编号：B143

PACPs：基于特征描述符和机器学习的抗癌肽预测模型

论文题目:PACPs:基于特征描述符和机器学习的抗癌肽预测模型

参赛学校:新疆大学

参赛成员(作者):李津津、廖艳红、王海云

指导老师:赵建平

目录

图索引

表索引

摘 要

目前，癌症是危害人类健康最重要的因素之一，已经成为了全球范围内数百万人口死亡的重要原因。传统的癌症的治疗方法包括靶向治疗、化疗和放射治疗等，在一定程度上能杀死病变癌细胞，但是同时也会杀死大量正常的细胞，带来严重的副作用。这些治疗手段费用昂贵且预后效果不佳，所以迫切需要开发新的定向清除癌细胞的新型抗癌药物。

近年来，利用现代生物技术合成的多肽药物因其适应性广、安全性高且疗效显著等特点，已成为癌症治疗药物研发的热点之一。目前，多肽药物已广泛应用于肺癌、乳腺癌、前列腺癌等疾病的治疗，具有广阔的发展前景。

在本研究中，本文建立了一种识别抗癌肽的预测模型——PACPs（Prediction of anticancer peptides）模型。该模型是由两层的堆叠框架构成。在第一层，本文将 21 种特征描述符和 13 种机器学习分类算法结合得到了 273 个基学习器。运用两步特征选择的方法，选择了 5 个基学习器：ET-CKSAAGP，ET-CTD，GDBT-CTD，GDBT-QSOrder，SVM-ASDC，构成了最优基学习器组合。在第二层，将 5 个基学习器输出的概率向量作为新的特征向量，来训练元学习器 LR，其输出作为抗癌肽识别最终的预测结果。基于堆叠方法，PACPs 将不同的特征表示符和机器学习分类算法进行了集成，从而整合了每个模型的优点，使 PACPs 模型成为了一个性能良好的预测模型。

PACPs 模型在训练集上的 AUC 值达到了 91.56%，在 2 个测试集上的 AUC 分别达到了 94.26%、90.18%，在独立测试集上达到了 80.71%。这些数据集的一致性能验证了 PACPs 模型的实用性。因此，本文认为 PACPs 模型在发现抗癌肽作为治疗癌症的新药方面具有推动作用。

【关键词】抗癌肽 机器学习 特征描述符 递归特征消除

Abstract

At present, cancer is one of the most important factors endangering human health, and has become an important cause of death for millions of people around the world. The traditional cancer treatment methods include targeted therapy, chemotherapy and radiotherapy, which can kill the diseased cancer cells to a certain extent, but also kill a large number of normal cells, resulting in serious side effects. Because of the high cost and poor prognosis of these treatments, it is urgent to develop new anticancer drugs for targeted removal of cancer cells.

In recent years, peptide drugs synthesized by modern biotechnology have become one of the hot spots in the research and development of cancer treatment drugs because of their wide adaptability, high safety and significant efficacy. At present, peptide drugs have been widely used in the treatment of lung cancer, breast cancer, prostate cancer and other diseases, and have broad development prospects.

In this study, we established a prediction model for recognition of anticancer peptides--PACPs (Prediction of anticancer peptides) model. The model is composed of two layers of stacked frame. In the first layer, we combine 21 feature descriptors with 13 machine learning classification algorithms to get 273 base learners. By using the two-step feature selection method, five base learners are selected: ET-CKSAAGP, ET-CTD, GDBT-CTD, GDBT-QSOrder and SVM-ASDC to form the optimal combination of base learners. In the second layer, the probability vectors output by the five base learners are used as new feature vectors to train the meta learner LR, whose output is used as the final prediction result of anti-cancer peptide recognition. Based on the stacking method, PACPs integrates different feature descriptors and machine

learning classification algorithm, which integrates the advantages of each model and makes PACPs model a good prediction model.

The AUC of PACPs model on training set is 91.56%, on two test sets is 94.26% and 90.18% respectively, and on independent test set is 80.71%. The consistency of these data sets verifies the practicability of PACPs model. Therefore, the PACPs model can promote the discovery of anticancer peptides as new drugs for cancer treatment.

KEYWORD: Anticancer peptide; Machine learning; feature descriptors; Recursive Feature Elimination

第 1 章　绪论

1.1 研究背景

癌症是一系列相关恶性肿瘤的统称, 通常是由正常细胞发生基因突变而引发的，其死亡率极高。据世界卫生组织国际癌症研究机构估计，2020 年全球新发癌症病例 1929 万例、死亡病例 996 万例，死亡率高达 50%以上，其中中国新发癌症 457 万人，占全球 23.7%，癌症死亡病例 300 万，占全球 30%。究其原因，中国人口基数大，癌症新发人数和死亡人数都比世界其他国家多。目前全球最常见的发病率和致死率高的癌症包括胃癌、肝癌、乳腺癌、肺癌等。随着 21 世纪全球癌症负担的不断增加，预防癌症成为公共卫生领域最具挑战性的问题之一。

手术、放疗、化疗和分子靶向药物是治疗癌症的主要手段，但这些传统方法不仅昂贵而且存在许多严重的副作用，如手术创伤性较大，复发率高，化疗可能会杀死正常细胞，对药物不敏感的部分肿瘤没有效果等。此外，这些抗癌类药物的联合使用也可能导致肿瘤的多药耐药性，加剧患者的免疫缺陷。

新的抗癌药物的开发对提高患者生存率至关重要。抗癌肽(ACPs)是抗菌肽(AMPs)中一类具有抗肿瘤活性的肽，可分为两类：一类是只消灭细菌和癌细胞，对正常细胞无影响，另一类是对细菌、癌细胞和正常细胞均有破坏作用。因此识别第一类抗癌肽具有重要的研究价值和广阔的应用前景。

随着大数据时代的到来，医疗和生物信息统计学越来越受到大众的关注，机器学习在医学研究领域应用越来越广泛，如病痛的诊断和疾病的鉴别、个性化的治疗和行为矫正、药物的发现和生产、临床和实验研究、放射学和放射治疗等。机器学习在这些领域的研究中不仅节约了大量的时间、人力和物料成本，而且提高了准确率和精度。基于上述背景，利用机器学习识别抗癌肽是非常必要的。

1.2 研究意义

随着生物信息学和生物技术的发展，临床上多肽类药物的应用越来越广泛，已应用于各类疾病，例如过敏性疾病、传染性疾病、自身免疫性疾病等。目前，在癌症患者治疗的早期诊断和中期治疗中，多肽类药物的应用也越来越广泛，在肺癌、乳腺癌、胃癌等高发癌症的诊断、治疗和预后预测中都具有很大潜力。

每一种活性肽都具有不同的结构，导致了其不同的作用机制。其中肿瘤抑制肽是与肿瘤相关的基因以及对肿瘤产生作用的调控因子特异相结合对多药抗性的肿瘤细胞系发挥抗癌作用的。它不仅可以直接抑制肿瘤的生长和转移、阻滞肿瘤细胞的周期、破坏细胞膜的结构促进肿瘤细胞死亡达到直接抗肿瘤作用，也可以通过激活机体的免疫系统、对功能蛋白进行干扰、抑制肿瘤血管形成等达到间接的抗肿瘤作用。多肽类药物的分子量小、毒性低、活性高、对人体的副作用较小，同时还具有靶向性、安全性、特异性等特点，使得肿瘤抑制肽在临床治疗上具有明显的优势，也使多肽类药物的研发成为癌症治疗的热点之一。

1.3 研究现状

肽的实验数据和计算数据呈指数增长，使得功能肽的识别成为一项费时、费力且依赖于多种因素的工作。机器学习提供了一系列工具和方法，如支持向量机、随机森林、极端随机树，使得研究人员可以利用高通量测序技术和机器学习算法开发更高精度预测功能肽的模型，以获得可靠的预测结果。

在过去的二十年中，基于 ML 的肽活性预测技术得到了极大的发展。已利用机器学习工具识别抗癌肽（ACPs）、抗高血压肽（AHTPs）、抗炎肽（AIPs）、抗疟肽、抗菌肽（AMPs）、抗寄生虫肽、抗病毒肽（AVPs）等。在目前的研究中，已经开发了几种基于 ML 的方法来从蛋白质序列中识别潜在的抗癌肽。

1.3.1 基于支持向量机的识别抗癌肽的模型

Hajisharifi 等[1](2013)提出了一个基于局部对齐核的 SVM 建立识别潜在的抗癌肽的模型，在他们构建的数据集（138 个 ACP 和 206 个非 ACP）上，该模型的精度和特异性分别为 89.7%和 92.68%。Vijayakumar 等[2](2015)提出了 ACPP 模型，使用 SVM 和蛋白质相关性度量（氨基酸组成信息、质心和分布度量的组合）作为输入特征，在他们构建的数据集（257 个 ACP 和 4019 个非 ACP）上，ACPP 的准确率达到 97.0%。Li、Wang 等[3](2016)使用新的混合特征（AAC、平均化学位移和 RAAC 的线性组合）来开发 SVM 模型，该方法在 iACP 数据集的训练集和独立测试集上的准确率分别为 93.6%和 89.3%。Vinothini Boopathi 等[23](2019)提出的 mACPpred 模型是一种用两步特征选择出七个特征编码（基于成分，物理化学性质和轮廓）并获得相应的最佳特征模型，结果的概率特征向量作为 SVM 的输入得到最终预测模型。在独立的 157 个 ACP 和 157 个非 ACP 测试集上 96.7%。四个文献都用 SVM 来训练抗癌肽的预测模型，得到的预测结果性能较好，说明 SVM 是一个用于抗癌肽预测的较好的模型。

