doi: 10.3969/j.issn.0490-6756.2020.01.022

基于支持向量机的癌细胞经典分泌蛋白与非经典分泌蛋白识别研究

余乐正[1,2], 柳凤娟[1], 李东海[1], 郭延芝[2], 李益洲[2]
(1. 贵州师范学院化学与材料学院, 贵阳 550018; 2. 四川大学化学学院, 成都 610065)

摘　要: 基于支持向量机算法,本文提出了一种能快速准确区分癌细胞经典分泌蛋白与非经典分泌蛋白的方法.通过严格的特征筛选,氨基酸组成、位置特异性得分矩阵和信号肽组成了最优特征集. 测试集检测结果表明,本方法对癌细胞经典分泌蛋白与非经典分泌蛋白具有较强的区分能力,可为寻找到不同种类癌症间通用的生物标志物提供理论参考.

关键词: 支持向量机; 癌症; 非经典分泌蛋白; 位置特异性得分矩阵; 信号肽
中图分类号: O604　　**文献标识码**: A　　**文章编号**: 0490-6756(2020)01-0152-05

A study on recognition of classically and non-classically secreted proteins from cancer cells based on support vector machine

YU Le-Zheng[1,2], *LIU Feng-Juan*[1], *LI Dong-Hai*[1], *GUO Yan-Zhi*[2], *LI Yi-Zhou*[2]
(1. School of Chemistry and Materials Science, Guizhou Education University, Guiyang 550018, China;
2. College of Chemistry, Sichuan University, Chengdu 610065, China)

Abstract: Based on support vector machine (SVM) algorithm, a fast and accurate method is proposed to distinguish the classically and non-classically secreted proteins from cancer cells. By a strict feature selection, the optimal feature set is obtained which consists of amino acid composition (AAC), position specificity score matrix (PSSM) and signal peptide (SP). The test results show that our method has strong ability to distinguish the non-classically secreted proteins (NCSPs) from the classically secreted proteins (CSPs) of cancer cells, which may provide theoretical reference for finding common biomarkers among different kinds of cancers.

Keywords: Support vector machine; Cancer; Non-classically secreted protein; Position specific scoring matrix; Signal peptide

1　引　言

恶性肿瘤(癌症)是当今对人类健康和生命威胁最大的疾病之一,并已成为我国人口死亡的首要原因[1]. 由于具有发展速度快、侵袭性强、易转移复发、预后差等特点,大多数癌症在晚期才被发现,导致治疗难度大,死亡率极高. 现代医学研究结果表明,癌症越早被发现,其治愈的几率就越高. 因此,实现对早期癌症的有效检测已成为治愈癌症、延长患者生命的关键[2]. 在癌症的发生发展过程

收稿日期: 2018-10-08
基金项目: 国家科技部和国家自然科学基金奖励补助资金(黔科合平台人才[2017]5790-07); 贵州省普通本科高等学校青年科技人才成长项目(黔教合 KY 字[2016]219); 贵州省科学技术基金一般项目(黔科合 J 字[2014]2134 号)
作者简介: 余乐正(1984-), 男, 四川井研人, 博士, 副教授, 研究方向为生物信息学、化学计量学. E-mail: xinyan_scu@126.com
通讯作者: 柳凤娟. E-mail: morose1984@163.com

中，肿瘤细胞会释放出一类反映癌症存在与生长的物质——肿瘤标志物. 肿瘤标志物可存在于血液、体液、细胞或组织中，主要包括 RNA，DNA，蛋白质等生物活性分子[3]. 通过对该类物质的快速准确检测，可为判断是否患有癌症、癌症类别、癌症分期、预后效果等提供实验依据. 由于不同发展阶段、不同种类的癌细胞分泌出的蛋白质类型和表达水平不尽相同，近年来分泌蛋白已成为肿瘤标志物的主要来源之一[4-9]. 例如，甲胎蛋白(AFP)、α-L-岩藻糖苷酶(AFU)、高尔基体蛋白 73(GP73)等已成为肝癌临床诊断的主要检测指标[10]，前列腺特异性抗原(PSA)则是前列腺癌最重要的早期检测指标[11].

