Computer Science and Application 计算机科学与应用, 2021, 11(3), 710-719
Published Online March 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.113073

Hans汉斯

使用核SVM和分割PSSM预测凋亡蛋白亚细胞位置

夏新男

云南大学，信息学院，云南 昆明
Email: xiaxinnan1@163.com

收稿日期：2021年2月25日；录用日期：2021年3月19日；发布日期：2021年3月30日

摘　要

凋亡蛋白与人类的一些疾病密切相关。准确的获得凋亡蛋白的亚细胞位置对理解疾病的发病机制和药物研发有至关重要的作用。目前，研究者们主要是通过蛋白质序列获取特征信息，从而对蛋白质亚细胞位置进行预测定位并获得了较好的结果。在本文中，我们首先改进了PSSM特征提取方法，对PSSM按行分块以获得凋亡蛋白序列的局部信息，我们称之为SePSSM，其次加入7种物理化学性质对氨基酸分类获取凋亡蛋白序列的全局信息。最终将得到的两种特征融合输入到使用不同核函数的SVM中进行预测定位，预测结果通过Jackknife检验得到。实验结果表明，对PSSM进行分割要优于无分隔，RBF核函数要优于其他核函数，融合特征在ZD98和ZW225数据集上获得了较好的效果，这表明我们的方法是有效的。

关键词

凋亡蛋白，PSSM，分割，物理化学性质，核SVM

Predicting Subcellular Localization of Apoptotic Proteins Using Kernel Svm and Segmentation Pssm Method

Xinnan Xia

Department of Computer Science and Engineering, Yunnan University, Kunming Yunnan
Email: xiaxinnan1@163.com

Received: Feb. 25th, 2021; accepted: Mar. 19th, 2021; published: Mar. 30th, 2021

文章引用: 夏新男. 使用核 SVM 和分割 PSSM 预测凋亡蛋白亚细胞位置[J]. 计算机科学与应用, 2021, 11(3): 710-719.
DOI: 10.12677/csa.2021.113073

Abstract

Apoptosis proteins are closely related to some human diseases. Accurate identification of the subcellular location of apoptosis proteins is crucial for understanding the pathogenesis of diseases and drug development. At present, researchers mainly obtain feature information from protein sequences to predict the subcellular location of proteins and obtain good results. In this paper, we first improved the feature extraction method of PSSM, segmented the PSSM matrix by row to obtain the local information of the apoptotic protein sequence, which is called SePSSM. Secondly, seven physicochemical properties were added to classify amino acids to obtain the global information of apoptotic protein sequence. Finally, the obtained two features are fused and input into SVM using different kernel functions for prediction, and the prediction results were obtained by Jackknife test. The experimental results show that PSSM method with segmentation is better than that without segmentation, the RBF kernel function is better than other kernel functions, and the fusion feature has achieved better results on the ZD98 and ZW225 datasets, which shows that our method is effective.

Keywords

Apoptosis Proteins, PSSM, Segmentation, Physicochemical Properties, Kernel SVM

1. 引言

细胞凋亡在生物体的生长发育及新陈代谢中起着非常重要的作用，而凋亡过程的紊乱可能与许多疾病如肿瘤、自身免疫性疾病的发生有直接或间接的关系[1] [2] [3] [4]。凋亡蛋白是指与细胞凋亡有关的蛋白质。研究表明，凋亡蛋白的功能与亚细胞位置密切相关[5]。因此对凋亡蛋白亚细胞位置的正确定位，能帮助我们理解凋亡蛋白功能、细胞凋亡机制和药物开发。然而，通过传统的生物实验方法来确定凋亡蛋白的位置既费时又费力，难以满足现在的科研需求[6]。因此，研究者开始借助计算机及其相关知识开发了许多有效且可靠的计算方法来替代或协助传统生物实验。