1.3.2 基于多分类器集成的抗癌肽预测模型

Akbar 等[5](2017)设计了一个 iACP-GAEnsC 模型，将混合特（Am-PseAAC、伪 GGDPC 和 RAAC 的线性组合）输入到五个不同的分类器（SVM、RF、KNN、概率神经网络和广义回归神经网络）中，然后用遗传算法和简单多数投票策略将这五个分类器进行集成，建立了相应的模型。其训练数据集上的准确率为 96.5%。Nalini Schaduangrat 等[22](2020)开发的 ACPred 是一种用于预测和表征肽抗癌活性的模型。通过使用 SVM、RF 与各种类别的肽特征来开发，并用 Jackknife 交叉验证测试在 138ACPs 和 205 非 ACPs 的数据集[1]上实现 95.61%的总体准确率。Sajid Ahmed 等[24](2020)提出的 ACP-MHCNN 是一种多头深度卷积神经网络模

型，由并行卷积组组成，这些组联合学习并组合了来自三种不同肽表示方法（BPF、物理化学性质和进化特征）的特征以准确识别 ACP。该模型在独立测试集 ACP-164 上 AUC 达到了 93%。以上模型都使用多个分类器对抗癌肽进行预测，相比一种分类器预测的性能有所提高。

1.3.3 基于多特征描述符与多分类器的其他肽预测模型

Leyi Wei 等[6](2018)提出 QSPred-FL 模型，对 99 个特征描述符（包括 9 个不同参数的特征编码算法）进行了全面比较，使用五种不同的机器学习算法进行群体感应肽（QSPs）的识别，用一个特征表示学习和特征选择方案来提取最具鉴别性的特征，并利用学习到的信息特征进行预测。该模型在 QSP400 数据集（200 个正样本和 200 个的负样本）上的准确性达到了 94.3%。Yannan Bin、Wei Zhang 等[7](2020)开发了一种识别神经肽（NPs）的模型 PredNeuroP。基于两层堆叠框架构建，在第一层，基于基学习器的 ACCS 和 PCC，筛选得到了 8 个最优的基学习器，在第二层上，将 8 个最优基学习器的输出用来训练第二层学习器 LR，其输出作为对 NPs 识别的预测结果。该模型在训练和测试数据集中的准确性分别为 89.3%和 87.2%。这两个模型使用多个特征描述符保留了更多的肽序列的信息，使用多个分类器集成了他们的优点，构建的模型性能有所提高。

本文基于以上研究现状，拟从建模、算法改进、实验分析、统计分析的角度进一步研究抗癌肽预测的问题，从而助力多肽药物的研发，改善癌症患者的治疗效果。

第2章　模型构建

本文提出了一个基于特征描述符和机器学习算法的预测抗癌肽的模型PACPs。其总体框架如图 1 所示。可以看出该模型的开发包括四个主要步骤：

第一步，数据收集和准备。建立基准数据集，划分训练集与测试集。第二步，特征提取。用特征描述符提取肽序列的特征，并结合分类器构建基分类器，把得到的概率向量作为新的特征。第三步，特征选择。通过平均 F1 得分和递归特征消除（RFE）的两步特征选择来选择最优的基学习器子集。第四步，模型构建。选择元学习器构建模型。

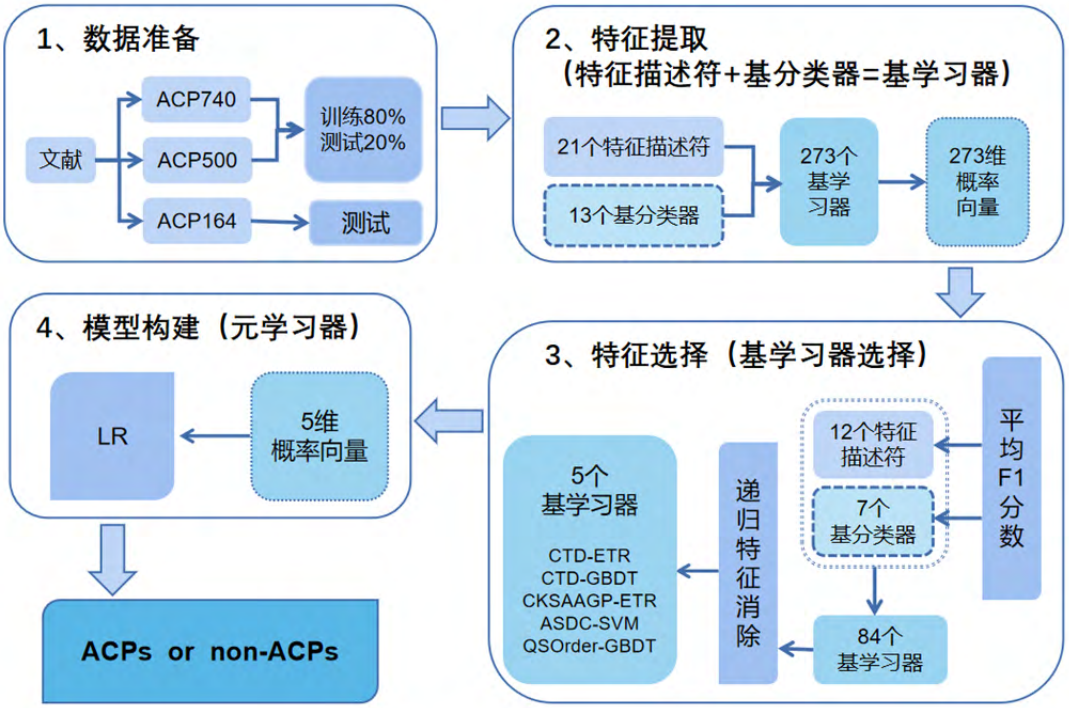

图 1　模型框架图

堆叠（Stacking）框架是一种代表性的整体学习方法，使用各种机器学习算法或基于元学习器（第二层学习器）的特征编码集成不同的基学习器（第一层学习器），以汇集每个模型的优势并构建表现良好的预测器。在本研究中，结合 21 种特征编码（即 AAC、CKSAAP、DPC、TPC、DDE、CTD、CTriad、KSCTriad、ASDC、DistancePair、GAAC、CKSAAGP、GDPC、GTPC、NMBroto、AC、

SOCNumber、QSOrder、PAAC、APAAC、PseKRAAC1）和 13 种机器学习算法（即 RF、GBDT、XGBoost、ET、LightGBM、SVM、LR、KNN、Decision Tree、NBC、SGD、LDA、QDA），在训练集中训练得到 273 个基学习器。在第一层，考虑到堆叠方法中基学习器的多样性和性能的重要性，基于基学习器的平均 F1 得分和递归特征消除方法来选择最优基学习器的组合，最终选择了 5 个基学习器（ET-CKSAAGP，ET-CTD，GDBT-CTD，GDBT-QSOrder，SVM-ASDC）并输出得到 5 维概率向量作为第二层输入进行训练，比较各分类器的性能，最终选择逻辑回归作为元学习器，其最终输出就是对抗癌肽的预测结果。

2.1 数据收集

2.1.1 多肽序列

肽由两个氨基酸脱水缩合而成，所形成的酰胺基为肽键，两个或以上肽键组成一个肽链，多个肽链进行多级折叠组成一个蛋白质分子，故蛋白质也被称为多肽。蛋白质作为人类生命活动的主要承担者，是由 20 种不同的氨基酸(A，C，D，E，F，G，H，I，K，L，M，N，P，Q，R，S，T，V，W，Y)经过一系列复杂的步骤所构成的大分子化合物。多肽的理化性质包括多肽的极性、疏水性、极化率、分配系数等。对于任意一个长度为 L 的多肽序列都可以用以下模型来表示：$P = S_1 S_2 S_3 S_4 S_5 \ldots\ldots S_L$，其中$S_1$代表组成多肽序列的第一个氨基酸残基，$S_2$代表第二个氨基酸残基，以此类推，$S_L$是第 L 个氨基酸残基。

```
>54|0
KWKSFLKTFKSAKKTVLHTALKAISS
>55|0
MNFQQRLQSLWTLARPFCPPLLATASQMQMVVLPCLGFTLLLWSQVSGAQGQEFHFGPCQVKGVVPQKLWEAF
WAVKDTMQAQDNITSARLLQQEVLQNVSDAESCYLVHTLLEFYLKTVFKNYHNRTVEVRTLKSFSTLANNFVLIVSQ
LQPSQENEMFSIRGDSAHRRFLLFRRAFKQLDVEAALTKALGEVDILLTWMQKFYKL
```

