

基于最佳特征集肿瘤凝集素的分类研究

周才貌, 田明珠, 黄玲娃

海口经济学院, 海南 海口 570100

DOI:10.61369/MRP.2025120043

摘 要 : 肿瘤凝集素与癌症关联密切, 可应用于临床诊断、治疗、药物输送及癌症靶向领域。因此, 提升其分类准确性对疾病研究具有重要意义, 能为深入理解和攻克癌症提供关键支持。本文构建了一种基于机器学习的肿瘤凝集素计算方法 (ZTL_M), 该方法通过 monoDiKGap 提取特征, 经 F-score 筛选得到最优特征集, 再利用多层感知器分类器完成识别。实验采用 5 倍交叉验证, 结果显示 ZTL_M 对肿瘤凝集素的识别准确率达 96.3%, 使用 monoDiKGap 可以提高模型识别肿瘤凝集素的能力, ZTL_M 方法比一些现有方法具有更好的性能。

关 键 词 : 肿瘤凝集素; monoDiKGap 方法; 特征选择; 多层感知器

Classification Research of Tumor Lectins Based on the Best Feature Set

Zhou Caimao, Tian Mingzhu, Huang Lingwa

Haikou University of Economics, Haikou, Hainan 570100

Abstract : Tumor lectins are closely related to cancer and can be applied in clinical diagnosis, treatment, drug delivery and cancer targeting fields. Therefore, enhancing the accuracy of its classification is of great significance to disease research and can provide crucial support for a deeper understanding and the conquest of cancer. This paper constructs a machine learning-based tumor lectin calculation method (ZTL_M). This method extracts features through monoDiKGap, obtains the optimal feature collection through F-score screening, and then completes the recognition using a multi-layer perceptron classifier. The experiment adopted 5x cross-validation. The results showed that the recognition accuracy of ZTL_M for tumor lectins reached 96.3%. The use of monoDiKGap can enhance the model's ability to recognize tumor lectins. The ZTL_M method has better performance than some existing methods.

Keywords : tumor anthocyanin; monoDiKGap method; feature selection; multi-layer perceptron

引言

凝集素是一类能特异识别并结合糖复合物 (存在于蛋白质或脂质上) 的蛋白质 / 糖蛋白, 含多个同糖结合位点, 可参与细胞识别、黏着及细胞间连接^[1]。其广泛分布于植物、无脊椎动物及高等动物中, 功能具有多样性^[2]: 植物凝集素是植物防御系统核心成分, 助力植物保护^[3]; 动物凝集素可识别并结合体液或细胞表面糖配体, 传递细胞信号^[4], 还在细胞增殖分化、内环境稳定、免疫应答及心脑血管血栓相关疾病等过程中发挥关键作用^[5]。基于这些特性, 凝集素可用于细胞生物学、免疫学等领域的研究, 进而造福人类^[6]。

肿瘤凝集素作为凝集素的一类, 与癌症密切相关, 参与肿瘤细胞相互作用及黏附、生长表达、分化转移等过程^[7,8]。其已应用于实验室研究、临床诊疗、药物输送及癌症靶向^[9], 因此从海量凝集素中识别肿瘤凝集素基因, 对癌症研究与防治意义重大^[10]。传统实验测定蛋白质序列、结构及功能的方式耗时耗力, 基于现有数据构建机器学习模型, 实现癌性与非癌性物质的精准识别, 成为致癌物质预测的重要途径。

目前, 有相当少的算法来区分肿瘤凝集素和非肿瘤凝集。近几年来, Kumar 等人开发了一种基于支持向量机的方法来区分肿瘤凝集素和非肿瘤凝集素, 5 折交叉验证的准确率是 69.09%^[11]。Lin 等人使用方差分析选择最优特征集得到最终的分类准确率为 75.19%^[2]。Qian 等人基于 g-AP 的特征提取方式来识别肿瘤凝集素, 使用支持向量机对其分类, 10 折交叉验证的准确率是 83.9%^[12]。虽然前人的做法取得了可喜的成果, 对这一领域的发展起到了一定的推动作用, 但是预测精度还没有达到令人满意的效果, 有待进一步提高准确率。