近年来，大量机器学习方法被开发用于识别不同的凋亡蛋白亚细胞位置，通常包括三个步骤：第一，从凋亡蛋白序列中提取包含不同种类蛋白质的信息作为凋亡蛋白亚细胞定位的特征向量，如信息增益(Increment of Diversity) [7]、位置特异性评分矩阵(Position Specific Scoring Matrix, PSSM) [8]、伪氨基酸组成(Pseudo Amino Acid Composition, PseAAC) [9]、氨基酸组成(Amino Acid Composition, AAC)和二肽组成(Dipeptide Composition) [10] [11]。近年来，随着特征提取方法的增多，以及计算机不断健壮的计算性能，许多研究者开始将多个特征进行融合以改进特征提取方法，如 Zhang 等人[12]通过将 Moran 自相关和互相关与 PSSM 集成在一起，提出了一种称为 MACC-PSSM 的新模型。又如刘等人[13]将氨基酸组成、二肽组成和自相关系数相结合来构建凋亡蛋白的特征表达模型。第二，将得到的特征向量输入到分类器中进行预测分类，在凋亡蛋白亚细胞定位中使用的分类器有协变判别函数法[5]、模糊 k-近邻[14]、支持向量机(Support Vector Machin, SVM) [15] [16]、集成分类器[17]等。第三，通过 Jackknife 检验、K 折交叉验证和独立集检验对分类器性能进行评估，以证明所提出方法的可靠性[18] [19] [20]。这些计算方法的使用可以大大加快凋亡蛋白亚细胞位置的研究。这些方法都是基于序列提取得到的特征，好的特征提取方法

对预测凋亡蛋白亚细胞位置是至关重要的，它能帮助我们提高预测准确率。

在本文中，为了能够更加准确的对凋亡蛋白亚细胞位置进行定位，我们考虑了凋亡蛋白序列的进化信息和序列信息。我们首先从序列中获取含有进化信息的 PSSM，然后以一个分割比例将 PSSM 矩阵按行分割为两个子矩阵，并以此构建一个新的特征，我们称之为分割 PSSM (Segmentation PSSM, SePSSM)。接下来我们对凋亡蛋白序列引入 7 种物化性质，并将此方法得到的特征与 SePSSM 特征进行线性融合。最后，我们将融合后的特征输入到含有四种不同核函数的支持向量机中，并通过 Jackknife 检验验证该方法的有效性。

2. 方法

2.1. 数据集

本研究使用了前人构建的两个基准数据集。由 Zhou 和 Doctor [5]构建的 ZD98 数据集包含 43 个细胞质蛋白(cytoplasm proteins, cy)、30 个膜蛋白(membrane proteins, me)、13 个线粒体蛋白(mitochondrial proteins, mi)和 12 个其他蛋白(other proteins, oth)。第二个数据集是由 Zhang 等人构建的 ZW225 [15]数据集，被分为 4 个亚细胞位置，包含 70 个细胞质蛋白(cytoplasm proteins, cy)、89 个膜蛋白(membrane proteins, me)、25 个线粒体蛋白(mitochondrial proteins, mi)和 41 个细胞核蛋白(nucleoid proteins, nu)。这两个数据集中的所有蛋白质序列均从 SWISS-PROT 中提取(http://www.ebi.ac.uk/swissprot/)得到[21]。虽然两个数据集的数量较少，但在以往的研究中被广泛使用。

2.2. 特征提取方法

2.2.1. 从 PSSM 中获取分割 PSSM(SePSSM)信息

首先我们使用 POSSUM 网页服务器(https://possum.erc.monash.edu/) [22]生成 PSSM，BLAST 程序中参数选择为 uniref50 数据库，3 次迭代，E-value 值为 0.001，得到一个 $L \times 20$ 的矩阵。

$$P_{PSSM} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,20} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ p_{L,1} & p_{L,2} & \cdots & p_{L,20} \end{bmatrix} \tag{1}$$

其中，L 表示蛋白质序列的长度，$p_{i,j}$ 表示蛋白质序列中氨基酸的进化信息。

接下来，为了从 PSSM 中获取更多重要信息，我们使用数学中的矩阵分块思想[23]，对 PSSM 矩阵进行按行分块，分割比例为 λ。因此一个 $L \times 20$ 的 PSSM 矩阵就被分割为 $L \times \lambda \times 20$ 和 $(L - L \times \lambda) \times 20$ 的两个子矩阵。由于只需将 PSSM 分割为两个子矩阵，因此对于分割比例 λ 的选取为 0 到 1 之间的任意数(0 和 1 都表示未分割)，在实验中每次实验按比例增加 0.1，因此 $\lambda \in [0.1, 0.9]$)。