图 2　多肽序列图

图 2 即为多肽序列的 fasta 格式，第一行以“ >54|0 ”作为多肽序列的名称，第二行为多肽序列的内容，由大写的 20 种不同英文字母表示代表的 20 种不同氨

基酸（表 1）组成，组合序列则代表一个多肽序列。

表 1　氨基酸类型

序号	氨基酸类型	字母表示	支链
1	丙氨酸	Ala(A)	疏水性
2	半胱氨酸	Cys(C)	亲水性
3	天冬氨酸	Asp(D)	酸性
4	谷氨酸	Glu(E)	酸性
5	苯丙氨酸	Phe(F)	疏水性
6	甘氨酸	Gly(G)	亲水性
7	组氨酸	His(H)	碱性
8	异亮氨酸	Iie(I)	疏水性
9	赖氨酸	Lys(K)	碱性
10	亮氨酸	Leu(L)	疏水性
11	甲硫氨酸	Met(M)	疏水性
12	天冬酰胺	sn(N)	亲水性
13	脯氨酸	Pro(P)	疏水性
14	谷氨酰胺	Gln(Q)	亲水性
15	精氨酸	Arg(R)	碱性
16	丝氨酸	Ser(S)	亲水性
17	苏氨酸	Thr(T)	亲水性
18	缬氨酸	Val(V)	疏水性
19	色氨酸	Trp(W)	疏水性
20	酪氨酸	Tyr(Y)	亲水性

2.1.2 数据集

相关研究成果[8][9][10][11]表明，建立良好且平衡的数据集对于构建一个稳健可靠的预测模型是至关重要的一部分。如[2][12][13]所述，每个数据集都包括正数据集和负数据集，且正负数据集之间没有重叠。在抗癌肽的研究中，正样本是实验验证的抗癌肽，负样本是不具有抗癌功能的抗菌肽（AMPs）[13]。本文使用三个独立的基准数据集来研究 PACPs 的有效性和通用性，这些基准数据集为 ACP740、ACP500 和 ACP164 数据集，如表 2 所示。本文把 ACP740、ACP500 分别划分训练集与测试集，将两个训练集合并来训练模型，两个测试集以及独立测试集 ACP164 用以对模型进行评估。本文构建训练数据集的目的是训练预测模型，构建独立测试数据集中的肽都没有出现在训练数据集中, 以确保对模型性能进行公平评估，验证预测模型的泛化能力。

表 2　数据集

样本集	总样本	正样本	负样本	用法
ACP740	740	376	364	80%训练集，20%测试集
ACP500	500	250	250	80%训练集，20%测试集
ACP164	164	82	82	独立测试集

（1）ACP740 数据集

ACP-DL[14]所构建的 ACP740 数据集，为了避免数据集的偏差，使用 CD-HIT[15]工具去除相似性超过 90%的肽序列，是非冗余数据集。此基准数据集可在 https://github.com/haichengyi/acp-dl 上公开获得。最终获得的 ACP740 数据集包括 740 个样本，其中 376 个为正样本，364 个为负样本。

（2）ACP500 数据集和 ACP164 数据集

ACPred-FL[16]构建了两个数据集：ACP500 数据集和 ACP164 数据集。同样为了避免数据集的偏差，都使用了 CD-HIT[15]工具去除相似性超过 90%的冗余肽序列，最终保留了 332 个 ACP 和 1023 个非 ACP。又因在平衡训练数据集中更容易获得更好的性能，随机选择了 250 个正样本和 250 个负样本来构建 ACP500，而 ACP164 则用剩余的 82 个正样本和 82 个随机选择的负样本。故这两个数据集也是没有重叠的非冗余数据集。ACP500 和 ACP164 数据集可从 http://server.malab.cn/ACPred-FL 获取。

2.1.3 三个数据集上抗癌肽氨基酸序列分析

本文在数据集 ACP740、ACP500 和 ACP164 上对所有肽序列中 20 种不同氨基酸的组成进行计数和比较（图 3），并且对所有肽序列的长度进行比较（图 4）。肽序列中 20 种不同氨基酸的组成比较中，发现某些残基如甘氨酸(G)、丝氨酸(S)、苯丙氨酸(F)、组氨酸(H)、半胱氨酸(C)、天冬酰胺(N)和酪氨酸(Y)在抗癌肽中比较丰富。而亮氨酸(L)、色氨酸(W)、赖氨酸(K)、丙氨酸(A)、精氨酸(R)和谷氨酰

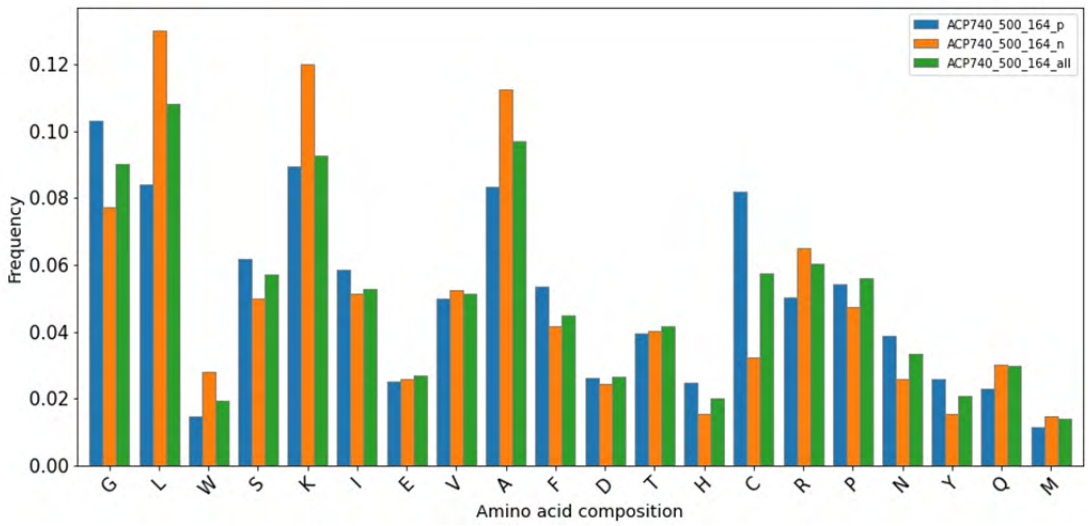

图 3　抗癌肽，非抗癌肽和总的肽序列的氨基酸组成的比较

胺(Q)在非抗癌肽中比较丰富。在肽序列的长度比较中，发现抗癌肽的序列长度为 11 至 97 之间，而非抗癌肽在 11 至 207 之间，抗癌肽主要集中在序列长度为 11 至 50 之间，而非抗癌肽主要集中在 10 至 35 之间，从图 4 中可看出抗癌肽往往比非抗癌肽具有更长的氨基酸序列。