根据是否含有 N 端信号肽，分泌蛋白可简单分为经典分泌蛋白(CSPs)和非经典分泌蛋白(NCSPs)两大类[12]. 通过经典分泌途径与非经典分泌途径，蛋白质均可被释放到癌细胞外，并参与癌细胞的相关生理过程. 已有研究证实，不同种类的癌细胞可分泌出相同的蛋白质，且这些蛋白质的分泌主要依赖于非经典分泌途径[13]. 因此，对癌细胞非经典分泌蛋白进行系统深入的研究，可为寻找到不同种类癌症间通用的肿瘤标志物提供理论参考. 基于蛋白质序列信息和支持向量机(SVM)算法，通过严格的特征筛选，本文构建了一个二元分类器以快速准确地识别癌细胞非经典分泌蛋白. 对于测试集，本方法总的预测准确率为 99.81%，表明本方法可作为一种辅助工具用于不同种类癌症间通用蛋白标志物的筛选.

2　材料与方法

2.1　材　料

本实验所用数据主要来自于人类癌症分泌蛋白质组数据库(HCSD)[14]. HCSD 已收录 13 种癌症的分泌蛋白数据，如肝癌、肺癌、乳腺癌、前列腺癌、胃癌、结直肠癌、鼻咽癌、宫颈癌、胶质母细胞瘤、膀胱癌、胰腺癌、卵巢癌、淋巴瘤等. 从该数据库中共得到 23 225 条癌细胞分泌蛋白，包括 5 263 条 CSPs 与 17 962 条 NCSPs. 此外，从前期工作中[8]，收集到 147 条 CSPs 与 102 条 NCSPs 作为独立测试集.

2.2　建模方法

作为现今最流行的机器学习算法之一，支持向量机已被广泛应用于解决各种分类问题. 由于采用了结构风险最小化准则，并具有坚实的理论支撑，支持向量机可较好地处理小样本、高维度、非线性、局部极小点等问题[15]. 在前期各类分泌蛋白的识别研究中[16-18]，支持向量机均表现出良好的应用效果，故本文也采用支持向量机来构建预测模型.

2.3　模型的性能评估参数

为客观准确地评估模型的实际预测性能，本文选取了以下 4 个评价参数：灵敏度(SE)，特异性(SP)，准确率(ACC)和马氏相关系数(MCC)[19].

$$SE=\frac{TP}{TP+FN} \tag{1}$$

$$SP=\frac{TN}{TN+FP} \tag{2}$$

$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \tag{4}$$

公式(1)～(4)中，TP 为真阳性，即正样本被准确识别的数量；FP 表示假阳性，即负样本被错误识别为正样本的数量；TN 表示真阴性，即负样本被准确识别的数量；FN 表示假阴性，即正样本被错误识别为负样本的数量.

3　实　验

3.1　训练集与测试集

为去除掉原始数据中冗余的序列信息，提高模型的稳定性，以相似度阈值为 25%，利用 CD-HIT Suite[20] 对原始数据进行处理后，共得到 761 条 CSPs 和 2 715 条 NCSPs. 随机提取其中的 70%作为训练集，剩余的 30%作为测试集[21]，故训练集最终由 533 条 CSPs 和 1 901 条 NCSPs 组成，而测试集则包含 228 条 CSPs 及 814 条 NCSPs.

3.2　特征提取与表征

除所用实验数据与建模方法外，特征筛选在蛋白质的分类预测研究中也发挥着非常重要的作用. 本研究分别采用氨基酸组成、自协方差变量、位置特异性得分矩阵以及信号肽来表征蛋白质中氨基酸的序列信息、邻接效应、进化信息及结构信息.

3.2.1　氨基酸组成　氨基酸组成(AAC)代表了

20 种常见氨基酸在蛋白质序列中出现的频率,每条蛋白质均被描述为一个 20 维的数字向量.