最后，我们基于 PsePSSM [24]的部分思想，根据公式(2)分别计算被分割后的两个矩阵每列的平均值，以得到每一条凋亡蛋白序列对于 20 种氨基酸的平均突变概率。

$$\overline{P_{n,j}} = \frac{1}{L_n} \sum_{i=1}^{L_n} p_{i,j} \left(j = 1, 2, \cdots, 20; n = 1, 2 \right) \tag{2}$$

其中，$L_n \left(n = 1, 2 \right)$ 为每个子矩阵的行数，即 $L_1 = \lfloor L \times \lambda \rfloor, L_2 = \lceil L - L \times \lambda \rceil, \left(\lambda \in [0.1, 0.9] \right)$。

最终我们得到一个 40 维的向量，我们将之称为分割 PSSM(SePSSM)。

$$\text{SePSSM} = \left(\overline{P_{1,1}}, \overline{P_{1,2}}, \cdots, \overline{P_{1,20}}, \overline{P_{2,1}}, \overline{P_{2,2}}, \cdots, \overline{P_{2,20}} \right)^{\mathrm{T}} \tag{3}$$

为了更好的理解 SePSSM，我们以一条长度为 1020 的凋亡蛋白为例，则将得到一个 1020 × 20 的 PSSM 矩阵，下图为分割示意图如图 1 所示。根据这种对 PSSM 矩阵的分割和处理，我们将得到 40 维的 SePSSM

特征。

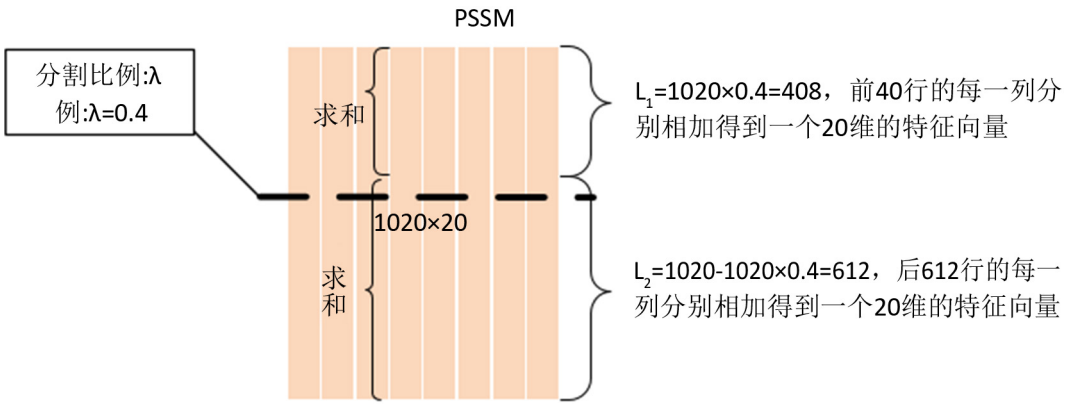

Figure 1. The segmentation diagram of SePSSM
图 1. SePSSM 分割图

2.2.2. 氨基酸的物理化学性质

氨基酸(AAs)的性质是由它们的侧链决定的，而这些侧链在形状、电荷和疏水性方面有所不同。因此，蛋白质可能具有不同的结构特征和生理功能。我们使用了原子吸收光谱的七种物理化学性质，包括归一化范德华体积、极性、溶剂可及性、变化、极化性、表面张力和二级结构[25]。根据每种物理化学性质，20 种氨基酸又被分为三组。这 7 种理化性质及其氨基酸的划分如表 1 所示。

Table 1. 7 kinds of physical and chemical properties and their division
表 1. 7 种物理化学性质及其划分

ID	物理化学性质	第一类	第二类	第三类
1	归一化范德华体积	GASCTPD	NVEQIL	MHKFRYW
2	极性	LIFWCMVY	PATGS	HQRKNED
3	溶剂可及性	ALFCGIVW	RKQEND	MPSTHY
4	变化	KR	ANCQGHILMFPSTWYV	DE
5	极化性	GASDT	CPNVEQIL	KMHFRYW
6	表面张力	GQDNAHR	KTSEC	ILMFPWYV
7	二级结构	EALMQKRH	VIYCWFT	GNPSD