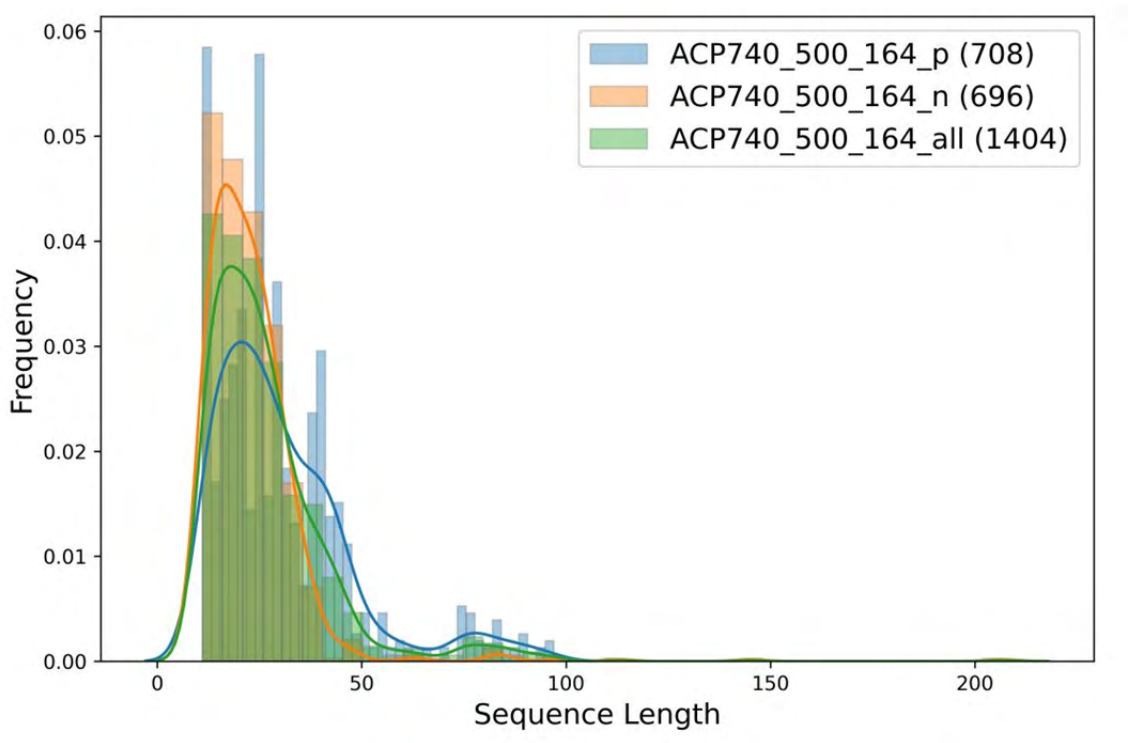

图 4　抗癌肽、非抗癌肽和总的肽序列的的氨基酸序列长度分布

2.2 特征提取（特征描述符+基分类器=基学习器）

2.2.1 特征描述符

将肽序列转换为合理的定长特征向量对于机器学习分类算法是非常重要的。而且对于每种特征编码策略，不同的参数设置可以提取不同的序列信息，本文选择了 21 类基于肽序列的特征描述符，如表 3 所示，这 21 类涵盖了大部分可用的特征描述符。这些特征描述符都是借助 iLearnPlus[17]平台实现的。

本文通过两步特征选择确定最优的基学习器子集，以下主要介绍最终模型所用到的特征描述符。

表 3　21 个特征描述符

描述符组	特征描述符	缩写
氨基酸组成	氨基酸组成	AAC
	k 间隔氨基酸对组成	CKSAAP
	Kmer 二肽组成	DPC
	二肽与预期平均值的偏差	DDE
	距离对和缩减字母的 PseAAC	DistancePair
	成分/过渡/分布	CTD
	自适应跳过二肽组成	ASDC
	Kmer 三肽组成	TPC
	联合三元组	CTriad
	联合 k 间隔三元组	KSCTriad
分组氨基酸组成	k 间隔氨基酸对的组成	CKSAAGP
	分组氨基酸组成	GAAC
	分组二肽组成	GDPC
	分组三肽组成	GTPC
自相关	归一化 Moreau-Broto	NMBroto
	自协方差	AC
准序	自协方差	AC
	准序列序描述符	QSOrder
伪氨基酸组成	两亲性伪氨基酸组成	PAAC
	两亲性伪氨基酸组成	APAAC
	伪 K 元组还原氨基酸组成	PseKRAAC

（1）组成/过渡/分布（CTD）

CTD 特征中组成(C)表示每个氨基酸的组成，过渡(T)表示一个氨基酸伴随另

一氨基酸残基的频率，分布(D)计算肽序列的分布。CTD 代表肽序列中特定结构或物理化学性质的氨基酸分布模式，包括疏水性、极性和溶剂可及性等 13 种物理化学性质。此特征描述符已成功应用于蛋白质结构预测和抗癌肽预测[17]。CTDC，CTDT 和 CTDD 分别由 39 维，39 维和 195 维特征向量表示，得到 CTD 特征由 273 维特征向量表示。

（2）自适应跳过二肽组成（ASDC）

ASDC 是 Kmer 二肽组成（DPC）的修改版，其充分考虑了存在于相邻残基之间和中间残基之间的相关信息。ASDC 已成功应用于抗癌肽的预测[16]、细胞穿透肽预测和血脑屏障肽[18]。对于给定的序列，ASDC 的特征向量[18]表示为：

$$\mathrm{ASDC} = \left(\mathrm{s}_1, \mathrm{s}_2, \ldots \mathrm{s}_\mathrm{q}, \ldots, \mathrm{s}_{400}\right),$$

$$\mathrm{s}_\mathrm{q} = \frac{\sum_{\mathrm{g}=1}^{\mathrm{L}-1} \mathrm{N}_\mathrm{q}^\mathrm{g}}{\sum_{\mathrm{p}=1}^{400} \sum_{\mathrm{g}=1}^{\mathrm{L}-1} \mathrm{N}_\mathrm{q}^\mathrm{g}}.$$

其中s_q表示所有可能二肽的出现频率小于等于 L-1 干预氨基酸，$\mathrm{N}_\mathrm{q}^\mathrm{g}$表示序列中的第 q 个 g-间隔二肽数。ASDC 特征由 400 维特征向量表示。

（3）k-间隔氨基酸基团对的组成（CKSAAGP）

CKSAAGP 是 k-间隔氨基酸对的组成（CKSAAP）描述符的变体，二肽组成（DPC）的进一步扩展。它计算由任何 k 残基分离的氨基酸基团对的频率，在蛋白质预测的几个研究[19]中作为代表肽序列的短基序的有效描述符。首先，20 个氨基酸残基按其物理化学性质分为五类。以 k=0 为例，有 25 个 0 间隔的群对（即 $g1g1$，$g1g2$，$g1g3$，…，$g5g5$）。因此，CKSAAGP 的特征向量可以定义为：

$$\left(\frac{N_{g1g1}}{N_{total}}, \frac{N_{g1g2}}{N_{total}}, \frac{N_{g1g3}}{N_{total}}, \ldots, \frac{N_{g5g5}}{N_{total}}\right)_{25}$$

每个描述符的值表示蛋白质或肽序列中相应残基对的组成。例如，如果残基对$g1g1$在蛋白质中出现 m 次，则残基对$g1g1$的组成等于 m 除以蛋白质中 0 间隔残基对的总数N_{total}。本文选取 iLearnPlus[17]中 CKSAAGP 的默认参数 k=3，故其特征由 100 维特征向量表示。

（4）准序列序描述符（QSOrder）

对于每种氨基酸类型，准序列顺序描述符可以定义为：

$$X_r = \begin{cases} \dfrac{f_r}{\sum_{r=1}^{20} f_r + w\sum_{d=1}^{nlag}\tau_d}, r = 1,2,\ldots,20 \\ \dfrac{w\tau_d - 20}{\sum_{r=1}^{20} f_r + w\sum_{d=1}^{nlag}\tau_d}, r = 21,22,\ldots,20+nlag \end{cases}$$

其中f_r是氨基酸类型r的归一化发生率，w是加权因子，$nlag$为lag的最大值。本文选取 iLearnPlus[17]中 QSOrder 的默认参数lag=2，w=0.05，故其特征由 44 维特征向量表示。

2.2.2 基分类器

本文选择了 13 个常见的机器学习分类算法，涵盖了大部分可用的机器学习分类算法，如表 4 所示。本文通过两步特征选择确定最优的基学习器子集，以下主要介绍最终模型所用到的机器学习分类算法。

表 4　常见机器学习分类算法

序号	机器学习分类算法	简称
1	随机森林	RF
2	梯度提升决策树	XGBoost
3	极端梯度增强	ET
4	极端随机树	LGBM
5	Light 梯度增强机	SVM
6	支持向量机	GBDT
7	逻辑回归	LR
8	K 近邻	NBC
9	决策树	KNN
10	朴素贝叶斯	LDA
11	随机梯度下降	DecisionTree
12	线性判别分析	SGD
13	二次判别分析	QDA

（1）梯度提升决策树

梯度提升决策树（Gradient Boosting Decision Tree，简称 GBDT）[20][25]是基于决策树的一种增强算法。其原理就是弱分类器（决策树）拟合残差，即在每一轮迭代中建立一棵决策树去拟合残差，使其残差在梯度方向上减少，再将该决策树与当前模型进行线性组合得到新模型，不断重复直到达到终止条件，得到最终的强学习器。

（2）极端随机树

极端随机树（Extremely randomized trees，简称 ET）算法类似于随机森林，都是由许多决策树[25]构成。其主要思想是：由于特征随机、参数随机、模型随机和分裂随机的存在，某棵决策树预测结果往往是不准确的，但多棵决策树组合得到的预测结果，就可以减少误差，达到更好的预测效果。ET 使用所有的训练样本得到每颗决策树，并且是完全随机地得到分叉值。

（3）支持向量机

支持向量机（Support Vector Machine，简称 SVM）[20][25]模型求解凸二次规划的最优化算法。其基本模型是定义在特征空间上的间隔最大的线性分类器，学习策略是间隔最大化，这使它有别于感知机，使用核技巧，故其实质是非线性分类器。

（4）逻辑回归

逻辑回归（Logistic Regression，简称 LR）[20][25]是一种广义线性回归分析模型，虽然被称为回归，但其本质是分类模型，并常用于二分类问题。LR 的基本思想是：假设数据服从某个分布，然后使用极大似然估计做参数的估计，这样就变成了以对数似然函数为目标函数的最优化问题。

2.2.3 基学习器

基学习器是由特征描述符编码提取序列信息，再输入到基分类器中学习得到预测概率向量，从而进一步提取序列信息的模型。如前文所述，本文使用了 21 种特征描述符来编码肽序列。对于每种特征编码策略，不同的参数设置可以提取不同的序列信息，本文采取了 iLearnPlus[17]平台中各个特征描述符的默认参数设

