改进CKSAAP结合RFE算法预测蛋白质棕榈酰化位点

汤亚东[1],谢 鹭[2],陈兰明[1]
1.上海海洋大学 食品科学与技术学院,上海 201306
2.上海生物信息技术研究中心,上海 201203

摘 要:蛋白质棕榈酰化是一种可逆的蛋白质翻译后修饰,在蛋白质稳定性和亚细胞定位等方面发挥重要作用。构建了一种预测蛋白质棕榈酰化位点的新模型(PSSM-CKSAAP-RFE)。采用蕴含进化信息的*k*-spaced氨基酸对组分方法表征蛋白质序列,通过递归特征消除法进行特征选择;基于上述特征训练支持向量机分类器,并采用夹克刀交叉验证法测试模型性能。研究结果显示,训练集和独立测试集的预测准确率、马修斯相关系数、特异性、敏感性和受试者工作特征曲线下面积分别为98.44%、0.94、98.95%、95.65%和0.990,以及98.41%、0.93、99.39%、92.31%和0.994,优于文献中报道的相关方法,为蛋白质棕榈酰化位点的预测提供了一种新模型。

关键词:蛋白质棕榈酰化位点;*k*-spaced氨基酸对组分;位置特异性得分矩阵;支持向量机;递归特征消除
文献标志码:A **中图分类号:**TP181 doi:10.3778/j.issn.1002-8331.1710-0297

汤亚东,谢鹭,陈兰明.改进CKSAAP结合RFE算法预测蛋白质棕榈酰化位点.计算机工程与应用,2019,55(5):143-148.
TANG Yadong, XIE Lu, CHEN Lanming. Identification of palmitoylation sites of proteins using modified CKSAAP combined with RFE method. Computer Engineering and Applications, 2019, 55(5):143-148.

Identification of Palmitoylation Sites of Proteins Using Modified CKSAAP Combined with RFE Method

TANG Yadong[1], XIE Lu[2], CHEN Lanming[1]
1.College of Food Science and Technology, Shanghai Ocean University, Shanghai 201306, China
2.Shanghai Center for Bioinformation Technology, Shanghai 201203, China

Abstract: Protein palmitoylation is reversible post-translational modification and plays important roles in protein stability, subcellular localization and many other functions. In this study, a new model to identify palmitoylation sites is constructed, designated as PSSM-CKSAAP-RFE. The evolutionary information of amino acid residues involved in tested proteins is represented by a Composition of *k*-Spaced Amino Acid Pairs(CKSAAP) method. Optional features are selected using a Recursive Feature Elimination(RFE) method. The Support Vector Machine(SVM) classifier is trained using the chosen features, and the performance of the model is examined using a Jackknife Cross Validation Test(JCVT). The resulting data shows that the values of accuracy, Matthews correlation coefficient, specificity, sensitivity and area under receiver operating characteristic curves(AUC) for the identification of palmitoylation sites are 98. 44%, 0.94, 98.95%, 95.65% and 0.990, as well as 98.41%, 0.93, 99.39%, 92.31% and 0.994 for the train dataset and test dataset, respectively, which are superior to previous methods in the literature. This study provides a new model for the identification of palmitoylation sites of proteins.

Key words: protein palmitoylation sites; composition of k-spaced amino acid pairs; position specific scoring matrix; support vector machine; recursive feature elimination

1 引言

蛋白质棕榈酰化修饰发生在真核细胞中蛋白质翻译后[1],即氨基酸序列中的半胱氨酸残基通过硫酯键与16碳饱和脂肪酸共价结合[2]。蛋白质通过棕榈酰化调

基金项目:国家自然科学基金面上项目(No.31671946)。
作者简介:汤亚东(1991—),男,硕士研究生,研究领域为生物信息学;谢鹭,女,博士,教授;陈兰明,女,博士,教授,E-mail:lmchen@shou.edu.cn。
收稿日期:2017-10-30 **修回日期:**2018-01-03 **文章编号:**1002-8331(2019)05-0143-06
CNKI **网络出版:**2018-04-20, http://kns.cnki.net/kcms/detail/11.2127.TP.20180420.1540.010.html

