAASP[24] 和 CAMP[25-27] 等，这些数据库数据量往往较大且含有多种类型的抗菌肽数据，并提供了抗菌肽的查询、序列比对、预测和分析等多种功能和工具. 专题数据库则针对某类专门的抗菌肽而建立，主要是为了研究特定类型的抗菌肽，如研究抗病毒肽的 AVPdb[28]、研究抗肿瘤肽的 CancerPPD[29]、研究抗寄生虫肽的 ParaPep[30]、研究防御素的 Defensins[31]，以及研究抗 HIV 病毒肽的 HIPdb[32] 等. 此外，也有根据抗菌肽来源而建立的数据库，例如专门收录蛙类来源抗菌肽的 DADP[33] 等.

与综合数据库相比，专题数据库由于限定了研究范畴因此收录的数据量相对较小，但是对抗菌肽的描述和分析则具有较强的针对性，更适用于对特定抗菌肽的研究. 我国目前关于抗菌肽的数据库还比较少，比较有代表性的是上海复旦大学遗传工程国家重点实验室建立的综合数据库 LAMP[34].

一些常见的抗菌肽数据库如表 1 所示，这些数据库存储了基本的抗菌肽数据，通常会记录抗菌肽的氨基酸序列、物理化学性质、抗菌功能等注释信

表 1 常见的抗菌肽数据库

数据库名称	创建时间	数据量	网址	概述	主要功能
APD3[23]	2015	2619	http://aps.unmc.edu/AP/main.php	综合数据库 美国	检索、预测，多肽设计、数据下载
DBAASP[24]	2014	4054	http://dbaasp.org/home.xhtml	综合数据库 法国	查询、排序搜索、理化性质计算、预测
$CAMP_{R3}$[27]	2015	10247	http://www.camp.bicnirrh.res.in	综合数据库 印度	检索、预测、序列比对、序列模式挖掘
LAMP[34]	2013	5547	http://biotechlab.fudan.edu.cn/database/lamp	综合数据库 中国	浏览、检索、Blast 序列比对
YADAMP[35]	2012	2133	http://www.yadamp.unisa.it	综合数据库 意大利	检索、抗菌性预测
DAMPD[36]	2011	1232	http://apps.sanbi.ac.za/dampd	综合数据库 南非	序列比对、理化性质查询、预测
ADAM[37]	2015	7007	http://bioinformatics.cs.ntou.edu.tw/adam	综合数据库 中国台湾	检索、浏览、预测、数据下载
AVPdb[28]	2013	2683	http://crdd.osdd.net/servers/avpdb	抗病毒肽数据库 印度	检索、Blast 比对、理化性质计算、预测
CancerPP[29]	2014	3491	http://crdd.osdd.net/raghava/cancerppd	抗肿瘤肽数据库 印度	检索、浏览、序列比对、数据下载
ParaPep[30]	2014	863	http://crdd.osdd.net/raghava/parapep	抗寄生虫肽数据库 印度	检索、浏览、相似性比对、数据下载
Defensins[33]	2006	350	http://defensins.bii.a-star.edu.sg	防御素数据库 新加坡	检索、数据分析、临床研究介绍
HIPdb[32]	2012	981	http://crdd.osdd.net/servers/hipdb	抗 HIV 病毒肽数据库 印度	检索、浏览、序列比对、数据下载
DADP [33]	2012	2571	http://split4.pmfst.hr/dadp	蛙类抗菌肽数据库 意大利	检索、数据分类

息，并对数据进行了初步的统计和分析. 一些数据库中还包含了序列比对工具来衡量目标多肽与已收录的抗菌肽之间的相似度，在样本较为有限的情况下，这些工具能够为寻找抗菌肽特征及其家族分类提供有用的信息. 部分数据库则提供了预测工具，但主要是用于鉴别抗菌肽和非抗菌肽. 不过，目前的抗菌肽数据库的功能还是较为简单，所包括的与序列分析和药物发现相关的工具还非常欠缺，这方面的工作还需要进一步增强. 随着预测方法的发展，整合各个实验室的数据，建立数据量和功能更加丰富的标准化数据库，将会为抗菌肽预测方法的深入研究提供更为有力的保障.

3　基于经验分析的抗菌肽预测方法

基于经验分析的方法是根据已知的经验规则或者模式对一类抗菌肽的某些生化属性和抗菌活性之间的关联进行统计或建模，利用验证的方式识别出该类抗菌肽. 这种方法主要利用同类抗菌肽样本进行训练，通常没有非抗菌肽和其他类别抗菌肽的参与，主要的预测方式是从待测多肽样本中识别并挑选出模型所描述的该类抗菌肽，一般没有通用的量化指标对这类方法的性能进行统一地评估和比较，下面介绍一些有代表性的方法.

3.1　基于序列比对的方法

序列比对是一种将两条或多条序列按照一定规律排列并进行对比的序列分析方法，其基本思想是找出待测序列和数据库(或训练集)中目标序列的相似性. 基于序列比对的抗菌肽预测是将多肽的氨基酸序列看成由基本字符组成的字符串，把待测序列同数据库中已收录的抗菌肽序列按照一定的规律排列在一起进行比较，并以字符的异同作为预测的依据. 通过序列比对可以搜索相似序列，并利用相似性进行同源性分析. 比对过程中需要在检测序列或目标序列中引入空位，以表示插入或删除，如图 4 所示.

```
Query:   179   ENGFRYIPFRIY----------QTTTER--------PFIQKLFRPVAADGQLHTLGDL   218
                 F+ IP RIY              TD+R          F  ++    A  G   T
Sbjct:   181   LESFKNIPLRIYTDDVRLHVHPETDFTDQRGRTKEEFGRFNGRIIDTCAQSGSFGTRIGA   240
```

图 4　序列比对，用字符表示相同的残基，“－”表示允许此处插入或删除残基以保证比对残基数目匹配，“＋”表示相似残基

序列比对的数学模型大体可以分为两类，一类是从全长序列出发，考虑序列整体相似性的整体比对，另一类是考虑序列部分区域相似性的局部比对. 抗菌肽大多是由较短序列片段组成的，其功能位点的序列具有相当大的保守性，局部相似性比对往往较之整体比对具有更高的灵敏度，其结果也更具有生物学意义，因此用于抗菌肽预测的序列比对方法通常采用局部比对.

目前最常用的序列比对工具主要包括基于局部相似性的比对搜索程序 FASTA[38]、BLASTP[39]，基于多次双序列两两比对的渐进多序列比对程序 CLUSTALW[40]，基于动态规划的 Smith-Waterman 算法[41]，以及基于谱隐马尔可夫模型的序列分析工具 HMMER[42] 等. 通过序列比对方法，检验待测多肽序列与数据库中的抗菌肽序列的相似性，就可以对抗菌肽进行预测.

抗菌肽的预测问题中主要采用的是双序列比对程序 BLAST 中用于蛋白质序列比对的算法 BLASTP. Wang 等人[43] 采用 BLASTP[44] 算法是比较有代表性的一种方法，通过比对待测序列与训练集中的短序列来发现最佳匹配序列来进行预测. 该方法先利用 BLASTP 程序进行 scanning 来确定匹配片段，序列的匹配程序由短序列(word)的联配得分总和来决定. 短序列的每个碱基均被计分：碱基对完全相同的得较大正值，不太匹配的得较小正值，完全不匹配的得负值，最后将各碱基对的分值相加，得分高的匹配序列称为高比值片段对(High-Scoring Segment Pairs，HSP)，最后根据总得分高低来判断序列间的相似程度. 对于一条待测多肽序列 $\boldsymbol{P}$ 和训练集$\{\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_n\}$，如果 $\boldsymbol{P}$ 和某一个训练样本 $\boldsymbol{P}_k$ 的 HSP 值(Score)满足式(1)，则认为 $\boldsymbol{P}$ 和 $\boldsymbol{P}_k$ 属于同一类别；若超过一个训练样本都满足式(1)，则 $\boldsymbol{P}$ 的类别在匹配的训练样本中随机进行分配.

$$\mathrm{Score}(\boldsymbol{P},\boldsymbol{P}_k)=\max\{\mathrm{Score}(\boldsymbol{P},\boldsymbol{P}_i)\mid i=1,2,\cdots,n\} \quad (1)$$

Ng 等人[45] 也采用类似的序列比对方法来预测抗菌肽，主要区别在于该方法先将训练集按类别分为正样本集合和负样本集合，然后将待测样本分别在两个集合上进行 BLASTP 序列比对得到两组中的最大 HSP 值，在哪个集合上得到的 HSP 值大，就表示待测样本与该集合上全体样本的相似程度更高，并推断该样本与该集合的类别相同.

上述两种序列比对方法比较依赖于训练集的规

模和类别丰富程度,对于某些特殊的待测样本会出现该样本与训练集中全部样本都不相似,即匹配度为零的情况,此时不能得到 HSP 值造成无法预测,只能再采用其他方法来处理,比如文献[43]和[45]分别采用了特征选择方法和 LZ 复杂度方法来应对这一情况,但都没序列比对方法的精度高.

一般来说,采用序列比对方法时,训练集中的抗菌肽种类和数量越多,出现无法预测的概率就越低,预测的精度也会越高.抗菌肽数据库由于收录了大量的抗菌肽数据,非常适合序列比对方法,因此抗菌肽数据库自身提供的预测工具大多基于序列比对方法而建立,通过比对数据库自身的抗菌肽序列来实现预测.比如 APD[23]、YADAMP[35]数据库的预测工具就采用了 BLAST 序列比对程序进行抗菌肽的预测,如果发现待测样本具有与抗菌肽序列相似的特征(比如某些疏水性残基有规律地出现在一些位点中),就推断该多肽具有很高的概率为抗菌肽.

此外,Xiao 等人[46]使用 ClustalW 多序列比对程序来鉴别从鸡肉组织中提取的 cathelicidin 族多肽的潜在抗菌性,他们将待测样本与全部已知的 cathelicidin 前体细胞中的氨基酸序列进行 ClustalW 算法比对.该方法先两两比对计算样本间氨基酸差异来得到各个样本之间的距离并获得距离矩阵,再利用邻接法(Neighbor-Joining)[47]构建引导树,根据引导树从最相近的两条序列开始,逐步引入临近的序列并反复重建比对,渐进地比对多个序列,最终成功鉴别出三条新的鸡类 cathelicidin 抗菌肽.

基于序列比对的预测方法简单直观,相对易于实现,但是如何给出一个合理优化的相似性度量准则目前还没有很好的标准,而且对于分歧较大的序列,预测的准确率以及算法的时间复杂度也都有待提高.另外,如果出现与训练数据匹配度极低的样本,该方法只能借助于其他方法来解决.