置，最后在特征池中得到 21 个特征描述符（表 3）。然后将特征池中的每个特征描述符输入到 13 机器学习分类器中，形成 273 个基学习器，再通过两步特征选择的方法选出 5 个基分类器构成的最优特征子集, 同时也得到了 5 个预测的概率向量作为新的 5 维特征向量输入到第二层的分类器中。

2.3 特征选择

本文采用了两步特征选择的方法, 从 21 个特征描述符和 13 个分类器构成的 273 个基学习器中选择最优基学习器子集。第一步运用平均 F1 得分对特征描述符和分类器分别进行了筛选。首先利用平均 F1 分数大于 70%的指标筛选出 12 个特征描述符和 7 个分类器，然后将特征描述符和分类器组合得到了 84 个基学习器。第二步采用了递归特征消除的方法进行最优基学习器子集的选择。最后得到了最优基学习器子集，包括 5 个基学习器：ET-CKSAAGP，ET-CTD，GDBT-CTD，GDBT-QSOrder，SVM-ASDC。

递归特征消除(Recursive Feature Elimination，RFE)[21]是封装器的一种贪心搜索算法。主要思想是利用训练集数据生成模型，再根据模型的特征权重，对特征进行取舍，消除权重不高的特征，从而得到数据集的特征子集。然后，对这个特征子集重复上述的过程，直到特征数量达到规定值位置。

2.4 分类器构建（元学习器）

如上所述，本文在第一层获得了 5 个最佳基础学习器，其输出的 5 维概率向量组合成新的特征向量。然后，把新的特征向量作为第二层上最终元学习器的输入进行训练。对于元学习器，本文沿用了基学习器中用平均 F1 分数筛选出的 7 种机器学习分类算法, 并根据其 8 个性能评估指标最终选择了性能最好的逻辑回归（LR）作为元学习器，其输出就是本文模型 PACPs 预测抗癌肽的最终结果。本文按照经验将模型得到预测为抗癌肽概率的分类点设置为 0.5，即如果肽的概率≥0.5，则该肽被识别为抗癌肽。

2.5 性能评估指标

本文选择了二分类任务中广泛使用的九个评价指标来评估抗癌肽预测算法模型的性能，包括准确性（ACC），F1 分数（F1），F2 分数（F2），几何平均值（GMean），灵敏度(SEN)，精度（PREC），特异性(SPEC)，马修相关系数(MCC)和 ROC 曲线下区域的面积（AUC）。评估指标定义公式如下:

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

$$\mathrm{F1}=\frac{2\times\mathrm{SEN}\times\mathrm{PREC}}{\mathrm{SEN}+\mathrm{PREC}},$$

$$\mathrm{F2}=\frac{5\times\mathrm{SEN}\times\mathrm{PREC}}{\mathrm{SEN}+4\times\mathrm{PREC}},$$

$$\mathrm{GMean}=\sqrt{\mathrm{SEN}\times\mathrm{SPEC}},$$

$$\mathrm{SEN}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},$$

$$\mathrm{SPEC}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},$$

$$\mathrm{PREC}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},$$

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

其中 TP，FP，TN，FN 分别表示正确预测的抗癌肽，非抗癌肽错误预测为抗癌肽，正确预测的非抗癌肽以及抗癌肽错误预测为非抗癌肽的数量，如表 5 所示。

表 5　二分类混淆矩阵

.		预测类别	
		Positive	Negative
真实类别	Positive	TP	FN
	Negative	FP	TN

第 3 章　结果与分析

在此次试验中, 本文在第一阶段基学习器的构建中先采用留出法来划分训练集和测试集，然后在训练集上采用了 10 折交叉验证法对模型进行调优，找到模型泛化性能最优的超参数。再在全部训练集上重新训练，使用独立测试集对基学习器性能进行度量。在第二阶段元学习器构建中，仍然使用第一阶段所划分的训练集训练元学习器，并在独立测试集上对模型整体性能进行度量。

3.1 基学习器的选择

堆叠学习模型的效果主要取决于基学习器的性能, 具有高精度和多样性的基学习器组合对最终预测模型的泛化精度非常重要。本文使用 21 个特征描述符和 13 个机器学习算法构建了 273 个基学习器。然后采用两步特征选择的方法，最后选出了 5 个基学习器构成了最优基学习器组合。

第一步，根据平均 F1 得分对特征描述符和机器学习算法分别进行选择。首先根据平均 F1 分数大于 70%的指标从 21 个特征描述符中筛选出了 12 个特征描述符，其结果如图 5 所示。

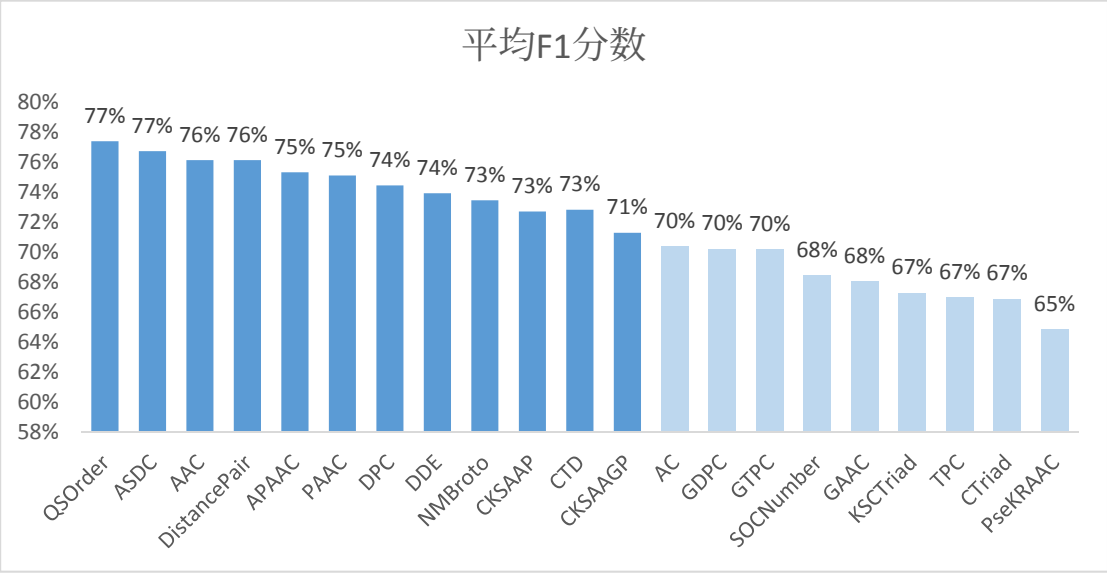

图 5　21 个特征描述符的平均 F1 分数

其次根据平均 F1 分数大于 70%的指标从 13 个分类器中筛选出了 7 个分类器，其结果如图 6 所示。

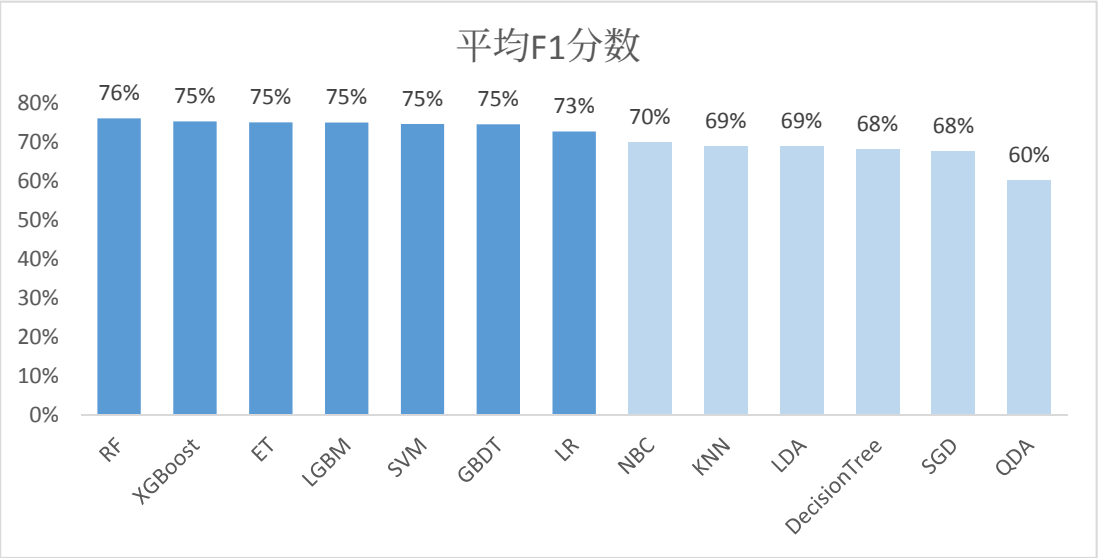

图 6　13 个机器学习分类器的平均 F1 得分

第二步，根据递归特征消除方法选出最优基学习器的特征子集。本文对常见的四种特征选择的方法：方差过滤、顺序搜索、SlectKBest、RFE 进行了横向比较，选出含有相同个数基学习器的最优特征子集。从 AUC 的结果（图 7）来看，基于 RFE 的模型优于其他三种模型，所以本文选择了效果相对较好的递归特征消除方法。