节蛋白质相互作用,及其运输、聚集和降解等[2-3],参与信号转导、有丝分裂和细胞凋亡等[4]。近年来,研究发现蛋白质棕榈酰化与很多疾病相关,例如结肠直肠癌、非小细胞肺癌、干细胞癌等[3],并可用作心血管疾病的生化标记物[5]。

虽然蛋白质组学和成像技术的发展加快了蛋白质棕榈酰化位点的鉴定,但是仅仅依靠实验技术鉴定修饰位点存在费时、费力等缺点,满足不了海量蛋白质序列棕榈酰化位点预测的实际需求。因此,开发有效、快速预测蛋白质棕榈酰化位点的生物信息学方法非常必要。迄今为止,国内外仅有少数预测蛋白质棕榈酰化位点的计算模型:2006年,Zhou等人首次应用聚类和评分策略(Clustering and Scoring Strategy,CSS)构建了蛋白质棕榈酰化位点预测模型CSS-Palm 1.0[6]。同年,Xue等人基于朴素贝叶斯(Naive Bayes)算法构建了计算模型NBA-Palm[7]。2008年,Ren等人通过改进的CSS算法将CSS-Palm 1.0升级为CSS-Palm 2.0,模型的预测性能有了明显的提高[8]。2009年,Wang等人采用*k*-spaced氨基酸对组分方法(a Composition of *k*-Spaced Amino Acid Pairs,CKSAAP)表征蛋白质序列,建立了CKSAAP-Palm预测模型[4]。2011年,Hu等人提出了基于氨基酸序列特征的预测算法IFS-Palm[3]。2013年,Shi等人采用多特征提取方法构建了WAP-Palm预测模型[2]。2014年,Kumari等人基于支持向量机(Support Vector Machine,SVM)分类器构建了PalmPred预测模型[1]。这些预测模型的马修斯相关系数和敏感性最高分别为0.71和79.23%[1]。因此,蛋白质棕榈酰化位点的预测性能仍有很大的提升空间。

本研究基于蛋白质序列的PSSM,采用CKSAAP方法表征蛋白质序列。PSSM反映了蛋白质序列的进化信息,CKSAAP方法反映了氨基酸对组分信息和局部"序"信息。本研究综合考虑了以上两种信息建立了新的预测模型PSSM-CKSAAP-RFE。基于训练集和测试集,显著提高了蛋白质棕榈酰化位点预测的准确率、特异性、敏感性和马修斯相关系数。

2 材料和方法

2.1 数据集

本研究利用Hu等人[3]构建的数据集HL151,该数据集用于构建IFS-Palm[3]和PalmPred[1]等预测模型,含有151条棕榈酰化蛋白质序列,共1 537个半胱氨酸残基。其中,实验证实的棕榈酰化位点234个,非棕榈酰化的半胱氨酸残基位点1 303个。本研究对数据集HL151中每一条蛋白质序列执行PSI-BLAST程序[4],其中一条蛋白质序列没有比对结果,故将此序列删除。此序列包括1个棕榈酰化位点,不包含非棕榈酰化位点。为了进一步验证新模型的性能,HL151数据集被分为训练集和独立测试集[1]。训练集中包括132条蛋白质序列,含有207个棕榈酰化位点和1 140个非棕榈酰化位点。独立数据集中包括18条蛋白质序列,含有26个棕榈酰化位点和163个非棕榈酰化位点。