3.2 基于定量构效关系的方法

定量构效关系(Quantitative Structure-Activity Relationships,QSAR)建模[48]是另一种常见的基于经验分析的预测方法,该方法是通过对一系列抗菌肽的结构或理化性质的定量描述,借助数学和统计学方法建立抗菌活性和 QSAR 描述子(即多肽分子表征)之间的量化模型,预测时输入待测样本的相关参数通过计算来求得相应指标数值,以此来确定变量之间相互依赖的定量关系,从而检验待测样本是否符合模型描述.

抗菌肽的 QSAR 预测方法基本上可分为以下 5 个步骤:

(1)选择一系列已知的抗菌肽;

(2)对抗菌肽进行生物活性的测定;

(3)进行抗菌肽结构的定量表征;

(4)建立数学模型,确定化学结构与生物活性之间的函数关系;

(5)对待测样本进行模型检验以预测其抗菌性.

抗菌肽的 QSAR 预测方法利用计算机对抗菌肽的信息进行数学分析,利用数学模式来描述抗菌肽分子结构的结构参数、理化参数与抗菌性质之间的相互关系.定量构效关系方法的核心在于如何建立 QSAR 模型,包括抗菌肽结构的表征方法、理论模型的推导方法和函数关系的建立等.常见的结构表征方法包括分子连接性方法、电拓扑状态指数方法、分子形状分析方法等;主要的建模方法则包括多元线性回归、主成分分析、偏最小二乘法等[49].

早期的 QSAR 方法大多局限于对单独的一类抗菌肽进行建模,采用的描述子比较简单但更有针对性.比如 Strøm 等人[50]用根据 20 个乳铁素抗菌肽与 α-螺旋和静电荷等相关的 12 种描述子建立了 QSAR 模型;类似地,Frecer[51]使用了 25 种不同的描述子建立了环形抗菌肽模型;而 Hilpert 等人[52]则使用 51 种描述子建立了短抗菌肽的 QSAR 模型;重庆理工大学的 Shu 等人[53]使用主成分分析法得到拓扑结构描述子,结合偏最小二乘法建立了关于牛科动物抗菌肽的 QSAR 模型.

上述这些模型根据各自的研究对象采用有针对性的生化指标即 QSAR 描述子,虽然对于特定类别的抗菌肽具有不错的预测能力,但只能反映同类抗菌肽的特性,预测的范围和规模十分有限.为此 Cherkasov 团队的一系列研究[15,54-55]采用了多肽可以通用的诱导描述子(Inductive Descriptors),以绝对电负性、共价半径、分子间距离等与诱导效应相关的原子规模信息作为参数,建立了多肽分子内和分子间相互作用与抗菌活性之间的 QSAR 模型.该方法考虑抗菌肽的化学中性分子的电负性特征,将分子中的带电原子球作为原子电容器来研究,这样抗菌肽性质描述参数就能够通过基本的原子参数来表示,如原子的电负性 X,共价半径 R 和分子间距离 r,这样抗菌肽分子间的相互作用就可以用带电原子间的相互作用模型来定量地描述,比如由 n 个原子组成的原子团 G 对第 j 个原子关于原子空间 Rs 和诱导因子 σ^* 的关系就可以由效应式(2)和(3)来计算:

$$Rs_{G\to j}=\alpha\sum_{i\subset G,i\neq j}^{n}\frac{R_i^2}{r_{i-j}^2} \tag{2}$$

$$\sigma_{G\to j}^{*}=\beta\sum_{i\subset G,i\neq j}^{n}\frac{(X_i^0-X_j^0)R_i^2}{r_{i-j}^2} \tag{3}$$

类似地，该方法引入 50 个抗菌肽分子的诱导描述子，并分别找到它们的电负性函数关系来建立 QSAR 模型，并以此刻画抗菌肽分子特性. 预测时，将待测多肽的相关参数代入模型，通过观察比对其电负性特征来判断其是否具有抗菌性.

定量构效关系方法是建立在实验基础上，从抗菌肽的分子结构和能量特性等要素出发，将抗菌肽的抗菌活性看做是其原子和基团间相互作用的外在表现，具有较高的预测精度. 但是由于抗菌肽物质结构相对复杂，导致 QSAR 模型的计算复杂度较高，并且受到了分离、纯化和合成等生化技术发展的制约，定量关系构效方法目前只能基于已确定的抗菌肽样本进行建模，通常只能应用于对特定类别的抗菌肽进行建模，无法用于大规模多类别的抗菌肽预测，而且模型的物理意义比较模糊.

3.3　基于模糊逻辑模型的方法

模糊逻辑模型是一种通过定义模糊集合和规则库，根据需要将因变量作为独立变量的一个函数，从而对因变量进行预测的方法.

Mikut 和 Hilpert[56] 提出了一种将模糊逻辑引入到分子描述子的表达中来分析抗菌肽的方法，并通过模糊规则来描述抗菌肽的性质. 该方法先将多肽分子的一些理化特性按照相关的数值展开为一个实值向量，例如对于长度为 n 的多肽 P，其第 l 种物化性质为亲水性，则其亲水性可以表示为一个由其各个氨基酸的亲水性指数组成的向量：

$$\boldsymbol{x}_l(\boldsymbol{P})=(\boldsymbol{x}_l(1,\boldsymbol{P}),\cdots,\boldsymbol{x}_l(n,\boldsymbol{P}))^{\mathrm{T}} \tag{4}$$

然后将该向量转化为模糊集 $\mu_{A_{l,i}}(n,\boldsymbol{P})\in[0,1]$ 上的隶属度值，计算时采用梯形隶属度函数，并将多肽中各氨基酸的隶属度按下式求出均值：

$$\mu_{A_{l,i}}(\boldsymbol{x}_l(\boldsymbol{P}))=\frac{1}{n}\sum_{n=1}^{n}\mu_{A_{l,i}}(\boldsymbol{x}_l(n,\boldsymbol{P})) \tag{5}$$

作为多肽 $\boldsymbol{P}$ 在第 l 个物化性质上的隶属度函数值. 该值介于 0 和 1 之间，值为 0 表示该序列中没有任何氨基酸具有该特性，而值为 1 则意味着全部氨基酸都具有这个属性. 因此对于一个给定长度的多肽，属性能够根据若干个氨基酸的函数来计算推断，这样就可以使用简单的规则来刻画抗菌肽的活性. 该方法对于区分有活性和无活性的多肽具有较好的预测准确度.

此外，Fernandes 等人[57] 也提出了一种基于模糊模型的抗菌肽分类方法，他们研究发现抗菌肽的一些物化性质与其抗菌性之间存在着模糊模式，因此他们通过一个模糊推断系统建立了与多肽两亲性相关的“if-then”规则来获取隶属度函数，从而得到输入-输出映射，从而对多肽做出鉴别. 进一步地，该团队将这一模糊模型同自适应神经网络相结合，建立了用于抗菌肽预测的自适应神经-模糊推理系统[58].

基于模糊逻辑的方法不需要建立精确的数学模型，模糊规则相对比较简单，易于实现. 但隶属度函数的建立缺乏系统的方法，主要依赖经验和试凑，难以总结统一的规则，对不同类型的抗菌肽样本往往需要构造新的模糊模型，方法的泛化性不强. 模糊逻辑的计算可以使用开源 MATLAB 工具箱 Gait-CAD(http://sourceforge.net/projects/gait-cad)来实现.

3.4　基于语言模型的方法

在分子生物学中，多肽中的每一种氨基酸都用一个相应的英文字母直观地表示，因此一个多肽序列也可以看做是一个按一定顺序排列的英文字符串. Loose 等人[16] 考虑到多肽序列在格式上较为类似于英文短语这一特点，针对抗菌肽预测问题建立了一种语言模型(图 5). 他们将氨基酸看做独立的字母，而肽链则是由氨基酸字母按照一定的语法规则排列而组成的句子，并采用 TEIRESIAS[59] 模式识别工具来进行预测. TEIRESIAS 是一种通用的两阶段组合的模式识别算法，其对于解决一些蛋白质族的模式发现问题效果显著. 为了寻找到与抗菌

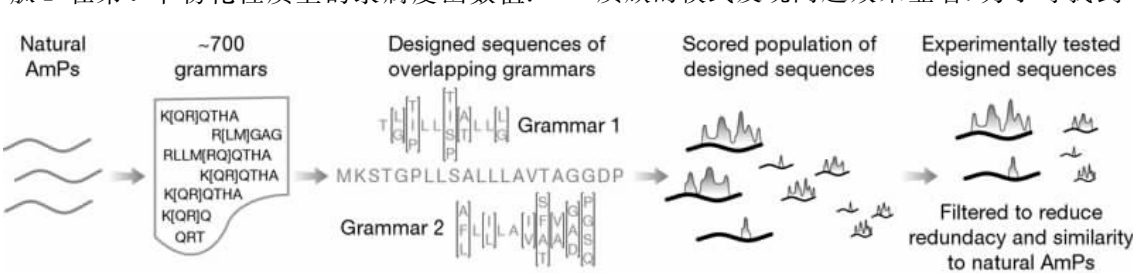


图 5　使用语言模型预测抗菌肽，根据抗菌肽资源库建立语法规则，并可以将其用于制造抗菌肽序列[16]

肽活性相关的语法规律，Loose 等人对 APD 数据库中的天然抗菌肽进行了模式分析，找出了 684 个语法规则，合并这些规则可以用来发现新的抗菌肽，相关成果发表在《Nature》中.

还有一种语言模型方法则是将不同的序列归纳成一套由 20 种天然氨基酸组成的简化字母表(RAAA)，即对抗菌肽序列进行聚类，并用来描述其多个特征，这种方法在预测蛋白质的结构类方面已经取得了成功[60]. Veltri 等人[61]采用了 GBMR4 字母表来建立蛋白质序列的模体特征，表里只包含 A、C、G、T 这 4 种氨基酸，并以此建立低维度的特征向量，再利用遗传算法进行抗菌肽的识别. GBMR4 表中字母与标准氨基酸的映射关系，及其所代表的氨基酸性质如表 2 所示.

表 2 GBMR4 字母表[61]

氨基酸	映射	注释
ADKERNTSQ	A	具有某些特殊转角的小品种氨基酸
CFLIVMYWH	C	非极性和/或芳香族氨基酸
G	G	柔性氨基酸
P	T	刚性氨基酸

Zuo 等人[62]则采用了结构化的字母表作为蛋白质区块(Proteins Blocks，PB)[63]来研究防御素抗菌肽. 蛋白质区块是一种由 16 个有代表性的结构化模体组成的字母表 RAAA，每个模体的氨基酸长度为 5，蛋白质 3D 结构也能够转化为这种蛋白质区块序列来处理[64]. 该团队基于防御素知识数据库，将多样性增量理论与 RAAA 结构化字母表相结合，对防御素的家族进行了预测分类，包括昆虫、植物、脊椎动物及余下其他四种，得到了 91% 的整体精度. 进一步地，他们又对脊椎动物的防御素进行了子家族分类，即 α-防御素、β-防御素以及 θ-防御素，并取得了 94% 的预测准确率.