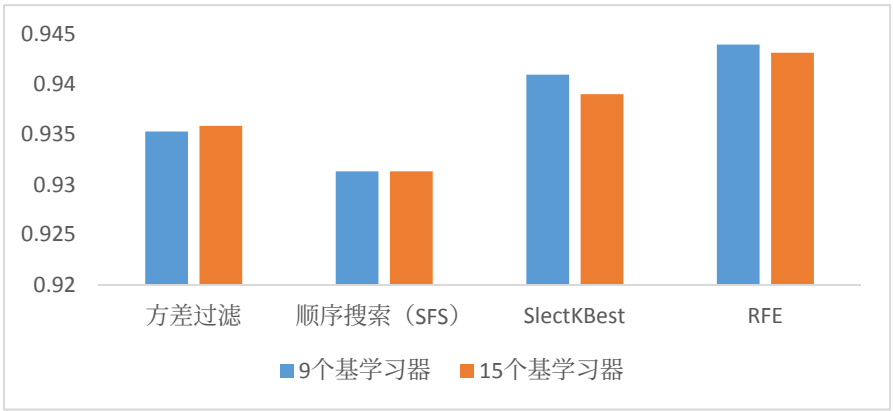

图 7　特征选择方法比较

本文对递归特征消除方法进行了进一步的优化，选择了带有交叉验证的递归特征消除（RFECV）方法来选择最优特征组合。通过不断的调整参数，发现当设

置所选用的模型为逻辑回归，迭代过程中每次移除的属性的数量为 8 个，交叉验证参数为 3，保留的特征数量为 5 时得出的预测效果相对较好，在 ACP740 数据集上 AUC 达到了 94.26%。因此本文运用 RFECV，选择了 5 个基学习器组成的最优特征子集，其中包含的基分类器为 ET-CKSAAGP，ET-CTD，GDBT-CTD，GDBT-QSOrder，SVM-ASDC。

3.2 第二层分类器的选择比较

通过第一层模型的构建，本文选择了最佳的 5 个基分类器，并将其得到的 5 维概率向量组成新的特征向量，用于第二层的元学习器的训练。对于元学习器的选择，本文仍然使用了第一层所选出来的 7 个机器学习算法，对 5 维概率向量进行了训练，得到的结果如图 8 所示。基于 AUC 来看，在 7 个分类器中，LR 的 AUC 值最高，此外通过阅读文献和收集资料可以发现，LR 已经过验证是第二层分类器中最好的选择，故本文选择了 LR 作为第二层分类器模型。

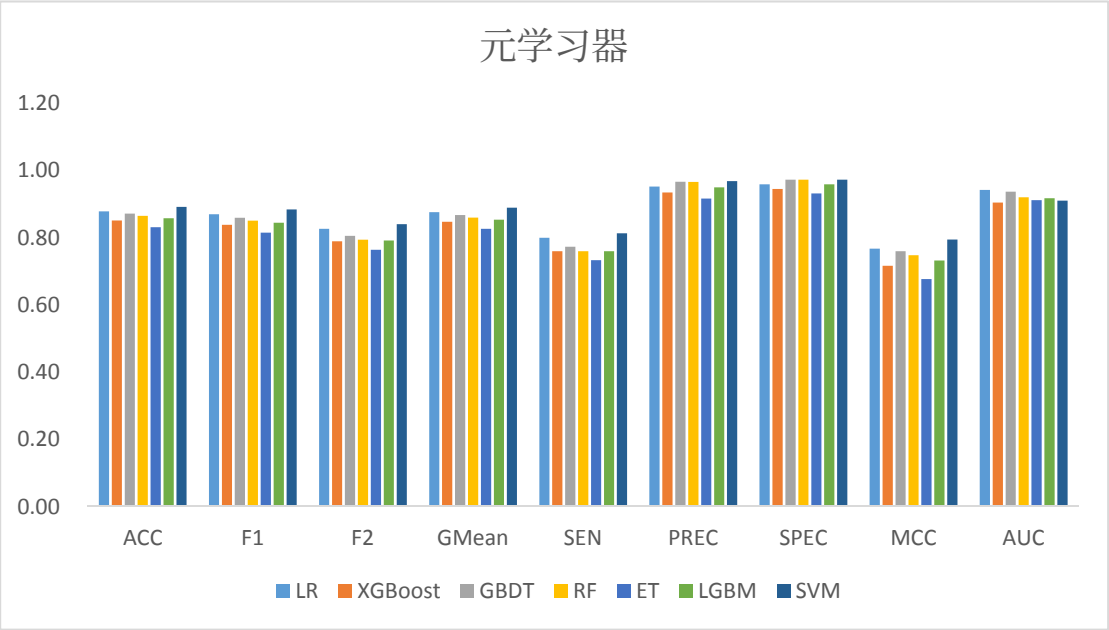

图 8　元学习器中分类算法的比较

3.3 PACPs 模型与基学习器的性能比较

通过两步特征选择的方法，基于 273 个基学习器，选择了 5 个最佳的基学习器，将其输出用来训练第二层分类器，构建预测抗癌肽的 PACPs 模型。为了验证 PACPs 模型的性能效果，本文在 ACP740 和 ACP500 的测试集以及独立测试集 ACP164，将 PACPs 模型的效果与 5 个特征基学习器的效果进行了比较分析。具有代表性的结果如图 9，10 示，在 ACP740 的测试集中，除去部分指标，PACPs 模型在 ACC、F1、GMean、SEN、MCC 和 AUC 的结果都优于 5 个基学习器的结果；在独立测试集 ACP164 中，PACPs 模型在所有性能度量指标的结果都比 5 个基学习器的结果好。综合来看，PACPs 模型在两个测试集样本上的分类效果理想，在独立测试集样本上的分类效果也达到了预期。此外，由于基学习器具有随机性，堆叠第二层分类器也会使得预测效果更稳定。