我们分别计算每一种物理化学性质的各类中所包含的氨基酸在蛋白质序列中出现的频率，因此每种物理化学性质将得到一个 3 维的特征向量。通过这七种物理化学性质，每一条凋亡蛋白序列我们将得到一个 7 × 3 = 21 维的特征向量。

2.3. 分类算法

提取特征后，可以使用各种分类算法来实现凋亡蛋白预测。在常用分类器中，由 Vapnik 提出的支持向量机表现出了很好的效果[26]。支持向量机(Support Vector Machine)是一种按监督学习方式对数据进行二元分类的广义线性分类器，核心原理是找到一个分类超平面，以最大化正样本和负样本之间的距离，且利用核函数使它们在高维空间中线性可分离[27]。SVM 最初是为二分类设计的，而蛋白质亚细胞位置预测通常是个多类分类问题。但可以通过一对剩余(One vs Rest, OVR)和一对一(One vs One, OVO)等策略

来实现 SVM 的多类分类，本文采用了前者 OVR 策略。由于选取不同的核函数，预测结果也会不同的。因此，本文将分别对四种常用核函数：径向基核函数(RBF)、线性核函数(linear)、sigmoid 核函数(sigmoid)和和多项式核函数(poly)进行实验，以寻找最佳的核函数进行分类预测。

2.4. 评价方法

在统计学中，常见的三种检验方法为：自检验，Jackknife 检验和独立集检验。Jackknife 检验能得到唯一的预测结果，被认为是最客观、最严格的检验方法，且被广泛用于评价蛋白质亚细胞定位等领域中的预测性能[28]。因此，在本文中，我们使用 Jackknife 检验方法和 5 种评价指标用于检验和评价我们所提出的预测凋亡蛋白亚细胞位置的预测模型方法。这 5 种评价方法是：敏感性(Sen)、F1 值、马修相关系数(MCC)、准确率(ACC)和总体准确率(OA) [29]。公式如下：

$$Sen_i = \frac{TP_i}{TP_i + FN_i} \tag{4}$$

$$F1_i = \frac{2 \times Pre_i \times Sen_i}{Pre_i + Sen_i} \tag{5}$$

$$MCC_i = \frac{TP_i \times TN_i - FP_i \times FN_i}{\sqrt{(TP_i + FP_i)(TP_i + FN_i)(FP_i + TN_i)(TN_i + FN_i)}} \tag{6}$$

$$ACC_i = \frac{TP_i + TN_i}{TP_i + FN_i + FP_i + TN_i} \tag{7}$$

$$OA = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} (TP_i + FN_i)} \quad (i = 1, 2, \cdots, c) \tag{8}$$

TP_i 表示属于第 i 类的样本被正确分到该类的数量，FN_i 表示属于第 i 类的样本没有正确分类到该类的数量，TN_i 表示被正确的分类到非 i 类的样本数，FP_i 表示非 i 类的样本错误的分类为 i 类的数量。Sen_i 表示数据集中第 i 类的预测准确率。$F1_i$ 表示对数据集中第 i 类的鲁棒性的度量，它是敏感性(Sen_i)和精度($Pre_i = TP_i/(TP_i + FP_i)$)的调和平均值，可以避免对某些指标的性能估计过高。MCC_i 综合了不同的参数，能从整体上评价预测算法性能。ACC_i 表示第 i 类在所有样本中预测正确的比率。OA_i 表示一个综合指标，它反映了总体的预测准确率，其中 c 为数据集中的类别数。以上 5 个指标的取值范围都在 0 到 1 之间，指标值越接近 1，表示分类性能越好，指标值越接近 0，表示分类器性能越差。

本文所提出的预测模型的流程图如图 2 所示，预测模型的详细描述如下：

- 获取序列。通过文献[5] [15]获得 ZW225 和 ZD98 两个数据集的氨基酸组成序列。
- 获得特征。通过 POSSUM 网页服务器获得两个数据集的 PSSM 矩阵，然后通过 SePSSM 方法(2.2.1 节)从 PSSM 中获得 40 维的 SePSSM 特征，之后再与 7 种物化性质(2.2.2 节)得到的 21 维特征进行融合，得到最终的 61 维特征。
- 预测分类。将得到的 61 维最终特征输入到含有不同核函数(RBF、poly、linear 和 sigmoid)的 SVM 中进行预测分类，并通过 Jackknife 检验得到最终的实验结果。