2.2 蛋白质序列的特征表示模型

2.2.1 *k*-spaced氨基酸对组分

每一个棕榈酰化位点可定义为以半胱氨酸残基为中心的肽段,其上游和下游分别包含n个氨基酸残基,肽段的长度称为窗口大小(window size),即$(2\times n+1)$。本研究利用Chen等人[9]的CKSAAP方法表征肽段序列;参考Tung等人[10]的方法,选取的窗口大小范围为11~31。从蛋白质序列中截取肽段,如果半胱氨酸残基一边的氨基酸个数少于n而出现空位,本研究用"X"填补空位。窗口长度确定后,k-spaced氨基酸对用如下形式表示:

$$p_i\{k\}p_j\ \ (i,j=1,2,\cdots,21)$$

其中p_i和p_j表示20种氨基酸和"X"中的任意一个,k表示p_i和p_j中间的氨基酸残基个数。因此,p_i和p_j表示哪种氨基酸均有21种可能。如果k确定,$p_i\{k\}p_j$表示一个有421种(21×21)可能的二肽。k值大小决定了构建模型的特征向量维数。为了保证构建模型的计算量不至过大,并且降低特征向量的冗余度[9],本研究设定k=0,1,2,3,4,因此,每一个肽段都可以用2 105维(421×5)特征向量表示。

2.2.2 基于PSSM的*k*-spaced氨基酸对组分

本研究运用PSI-BLAST程序[11]比对NCBI(National Center for Biotechnology Information)中的非冗余蛋白质数据库NR(Non-Redundant)得到待测蛋白质序列的PSSM矩阵。E值、循环比对次数分别设为0.001和3,其余参数均为默认值。每一条长度为L的蛋白质序列得到一个对应的$L\times 20$维的PSSM矩阵。每一个肽段对应的矩阵从蛋白质序列对应的PSSM矩阵中提取得到。如果肽段中半胱氨酸残基一边的氨基酸个数少于n,提取PSSM矩阵时会出现空位,本研究用"0"填补空位。为了避免数据间因差异过大导致预测性能下降,本研究把PSSM矩阵中的元素映射到(0,1)之间,映射公式为$f(x)=1/(1+e^{-x})$[10],其中x表示PSSM矩阵中的原始值。

为了分析氨基酸之间"序"信息对模型性能的影响,基于PSSM矩阵提取肽段的*k*-spaced特征。改进的*k*-spaced特征定义如下:

$$f_{i,j,k}=\sum_{s=1}^{L-k-1}p_{s,i}\times p_{s+k+1}\ \ (1\leqslant i,j\leqslant 20)$$

式中,L表示蛋白质序列的长度;k表示氨基酸i和j之间的间隔,当k=0时,即两个氨基酸相邻;s表示蛋白质序列中氨基酸的下标,$s=1$表示从第一个氨基酸开

始。最后，用三维矩阵 $\left[f_{i,j,k}\right]$ 中提取自PSSM矩阵的元素表示查询序列。当 k=0，1，2，3，4时，每条肽段都将对应一个400×(k+1)维的特征向量。

2.3　支持向量机与递归特征消除

蛋白质棕榈酰化位点的预测属于二分类问题，SVM通过核函数将二维数据映射到高维空间，并找到最优超平面(hyper-plane)最大程度的将两类数据分开。本研究采用SVM工具(LIBSVM)[12]构建分类模型，并使用径向基函数(Radial Basis Function，RBF)作为核函数。其中，参数 C 和 γ 通过grid[12]方法获得，C 决定了超平面的平滑度，γ 决定了数据映射到高维空间后的分布情况。RBF公式[12]如下：

$$K(P_p,P_n)=\mathrm{e}^{\left(-\gamma-\gamma\|P_p-P_n\|^2\right)}$$

式中，P_p 和 P_n 分别表示蛋白质棕榈酰化和非棕榈酰化位点，γ 为RBF自带参数。

本研究采用Guyon等人[13]的递归特征消除法(Recursive Feature Elimination，RFE)进行特征选择，该方法是用于“降维”的有效方法。RFE算法基于SVM训练时产生的权向量生成排序系数，每次删除一个特征，并将其余特征组成一个新的训练集，依次循环，最终得到特征排列顺序表[14]。本研究基于RFE算法对特征向量进行排序，选取最优 K 维特征子集构建分类模型，K 表示上述排序列表中最前面的特征维数。