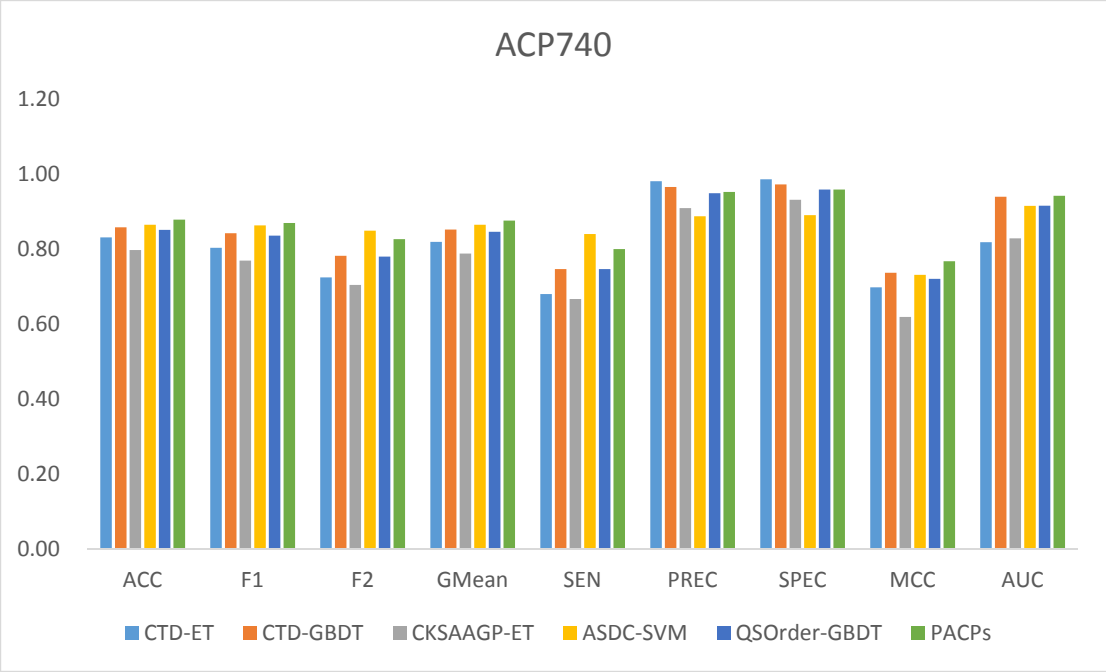

图 9　ACP740 数据集上的性能比较

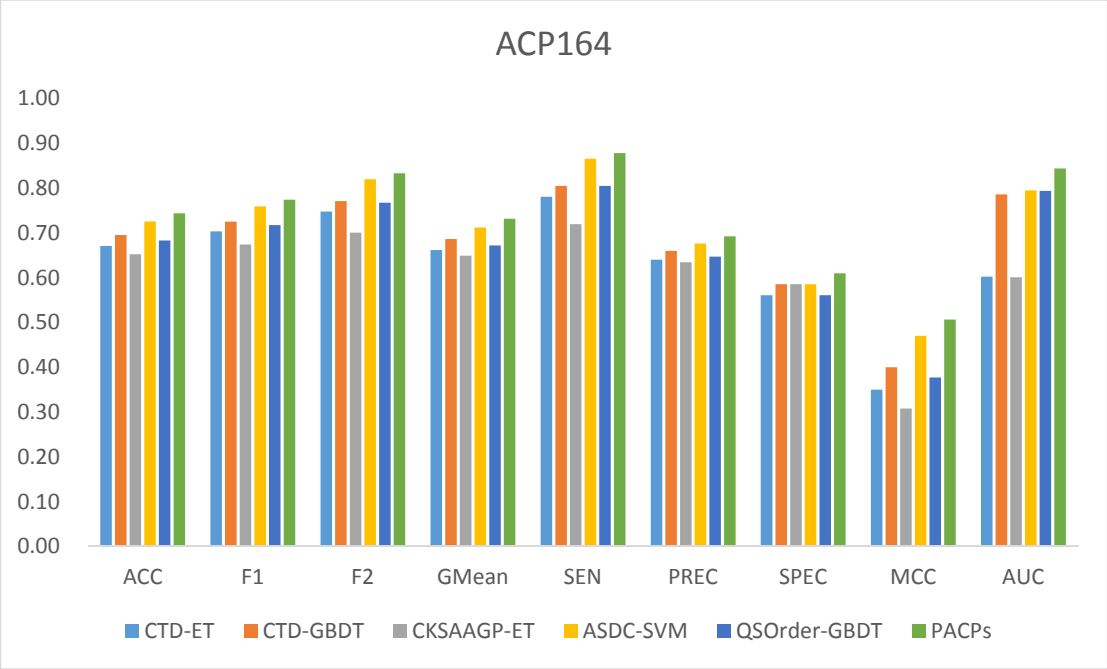

图 10　ACP164 数据集上的性能比较

3.4 模型之间的比较

本文将提出的模型 PACPs 在 ACP740 和 ACP500 的测试集，以及独立测试集 ACP164 上, 与预测抗癌肽的模型 ACP-MHCNN[24]、ACPred[22]和 mACPpred[23]进行了复现比较试验。

ACP-MHCNN[24]是一种多头深度卷积神经网络模型，用于以交互方式从不同信息源中提取和组合判别基于序列，物理化学和进化的特征来鉴定 ACP。ACPred[22]是一种基于 SVM、RF 与各种类别的肽特征预测和表征肽抗癌活性的模型。mACPpred[23]是一种用两步特征选择出七个特征编码并获得相应的最佳特征模型，其结果的概率向量作为 SVM 的输入得到最终预测模型。

所得到的结果如表 6 所示。PACPs 在 ACP740 测试集上 AUC、PREC、MCC、SPEC 和 ACC 表现都优于其他三个预测抗癌肽的模型，在独立测试集 ACP164 上 AUC、MCC、SEN 和 ACC 表现都优于其他三个预测抗癌肽的模型, 在 ACP500

测试集上的 AUC 表现也优于其他三个预测抗癌肽的模型。说明了本文所提出的模型性能良好，可以用于抗癌肽的识别。

表 6　不同模型的性能比较

数据集	模型	ACC	SEN	SPEC	MCC	PREC	AUC
ACP740	ACP-MHCNN	0.83	0.83	0.82	0.65	0.83	0.90
	ACPred	0.77	0.90	0.63	0.55	0.72	0.77
	mACPpred	0.63	0.79	0.45	0.26	0.60	0.67
	ACPsP	**0.88**	0.80	**0.96**	**0.77**	**0.95**	**0.94**
ACP500	ACP-MHCNN	0.85	0.90	0.79	0.70	0.82	0.89
	ACPred	0.49	0.87	0.12	-0.02	0.50	0.37
	mACPpred	0.39	0.72	0.05	-0.31	0.43	0.29
	ACPsP	0.79	0.87	0.72	0.59	0.73	**0.90**
ACP164	ACP-MHCNN	0.68	0.65	0.70	0.38	0.74	0.73
	ACPred	0.52	0.89	0.15	0.05	0.51	0.41
	mACPpred	0.40	0.67	0.13	-0.23	0.44	0.30
	ACPsP	**0.74**	**0.88**	0.61	**0.51**	0.69	**0.84**

第 4 章　总结与展望

4.1 主要结论

本文建立了一种识别抗癌肽的预测模型 PACPs。该模型是由两层的堆叠框架构成的。在第一层，本文将 21 个特征描述符和 13 个机器学习分类器结合得到了 273 个基分类器。运用两步特征选择的方法，选择了最优基学习器组合。首先根据平均 F1 分数大于 70%的指标，选出了 12 个特征描述符和 7 个分类器。然后运用了带有交叉验证的递归特征消除的方法（RFECV）选出了 5 个基学习器：ET-CKSAAGP，ET-CTD，GDBT-CTD，GDBT-QSOrder，SVM-ASDC，构成了最优基学习器组合。第二层，将 5 个基学习器学习者的输出组成新的特征向量，来训练第二层的分类器 LR，其输出是对抗癌肽的最终预测结果。基于叠加的方法，PACPs 将不同的特征表示符和机器学习分类器进行了集成，从而整合了每个模型的有点，使 PACPs 模型成为了一个性能良好的预测模型。

目前抗癌肽疗法已广泛应用于癌症治疗的不同阶段，但抗癌肽的鉴定仍然高度受限于实验室，费用昂贵和周期漫长。PACPs 模型有助于抗癌肽的预测，从而推动肽类药物的研发设计，间接改善癌症患者的治疗效果。

4.2 模型反思与评价

4.2.1 模型的评价

基于抗癌肽的数据集，结合特征描述符和机器学习算法构建的 PACPs 模型，相较于其他模型，采用了更多的特征描述符和机器学习算法，使基学习器的性能优点得到了更好的整合。根据测试集的评估结果可以看出，在大的数据集上预测的精度较高，在小的数据集上也达到了不错的效果。

4.2.2 模型的反思

基于深度学习所构建模型的结果缺乏可解释性和可再现性, 这可能会限制它们的效用。在肽中使用机器学习算法预测很少，还有重要的已知缺点。其中用于训练模型的数据集应该精确、策划和高质量。

由于所涉及的高实验成本和复杂性，数据的收集、共享和分发具有挑战性，这导致了整个肽序列空间的覆盖不完整。数据的格式和质量通常在构建模型之前需要预清洗或预处理，这带来了额外的挑战。本文模型所使用的数据集采用的是已经过的处理的数据集，PACPs 缺乏对数据的处理过程，还有待改进。

模型的比较检验中, 在相同的数据集中, PACPs 在数据集 ACP740 和 ACP164 上表现较好，在数据集 ACP500 的性能还存在不足，有待提升。

模型的实现代码还存在优化空间, 本文提出的模型对序列信息的转化是借助 iLearnPlus[17]平台实现的，没有对特征描述符进行专门的函数定义；采用的特征描述符和机器学习算法都是基于默认的参数。下一步的工作可以通过定义特征描述符使代码更加完整，调整合适的参数提高模型的性能。

4.2.3 模型的应用与推广

本文基于多种特征描述符和机器学习所建立的模型框架, 可以推广应用于其他肽的鉴别预测，例如抗冠状病毒肽、神经肽、群体感应肽、抗高血压肽等。用肽序列数据与机器学习算法组合形成基学习器并进行训练, 再用特征选择的方法挑选出能鉴别肽序列特征的基学习器，结合元学习器构建肽的鉴别模型，提高预测肽的准确性，推动关于肽类药物的研发设计，间接改善患者的治疗效果。

参考文献

[1] Hajisharifi Z, Piryaiee M, Mohammad Beigi M, Behbahani M, Mohabatkar H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. J Theor Biol. 2014 Jan 21;341:34-40.

[2] Vijayakumar S, Lakshmi P. ACPP: a web server for prediction and design of anti‐cancer peptides. Int J Pept Res Ther. 2015;21(1):99‐106.

[3] Li F‐M, Wang X‐Q. Identifying anticancer peptides by using improved hybrid compositions. Sci Rep. 2016;6:33910.

[4] Khan F, Akbar S, Basit A, Khan I, Akhlaq H Identification of anticancer peptides using optimal feature space of Chou's split amino acid composition and support vector machine. Paper presented at: Proceedings of the 2017 4th International Conference on Biomedical and Bioinformatics Engineering 2017.

[5] Akbar S, Hayat M, Iqbal M, Jan MA. iACP‐GAEnsC: evolutionary genetic algorithm based Ensemble classification of anticancer peptides by utilizing hybrid feature space. Artif Intell Med. 2017;79:62‐70.

[6] Wei L, Hu J, Li F, Song J, Su R, Zou Q. Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. Brief Bioinform. 2018 Oct 31.