基于语言模型的方法能够较好地将抗菌肽序列抽离出来建立简单有效的语法规则，由于采用统一的字母表和语法规则，十分适合计算机自动执行预测任务. 但该方法比较依赖训练集数据中的现有语义模式，对于语法未知的新样本预测和挖掘能力较弱，很难识别出新类型的抗菌肽.

3.5 其他经验分析方法

除了上述几类常见的方法外，有些研究者也提出了其他的一些基于经验分析的方法.

一些研究人员利用分子动力学仿真方法进行抗菌肽的预测，例如对菌膜和抗菌肽磷脂双分子层的相互作用进行仿真. 目前有很多用于分子动态仿真的免费和商业软件，比如免费程序 CHARMM[65](http://www.charmm.org)就被广泛地用于模拟抗菌肽脂双层或去污微团的相互作用；而另一个免费程序包 GROMACS[66](http://www.gromacs.org)经常被用在分子动态研究中用来生成分子轨迹；Discovery Studio(http://accelrys.com/products/discovery-studio)是一套集成了一系列分析应用程序的商业软件，能够为序列分析、QSAR 建模、分子建模和计算机仿真提供解决方案. 分子动力学仿真方法能够模拟生物实验条件，并能自由地调整模拟的条件和参数，从而动态地观察和显示各种结果. 但该方法只能进行特定的分子动力学仿真，不能全面地考察抗菌肽特性，预测性能不强. 文献[67]对抗菌肽分子动态的计算机仿真方法进行了较为详尽的介绍.

此外，Nagarajan 等人[68]提出了一种基于傅里叶变换和欧几里德度量的编码方式，通过傅里叶变换将多肽的 5 个与抗菌活性相关的生化指标变换到频域空间来得到每条多肽的功率谱，通过分析发现抗菌肽的功率谱在频域空间上具有独特的峰值，并据此来观察和比较待测样本的功率谱，从而对其抗菌性进行预测；Yount 和 Yeaman[69]则提取出多肽的多维度特征作为关键的模式，通过识别 3D 结构下与抗菌活性相关的序列和模体来进行预测；而 Jenssen 等人[70]基于多元统计、主成分分析和偏最小二乘法建立了对抗菌肽的预测模型；四川大学的杨莉等人[71]通过计算每个氨基酸在每个位置时对整段抗菌肽活性的贡献值，并进行抗菌活性标准化换算来建立预测模型，从而可以对随机获得的多肽进行抗菌活性预测. 一些主要的基于经验分析的预测方法如表 3 所示.

基于经验分析的方法对模型所反映的那类抗菌肽具有较好的预测精度，但由于预测的对象比较封闭，这类方法往往不能迁移到对其他类别的抗菌肽和非抗菌肽的识别上，因此很难判断未被识别出的多肽样本确实不是抗菌肽，还是该样本实际上是模型无法描述的另一种抗菌肽. 另外，基于经验分析的方法在建立模型时大多采用如多元线性回归或主成分分析这类线性方法，虽然能够比较直观地反映多肽的生化特征与抗菌活性之间的线性关系，但是却难以发现诸如分子间相互作用等具有非叠加性质的非线性关系.

表 3　几种主要的基于经验分析的预测方法

名称	描述	适用范围	优点	局限性
序列比对	将待测样本与已知样本的序列进行字符比对，通过比较其相似程度来预测	训练样本数量充足且氨基酸序列已测定	简单直观，易于实现	较为依赖训练数据，算法效率不高，准确率偏低，相似性度量准则无统一标准
定量构效关系模型	对抗菌肽的结构和理化性质进行定量描述，并建立抗菌活性和多肽分子之间的量化模型，通过检验待测样本是否符合模型进行预测	分子特征和物化性质明确的抗菌肽	计算量小，预测能力较强	物理意义模糊，变量间作用模式难以给出，预测范围小
模糊逻辑模型	通过模糊规则来描述抗菌肽的性质，通过隶属程度来推断多肽的抗菌性	物化性质指标可以定量计算的样本	无需建立精确模型，模糊规则较为简单	建立隶属度函数缺乏系统方法，泛化性不强
语言模型	将抗菌肽序列看做英文短语，通过建立字母表和语法规则，利用模式识别方法做出预测	氨基酸残基序列排布确定的多肽	形式统一，适用于计算机自动处理	依赖现有知识，对未知规则的样本挖掘能力较弱
分子动力学仿真	利用抗菌肽分子中的一些相互作用，建立动力学仿真模型来预测抗菌肽	分子动力学特性已知的抗菌肽	能够模拟生物实验，自由调整模拟条件进行动态观测	仿真对象较为单一，预测性能不强

4　基于机器学习的抗菌肽预测方法

机器学习是通过对抗菌肽的已知数据和已有经验进行学习，采用推理、归纳、综合或模型拟合等方式，从中找出规律并将其推广到未知数据的方法. 这类方法不仅可以发现抗菌活性同生化属性之间的线性关系，还可以挖掘内在的非线性关联，对于模型复杂且缺乏一般性理论的抗菌肽预测问题来说非常适合.

用于抗菌肽预测的机器学习方法主要采取的是有监督学习，这种学习方式将抗菌肽的预测转化为一种分类问题来处理. 该方法根据已知样本的类型将多肽划分成不同的类，然后通过建立有效的分类器，对待测样本最可能归属的类别进行合理的推断，从而达到预测的目的.

如图 6 所示，机器学习方法的基本流程主要包括学习、预测和验证三个阶段，在学习阶段，先要构建多肽数据集，数据集中既包括抗菌肽样本也包括非抗菌肽样本，再利用特征提取方法抽取出抗菌肽的模式特征，用数学描述来表达抗菌肽的特质，并在此基础上进行分类算法设计和相关参数学习从而训练得到相应的预测模型；在预测阶段，将待测多肽样本提交给训练好的预测模型，通过计算机处理输出相应的预测结果；在验证阶段，根据不同的评价指标检验所设计的分类器性能，如精度、时间成本和算法泛化性等，并通过优化调整相关分类器参数，以得到相对满意的结果.

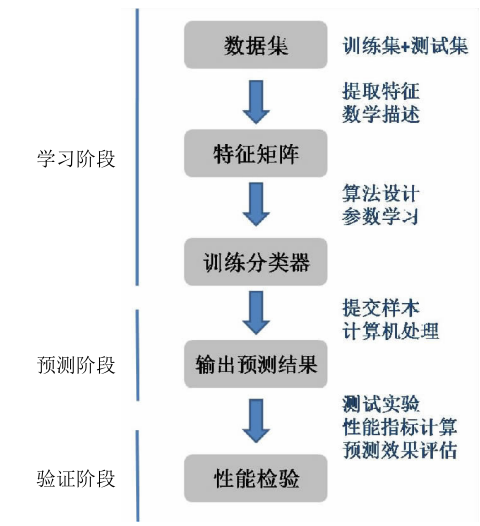


图 6　基于机器学习的抗菌肽预测方法流程

抗菌肽数据包含了很多文字化注释信息，如序列排布情况、理化性质、抗菌性类型等，无法直接被机器学习方法所使用. 因此为了能用分类算法对抗菌肽数据进行预测，要先将这类数据进行序列编码，也就是用数学方法对抗菌肽的特征进行定量的描述. 主要的特征提取方法是从抗菌肽的一级、二级结构以及理化性质活性关系中提取相关特征信息，如氨基酸组分、氨基酸残基理化性质、基因本体等，从而实现对抗菌肽生物特征的量化. 特征提取的结果一般是将一条多肽的特征 $\boldsymbol{X}$ 用 n 维向量 $\boldsymbol{X}=\{x_1, x_2, \cdots, x_n\}$ 来表示，每一维分量都能够刻画多肽的一个特征，例如抗菌肽的氨基酸组分特征就可以简单地表示成一个 20 维的向量，其中每一个分量的值为一种氨基酸在该抗菌肽序列中的出现次数. 而多肽的类别标签 $\boldsymbol{Y}$ 可以根据实际问题中样本类型来分别定义为不同的值，类别相同的样本有相同的标签值. 这样，每一个多肽样本就可以由其相应的特征向量 $\boldsymbol{X}$ 和类别 $\boldsymbol{Y}$ 来唯一的表示，训练集中每个数据所对应的$(\boldsymbol{X},\boldsymbol{Y})$用于学习分类器. 在进行预测时，只要将待测样本的特征值 $\boldsymbol{X}^*$ 输入到训练得到的分类器

中，分类器就会输出预测的 $\boldsymbol{Y}^*$ 值，并将待测样本预测为其 $\boldsymbol{Y}^*$ 值所对应的类别.

由于不同的预测任务下数据的模式和特性有所不同，通常都要有针对性地选择和设计分类算法，而分类算法的性能往往直接决定了预测的最终效果，因此在整个预测过程中，如何设计具有高性能的分类算法是整个预测方法的最核心问题. 在国内外的研究中出现了很多基于机器学习的抗菌肽预测算法，从单一的分类算法到多算法集成与综合学习，从基于单标签学习的抗菌肽鉴别研究拓展到多标签学习的抗菌肽活性预测，算法的准确度不断提高，应用范围也逐步扩大，取得了不少优秀的研究成果，下面对一些有代表性的成果加以简介.

4.1 基于二分类学习的抗菌肽鉴别

抗菌肽预测研究一个主要的内容是抗菌肽的鉴别，也就是判断一条未知的多肽是否为抗菌肽，属于有监督学习中的二分类学习问题. 这种学习方式在预测时将全部多肽样本划分为两类：抗菌肽和非抗菌肽，前者标记为正类样本后者标记为负类样本，并以此为前提对全部样本进行学习. 在预测时若分类器将待测样本划分为正类则判断该多肽样本为抗菌肽，否则就推断其为非抗菌肽. 此时，多肽的类别标签 y 为一个数值，表示该样本是否具有抗菌性，比如若该条多肽为抗菌肽，可以令 $y=1$，若其为非抗菌肽，令 $y=-1$. 相应的，在预测时若输出的标签值为 1 就推断该多肽具有抗菌性，否则认为其为非抗菌肽，这样就得到了预测结果.

早期抗菌肽预测的机器学习方法研究，主要是将一些经典的有监督学习算法直接应用于抗菌肽序列，常见的算法包括人工神经网络、支持向量机和随机森林等.

4.1.1 人工神经网络

人工神经网络(Artificial Neural Network，ANN)是一种模仿大脑的神经元之间传递和处理信息的数学模型，是比较早地被用于识别抗菌肽的一种机器学习方法[13].