3. 结果和分析

3.1. SePSSM 特征的分割比例

由于 SePSSM 特征是通过分割 PSSM 矩阵而得到的，而选择不同的分割比例 λ 会得到不同的预测结果。为了得到最佳的结果，我们先对 SePSSM 特征进行分割比例的选择，每个分割比例在两个数据集上

的总体准确率如下表 2。为了证明分割的有效性，我们加入了不分割的实验结果，当 λ 取 1 时，表示 PSSM 矩阵未进行分割。

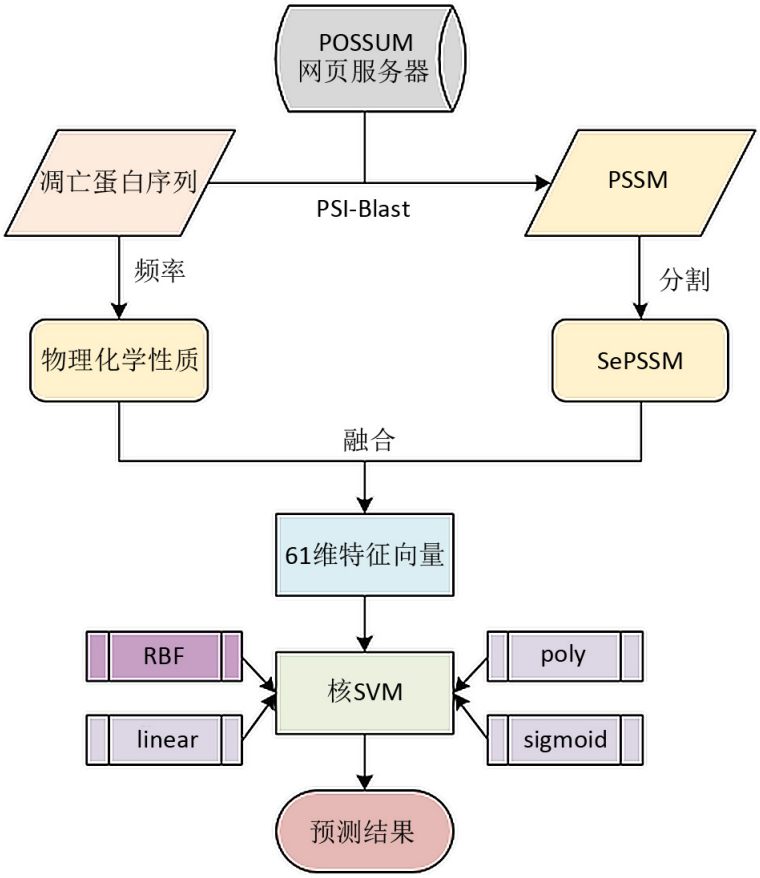

Figure 2. Flowchart of the proposed method
图 2. 提出的方法的框架

Table 2. The OA(%) results of the two datasets at different segmentation ratios
表 2. 两个数据集在不同分割比例的 OA(%)结果

λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
ZW225	77.2	77.7	79.5	80.8	81.7	75.9	75.4	77.2	79.9	72.8
ZD98	85.7	86.7	90.8	90.8	92.9	92.9	91.8	91.8	91.8	87.8

从表 2 我们可以看出，随着λ的增大，ZW225 数据集的 OA 值一直在波动，在 λ 为 0.5 时 SePSSM 得到最好值 81.7%；而 ZD98 数据集的结果先增大后减小，保持不变之后又减小，在 λ 为 0.5 和 0.6 时 SePSSM 结果最好，最好值为 92.9%。且从结果中可以看出，对 PSSM 矩阵进行适当的分割要比未分割的结果好，这是因为分割之后将数据细化，获得了更多有用的特征信息，而这些信息更能对每个类进行区分，从而有利于进行预测分类。当分割比例为 0.5 时在 ZW225 和 ZD98 两个数据集上得到的结果要比未分割或其他分割比例的结果好。因此，本文中的分割比例取 0.5。