本研究构建了 ZTL_M 模型, 用于识别蛋白质序列中的肿瘤凝集素。研究先通过 CC-PSSM、DR、monoDiKGap 三种方法提取蛋白质序列特征, 再用 F-score 算法筛选特征——选出与类别相关性强、自身冗余度低的最优特征集, 既缩短实验周期、降低时间成本, 又能提升分类准确率。随后, 将优化后的特征集分别搭配 LR、RF 和 MLP 三种分类器进行测试。结果显示, monoDiKGap 特征提取方法

项目信息: 海口经济学院校级科学项目: 基于最佳特征集的肿瘤凝集素分类研究, 项目编号: HJKY(YB)24-46。

作者简介: 周才貌 (1997.01—), 海南海口人, 研究生, 助教, 研究方向: 生物信息学。

与 MLP 分类器的组合表现最佳，在 5 倍交叉验证和独立测试中均取得高精度，充分验证了 ZTL_M 模型的可行性。图 1 完整呈现了该模型的构建全过程。

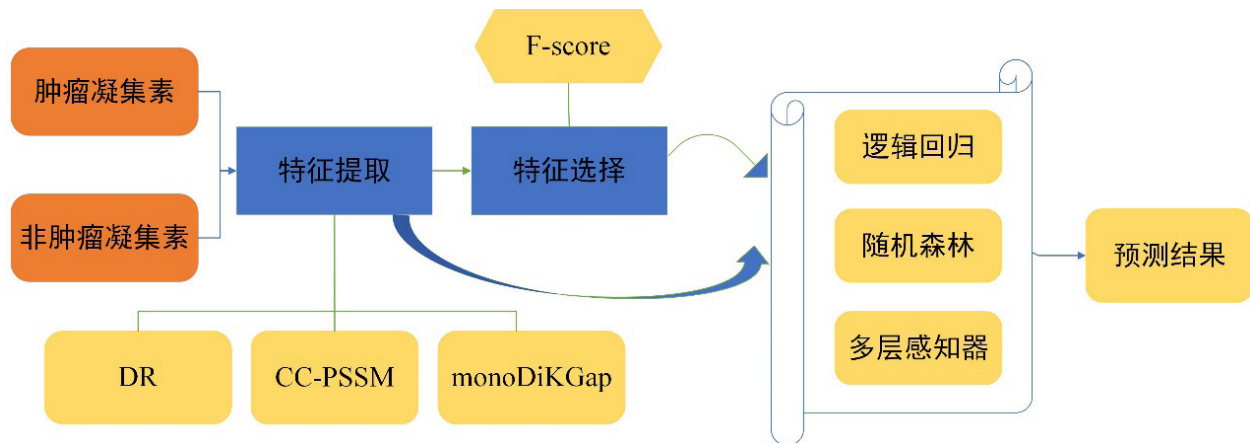


图 1 预测肿瘤凝集素的总体框架

一、材料和方法

（一）基准数据集

获取一个高质量的基准数据集，是建立可靠和准确模型的重要基础^[13]。本文的基准数据集来源于 CancerLectinDB 数据库，它是由 Song 和 Pan 等人提供的，这来自名为 CaledPred 的网络服务器。因此，我们得到了有关肿瘤凝集素的高质量数据，这用于建模和后续比较。它包含 178 个肿瘤凝集素和 226 个非肿瘤凝集素。我们从中随机选用 80% 的数据集采用 5 倍交叉验证来训练模型，剩余 20% 的数据集来验证我们的模型。