神经网络用于抗菌肽预测的一个优势在于，它能够通过改变内部的激励函数来处理不同特征类型的样本，从而具有较强适应能力. 在进行预测时，为了能充分学习到现有抗菌肽的特征信息，通常会从综合数据库里选择各种来源不同、功能各异的抗菌肽用作训练，由于各类抗菌肽的特性差别较大，这就需要预测模型能够适应不同的复杂样本，因此神经网络十分适用于抗菌肽的预测问题. 神经网络的另一个优点则是其具有较强的容错和容差能力，在训练的过程中能够取得稳健的误差估计. 训练分类器所使用的抗菌肽数据一般都来自于数据库，其中的一些抗菌肽数据很可能含有错误信息，这些错误有些来自于实验本身或人工注释的失误，也有一些是因为缺少必要的实验而造成信息残缺，因此对于含有噪声的抗菌肽数据的学习，神经网络具有很不错的效果.

在应用神经网络对抗菌肽进行预测时，模型的输入就是多肽样本的特征向量，模型的输出为样本的类别标签. 模型的结构和参数的选择则没有固定的模式，现有的方法主要还是根据经验来确定具体的神经网络算法. Torrent 等人[72]基于多肽序列的理化性质特征提出了一种共轭梯度 BP 神经网络分类模型，该网络的隐含层包含 50 个神经结点，而利用共轭梯度法保证了算法的全局收敛性，同时也加快了收敛速度，该方法不仅取得 90% 的预测精度，还发现了序列的结构和聚合参数对识别抗菌肽具有重要影响. Holton 等人[73]则利用一种新型的 $N-1$ 神经网络对序列中模体的位置和长度特征进行了学习，$N-1$ 神经网络的输入为长度为 N 的多肽序列，输出为多肽样本的抗菌性，整个网络是由 N 个序列-特征级联神经网络和一个特征-输出网络组成，每个网络均采用双层反馈神经网络. $N-1$ 神经网络并不是计算氨基酸出现的频率，而是利用整个模体特征，并且考虑残基在序列中的位置次序关系，将过拟合的风险降到最小，该团队利用这一方法对细胞穿透肽进行了预测并建立预测工具 CPPpred，取得了较好的效果. Soltani 等人[74]利用了带有最优参数的多层感知器神经网络方法来决定 QRSA 模型的关键参数，从而帮助识别抗真菌肽的膜结构特征；而 Fjell 等人[13]则利用 QRSA 描述符训练神经网络模型，并在 1433 个随机选出的抗菌肽上得到 94% 的识别准确率.

神经网络具有较强的非线性拟合能力和并行分布处理能力，且学习规则简单，便于通过计算机执行预测任务. 然而在训练神经网络时需要大量的参数，并且网络拓扑结构和参数初值的选取只能通过经验拼凑. 另外，神经网络对自身的推理过程和依据缺乏解释能力，不能观察之间的学习过程，无法对抗菌肽内在关联进行有效地挖掘.

4.1.2 支持向量机

支持向量机(Support Vector Machine，SVM)是一种建立在统计学习理论的 VC 维理论和结构风

险最小原理基础上的二分类模型，它可以自动寻找对分类帮助较大的支持向量，并最大化两类之间的间隔，通过调整核函数能够实现线性和非线性分类问题，并能利用内积核等方式处理高维问题，具有较好的推广性能[75]. 抗菌肽数据通常具有数据量大、非线性、含噪声的特点，而且由于抗菌肽的特征类型复杂，在对其进行特征提取时往往会提取出高维特征向量，更增加了分类问题的复杂性. 而支持向量机对非线性和高维模式具有很突出的优点，因此有一些学者利用支持向量机方法来处理抗菌肽预测问题.

Rondón-Villarreal 等人[76]就提出了一种新的 p-谱核支持向量机算法来解决抗菌肽的预测问题，该方法的特点在于采用了谱核函数[77]的概念，根据抗菌肽序列排列特征将其看做字符串来处理. 在比较两个字符序列串的相似程度时，一个直观的办法就是计算它们有多少(长度为 p 的)的子字符串是相同的，因此字符序列谱被定义为，对于一个字符序列 $\boldsymbol{s}$，其 p-谱为 $\boldsymbol{s}$ 中长度为 p 的子字符串所出现的频次，而两条序列 p-谱的内积就为 p-谱核函数 k：

$$k_p(\boldsymbol{s},\boldsymbol{t})=\langle\phi^p(\boldsymbol{s}),\phi^p(\boldsymbol{t})\rangle=\sum_{\boldsymbol{u}\in Q_p}\phi_{\boldsymbol{u}}^p(\boldsymbol{s})\phi_{\boldsymbol{u}}^p(\boldsymbol{t}) \quad (6)$$

$$\phi:\boldsymbol{s}\mapsto(\phi_{\boldsymbol{u}}(\boldsymbol{s}))_{\boldsymbol{u}\in Q_p}\in\boldsymbol{F} \quad (7)$$

其中对于嵌入空间 $\boldsymbol{F}$，p-谱 $\phi_{\boldsymbol{u}}(\boldsymbol{s})$ 表示长度为 p 的子字符串 $\boldsymbol{u}$ 在序列 $\boldsymbol{s}$ 中出现的次数，$\boldsymbol{Q}_p$ 为 $\boldsymbol{s}$ 中全部长度为 p 的子字符串集合. 相应的也可以根据上述定义计算 p-谱核矩阵，例如对于序列“bar”、“bat”、“car”和“cat”，它们的 2-谱和 2-谱核矩阵分别如表 4 和表 5 所示.

表 4　序列的 2-谱示例[77]

ϕ	ar	at	ba	ca
bar	1	0	1	0
bat	0	1	1	0
car	1	0	0	1
cat	0	1	0	1

表 5　序列的 2-谱核矩阵示例[77]

K	bar	bat	car	cat
bar	2	1	1	0
bat	1	2	0	1
car	1	0	2	1
cat	0	1	1	2

对于抗菌肽预测问题，该支持向量机的 p-谱核函数为

$$k_p(\boldsymbol{s},\boldsymbol{t})=\sum_{i=1}^{|\boldsymbol{s}|-p+1}\sum_{j=1}^{|\boldsymbol{t}|-p+1}\phi(\boldsymbol{s}(i:i+p-1),\boldsymbol{t}(j:j+p-1)) \quad (8)$$

其中 $k_p(\boldsymbol{s},\boldsymbol{t})$ 表示抗菌肽氨基酸序列 $\boldsymbol{s}$ 和 $\boldsymbol{t}$ 的 p-谱核，$|\boldsymbol{s}|$ 是序列 $\boldsymbol{s}$ 的长度，$\boldsymbol{s}(i:i+p-1)$ 是 $\boldsymbol{s}$ 中从第 i 个位置到第 $i+1-p$ 个位置的子序列，函数 $\phi(\boldsymbol{a},\boldsymbol{b})$ 的定义如下：

$$\phi(\boldsymbol{a},\boldsymbol{b})=\begin{cases}1, & \text{若 } \boldsymbol{a}=\boldsymbol{b}\\ 0, & \text{否则}\end{cases} \quad (9)$$

图 7 给出了使用基于 p-谱核的支持向量机预测抗菌肽的过程，该方法在 1200 个多肽样本的 10 交叉验证下取得 88.33%的准确率.

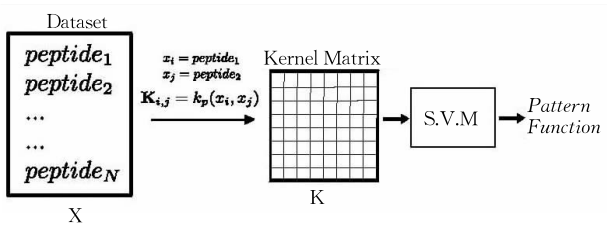


图 7　基于 p-谱核的支持向量机抗菌肽预测方法[76]

Lata 等人提出的抗菌肽预测方法 AntiBP[15]也采用了支持向量机分类器，他们从 APD 数据库中提取了 436 个抗菌肽并随机收集了等量的非抗菌肽建立了数据集，基于肽链 N 端和 C 端残基的氨基酸序列排列特征，分别对支持向量机、神经网络以及量化矩阵三种算法进行了测试和比较，最后支持向量机算法取得了最高的预测准确率(92.11%)；Porto 等人[78-79]利用支持向量机算法发现了多肽的半胱氨酸结模体特征(cysteine knot motifs)与多肽抗菌活性之间具有重要关联，并采用了三种不同核函数在稳定半胱氨酸抗菌肽数据集上分别进行了测试，其中最高的预测精度达到了 90%；而 Vijayakumar 和 Lakshmi[80]同样利用支持向量机算法建立了针对抗癌肽的预测工具 ACPP，在独立测试集上取得了 96%的准确率；Khosravian 等人[81]则基于多肽序列的伪氨基酸组分特征，分别使用支持向量机与多层感知器前馈神经网络对抗菌肽进行预测，实验结果表明支持向量机算法的预测效果更为出色，准确率能够达到 95.51%；而 Poorinmohammad 等人[82]采用支持向量机分类算法对抵抗 HIV 病毒的抗菌肽进行预测，通过对抗菌肽氨基酸残基序列的伪氨基酸组分(Pseudo Amino Acid Composition，PseAAC)特征进行学习，该方法取得了 96.76%的预测准确率.

对于抗菌肽预测问题来说，支持向量机的通用性和预测效果还是不错的，但是由于支持向量机要借助二次规划来求解，在训练样本数量很大的情况下时间和空间成本较大，在测试阶段相当耗费时间. 支持向量机对于预测问题没有通用解决方案，只能

靠尝试选择不同的核函数来处理,有时会存在过拟合的情况.

4.1.3 随机森林

随机森林(Random Forests,RF)是采用随机的方式将多个决策树组成森林的一种集成算法,是Bagging算法的一个扩展变体.基于统计学习理论,该方法利用Bootstrap重采样方法从原始样本中抽取多组样本,并在每组样本上训练决策树作为基分类器,每棵决策树只随机地用到训练数据中的一部分信息[83].而在生成每棵决策树的时候,每个结点也只是在随机选出的部分结点中产生,结点按照完全分裂的方式进行分裂,直到不能分裂为止[84].在分类时,当有一个新的输入样本进入随机森林,就让每一棵决策树分别对所给样本的类别进行预测.基于Bagging集成方式,每棵决策树在各自子数据集上并行工作,再通过投票机制将各个决策树的判定进行汇总作为最终的预测结果.

随机森林算法的训练速度快、实现简单,计算开销小,而且能够处理很高维度特征的数据,该方法在分类时能够评估输入变量的重要性.与Bagging算法相比,随机森林中基分类器的多样性不仅来自于样本扰动,还来自于特征扰动,这就能够通过增加个体决策树之间的差异度来进一步提升集成后的泛化性能.尤其是该算法可以检测到特征之间的相互影响,因此较为适合抗菌肽的预测问题.由于随机森林算法是根据训练数据的属性,一层层构建决策树,因此在训练阶段,通常要先要确立有效的抗菌肽特征属性,如肽链长度、排列顺序、亲水性等,然后随机地从这些属性中选取部分属性,通过如信息增益等策略来选择一个属性作为该结点的分裂属性.此外,将抗菌肽的二级结构属性如肽链的旋转特征加入到随机森林的训练中,能有效地提高预测性能.另外,为了提高随机森林的预测效果,往往还需要剔除训练集中序列相似度高的多肽样本.