2.4　交叉验证与性能评价

本研究参考Si等人[15]的方法，以准确率(Accuracy，Acc)、Matthews相关系数(MCC)、敏感性(Sensitivity，S_n)、特异性(Specificity，S_p)和受试者工作特征曲线下面积(AUC)[15]为评价指标，定义如下：

$$S_n=\frac{TP}{TP+FN}$$

$$S_p=\frac{TN}{TN+TP}$$

$$Acc=\frac{TP+TN}{TP+TN+FP+FN}$$

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

式中，TP、TN、FP、FN 分别表示真阳性(true positive)、真阴性(true negative)、假阳性(false positive)、假阴性(false negative)的样本数。迄今为止，MCC和AUC是精确评估预测模型整体性能的最重要的两个指标[16]。MCC和AUC的值越大，表示预测模型的整体性能越好。

3　结果与分析

3.1　新模型基于训练集预测性能的评估

鉴于窗口大小的选择直接影响预测性能，本研究首先分析了不同窗口大小对新模型(PSSM-CKSAAP)预测性能的影响。夹克刀交叉验证法(Jackknife Cross Validation Test，JCVT)检测结果显示，当窗口大小为19时，MCC值最大(0.76)(图1)。因此，本研究确定窗口大小为19。基于此窗口大小的肽段对应2 000维(400×5)特征。采用RFE算法选取最优的特征子集，基于 K=[10，20，30，…，600]维特征子集，以SVM为分类器，采用JCVT检测其预测性能，结果如图2所示。当 K=370时，新模型的预测准确率最高达到98.66%。因此，本研究采用的窗口大小为19，选取前370维最优特征子集构建蛋白质棕榈酰化位点鉴定模型(PSSM-CKSAAP-RFE)。

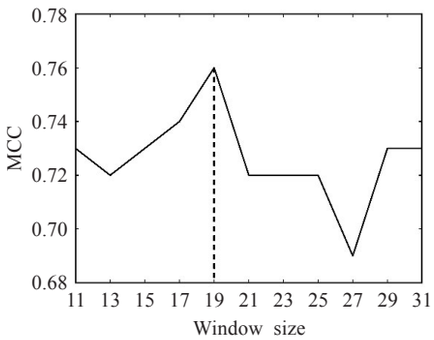

图1　不同窗口大小对PSSM-CKSAAP预测性能的影响

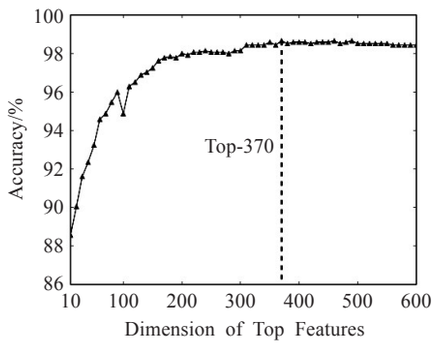

图2　训练集最优特征子集对模型性能的影响

本研究基于训练集比较了CKSAAP、PSSM-CKSAAP和PSSM-CKSAAP-RFE方法的预测性能。采用CKSAAP方法，将不同肽段转化为相应的特征向量，最终每个肽段均用2 105(421×5)维特征向量表示。以SVM为分类器，通过JCVT检测其预测性能。结果显示，当窗口大小为27时，MCC值最大(0.67)(图3)。因此，对于CKSAAP方法，确定的窗口大小为27。如表1所示，PSSM-CKSAAP方法优于CKSAAP方法的预测性能，而PSSM-CKSAAP-RFE方法取得了最优的预测效果，其预测准确率、Matthews相关系数、特异性、敏感性和AUC分别比CKSAAP方法提高了约7、30、1、30和8个百分点(表1)。三种方法的ROC曲线如图4所示，其中PSSM-CKSAAP-RFE方法的AUC值最大(0.990)，其次为PSSM-CKSAAP(0.938)，CKSAAP方法的AUC值最小(0.909)，表明PSSM-CKSAAP-RFE模型的预测性能最佳。