（二）提取特征方法

特征提取是生物序列分析的首要步骤。机器学习分类中，先将蛋白质序列转化为表征其特征的数值向量，再经特征选择（降维），最后选用合适分类方法，以优化分类效果。

1. monoDiKGap

MonoDiKGap 是基于氨基酸序列排列的特征提取方法^[14]，广泛用于蛋白质等生物分类预测，核心是捕捉相邻残基相互作用的有效特征。蛋白质氨基酸数量多易导致特征冗余引发维数灾难，还可能因信息丢失造成缺失，该方法通过 KGap 参数解决此问题。其适用范围涵盖蛋白质、DNA 和 RNA，DNA/RNA 的 KGap 取值 1-5，蛋白质为 1-10，且 KGap 值越大特征越多。以下以蛋白质序列为例，详细分析特征数量与样式随 KGap 值的变化：

当 KGap=1 时，此时的特征结构为 X_{XX}，得到了 20×20×20×1 维的特征向量。当 KGap=2 时，此时的特征结构为 X_{XX} 和 X_{_XX}，得到了 20×20×20×2 维的特征向量。当 KGap=3 时，此时的特征结构为 X_{XX}、X_{_XX} 和 X_{__XX}，得到了 20×20×20×3 维的特征向量。其中 X 是 20 种标准氨基酸：

$$X=\{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}$$

本文我们选取 KGap 设置为 3，monoDiKGap 方法提取出的全部特征集为 24000 维特征，最后通过 AdaBoost 模型生成最佳特征集为 337 维。

2. 特征选择

特征选择可克服模型过拟合，筛选最优关键特征以优化建模^[15]。多数降维算法侧重特征与目标类的相关性，却忽视特征冗余——高度相关特征无法叠加提升分类效果，还可能耗费资源。本文采用 F-score 方法（衡量特征类间区分度^[16]），通过特征评分筛选出维度精简、表达力最优的特征向量。

（三）机器学习分类算法

为了识别序列是否是肿瘤凝集素，把分类被视为二元分类问题^[17]。我们选取了几种常用的机器学习方法进行分类，分别是多层感知器 (MLP)、逻辑回归 (LR) 和随机森林 (RF)。

（四）模型评价标准

为了能够有效评估所提出方法的性能，在本文中我们采用评价指标分别是

敏感度 (Sn)，特异度 (Sp)，准确度 (ACC) 和精确度 (Precision)。对应的公式定义如 (4) 所示：

$$\begin{cases} Sn = \frac{Tp}{Tp + Fn} \\ Sp = \frac{Tn}{Tn + Fp} \\ Precision = \frac{Tp}{Tn + Fp} \\ ACC = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \end{cases} \quad (4)$$

Tp 是指肿瘤凝集素样本的准确预测。Fp 是指非肿瘤凝集素作

为肿瘤凝集素的预测。Tn是指非肿瘤凝集素的准确预测。Fn是指肿瘤凝集素作为非肿瘤凝集素的预测。

二、结果与讨论

（一）不同特征和分类器的性能比较

为找到更优的肿瘤凝集素识别方案，我们对比 DR、CC-PSSM、monoDiKGap 三种特征提取方法，搭配不同分类器通过 5 倍交叉验证测试效果。结果显示，DR 和 CC-PSSM 的分类准确率均未突破 0.7，识别表现不佳。而 monoDiKGap 提取的 337 维最佳特征子集，与 MLP 分类器结合时效果最优，Sn、Sp、Precision、ACC 各项指标分别达 0.951、0.966、0.954、0.957，准确率较 LR 高 0.016、较 RF 高 0.164。这表明 monoDiKGap 提取的特征更适配肿瘤凝集素识别，也印证了选择合适的特征提取方法对提升分类效果的关键作用。

表 1 不同特征提取方法与分类器的分类效果对比					
方法	分类算法	Sn	Sp	Precision	ACC
DR	LR	0.522	0.652	0.544	0.595
	RF	0.459	0.735	0.580	0.613
	MLP	0.564	0.658	0.571	0.616
CC-PSSM	LR	0.562	0.450	0.562	0.562
	RF	0.589	0.450	0.582	0.589
	MLP	0.572	0.442	0.571	0.572
monoDiKGap	LR	0.909	0.967	0.956	0.941
	RF	0.775	0.807	0.759	0.793
	MLP	0.951	0.966	0.954	0.957