Chang和Yang[85]就利用随机森林算法,针对抗病毒肽的预测问题建立了聚合特征和二级结构特征的物化性质分类模型,该方法采用肽链长度、净电荷数、不稳定指数、脂肪族氨基酸指数以及亲水性等五种基本理化性质,结合氨基酸组分特征属性来训练随机森林模型,取得了90%的预测准确率;Karnik等人[86]则结合了随机森林算法和递归量化分析方法对防御素抗菌肽建立了预测模型,获取了防御素的模式特征,并在238条非冗余序列数据上采用10-折交叉验证,得到了78.12%的预测准确率.而Thomas等人[25]使用CAMP数据库收录的2578个抗菌肽以及4011个在UniProt数据库中随机选取的多肽序列组成了数据集,组合了包括氨基酸的分解、物化性质以及结构特性等275个特征,分别检验了随机森林、支持向量机和判别分析方法的预测性能,最终随机森林以93.2%的准确率取得了最好的效果.

相对于神经网络和支持向量机算法,随机森林对训练数据属性的选取要求较高,对于有不同级别属性的数据,级别划分较多的属性会对随机森林产生更大的影响.然而抗菌肽的各类特征属性对算法的影响是很难预先评估的,不同类型的抗菌肽具有各自的特性,如何选取有效的抗菌肽属性还没有理论性的标准,而这也导致用随机森林算法进行预测的效果不太稳定.另外,随机森林处理数据缺失情况的能力也有待提高.

4.1.4 其他方法

除了上述几种机器学习算法,遗传算法[87]、高斯核回归[88]、集成算法[89]以及有限状态机[90]等方法也被一些研究者应用在抗菌肽预测中,取得了一定的效果.

另外,还有一些研究将机器学习方法同基于经验分析的计算方法相结合.中国科学院天津工业生物技术研究所的王萍和上海大学系统生物技术研究所的蔡煜东教授等人合作[43]提出了一种分段混合预测方法,预测时首先使用前面提到的BLASTP序列比对方法来挑选出能确定的抗菌肽,而后再将剩余的待测样本用基于特征选择的最近邻分类算法进行二次预测,该方法在新的基准数据集上的留一法测试获得了80.23%的预测成功率;而Ng等人[45]也采取类似的分段预测方式,但分类器改用了支持向量机,取得了87.59%的预测精度;Avram等人[91]将偏最小二乘法与3D-QSAR模型相结合,用来预测蜂毒肽及其衍生物的突变;Taboureau等人[92]也利用偏最小二乘法来提高QSAR模型的效果,并以此分析和设计诺弗斯匹林(novispirin)抗菌肽;Jaén-Oltrah和Tomás-Vert等人[93-94]在鉴别抗菌肽时提出了一种新的拓扑描述子,并基于采用神经网络方法对数据进行训练以获得相应的QSAR模型,对于抗菌肽和非抗菌肽的预测正确率分别达到了93.61%和95.92%;Fernandes等人[58]则将神经网络和模糊理论结合在一起,每个隐含层都采用模糊规则得到输出值,从而建立了用于预测抗菌肽的自适应神经-模糊推理系统ANFIS,取得了96.7%

的预测准确率.

表 6 统计了上述各文献中所采用的机器学习方法在预测不同抗菌肽时的效果，包括预测的抗菌肽类型、数据数量以及预测准确率，以方便读者查阅.

表 6　各机器学习算法预测效果统计

预测算法	文献	抗菌肽类型	样本数量	准确率/%
ANN	[72]	综合抗菌肽	1074	90.00
ANN	[73]	细胞穿透肽	174	82.98
ANN	[74]	抗真菌肽	58	88.97
ANN	[13]	综合抗菌肽	1433	94.00
SVM	[76]	综合抗菌肽	1200	88.33
SVM	[15]	综合抗菌肽	872	92.11
SVM	[79]	综合抗菌肽	600	90.00
SVM	[80]	抗癌肽	4276	96.00
SVM	[81]	抗细菌肽	9946	95.51
SVM	[82]	抗 HIV-1 肽	1051	96.76
RF	[85]	抗病毒肽	1660	90.00
RF	[86]	防卫素	238	78.12
RF	[25]	综合抗菌肽	6859	93.20
遗传算法	[87]	抗细菌肽	3175	95.00
高斯核回归	[88]	大肠杆菌抑制肽	115	86.59
集成算法	[89]	抗病毒肽	654	93.26
有限状态机	[90]	综合抗菌肽	2086	93.10
分段算法	[44]	综合抗菌肽	9731	80.23
分段算法	[46]	综合抗菌肽	9731	87.59
ANFIS	[63]	综合抗菌肽	231	96.70

4.2　基于多分类/多标签学习的抗菌肽预测

4.2.1　多分类学习方法

抗菌肽的鉴别通常只是将多肽分为抗菌肽和非抗菌肽两类，而最新的一些研究则开始着眼于多分类的情况，也就是抗菌肽样本的属性不再只是简单的两个类别，而是扩展到了多个类别，预测时要将待测样本归类到多类中的一类，例如有些预测任务需要把抗菌肽根据其抗菌功能或来源的不同进一步划分，就属于多分类的情况. 此时抗菌肽样本的类别 $\boldsymbol{Y}$ 的取值就有至少 3 种不同的情况，但这种情况下每个样本标签 $\boldsymbol{Y}$ 的取值只能是其中一种，不同类之间是互斥的，也就是一个样本只能属于一类，不能同时属于多类.

相关研究始于 2010 年，Lata 等人对先前的二分类预测系统 AntiBP 进行了改进，利用支持向量机算法和多肽 N 端与 C 端前 15 个氨基酸残基序列特征，提出了第一个具有多分类能力的新预测器 AntiBP2[95]. 由于单个支持向量机只能应对二分类情况，该方法使用“一类对其余”的策略使得算法能够处理多分类的情况，不仅取得了 92.14% 的总体预测准确率，还能够将抗菌肽按来源高精度地分为细菌来源、蛙类来源、昆虫来源、哺乳动物来源和植物来源 5 类，另外该方法还可以将每类来源的抗菌肽按家族再次分类；2012 年，Joseph 等人[96]基于多肽的序列特征用随机森林算法和支持向量机，也采用“一类对其余”的策略建立了多分类预测工具 ClassAMP，该方法能够区分抗细菌肽、抗真菌肽和抗病毒肽三类不同功能的抗菌肽，并分别取得了 97%、57% 和 87% 的预测准确率.

4.2.2　多标签学习方法

随着抗菌肽研究的不断深入和多效抗菌肽的出现，预测任务的要求也变得越来越高，传统的预测方法已经难以适应新的预测问题. 随着抗菌肽的类型不断丰富，越来越多的抗菌肽被发现具有多效抗菌活性，比如在文献[97]建立的包含 878 条抗菌肽的数据集中，具有 2 种以上不同抗菌活性的多效抗菌肽就有 424 条. 与普通抗菌肽相比，这些特殊的抗菌肽作用机理更为复杂，具有极大的研究价值，因此对多效抗菌肽抗菌活性的预测成为目前一个重要的研究内容. 然而常规的二分类或多分类预测模型却无法用于处理多效抗菌肽，这是由于多效抗菌肽的样本序列和抗菌活性是“一对多”的关系，这在数学上并不是一种映射关系，这是基于“一对一”映射关系的传统预测方法所无法解决的，因此需要采用专门针对这类问题的多标签学习方法来处理.

多标签学习是一种新型的机器学习方法，可以处理具有“一对多”分类属性的数据，这种方法在进行预测时将抗菌肽的每种抗菌活性视为一类标签，允许一个样本同时具有多个标签. 例如，若训练集中的样本共具有 m 种抗菌类别，则对于一个抗菌肽样本 $(\boldsymbol{X},\boldsymbol{Y})$，它可能同时具有 m 类抗菌性中的一种或多种，则其类别标签 $\boldsymbol{Y}$ 可以定义为一个多值向量 $\boldsymbol{Y}=\{y_1,y_2,\cdots,y_m\}$，其中每一个 y_i 对应着第 i 种抗菌活性 $(1\leq i\leq m)$，若该样本具有该抗菌活性，则可以令 $y_i=1$，若不具有该活性，则令 $y_i=-1$.

需要说明的是，与传统的单标签学习（二分类和多分类）相比多标签学习是一个更一般化的方法，单标签样本可以看做是多标签样本的一个特例，所以对多效抗菌肽的预测其实就包含了对单效抗菌肽在内全部类型抗菌肽的预测，是一种更广义的提法. 因此，对多效抗菌肽抗菌活性的高精度预测，其本质就是要基于多标签学习方法建立一个能精确描述抗菌肽的多效抗菌活性，并能有效反映标签之间关联关系的预测模型.

国内外关于多效抗菌肽的预测研究近年来才刚刚开展，现有的成果相对较少. 2013 年，景德镇陶瓷大学的肖绚教授和美国 Gordon 生命科学研究院的

Chou Kuo-Cheng 教授等人合作，首先对这一问题进行了研究，他们利用多肽的伪氨基酸组分特征和模糊 K 近邻算法(Fuzzy K Nearest Neighbor，FKNN)设计了一个两级分类器 iAMP-2L[97]，可以对多效抗菌肽进行预测，该方法先利用二分类算法判断待测多肽样本是否为抗菌肽，然后再将判断为抗菌肽的样本转入第二级预测，利用多标签分类器预测抗菌肽具有哪些抗菌活性，iAMP-2L 对于抗菌肽的鉴别准确率为 86.32%，对于 5 类抗菌活性的预测取得了最高为 60.79%的成功率；而后，该团队在上述的两阶段方法的基础上提出了改进方法[98]，一方面他们改用了抗菌肽序列的理化性质矩阵(Physical-Chemical Property Matrix，PCM)作为特征编码，另一方面在第二阶段时先利用 K 近邻方法判断待检测的抗菌肽的抗菌活性类别数，再利用 FKNN 方法对具体的抗菌活性各类别进行标记，从而将预测精度提高到 65%以上. 2016 年，景德镇陶瓷大学的邹洪亮则融合了氨基酸组份、二肽组份及多肽组份等多种序列特征，采用 LIFT 多标签学习算法[99]，取得了 69.95%的预测准确率.