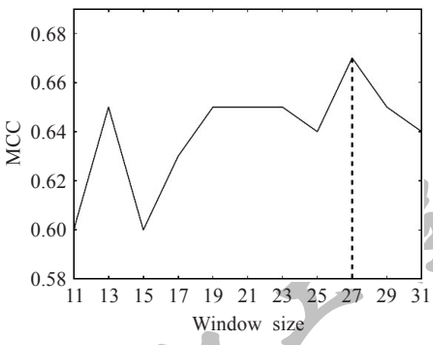

图3 不同窗口大小对CKSAAP预测性能的影响

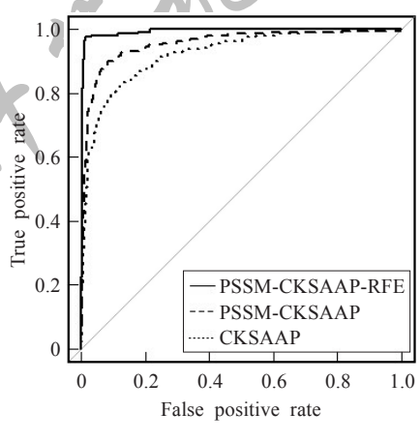

图4 基于训练集三种方法的ROC曲线

3.2 新模型基于独立测试集的预测性能的评估

为了进一步检测新模型的预测性能,本研究也采用了独立测试集进行检测。如表2所示,与训练集一致,PSSM-CKSAAP-RFE方法取得了最优的预测结果,其预测准确率、MCC、特异性、敏感性和AUC分别比CKSAAP方法提高了约6、30、1、35和16个百分点(表2)。PSSM-CKSAAP-RFE方法的预测敏感性(92.31%)也明显优于PSSM-CKSAAP方法(61.54%)。三种方法在独立测试集上的ROC曲线如图5所示。

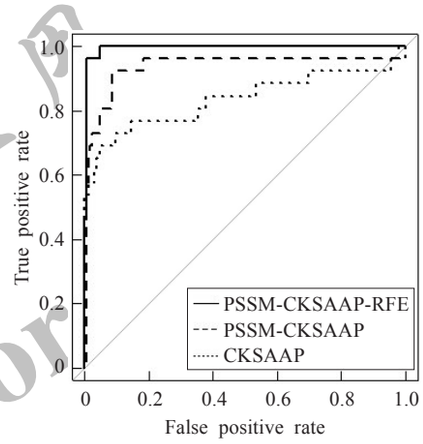

图5 基于测试集三种方法的ROC曲线

3.3 新模型与现有方法的比较

基于相同的数据集和JCVT验证方法,本研究比较了新模型(PSSM-CKSAAP-RFE)与文献中报道模型的预测性能。结果显示,基于训练集,新模型PSSM-CKSAAP-RFE预测性能明显优于IFS-Palm[3]和PalmPred[1]模型。IFS-Palm仅基于蛋白质序列提取特征,采用增量特征选择方法选择最优特征并构建棕榈酰化位点预测模型。PalmPred仅基于蛋白质序列的PSSM提取特征,并构建预测模型,反应了PSSM特征用于预测棕榈酰化位点的有效性。与其中预测性能较好的PalmPred相比较,新模型的预测准确率、Matthews相关系数、特异性和敏感性分别提高了约7、23、4和15个百分点,新模型的AUC为0.990(表3)。新模型整体性能的显著提高表明,包含了蛋白质序列的进化信息、氨基酸对组分和"序"信息的改进CKSAAP组合特征比其中任何一种单一特征对棕榈酰化位点更有鉴别作用。

基于测试集,与其他三种模型比较。结果显示,新模型的预测性能明显优于模型CKSAAP-Palm,整体上优于模型IFS-Palm,略低于模型PalmPred。基于测试