3.2.2 *自协方差变量* 在蛋白质的分类研究中,自互协方差(ACC)常用于计算蛋白质序列中氨基酸残基间的邻接效应. 自互协方差共包含两种变量,即相同描述符间产生的自协方差变量(AC)与不同描述符间形成的互协方差变量(CC). 由于自协方差变量的维数远小于互协方差变量的,且前者对邻接效应的贡献度远大于后者[22],故本文只采用自协方差变量来表征氨基酸残基间的邻接效应. 此外,前面的研究工作[23]已对自协方差变量的相关计算公式进行了详细描述,此处不再赘述. 由于本研究选用了疏水性、等电点、极性、转移自由能、侧链体积等 5 个理化性质,且氨基酸间的最大距离取值为 5,故每条蛋白质最终被转化为一个 25 维的数字向量.

3.2.3 *位置特异性得分矩阵* 由于能有效表征蛋白质序列中氨基酸残基的进化信息[24],位置特异性得分矩阵(PSSM)已被广泛应用于各种蛋白质的分类研究. 利用 PSI-BLAST 程序(期望值阈值为 10^{-3})对 Swiss-Prot 数据库进行搜索,并经 3 次迭代后,获得了每条蛋白质的位置特异性得分矩阵. 通过相关公式[23]对这些矩阵进行统一处理后,每条蛋白质均被转换为一个 20 维的数字向量.

3.2.4 *信号肽* 是否含有 N 端信号肽是经典分泌蛋白与非经典分泌蛋白结构间最显著的差异,故信号肽已成为区分两者的一个重要特征. 作为目前预测能力最强、应用范围最广的信号肽识别软件,SignalP 4.1[25]被用于蛋白质 N 端信号肽的识别,并通过 *D-score* 值予以表征.

3.3 蛋白质替代模型

基于上述特征,本文共建立了 7 个蛋白质替代模型:模型 1 仅含氨基酸组成(AAC);模型 2 仅含位置特异性得分矩阵(PSSM);模型 3 为氨基酸组成与自协方差变量融合形成的伪氨基酸组成(PseAAC);模型 4 为氨基酸组成与位置特异性得分矩阵融合形成的伪位置特异性得分矩阵(PsePSSM);模型 5 由氨基酸组成与信号肽融合而成;模型 6 由伪氨基酸组成与信号肽融合而成;模型 7 由伪位置特异性得分矩阵与信号肽融合而成.

3.4 模型的构建

本文最终的支持向量机预测模型是通过 libsvm 3.12 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/)工具箱建立起来的. 选择径向基函数(RBF)为模型核函数,并利用网格搜索法对模型的正则化参数 C 和核函数参数 γ 进行优化. 此外,作为最客观的模型性能检测方法之一[26],留一法(Jackknife test)被用于构建最终的预测模型.

4 结果与讨论

4.1 特征筛选及替代模型的确定

基于 3.3 节描述的 7 个蛋白质替代模型,本文共构建了 7 个支持向量机预测模型,相关训练结果均列于表 1 中.

表 1 不同蛋白质替代模型对训练结果的影响

Tab.1 Performance of different protein substitution models

模型	C	γ	准确率
模型 1	8.0	0.5	85.209 5
模型 2	8.0	0.5	88.783 9
模型 3	2.0	0.5	87.921 1
模型 4	2.0	0.5	91.166 8
模型 5	0.5	0.031 25	99.752 5
模型 6	32	0.007 812 5	99.671 3
模型 7	2.0	0.5	99.671 3

根据模型 1 与模型 2 的训练结果,PSSM 对蛋白质的表征能力略优于 AAC,表明 PSSM 的确能较好地反映蛋白质序列中氨基酸残基的进化信息. 模型 3、模型 4 的训练结果表明,AC 和 PSSM 的加入的确能有效提高模型的预测性能,且 PSSM 所包含的信息量多于 AC 的. 比较前 4 个模型与后 3 个模型的训练结果,信号肽的加入使得模型 5~7 的预测性能均有较大幅度的提升,表明信号肽在 CSP 与 NCSP 的分类研究中的确发挥着重要作用. 同时,正是由于信号肽对 CSP 和 NCSP 过于强大的区分能力,使其掩盖了蛋白质替代模型 PseAAC 与 PsePSSM 之间的性能差异. 虽然模型 5 的预测准确率最高,但模型 7 的优化参数最为合理,包含的信息量更多,且两者之间的预测准确率相差很小,故本文选择模型 7 作为最终的蛋白质替代模型.