5 预测方法的性能检验

为了检验预测方法的效果，需要对其性能进行客观地检验，采用合理有效的评估方法和评价指标就显得十分重要. 性能检验结果既评估了一个方法预测性能好坏的标准，又是与其他方法进行比较的依据. 由于基于经验分析的方法主要是利用验证的方式识别抗菌肽，一般没有通用的量化指标对这类方法的性能进行统一地评估和比较，因此下面主要介绍基于机器学习方法的性能检验方法.

性能检验首先需要可行的实验评估方法，即如何在已有的数据集上选择训练集和测试集，以完成预测算法的训练和测试. 训练集和测试集中的样本从多肽样本真实分布中独立同分布采样而得，既包括抗菌肽也包括非抗菌肽，且具有较低的序列相似性以避免同源性偏差. 用于抗菌肽预测的检验方法主要包括机器学习中被人熟知的留一法、交叉验证法和独立测试集验证法. 留一法作为交叉验证法的一个特例，不受随机样本划分方式的影响，是非常客观的检验方法，但是由于计算量较大，比较适合数据集规模较小的情况，因此在数据集规模较大的情况下，通常都采用交叉验证法来进行验证. 而独立测试集验证通常作为前两种方法的补充，能够较为有效地考察方法的泛化性能.

基于相应的实验评估方法，预测算法能够得到对测试集样本的预测结果，结合测试集样本的真实情况，能够得到下面六个统计参数：测试样本的真实标签集 Y、预测算法推断得到的测试样本标签集 Z、测试集中被预测正确的抗菌肽个数 TP、被预测正确的非抗菌肽个数 TN、被预测错误的抗菌肽个数 FP 以及被预测错误的非抗菌肽个数 FN.

实验评估方法确定的是如何对预测算法进行训练和测试，为了根据测试出的结果从不同的角度对预测算法进行评判还需要不同的评价指标. 评价指标反映了预测任务需求，在对比不同预测方法时，使用不同的评价指标往往会导致不同的评判结果，这意味着预测方法的好坏是相对的，哪种方法是好的不仅取决于数据和算法，还取决于任务需求. 对于抗菌肽的鉴别问题，首要考虑的是预测的精度，常见的评价指标如下：

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN},$$

$$Precision=\frac{TP}{TP+FP},$$

$$Recall=\frac{TP}{TP+FN},$$

$$MCC=\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

其中，$Accuracy$ 指标评价了全部样本中预测正确的比例，反映了算法对于抗菌肽和非抗菌肽整体的预测效果；$Precision$ 和 $Recall$ 指标则更关注对抗菌肽的识别效果，分别衡量的是真实抗菌肽样本中被预测正确的比例(强调预测的质量)和被预测为抗菌肽的样本中预测正确的比例(强调预测的数量)，预测算法在这两个指标上的值通常是此消彼长的关系；MCC 即 Matthews 相关系数，常用于考察数据不平衡下的预测效果，MCC 体现了测试集中样本的真实标记和预测标记之间的相关性，其值越高说明预测标记与真实标记越接近.

而对于多效抗菌肽的预测问题，由于要考虑多个抗菌活性标签的情况，在评价时除了样本全局指标，通常还要使用多标签评价指标. 对于测试集 $\boldsymbol{S}=\{(x_1,\boldsymbol{Y}_1),(x_2,\boldsymbol{Y}_2),\cdots,(x_p,\boldsymbol{Y}_p)\}$，其标签类别数目为 L，常见的多标签评价指标如下：

$$Hamming\ Loss:\ hloss_s(h)=\frac{1}{p}\sum_{i=1}^{p}\frac{1}{L}\left|h(x_i,\boldsymbol{Y}_i)\right|,$$

其中 $h(x_i,\boldsymbol{Y}_i)$ 表示该测试样本预测标签与实际标签不符的标签个数. 该指标评价预测标签集合与实际标签集合之间的匹配错误率，反映了算法能否较好的利用示例与标签间的关系，值越小表示算法性能越好.

$$Absolute\text{-}True=\frac{1}{p}\sum_{i=1}^{p}g(x_i,\boldsymbol{Y}_i),$$

其中，若算法对于测试样本 $(x_i,\boldsymbol{Y}_i)$ 在每个标签上的预测结果与真实标签完全一致时 $g(x_i,\boldsymbol{Y}_i)=1$，否则 $g(x_i,\boldsymbol{Y}_i)=0$. 该评价指标严格评估了预测的标签集与真实标签集的精确匹配度.

此外，由于多效抗菌肽样本具有多个标签，因此在计算多效抗菌肽的预测精度时，*Accuracy*、*Precision* 和 *Recall* 的计算方法相应变为

$$Accuracy_{ML}=\frac{1}{p}\sum_{i=1}^{p}\frac{|\boldsymbol{Y}_i\cap\boldsymbol{Z}_i|}{|\boldsymbol{Y}_i\cup\boldsymbol{Z}_i|},$$

$$Precision_{ML}=\frac{1}{p}\sum_{i=1}^{p}\frac{|\boldsymbol{Y}_i\cap\boldsymbol{Z}_i|}{|\boldsymbol{Z}_i|},$$

$$Recall_{ML}=\frac{1}{p}\sum_{i=1}^{p}\frac{|\boldsymbol{Y}_i\cap\boldsymbol{Z}_i|}{|\boldsymbol{Y}_i|}.$$

这里，$\boldsymbol{Z}_i$ 表示测试样本 $(x_i,\boldsymbol{Y}_i)$ 的预测标签.

在抗菌肽预测中，由于数据类型、数据规模、任务要求以及预测方法的不同，具体采用的评估方法和评价指标也有所不同，这里给出的只是实际研究中经常用到的一些性能检验手段. 一般来说，抗菌肽鉴别问题通常比较关注鉴别的准确率，而抗菌肽功能预测问题则由于存在多效抗菌肽的情况，还需要考察算法的多标签学习特性.

6　问题与展望

总体来讲，基于计算方法的抗菌肽预测研究还是处于起步阶段，各类方法的研究深度还有待提高. 基于经验分析的方法，比较依赖于现有资料的累积，对新知识和内在规律的挖掘力不足，而基于机器学习的方法，大都是直接套用现成的算法，并非是对抗菌肽预测的特点有针对性地加以分析和解决. 因此，基于计算方法对抗菌肽进行预测仍有诸多问题亟待解决.

(1) 从受“污染”数据中学习

基于计算方法的抗菌肽预测离不开数据，抗菌肽数据库中所收录的数据往往会受到“污染”，不可避免地存在着一些错误，主要的原因包括实验误差、错误生物解释、人工注释误差以及某些确实存在的性质尚未被发现或未经实验证实而造成现有数据信息残缺等. 如果训练数据错误较多或者某些重要数据的信息有误，这对于学习算法尤其是随机森林这类对于数据较为敏感的机器学习方法的性能影响很大. 然而大多数生物信息学研究人员都很少考虑数据的来源和质量对预测效果的影响，况且这些研究者也确实很难从生物数据本身的层面来解决这类问题，目前只能通过算法来克服. 如何在噪声数据和不完整数据中学习到真实有效的信息，发展具有较强鲁棒性和广泛适应性的学习方法是未来需要深入研究的一个重要问题.

(2) 预测效果与方法解释的问题

在现有的抗菌肽预测问题中，基于经验分析的方法通常能够根据对现有抗菌肽资料的解释和分析，并找到抗菌肽具有的一些生物学特点和某些规律，从而建立一些有意义的模型从而对新样本进行检验，但这类方法的预测效果和预测范围较为有限，预测效率也相对偏低. 而基于机器学习的方法，主要着眼点在于对预测结果的提升，其预测效果比基于经验分析的方法要好的多，但主要的一些方法如神经网络等是黑箱方法，往往只是从学习过程中得到结果，而无法对过程进行有效的解释，比较难于理解，这也是基于机器学习的预测方法需要考虑的一个问题. 在机器学习中引入统计学方法目前来看应该是个比较好的途径，例如高斯过程学习算法作为贝叶斯框架下的一种概率方法，就值得抗菌肽预测方法研究者加以关注.

(3) 不平衡数据的学习

根据数据库的统计信息来看，抗菌肽无论是根据抗菌属性还是样本来源，其分布都是很不平衡的. 例如，ADP 数据库[23] 收录的 14 类抗菌肽中，数量最多的抗细菌肽有 2263 条，而数量最少的抗原生生物肽只有 4 条，两者相差 550 多倍，这就造成了预测时的数据不平衡问题. 对于这种情况，如果按照常规方法来学习，由于少数类样本包含的信息十分有限，从而难以确定少数类数据的分布，造成少数类的识别困难. 而且，算法为了保证高准确率，在分类时还会偏向于预测结果为比例更大的样本，极端情况下甚至直接将全部待测样本都预测为多数类，并仍能取得很高的正确率. 但是在实际中，少数类抗菌肽因为其特殊性往往更加需要被精确的识别和预测，然而目前的研究成果都没考虑到这一问题.

(4) 多效抗菌肽的预测

从目前的研究发展来看，基于计算方法的抗菌

肽预测研究应该向通用性、实用性方向发展,因此未来算法应该具有能处理多效抗菌肽的能力,基于多标签学习的方法潜力较大.但是目前该研究仍处于起步阶段,见诸于报道的成果较少,现有算法只对这一问题进行了初步的研究,预测效果十分有限.与其他领域的多标签学习问题相比,多效抗菌肽的预测问题是比较特殊的.多标签学习的初衷是为了解决歧义性问题,样本通常都是具有一个或多个正标签的正类样本,一般都不会是负类样本,例如在蛋白质亚细胞定位问题中,每种亚细胞位置都对应着一种正类标签,蛋白质在合成后都要被运输到某一个或多个亚细胞位置才能行使功能,不存在不被定位的蛋白质.而抗菌肽的预测问题则完全不同,一条多肽完全可以是不具有任何抗菌活性的非抗菌肽,其本质是个具有负类样本的预测问题,这是传统的多标签学习所无法直接处理的,因此如何将非抗菌肽合理地纳入到预测模型中是采用多标签学习对多效抗菌肽进行预测的一个关键问题.

文献[97]和[98]采用的应对策略是采用两级预测方式直接避开这一矛盾,首先先将抗菌肽初步挑选出来之后,再专门对这些抗菌肽正样本进行常规的多标签学习处理.但是两级预测需要两次学习和分类,一方面降低了预测效率,另一方面会将两次预测的误差叠加导致预测精度下降.而另一个较为容易实现的策略是将负类样本看成一类具有特殊标签的正样本来处理,但由于非抗菌肽不具有抗菌活性,具有这类特殊正标签的样本就不能同时含有其他类别的正标签,人为地割裂了标签间的关联,无法深层次地挖掘潜在的规律,增大了学习难度导致预测模型精度一般.而抗菌肽各个抗菌活性之间的关联关系信息,不仅对于提高多效抗菌肽预测精度大有帮助,而且对于抗菌肽作用机理研究也是十分重要的内容,因此如何有效挖掘抗菌肽各类抗菌活性的关联性很值得进行深入的研究.