[7] Bin Y, Zhang W, Tang W, Dai R, Li M, Zhu Q, Xia J. Prediction of Neuropeptides from Sequence Information Using Ensemble Classifier and Hybrid Features. J Proteome Res. 2020 Sep 4;19(9):3732-3740.

[8] Wei,L. et al. (2017a) SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penETating peptides. BMC Genomics, 18, 1 – 11.

[9] Wei,L. et al. (2017b) Fast prediction of methylation sites using sequence-based

feature selection technique. IEEE/ACM Trans. Comput. Biol. Bioinform.

[10] Wei,L. et al. (2017c) CPPred-RF: a sequence-based predictor for identifying cell-penETating peptides and their uptake efficiency. J. Proteome Res., 16, 2044.

[11] Xing,P. et al. (2017) Identifying N6-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine. Sci. Rep., 7,46757.

[12]Chen,W. et al. (2016) iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget, 7, 16895.

[13] Tyagi,A. et al. (2015) CancerPPD: a database of anticancer peptides and proteins. Nucleic Acids Res., 43, D837.

[14] Hai-Cheng Yi. et al. (2019) . Acp-dl: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. Molecular therapy. Nucleic acids, pages 1 – 9, 2019.

[15] Li W, Godzik A. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658 – 9.

[16] Wei L, Zhou C, Chen H, et al. (2018) ACPred-FL: a sequencebased predictor based on effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics. Bioinformatics, 34(23), 2018, 4007 – 4016.

[17] Zhen Chen, et al.(2021). iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. Nucleic Acids Research, 2021 1 – 19.

[18] Ruyu Dai, et al.(2021). BBPpred: Sequence-Based Prediction of Blood-Brain Barrier Peptides with Feature Representation Learning and Logistic Regression. Journal of Chemical Information and Modeling. J. Chem. Inf. Model. 2021, 61, 525−534

[19] Yuxuan Pang, et al.(2021). Identifying anti-coronavirus peptides by incorporating different negative datasets and imbalanced learning strategies. Briefings in Bioinformatics, 00(00), 2021, 1 – 11.

[20] 李航.统计学习方法.北京:清华大学出版社, 2012 年.

[21] 齐伟.数据准备和特征工程.电子工业出版社, 2020 年.

[22] Nalini Schaduangrat, et al.(2019). ACPred: A Computational Tool for the Prediction and Analysis of Anticancer Peptides. Molecules 2019, 24, 1973.

[23] Vinothini Boopathi, et al.(2019). mACPpred: A Support Vector Machine-Based Meta-Predictor for Identification of Anticancer Peptides. Int. J. Mol. Sci. 2019, 20, 1964.

[24]Sajid Ahmed, et al. (2020).ACP-MHCNN: An Accurate Multi-Headed Deep-Convolutional Neural Network to Predict Anticancer peptides. bioRxiv preprint.

[25] 周志华.机器学习.清华大学出版社, 2016 年.

附录

1. 21 个特征描述符的平均 F1 得分

序号	特征描述符	缩写	平均 F1 分数
1	准序列序描述符	QSOrder	77%
2	自适应跳过二肽组成	ASDC	77%
3	氨基酸组成	AAC	76%
4	距离对和缩减字母的 PseAAC	DistancePair	76%
5	两亲性伪氨基酸组成	APAAC	75%
6	两亲性伪氨基酸组成	PAAC	75%
7	Kmer 二肽组成	DPC	74%
8	二肽与预期平均值的偏差	DDE	74%
9	归一化 Moreau-Broto	NMBroto	73%
10	k 间隔氨基酸对组成	CKSAAP	73%
11	成分/过渡/分布	CTD	73%
12	k 间隔氨基酸对的组成	CKSAAGP	71%
13	自协方差	AC	70%
14	分组二肽组成	GDPC	70%
15	分组三肽组成	GTPC	70%
16	顺序耦合号	SOCNumber	68%
17	分组氨基酸组成	GAAC	68%
18	联合 k 间隔三元组	KSCTriad	67%
19	Kmer 三肽组成	TPC	67%
20	联合三元组	CTriad	67%
21	伪 K 元组还原氨基酸组成	PseKRAAC	65%

2. 7 个机器学习算法的平均 F1 得分

序号	机器学习分类算法	简称	平均 F1 分数
1	随机森林	RF	76%
2	梯度提升决策树	XGBoost	75%
3	极端梯度增强	ET	75%
4	极端随机树	LGBM	75%
5	Light 梯度增强机	SVM	75%
6	支持向量机	GBDT	75%
7	逻辑回归	LR	73%
8	K 近邻	NBC	70%
9	决策树	KNN	69%
10	朴素贝叶斯	LDA	69%
11	随机梯度下降	DecisionTree	68%
12	线性判别分析	SGD	68%
13	二次判别分析	QDA	60%

3. 在选择同一数量基学习器下，不能特征选择方法得到的预测结果的 AUC 值

方法	9 个基学习器	15 个基学习器
方差过滤	0.935292685	0.935861005
顺序搜索（SFS）	0.931314443	0.931314443
SlectKBest	0.940975887	0.939027361
RFE	0.943979865	0.943167979

4. 基学习器与 PACPs 模型的预测结果的性能比较

样本	基学习器	ACC	F1	F2	GMean	SEN	PREC	SPEC	MCC	AUC
ACP 740	CTD-ET	0.83	0.80	0.72	0.82	0.68	0.98	0.99	0.70	0.818
	CTD-GBDT	0.86	0.84	0.78	0.85	0.75	0.97	0.97	0.74	0.940
	CKSAAGP-ET	0.80	0.77	0.70	0.79	0.67	0.91	0.93	0.62	0.829
	ASDC-SVM	0.86	0.86	0.85	0.86	0.84	0.89	0.89	0.73	0.915
	QSOrder-GBDT	0.85	0.84	0.78	0.85	0.75	0.95	0.96	0.72	0.916
	PACPs	0.88	0.87	0.83	0.88	0.80	0.95	0.96	0.77	0.942
ACP 500	CTD-ET	0.81	0.81	0.86	0.81	0.89	0.75	0.74	0.63	0.863
	CTD-GBDT	0.83	0.84	0.91	0.83	0.96	0.75	0.72	0.69	0.928
	CKSAAGP-ET	0.83	0.83	0.88	0.83	0.91	0.76	0.76	0.67	0.864
	ASDC-SVM	0.78	0.79	0.86	0.78	0.91	0.70	0.67	0.59	0.875
	QSOrder-GBDT	0.79	0.79	0.84	0.79	0.87	0.73	0.72	0.59	0.902
	PACPs	0.79	0.79	0.84	0.79	0.87	0.73	0.72	0.59	0.902
ACP 164	CTD-ET	0.67	0.70	0.75	0.66	0.78	0.64	0.56	0.35	0.602
	CTD-GBDT	0.70	0.73	0.77	0.69	0.80	0.66	0.59	0.40	0.786
	CKSAAGP-ET	0.65	0.67	0.70	0.65	0.72	0.63	0.59	0.31	0.601
	ASDC-SVM	0.73	0.76	0.82	0.71	0.87	0.68	0.59	0.47	0.795
	QSOrder-GBDT	0.68	0.72	0.77	0.67	0.80	0.65	0.56	0.38	0.794
	PACPs	0.74	0.77	0.83	0.73	0.88	0.69	0.61	0.51	0.844

5. 元学习器采用的机器学习算法的比较

分类器	ACC	F1	F2	GMean	SEN	PREC	SPEC	MCC	AUC
LR	0.88	0.87	0.83	0.88	0.80	0.95	0.96	0.77	0.942
XGBoost	0.85	0.84	0.79	0.85	0.76	0.93	0.95	0.72	0.905
GBDT	0.87	0.86	0.81	0.87	0.77	0.97	0.97	0.76	0.937
RF	0.86	0.85	0.79	0.86	0.76	0.97	0.97	0.75	0.920
ET	0.83	0.81	0.76	0.83	0.73	0.92	0.93	0.68	0.912
LGBM	0.86	0.84	0.79	0.85	0.76	0.95	0.96	0.73	0.918
SVM	0.89	0.88	0.84	0.89	0.81	0.97	0.97	0.79	0.910

致谢

从选题、立论、建模到撰写的整个过程中，我们都学到了很多东西，不仅把学到的专业知识用于生活中具体的实践上，而且锻炼了遇到问题时分析和解决问题的能力。

首先，要感谢指导老师，从论文的选题、开题再到建模和论文的完成，都离不开指导老师的悉心指导。其次要感谢小组每位成员，在建模过程中遇到困难都坚持去解决化解困难，小组团结一致最终完成了此次建模比赛。然后还要感谢的导师组的所有的小伙伴们，每次讨论课的积极分享与热烈讨论都扩展了本文的思维。最后，对大学生统计建模组委会和对参加本论文评阅和对本论文提出宝贵意见的所有老师和同学表示诚挚的谢意!