4.2 模型的实际应用

利用 3.1 节构建的测试集,对模型 5~7 的实际预测性能进行了比较,相关测试结果均列于表 2 中.

表 2　不同 SVM 模型对测试集的预测结果

Tab. 2　Prediction results of different SVM models obtained by analyzing the test sets

蛋白质类型	CSPs	NCSPs	合计
测试集数据	228	814	1 042
模型 5			
准确预测数	228	805	1 033
准确率（%）	100	98.98	99.14
模型 6			
准确预测数	228	807	1 035
准确率（%）	100	99.14	99.33
模型 7			
准确预测数	226	814	1 040
准确率（%）	99.12	100	99.81

如表 2 所示，虽然模型 5、模型 6 准确识别出所有 228 条 CSPs，但它们对 NCSPs 的预测性能均弱于模型 7. 模型 7 不仅准确识别出测试集中所有 814 条 NCSPs，其对癌细胞分泌蛋白总的预测准确率与 MCC 值也最高（99.81%与 99.44%），表明以模型 7 为最终的蛋白质替代模型是正确的.

为进一步比较模型 5～7 的实际预测性能，通过 2.1 节提到的独立测试集再次进行了检测. 模型 5～7 均准确识别出所有 147 条 CSPs，且模型 5 将 2 条 NCSPs 错误预测为 CSPs，而模型 6 和模型 7 仅错误预测 1 条 NCSP. 进一步的研究发现，三个模型均错误预测的蛋白质（Q86UK5）在 UniProt 数据库中被标注为膜蛋白，SignalP 4.1 也预测其为膜蛋白. 由于该蛋白质的 *D-score* 值为 0.438，与 SignalP 4.1 的默认值（0.45）极为接近，这可能使得三个预测器均将其错误识别为 CSP. 这一结果表明在区分经典分泌蛋白和非经典分泌蛋白时，还应注意区分分泌蛋白与膜蛋白.

5　结　论

经仔细分析癌细胞经典分泌蛋白与非经典分泌蛋白的各种特征，本文基于支持向量机算法构建了一个二元分类器以快速准确地识别癌细胞非经典分泌蛋白. 研究结果表明，本方法对癌细胞非经典分泌蛋白具有较好的预测性能，可作为一种辅助工具用于筛选不同种类癌症间通用的蛋白标志物. 后续研究将尝试构建一个可快速准确区分不同种类癌细胞分泌蛋白的多元分类预测器，从而为寻找到每类癌症的特异性肿瘤标志物提供理论参考.

参考文献：

[1] Chen W, Zheng R, Baade P D, *et al*. Cancer statistics in China, 2015 [J]. CA Cancer J Clin, 2016, 66: 115.

[2] Chen W, Sun K, Zheng R, *et al*. Cancer incidence and mortality in China, 2014 [J]. Chin J Cancer Res 2018, 30: 1.

[3] Paul D, Kumar A, Gajbhiye A, *et al*. Mass spectrometry-based proteomics in molecular diagnostics: discovery of cancer biomarkers using tissue culture [J]. Biomed Res Int, 2013, 2013: 783131.

[4] Makridakis M, Vlahou A. Secretome proteomics for discovery of cancer biomarkers [J]. J Proteomics, 2010, 73: 2291.

[5] Coghlin C, Murray G I. Progress in the development of protein biomarkers of oesophageal and gastric cancers [J]. Proteomics Clin Appl, 2016, 10: 532.