(5)多肽样本的价值评估方法

在预测抗菌肽的过程中,如何利用计算方法还对多肽样本的价值进行评估,来进一步帮助提高抗菌肽的预测效果,也是一个值得关注的问题.对于初始样本数据的效用进行评估,可以帮助构建有效的训练样本集有助于提高预测精度;而对于已知抗菌肽样本的评估,可以帮助衡量和比较各样本所具有抗菌活性的强弱以及其生物特征对抗菌活性影响的程度;对于未知的多肽样本的评估,则可以指导生物实验方法优先处理那些最可能具有抗菌活性的样本以降低标注成本.因此,发展对多肽样本的计算评估方法,作为探索抗菌肽作用机理的另一种有效手段,对于提高预测精度和帮助衡量样本价值等方面极具意义,也将是抗菌肽预测和分析方法中的一项重要研究内容.

(6)大数据水平下的预测方法

计算方法最大的优势在于预测的效率高、速度快,随着蛋白质组学的迅速发展,新的多肽样本数量也将急剧膨胀,数据的类型更多、维度更大、非结构化的程度也会更高,人们对于高通量大规模的预测要求也会越来越迫切.如何从海量的多肽样本里发现抗菌肽,如何在大数据层面设计算法来满足对抗菌肽信息的挖掘,也必将会成为将来的一个研究热点.

7 结 论

发展基于计算的抗菌肽预测方法不仅能够大幅降低预测成本、提高预测效率和规模,其对于帮助明确抗菌肽的作用机制、指导抗菌肽药物的人工改造和设计也具有极大的价值,对于早日解决我国严重的抗生素滥用问题具有十分重要的意义.可以预见,基于计算的预测方法对于了解抗菌肽的生物机理、探索和发现抗菌肽及其相关规律等方面会发挥越来越重要的作用.本文对目前国内外主要的抗菌肽预测方法进行了总结和阐述,讨论了这些方法的特点和差异性,并对该领域的研究发展方向进一步地分析和展望,希望能为相关研究人员提供一定的借鉴.

致 谢 本文得到国家自然科学基金委员会、中国博士后基金委员会等机构的支持,在此深表感谢!

参 考 文 献

[1] Koczulla A R, Bals R. Antimicrobial peptides: Current status and therapeutic potential. Drugs, 2003, 63(4): 389-406

[2] Brogden K A. Antimicrobial peptides: Pore formers or metabolic inhibitors in bacteria. Nature Reviews Microbiology, 2005, 3(3): 238-250

[3] Natsuga K, Cipolat S, Watt F M. Increased bacterial load and expression of antimicrobial peptides in skin of barrier-deficient mice with reduced cancer susceptibility. Journal of Investigative Dermatology, 2016, 136(1): 99-106

[4] Matanic V, Castilla V. Antiviral activity of antimicrobial cationic peptides against Junin virus and herpes simplex virus. International Journal of Antimicrobial Agents, 2004, 23(4): 382-389

[5] Login F H, Balmand S, et al. Antimicrobial peptides keep insect endosymbionts under control. Science, 2011, 334(9054): 362-365

[6] Yeaman MR, Yount NY. Mechanisms of antimicrobial peptide action and resistance. Pharmacological reviews, 2003, 55(1): 27-55

[7] Carmona-Ribeiro A M, de Melo-Carasco L D. Novel formulations for antimicrobial peptides. International Journal of Molecular Sciences, 2014, 15(10): 18040-18083

[8] Mangoni M L, Bhunia A. Antimicrobial peptides in medicinal chemistry: Advances and applications. Current Topics in Medicinal Chemistry, 2016, 16(1): 2-3

[9] Torrent M, Nogues M V, Boix E. Discovering new in silico tools for antimicrobial peptide prediction. Current Drug Targets, 2012, 13(9): 1148-1157

[10] Hein K Z, Takahashi H, et al. Disulphide-reduced psoriasin is a human apoptosis-inducing broad-spectrum fungicide. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(42): 13039-13044

[11] Schroeder B O, Wu Z, et al. Reduction of disulphide bonds unmasks potent antimicrobial activity of human β-defensin 1. Nature, 2011, 469(7330): 419-423

[12] Gautam A, Sharma A, et al. Development of antimicrobial peptide prediction tool for aquaculture industries. Probiotics Antimicrobial Proteins, 2016, 8(3): 141-149

[13] Fjell C D, Jenssen H, et al. Identification of novel antibacterial peptides by chemoinformatics and machine learning. Journal of Medicinal Chemistry, 2009, 52(7): 2006-2015

[14] Zhou Xiao-Fu, Miao Lu, et al. Bioinformatics forecast and analysis of plant antimicrobial peptides. Biotechnology, 2014, 24(3): 91-95(in Chinese)
(周晓馥，苗璐等. 利用生物信息学对植物抗菌肽的预测与分析. 生物技术，2014，24(3)：91-95)

[15] Lata S, Sharma B K, Raghava G P S. Analysis and prediction of antibacterial peptides. BMC Bioinformatics, 2007, 8(1): 263

[16] Loose C, Jensen K, Rigoutso I, Stephanopoulos G. A linguistic model for the rational design of antimicrobial peptides. Nature, 2006, 443(7113): 867-869

[17] Kindrachuk J, Napper S. Structure-activity relationships of multifunctional host defence peptides. Mini-Reviews in Medicinal Chemistry, 2010, 10(7): 596-614

[18] Fjell C D, Hiss J A, Hancock R E W, Schneider G. Designing antimicrobial peptides: Form follows function. Nature Reviews Drug Discovery, 2012, 11(1): 37-51

[19] Apweiler R, Bairoch A, et al. UniProt: The universal protein knowledgebase. Nucleic Acids Research, 2004, 32(D1): D115-D119

[20] Berman H M, Westbrook J, et al. The protein data bank. Nucleic Acids Research, 2000, 28(1): 235-242

[21] Wang Z, Wang G S. APD: The antimicrobial peptide database. Nucleic Acids Research, 2004, 32(D1): D590-D592

[22] Wang G S, Li X, Wang Z. APD2: The updated antimicrobial peptide database and its application in peptide design. Nucleic Acids Research, 2009, 37(D1): D933-D937

[23] Wang G S, Li X, Wang Z. APD3: The antimicrobial peptide database as a tool for research and education. Nucleic Acids Research, 2015, 44(D1): D1807-D1093

[24] Gogoladze G, Grigolava M, et al. DBAASP: Database of antimicrobial activity and structure of peptides. FEMS Microbiology Letters, 2014, 357(1): 63-68

[25] Thomas S, Karnik S, Barai R S, et al. CAMP: A useful resource for research on antimicrobial peptides. Nucleic Acids Research, 2010, 38(D1): D774-D780

[26] Waghu F H, Gopi L, Barai R S, et al. CAMP: Collection of sequences and structures of antimicrobial peptides. Nucleic Acids Research, 2014, 42(D1): D1154-D1158

[27] Waghu F H, Barai R S, Gurung P, Idicula-Thomas S. $CAMP_{R3}$: A database on sequences, structures and signatures of antimicrobial peptides. Nucleic Acids Research, 2015, 44(D1): D1094-D1097

[28] Qureshi A, Thakur N, Tandon H, Kumar M. AVPdb: A database of experimentally validated antiviral peptides targeting medically important viruses. Nucleic Acids Research, 2015, 43(D1): D837-D843

[29] Tyagi A, Tuknait A, et al. CancerPPD: A database of anticancer peptides and proteins. Nucleic Acids Research, 2009, 37: D963-D968

[30] Mehta D, Anand P, et al. ParaPep: A web resource for experimentally validated antiparasitic peptide sequences and their structures. Database, 2014, 2014: bau051

[31] Seebah S, Suresh A, et al. Defensins knowledgebase: A manually curated database and information source focused on the defensins family of antimicrobial peptides. Nucleic Acids Research, 2007, 35(D1): D265-D268

[32] Qureshi A, Thakur N, Kumar M. HIPdb: A database of experimentally validated HIV inhibiting peptides. PLoS One, 2013, 8(1): e54908

[33] Novkovic M, Simunic J. DADP: The database of anuran defense peptides. Bioinformatics, 2012, 28(10): 1406-1407

[34] Zhao X W, Wu H Y, Lu H R, et al. LAMP: A database linking antimicrobial peptides. PLoS One, 2013, 8(6): e66557

[35] Piotto S P, Sessa L, Concilio S, Iannelli P. YADAMP: Yet another database of antimicrobial peptides. International Journal of Antimicrobial Agents, 2012, 39(4): 346-351

[36] Sundararajan V S, Gabere M N, et al. DAMPD: A manually curated antimicrobial peptide database. Nucleic Acids Research, 2012, 40(D1): D1108-D1112

[37] Lee H T, Lee C C, et al. A large-scale structural classification of antimicrobial peptides. Biomedical Research International, 2015, 2015(1): 475062

[38] Person W R, Lipman D J. Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences, 1988, 85(8): 2444-2448

[39] Altschul S F, Madden T L, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research, 1997, 25(17): 3389-3402

[40] Larkin M A, Blackshields G, et al. Clustal W and Clustal X version 2.0. Bioinformatics, 2007, 23(21): 2947-2948

[41] Smith T F, Waterman M S. Identification of common molecular subsequences. Journal of Molecular Biology, 1981, 147(1): 195-197

[42] Finn R D, Clements J, Eddy S R. HMMER web server: Interactive sequence similarity searching. Nucleic Acids Research, 2011, 39(S2): W29-W37

[43] Wang P, Hu L L, et al. Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. PLoS One, 2011, 6(4): e18476

[44] Altschul S F. Evaluating the statistical significance of multiple distinct local alignments//Suhai S, ed. Theoretical and Computational Methods in Genome Research. New York, USA: Plenum, 1997: 1-14

[45] Ng X Y, Rosdi B A, Shahrudin S. Prediction of antimicrobial peptides based on sequence alignment and support vector machine-pairwise algorithm utilizing LZ-Complexity. BioMed Research International, 2015, 2015(1): 212-715

[46] Xiao Y J, Cai Y B, et al. Identification and functional characterization of three chicken cathelicidins with potent antimicrobial activity. The Journal of Biological Chemistry, 2006, 281(5): 2858-2867

[47] Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 1987, 4(6): 406-425

[48] Kubinyi H, Folkers G, Martin Y C, et al. 3D QSAR in Drug Design. Dordrecht, Holland: Kluwer Academic Publisher, 2002

[49] Zhou Xi-Bin, et al. Research progress and application of some QSAR modeling approach in chemistry. Computers and Applied Chemistry, 2011, 28(6): 761-764(in Chinese)
(周喜斌等. 几种 QSAR 建模方法在化学中的应用与研究进展. 计算机与应用化学, 2011, 28(6): 761-764)