表1 基于训练集不同方法预测性能比较

预测算法	准确率 Acc/%	Matthews相关系数 MCC	特异性 S_p/%	敏感性 S_n/%	曲线下面积 AUC
CKSAAP	91.98	0.67	97.54	61.35	0.909
PSSM-CKSAAP	94.06	0.76	97.63	74.40	0.938
PSSM-CKSAAP-RFE	98.44	0.94	98.95	95.65	0.990

表2 基于测试集不同方法预测性能比较

预测算法	准确率 Acc/%	Matthews相关系数 MCC	特异性 S_p/%	敏感性 S_n/%	曲线下面积 AUC
CKSAAP	92.59	0.65	98.16	57.69	0.834
PSSM-CKSAAP	93.12	0.68	98.16	61.54	0.907
PSSM-CKSAAP-RFE	98.41	0.93	99.39	92.31	0.994

表3 不同预测方法基于训练集的预测结果比较

预测算法	准确率 Acc/%	Matthews相关系数 MCC	特异性 S_p/%	敏感性 S_n/%	曲线下面积 AUC
IFS-Palm[3]	90.65	0.64	94.65	68.60	—
PalmPred[1]	91.98	0.71	94.30	79.23	—
PSSM-CKSAAP-RFE	98.44	0.94	98.95	95.65	0.990

注:—表示文献中没有提供该数据。

表4　不同预测方法基于测试集的预测结果比较

预测算法	准确率 Acc/%	Matthews相关系数 MCC	特异性 S_p/%	敏感性 S_n/%	曲线下面积 AUC
CKSAAP-Palm[4]	83.16	0.43	86.50	62.96	—
IFS-Palm[3]	97.89	0.91	98.77	92.59	—
PalmPred[1]	98.42	0.94	98.77	96.30	—
PSSM-CKSAAP-RFE	98.41	0.93	99.39	92.31	0.994

注：—表示文献中没有提供该数据。

集，新模型的AUC为0.994（表4）。与模型PalmPred比较，新模型的预测准确率和MCC基本相同，但是特异性提高了约0.6个百分点，而敏感性则降低了4个百分点。

4　讨论

传统的CKSAAP方法由Chen等人[9]于2007年建立，用于预测蛋白质灵活化区域。鉴于该方法考虑了蛋白质序列中氨基酸之间的相互作用信息，因此在蛋白质表位预测（epitope prediction）中应用广泛[4,10,17-20]，例如蛹化位点（pupylation sites）、柔性/刚性区域（flexible/rigid region）、O-糖基化位点（O-glycosylation sites）、泛素化位点（uiquitination sites）、棕榈酰化位点、甲基化位点和磷酸化位点的预测，并取得了良好的效果。同样，PSSM的成功应用表明进化信息提供了比蛋白质序列本身更多的信息[21]。然而这两种方法在预测过程中，仅仅考虑了蛋白质序列的一种特征。本研究结合PSSM和CKSAAP构建了一种预测蛋白质棕榈酰化位点的新模型，工作流程图如图6所示。蛋白质棕榈酰化在细胞的动力学过程和信号通路中起着重要的作用。棕榈酰化不仅能够提高蛋白质的疏水性从而促进蛋白质与膜的结合[6]，而且可以调控细胞内的转运、蛋白质-蛋白质相互作用以及蛋白质的活性等[3]。

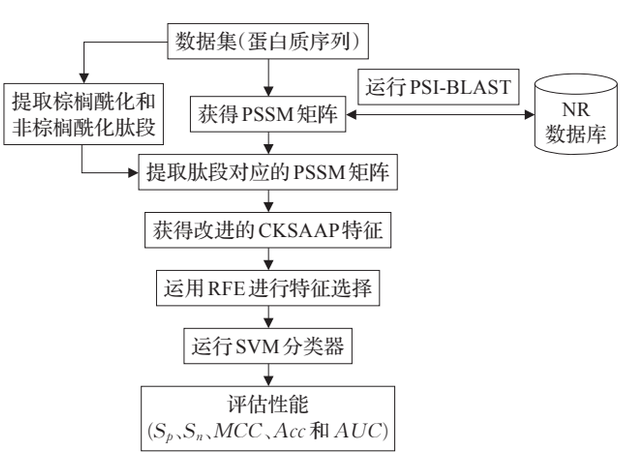

图6　PSSM-CKSAAP-RFE模型的工作流程图

JCVT广泛应用于模型预测性能的评估[22]。在本研究中，对于每一个数据集，该方法每次选取一个样本作测试，其余的样本作训练，依次循环，直至每个样本都做过测试，最后得到所有结果的均值。