（二）比较特征选择在不同分类器下的性能

为减少特征冗余、用更少特征实现精准分类，本文验证了特征选择对模型建立的重要性。我们采用 F-score 方法，对 DR、CC-PSSM、monoDiKGap 三种方法提取的特征进行降维，再搭配 LR、RF、MLP 分类器，通过 5 倍交叉验证对比性能。表 2 数据显示，降维后 monoDiKGap 仍是识别肿瘤凝集素的最佳特征，尤其搭配 MLP 分类器时表现最优。其经 F-score 降维后得到 230 维特征，数量精简且表现力强，分类准确率达 96.3%，性能优于 DR 和 CC-PSSM。实验表明，特征维数减少并未导致准确率下降，少而精的最优特征子集反而实现了最佳分类效果，进一步凸显了其对 ZTL_M 模型的重要性。

（三）与其他分类器的比较

为了进一步验证进行公平比较，我们在同一基准数据集上进行其他分类器的性能，分别是 SVM、NaiveBayes (NB) 和 Bagging。我们采用 monoDiKGap 提取的 337 个特征子集，经过 F-score 方法得到最优的 230 维特征向量，在 KNN、NB 和 Bagging 下进行性能评估。最终的预测结果列举在表 3 中。表 3 中的数据更加说明了最优特征子集在 MLP 分类器下具有很高的预测效

果，并能够准确区分非肿瘤凝集素和肿瘤凝集素。

（四）与其他先进方法的性能比较

为客观验证 ZTL_M 模型的性能优势，我们采用与现有方法一致的基准数据集，将其与四种现有方法进行性能对比。ZTL_M 模型基于 monoDiKGap 特征表示方法提取特征，搭配 MLP 分类器完成肿瘤凝集素识别。表 4 结果显示，该模型在各项核心指标中表现最优，ACC 达 96.3%、SN 为 97.2%、SP 为 95.6%，均优于其他对比方法。这表明 ZTL_M 模型能更高效地区分肿瘤凝集素与非肿瘤凝集素，同时印证了基于 monoDiKGap 特征的模型，在 5 倍交叉验证和独立测试中均具备较高精度，凸显了其性能优越性。

表 2 特性选择方法在不同分类器的性能比较						
方法	特征选择	分类算法	Sn	Sp	Precision	ACC
DR	F-score	LR	0.508	0.636	0.530	0.580
		RF	0.423	0.735	0.560	0.598
		MLP	0.557	0.686	0.590	0.629
CC-PSSM	F-score	LR	0.594	0.413	0.596	0.594
		RF	0.567	0.476	0.557	0.567
		MLP	0.614	0.407	0.611	0.614
mono-DiK-Gap	F-score	LR	0.930	0.945	0.933	0.938
		RF	0.747	0.790	0.737	0.771
		MLP	0.972	0.956	0.949	0.963

表 3 monoDiKGap 在其他分类器下的性能比较							
方法	特征选择	分类算法	Sn	Sp	Precision	MCC	ACC
mono-DiK-Gap	F-score	SVM	0.831	0.829	0.802	0.664	0.830
		NB	0.797	0.917	0.884	0.726	0.864
		Bagging	0.760	0.663	0.642	0.421	0.706
		MLP	0.972	0.956	0.949	0.927	0.963

表 4 与其他方法的性能比较			
方法	Sn (%)	Sp (%)	ACC (%)
Kumar et al.	68.0	69.9	69.1
Lin et al.	69.1	80.1	75.2
Hong-Yan Lai et al. ^[18]	75.3	80.5	77.5
Lili Qian et al.	83.2	80.9	83.9
TL_MP (this paper)	97.2	95.6	96.3

三、结论

总而言之，我们构建了基于机器学习的肿瘤凝集素计算模型 ZTL_M。该模型采用 monoDiKGap 蛋白质序列特征表示方法，搭配 MLP 分类器，初始便取得了 SN、SP 和 ACC。后续通过 F-score 特征选择方法降维后，模型预测性能进一步提升，且特征维数减少并未导致准确率下降，反而以更少特征实现了最优预