[50] Strøm M B, Stensen W, Svendsen J S, Rekdal Ø. Increased antibacterial activity of 15-residue murine lactoferricin derivatives. The Journal of Peptide Research, 2001, 57(2): 127-139

[51] Frecer V. QSAR analysis of antimicrobial and haemolytic effects of cyclic cationic antimicrobial peptides derived from protegrin-1. Bioorganic & Medicinal Chemistry, 2006, 14(17): 6065-6074

[52] Hilpert K, Elliott M R, et al. Sequence requirements and an optimization strategy for short antimicrobial peptides. Chemistry and Biology, 2006, 13(10): 1101-1107

[53] Shu M, Yu R, et al. Predicting the activity of antimicrobial peptides with amino acid topological information. Medicinal Chemistry, 2013, 9(1): 32-44

[54] Cherkasov A, Jankovic B. Application of 'Inductive' QSAR descriptors for quantification of antibacterial activity of cationic polypeptides. Molecules, 2004, 9(12): 1034-1052

[55] Cherkasov A. Inductive QSAR descriptors. Distinguishing compounds with antibacterial activity by artificial neural networks. International Journal of Molecular Sciences, 2005, 6(1): 63-86

[56] Mikut R, Hilpert, K. Interpretable features for the activity prediction of short antimicrobial peptides using fuzzy logic. International Journal of Peptide Research and Therapeutics, 2009, 15(2): 129-137

[57] Fernandes F C, Porto W F, Franco O L. A wide antimicrobial peptides search method using fuzzy modeling. Lecture Notes in Computer Science, 2009, 5676(1): 147-150

[58] Fernandes F C, Rigden D J, Franco O L. Prediction of antimicrobial peptides based on the adaptive neuro-fuzzy inference system application. Peptide Science, 2012, 98(4): 280-287

[59] Rigoutsos I, Floratos A. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics, 1998, 14(1): 55-67

[60] Li Q Z, Lu Z Q. The prediction of the structural class of protein: Application of the measure of diversity. Journal of Theoretical Biology, 2001, 213(3): 493-502

[61] Veltri D, Kamath U, Shehu A. A novel method to improve recognition of antimicrobial peptides through distal sequence-based features//Proceedings of the 2014 IEEE International Conference on Bioinformatics and Biomedicine. Belfast, UK, 2014: 371-378

[62] Zuo Y C, Li Q Z. Using reduced amino acid composition to predict defensin family and subfamily: Integrating similarity measure and structural alphabet. Peptides, 2009, 30(10): 1788-1793

[63] de Brevern A G. New assessment of a structural alphabet. In Silico Biology, 2005, 5(3): 283-289

[64] Etchebest C, Benros C, et al. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. European Biophysics Journal, 2007, 36(8): 1059-1069

[65] Brooks B R, Iii C L B, et al. CHARMM: The biomolecular simulation program. Journal of Computational Chemistry, 2009, 30(10): 1545-1614

[66] Hess B, Kutzner C, et al. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. Journal of Chemical Theory and Computation, 2008, 4(3): 435-447

[67] Mátyus E, Kandt C, Tieleman D P. Computer simulation of antimicrobial peptides. Current Medicinal Chemistry, 2007, 14(26): 2789-2798

[68] Nagarajan V, Kaushik N, et al. A fourier transformation based method to mine peptide space for antimicrobial activity. BMC Bioinformatics, 2006, 7(Suppl 2): S2

[69] Yount N Y, Yeaman M R. Multidimensional signatures in antimicrobial peptides. PNAS, 2004, 101(1): 7363-7368

[70] Jenssen H, Lejon T, Hilpert K, et al. Evaluating different descriptors for model design of antimicrobial peptides with enhanced activity toward P. aeruginosa. Chemical Biology and Drug Design, 2007, 70(2): 134-142

[71] Yang Li, Wang Zhen-Ling, He Gu, Wei Yu-Quan. Prediction of antimicrobial peptide activity and antimicrobial peptides, China, 201410241068.9, 2014.5(in Chinese)
(杨莉，王震玲，何谷，魏于全. 抗菌肽抗菌活性预测方法及抗菌肽，中国，201410241068.9，2014.5)

[72] Torrent M, Andreu D, Nogues V M, Boix E. Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. PLoS One, 2011, 6(2): e16968

[73] Holton T A, Pollastri G, Shields D C, Mooney C. CPPpred: Prediction of cell penetrating peptides. Bioinformatics, 2013, 29(23): 3094-3096

[74] Soltani S, Keymanesh K, Sardari S. Evaluation of structural features of membrane acting antifungal peptides by artificial neural network. Journal of Biological Sciences, 2008, 8(5): 834-845

[75] Hsu C W, Lin C J. A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks, 2002, 13(2): 415-425

[76] Rondón-Villarreal P, Sierra D A, Torres R. Classification of antimicrobial peptides by using the p-spectrum kernel and support vector machines. Advances in Intelligent Systems and Computing, 2014, 232(1): 155-160

[77] Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. Cambridge, UK: Cambridge University Press, 2004

[78] Porto W F, Fernandes F C, Franco O L. An SVM model based on physicochemical properties to predict antimicrobial activity from protein sequences with cysteine knot motifs. Lecture Notes in Computer Science, 2010, 6268(1): 59-62

[79] Porto W F, Pires A S, Franco L. CS-AMPPred: An updated SVM model for antimicrobial activity prediction in cysteine-stabilized peptides. PLoS One, 2012, 7(12): e51444

[80] Vijayakumar S, Lakshmi P T V. ACPP: A web server for prediction and design of anti-cancer peptides. International Journal of Peptide Research and Therapeutic, 2015, 21(1): 99-106

[81] Khosravian M, Faramarzi F K, Beigi M M, et al. Predicting antibacterial peptides by the concept of chou's pseudo-amino acid composition and machine learning methods. Protein and Peptide Letters, 2013, 20(2): 180-186

[82] Poorinmohammad N, Mohabatkar H, et al. Computational prediction of anti HIV-1 peptides and in vitro evaluation of anti HIV-1 activity of HIV-1 P24-derived peptides. Journal of Peptide Science, 2015, 21(1): 10-16

[83] Breiman L. Random forests. Machine Learning, 2001, 45(1): 5-32

[84] Zhou Z H. Ensemble Methods: Foundations and Algorithms. Boca Raton, USA: CRC Press, 2012

[85] Chang K Y, Yang J R. Analysis and prediction of highly effective antiviral peptides based on random forests. PLoS One, 2013, 8(8): e70166

[86] Karnik S, Prasad A, et al. Identification of defensins employing recurrence quantification analysis and random forest classifiers. Lecture Notes in Computer Science, 2009, 5909(1): 152-157

[87] Veltri D, Kamath U, Shehu A. Improving recognition of antimicrobial peptides and target selectivity through machine learning and genetic programming. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017, 14(2): 300-313

[88] Xiao X, You Z B. Predicting minimum inhibitory concentration of antimicrobial peptides by the pseudo-amino acid composition and gaussian kernel regression//Proceedings of the 2015 8th International Conference on BioMedical Engineering and Informatics. Shenyang, China, 2015: 301-305

[89] Zare M, Mohabatkar H, et al. Using Chou's pseudo amino acid composition and machine learning method to predict the antiviral peptides. The Open Bioinformatics Journal, 2015, 9(1): 13-19

[90] Whelan C, Roark B, Sonmez K. Designing antimicrobial peptides with weighted finite-state transducers//Proceedings of the 2014 IEEE International Conference on Bioinformatics and Biomedicine. Buenos Aires, Argentina, 2010: 764-767

[91] Avram S, Mihailescu D, Borcan F, Milac A L. Prediction of improved antimicrobial mastoparan derivatives by 3D-QSAR-CoMSIA/CoMFA and computational mutagenesis. Monatshefte Fuer Chemie/Chemical Monthly, 2012, 143(4): 535-543

[92] Taboureau O, Olsen O H, et al. Design of novispirin antimicrobial peptides by quantitative structure-activity relationship. Chemical Biology and Drug Design, 2006, 68(1): 48-57

[93] Jaén-Oltra J, Salabert-Salvador M T, et al. Artificial neural network applied to prediction of fluorquinolone antibacterial activity by topological methods. Journal of Medicinal Chemistry, 2000, 43(6): 1143-1148

[94] Tomás-Vert F, Pérez-Giménez F, et al. Artificial neural network applied to the discrimination of antibacterial activity by topological methods. Journal of Molecular Structure: THEOCHEM, 2000, 504(1-3): 249-259

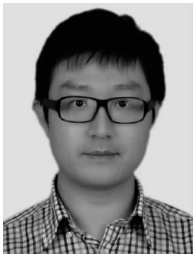
[95] Lata S, Mishra N K, Raghava G P S. AntiBP2: Improved version of antibacterial peptide prediction. BMC Bioinformatics, 2010, 11(S1): S19

[96] Joseph S, Karnik S, Nilawe P, Jayaraman V K, Idicula-Thomas S. ClassAMP: A prediction tool for classification of antimicrobial peptides. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2012, 9(5): 1535-1538

[97] Xiao X, Wang P, Lin W Z, et al. iAMP-2L a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Analytical Biochemistry, 2013, 436(2): 168-177

[98] Wang P, Xiao X. Multi-label classifier design for predicting the functional types of antimicrobial peptides. Advanced Materials Research, 2013, 718-720(2): 293-298

[99] Zou H L. A new multi-label classifier for identifying the functional types of singleplex and multiplex antimicrobial peptides. International Journal of Peptide Research and Therapeutics, 2016, 22(2): 281-287



CAO Jun-Zhe, born in 1984, Ph.D., lecturer. His research interests include machine learning, data mining and bioinformatics.

GU Hong, born in 1961, Ph.D., professor. His research interests include machine learning, bioinformatics and big data.

Background

Prediction of antimicrobial peptides based on computational methods belongs to bioinformatics and artificial intelligence. Being different from biological experiment approaches, computational prediction analyzes APMs data by using math and computer technology to make decisions automatically, therefore, it has many advantages such as higher efficiency, faster processing speed, lower cost and better performance under large scale data. As a crossing research of above two fields, this study arouses people's great interests in recent years.

In this work, we offer a detailed survey of prediction based on computational method for antimicrobial peptides. Antimicrobial peptides data and algorithm are two key parts for this kind of approach, thus we summarize popular antimicrobial peptides databases, as well as state-of-the-art predicting ways which mainly include methods based on empirical analysis and machine learning. In the end, this paper discusses the development of APM prediction based on computation, points the remaining problems and proposes the meaningful research directions in future.

This work is supported by the National Natural Science Foundation of China (61502074), Project Funded by the China Postdoctoral Science Foundation (2016M591430) and the Dalian University of Technology Fundamental Research Fund (DUT17RC(4)09).