新模型PSSM-CKSAAP-RFE基于最优特征子集和SVM，在训练集、测试集中的JCTV的检验结果显示，预测蛋白质棕榈酰化位点的准确率、敏感性、特异性、MCC和AUC分别为98.44%、95.65%、98.95%、0.94和0.990以及98.41%、92.31%、99.39%、0.93和0.994。与文献中已报道的预测模型比较，新模型在训练集上的预测效果显著提高，MCC和敏感性提升尤为明显，分别提高了23、15个百分点。MCC是精确评估模型整体性能的重要指标之一，敏感性则反应了模型的预测准确率，这两个指标的显著提高，反应了模型性能的整体提升。可能原因在于：(1)PSSM蕴含了蛋白质序列的进化信息，CKSAAP反应了氨基酸之间的相互作用信息，基于PSSM的CKSAAP特征结合了两种信息，对蛋白质序列中蕴含的信息进行了更深入的挖掘。(2)采用RFE法选择最优特征子集，进一步提高了模型的预测性能。在测试集上的预测效果优于大多数模型，略低于模型PalmPred，可能原因在于：训练集中包含的数据量较少。训练集和测试集中分别包含132和18条蛋白质序列，训练集的数据量明显大于测试集。与其他模型比较，新模型基于的数据量越大，越能体现其优越性。数据量越大，蛋白质序列中蕴含的内在作用信息越能被充分挖掘。因此，新模型在训练集上性能显著提高，在测试集上虽取得了较好的预测效果，但没有体现其优越性。综上表明，本研究为蛋白质棕榈酰化位点的预测提供了可行、可靠的技术支撑。

5　结论

本研究基于蛋白质序列预测蛋白质棕榈酰化位点，取得了良好的效果，证明了新方法的可行性、有效性，补充和发展了现有的蛋白质棕榈酰化位点的预测方法。

参考文献：

[1] Raghava G P S，Kumari B，Kumar R，et al.PalmPred：an SVM based palmitoylation prediction method using sequence profile information[J].PLoS One，2014，9：e89246.

[2] Shi S P，Sun X Y，Qiu J D，et al.The prediction of palmitoylation site locations using a multiple feature extraction method[J].Journal of Molecular Graphics & Modelling，2013，40：125-130.

[3] Hu L L，Wan S B，Niu S，et al.Prediction and analysis of protein palmitoylation sites[J].Biochimie，2011，93：