[6] Huang Y, Zhu H. Protein array-based approaches for biomarker discovery in cancer [J]. Genomics Proteomics Bioinformatics, 2017, 15: 73.

[7] Zamay T N, Zamay G S, Kolovskaya O S, *et al*. Current and prospective protein biomarkers of lung cancer [J]. Cancers, 2017, 9: 155.

[8] 余乐正，柳凤娟，吴正雨，等. 分泌蛋白质组学在肿瘤标志物中的研究进展[J]. 生物技术通报，2017，33：12.

[9] Lomnytska M, Pinto R, Becker S, *et al*. Platelet protein biomarker panel for ovarian cancer diagnosis [J]. Biomarker Res, 2018, 6: 2.

[10] 郝磊，郝坤. 血清 AFP、AFU、CEA、GP73 及糖链抗原系列联合检测对于早期原发性肝癌的诊断价值[J]. 实用癌症杂志，2017，32：1609.

[11] Cohen J D, Li L, Wang Y, *et al*. Detection and localization of surgically resectable cancers with a multi-analyte blood test [J]. Science, 2018, 359: 926.

[12] Bendtsen J D, Jensen L J, Blom N, *et al*. Feature-based prediction of non-classical and leaderless protein secretion [J]. Protein Eng Des Sel, 2004, 17: 349.

[13] Villarreal L, Méndez O, Salvans C, *et al*. Unconventional secretion is a major contributor of cancer cell line secretomes [J]. Mol Cell Proteomics, 2013, 12: 1046.

[14] Feizi A, Banaei-Esfahani A, Nielsen J. HCSD: the human cancer secretome database [J]. Database, 2015, 2015: bav051.

[15] 彭继慎，于精哲，夏乃钦. 基于支持向量机的传感器的非线性校正[J]. 计算机测量与控制，2011，

19: 243.

[16] Yu L, Guo Y, Zhang Z, *et al*. SecretP: a new method for predicting mammalian secreted proteins [J]. Peptides, 2010, 31: 574.

[17] Yu L, Guo Y, Li Y, *et al*. SecretP: identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition [J]. J Theor Biol, 2010, 267: 1.

[18] Yu L, Luo J, Guo Y, *et al*. In silico identification of Gram-negative bacterial secreted proteins from primary sequence [J]. Comput Biol Med, 2013, 43: 1177.

[19] Matthews B W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme [J]. Biochim Biophys Acta, 1975, 405: 442.

[20] Huang Y, Niu B, Gao Y, *et al*. CD-HIT suite: a web server for clustering and comparing biological sequences [J]. Bioinformatics, 2010, 26: 680.

[21] Luo J, Yu L, Guo Y, *et al*. Functional classification of secreted proteins by position specific scoring matrix and auto covariance [J]. Chemometr Intell Lab, 2012, 110: 163.

[22] Guo Y, Yu L, Wen Z, *et al*. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences [J]. Nucleic Acids Res, 2008, 36: 3025.

[23] 余乐正, 赵柳青, 陈曼, 等. 基于 SVM 的革兰氏阴性菌分泌系统蛋白识别方法[J]. 四川大学学报: 自然科学版, 2016, 53: 443.

[24] Yu B, Li S, Qiu W, *et al*. Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction [J]. BMC Genomics, 2018, 19: 478.

[25] Nielsen, H. Predicting secretory proteins withSignalP [J]. Methods Mol Biol, 2017, 1611: 59.

[26] Chou K C, Zhang C T. Prediction of protein structural classes [J]. Crit Rev Biochem Mol Biol, 1995, 30: 275.

引用本文格式:

中 文: 余乐正,柳凤娟,李东海,等. 基于支持向量机的癌细胞经典分泌蛋白与非经典分泌蛋白识别研究[J]. 四川大学学报: 自然科学版, 2020, 57: 152.

英 文: Yu L Z, Liu F J, Li D H, *et al*. A study on recognition of classically and non-classically secreted proteins from cancer cells based on support vector [J]. J Sichuan Univ: Nat Sci Ed, 2020, 57: 152.