测效果。我们将 ZTL_M 与四种现有方法在相同基准数据集上对比, 经5倍交叉验证和独立测试验证, 其各项性能指标均优于对比方法, 充分证明 ZTL_M 模型有效提升了肿瘤凝集素的识别能力, 表现出显著优势。

参考文献

- [1]Sarhadi, V KArmengol G. Molecular biomarkers in cancer [J]. *Biomolecules*, 2022, 12(8): 1021.
- [2]Lin, H, Liu W X, He J, et al. Predicting cancerlectins by the optimal g-gap dipeptides [J]. *Scientific Reports*, 2015, 5(1): 16964.
- [3]Lavanya, V, Bommanabonia A K, Ahmed N, et al. Immunomodulatory effects of jacalin, a dietary plant lectin on the peripheral blood mononuclear cells (pbmcs) [J]. *Applied Biochemistry and Biotechnology*, 2022, 194(1): 587.
- [4]Shatz-Azoulay, H, Vinik Y, Isaac R, et al. The animal lectin galectin-8 promotes cytokine expression and metastatic tumor growth in mice [J]. *Scientific Reports*, 2020, 10(1): 7375.
- [5]Hernández, E, Sánchez-Maldonado C, Chávez M, et al. The therapeutic potential of galectin-1 and galectin-3 in the treatment of neurodegenerative diseases [J]. *Expert Review of Neurotherapeutics*, 2020, 20(1): 1.
- [6]Duan, L, Zangiabadi MZhao Y. Synthetic lectins for selective binding of glycoproteins in water [J]. *Chemical Communications*, 2020, 56(70).
- [7]Mane, V, Arakera S B, Pingle S, et al. Lectin as an anticancer therapeutic agent [M]. *Handbook of research on natural products and their bioactive compounds as cancer therapeutics*. IGI Global. 2022: 384.
- [8]Cummings, R D. The mannose receptor ligands and the macrophage glycome [J]. *Current Opinion in Structural Biology*, 2022, 75: 102394.
- [9]FREEMAN, H J. Role of lectins in gastrointestinal disorders [J]. *Herbs, Spices, and Medicinal Plants for Human Gastrointestinal Disorders: Health Benefits and Safety*, 2022.
- [10]Ali, F, Ghulam A, Maher Z A, et al. Deep-pcl: A deep learning model for prediction of cancerlectins and non cancerlectins using optimized integrated features [J]. *Chemo-metrics and Intelligent Laboratory Systems*, 2022, 221: 104484.
- [11]Kumar, R, Panwar B, Chauhan J S, et al. Analysis and prediction of cancerlectins using evolutionary and domain information [J]. *Bmc Research Notes*, 2011, 4(1): 237.
- [12]Qian, L, Wen YHan G. Identification of cancerlectins using support vector machines with fusion of g-gap dipeptide [J]. *Frontiers in Genetics*, 2020, 11.
- [13]Su, W, Liu M L, Yang Y H, et al. Ppd: A manually curated database for experimentally verified prokaryotic promoters [J]. *Journal of Molecular Biology*, 2021, 433(11): 166860.
- [14]Rafsanjani, Muhammod, Sajid, et al. Pyfeat: A python-based effective feature generation tool for DNA, rna, and protein sequences [J]. *Bioinformatics*, 2019.
- [15]Basith, S, Manavalan B, Shin T H, et al. Sdm6a: A web-based integrative machine-learning framework for predicting 6ma sites in the rice genome [J]. *Molecular Therapy Nucleic Acids*, 2019.
- [16]Hua, T, Ya-Wei Z, Ping Z, et al. Hbpred: A tool to identify growth hormone-binding proteins [J]. *Int J Biol*, 2018, 14(8): 957.
- [17]Ding, Y, Tang JGuo F. Identification of drug - target interactions via fuzzy bipartite local model [J]. *Neural Computing and Applications*, 2020, 32(D1): 1.
- [18]Hong-Yan, Lai, Xin-Xin, et al. Sequence-based predictive modeling to identify cancerlectins [J]. *Oncotarget*, 2017.