489-496.
[4] Wang X B,Wu L Y,Wang Y C,et al.Prediction of palmitoylation sites using the composition of *k*-spaced amino acid pairs[J].Protein Engineering,Design & Selection,2009,22:707-712.
[5] Ferri N,Paoletti R,Corsini A,Lipid-modified proteins as biomarkers for cardiovascular disease:a review[J].Biomarkers:Biochemical Indicators of Exposure,Response,and Susceptibility to Chemicals,2005,10:219-237.
[6] Zhou F,Xue Y,Yao X,et al.CSS-Palm:palmitoylation site prediction with a Clustering and Scoring Strategy (CSS)[J].Bioinformatics,2006,22:894-896.
[7] Xue Y,Chen H,Jin C,et al.NBA-Palm:prediction of palmitoylation site implemented in Naive Bayes algorithm[J].BMC Bioinformatics,2006,7:458.
[8] Ren J,Wen L,Gao X,et al.CSS-Palm 2.0:an updated software for palmitoylation sites prediction[J].Protein Engineering,Design & Selection,2008,21:639-644.
[9] Chen K,Kurgan L A,Ruan J,Prediction of flexible/rigid regions from protein sequences using *k*-spaced amino acid pairs[J].BMC Structural Biology,2007,7:25.
[10] Tung C W.Prediction of pupylation sites using the composition of *k*-spaced amino acid pairs[J].Journal of Theoretical Biology,2013,336:11-17.
[11] Altschul S F,Madden T L,Schaffer A A,et al.Gapped blast and PSI-BLAST:a new generation of protein database search programs[J].Nucleic Acids Research,1997,17(25):3389-3402.
[12] Chang C C,Lin C J.Libsvm:a library for supporet vector machines[J].ACM Transactions on Intelligent Systems and Technology,2011,2(3):1-27.
[13] Guyon I,Jason W,Stephen B.Gene selection for cancer classification using support vector machines[J].Machine Learning,2002,46:389-422.
[14] 李烨,王永丽,贺国平.基于支持向量机的结肠癌信息基因提取[J].山东科技大学学报(自然科学版),2012,31(3):84-89.
[15] Si J,Zhao R,Wu R,An overview of the prediction of protein DNA-binding sites[J].International Journal of Molecular Sciences,2015,16:5194-5215.
[16] Zou Q,Xie S,Lin Z,et al.Finding the best classification threshold in imbalanced classification[J].Big Data Research,2016,5:2-8.
[17] Chen Z,Chen Y Z,Wang X F,et al.Prediction of ubiquitination sites by using the composition of *k*-spaced amino acid pairs[J].PLoS One,2011,6:e22930.
[18] Hasan M M,Zhou Y,Lu X,et al.Computational identification of protein pupylation sites by using profile-based composition of *k*-spaced amino acid pairs[J].PLoS One,2015,10:e0129635.
[19] Bui V M,Weng S L,Lu C T,et al.SOHSite:incorporating evolutionary information and physicochemical properties to identify protein S-sulfenylation sites[J].BMC Genomics,2012,17:1-9.
[20] Zangooei M H,Jalili S,Protein secondary structure prediction using DWKF based on SVR-NSGAII[J].Neurocomputing,2012,94:87-101.
[21] Liu T,Qin Y,Wang Y,et al.Prediction of protein structural class based on gapped-dipeptides and a recursive feature selection approach[J].International Journal of Molecular Sciences,2015,17.
[22] Meher P K,Sahu T K,Banchariya A,et al.DIRProt:a computational approach for discriminating insecticide resistant proteins from non-resistant proteins[J].BMC Bioinformatics,2017,18:190.

(上接第7页)

[11] Ainen I K A,Anti P F.Dynamic local search algorithm for the clustering problem[R].Joensuu,Finland:University of Joensuu.Department of Computer Science,2002:6-8.
[12] Mitchel A.The ESRI guide to GIS analysis,volume 2:Spartial measurements and statistics[M].[S.l.]:Esri Press,2005:88-93.
[13] Lu G Y,Wong D W.An adaptive inverse-distance weighting spatial interpolation technique[J].Computers & Geosciences,2008,34(9):1044-1055.
[14] 张文元,谈国新,朱相舟.停留点空间聚类在景区热点分析中的应用[J].计算机工程与应用,2018,54(4):263-270.
[15] Gionis A,Mannila H,Tsaparas P.Clustering aggregation[J].ACM Transactions on Knowledge Discovery from Data,2007,1(1):1-30.
[16] Zahn C T.Graph-theoretical methods for detecting and describing gestalt clusters[J].IEEE Trans on Computers,1971:68-86.
[17] Veenman C J,Reinders M J T,Backer E.A maximum variance cluster algorithm[J].IEEE Transactions on Pattern Analysis & Machine Intelligence,2002,24(9):1273-1280.
[18] 中国科学院遥感与数字地球研究所.SatSee-Fire[EB/OL].[2018-10-22].http://satsee.radi.ac.cn:8080/index.html.
[19] 冯少荣,肖文俊.DBSCAN聚类算法的研究与改进[J].中国矿业大学学报,2008(1):105-111.