3.2. 不同特征的性能

为了验证我们所提出的 SePSSM 的性能，我们计算了 SePSSM、PsePSSM [24]、7 种理化性质(PhyChe)

和融合特征分别在 ZW225 和 ZD98 两个数据集上的总体准确率，所有结果在同一条件下通过 Jackknife 检验得到，结果如下表 3。

Table 3. The OA(%) of different feature methods on two datasets
表 3. 不同特征方法在两个数据集上的总体准确率(%)

数据集	PsePSSM	SePSSM	PhyChe	SePSSM+ PhyChe
ZW225	74.1	81.7	80.4	94.6
ZD98	91.8	92.9	68.4	96.9

由于 PsePSSM 中含有一个决定特征维度的参数，在序列较长时该参数的选择将会耗费较多的计算时间。而从表 3 可以看出，SePSSM 在 PSSM 上进行分割就能达到一个较好的预测效果，且在 ZW225 数据集上比 PsePSSM 高 7.6%。这表明我们提出的 SePSSM 的性能要优于 PsePSSM，且计算方法省去了 PsePSSM 计算其他特征的时间，这将更加简便。将 SePSSM 和 7 种理化性质融合的预测结果要优于单一的特征，达到了一个较好的预测值，分别为 94.6%和 96.9%。

3.3. 不同核函数对结果的影响

在使用支持向量机进行分类的时候，可以对 SVM 的核函数进行选择，选取不同的核函数会对结果产生较大影响。因此本文在融合 SePSSM 和 PhyChe 对凋亡蛋白序列进行特征提取，SePSSM 的分割比例为 0.5。选取不同的核函数在两个数据集上的准确率如下表 4 和表 5。

Table 4. The prediction results of ZW225with different kernel functions
表 4. ZW225 在不同核函数下的预测结果

亚细胞位置	RBF	linear	poly	sigmoid
cy	98.2	97.3	96.4	96.0
me	96.0	95.5	88.8	92.4
mi	97.8	97.8	92.9	96.0
nu	97.3	97.8	94.2	95.1
OA	94.6	94.2	86.2	89.7

Table 5. The prediction results of ZD98 with different kernel functions
表 5. ZD98 在不同核函数下的预测结果

亚细胞位置	RBF	linear	poly	sigmoid
cy	98.0	97.0	90.8	90.8
me	98.0	98.0	91.8	93.9
mi	99.0	99.0	99.0	99.0
oth	99.0	98.0	93.9	96.0
OA	96.9	95.9	87.8	89.8

从表 4 和表 5 可以看出，对 ZW225 和 ZD98 两个数据集，用 SVM 进行分类时，采用 RBF 核函数和 linear 核函数的预测效果要优于 poly 和 sigmoid 两种核函数，且采用 RBF 核函数预测总体准确率要略高于 linear 核函数，sigmoid 核函数得到的效果最差。因此，本文选用 RBF 核函数作为 SVM 算法的核函数。

3.4. 我们提出的方法性能

本文通过融合 PsePSSM 和 PhyChe 两种特征，SVM 的核函数选择 RBF，通过 Jackknife 检验对 ZW225 和 ZD98 数据集进行验证，得到的结果如表 6 所示。从表 6 可以看出，我们的方法对 ZW225 和 ZD98 数据集的 OA 分别达到了 94.6%和 96.9%。实验结果表明，该方法能够有效预测凋亡蛋白的亚细胞位置。

Table 6. The predictive performance of datasets ZW225 and ZD98 protein subcellular localization on the Jackknife test
表 6. ZW225 和 ZD98 数据集的蛋白质亚细胞定位在 Jackknife 检验下的预测性能

数据集	亚细胞位置	Sen(%)	F1(%)	MCC	OA(%)
ZW225	cy	98.6	97.2	0.96	94.6
	me	96.6	95.0	0.92	
	mi	84.0	89.4	0.88	
	nu	90.2	92.5	0.91	
ZD98	cy	97.7	97.7	0.96	96.9
	me	96.7	96.7	0.95	
	mi	100	96.3	0.96	
	oth	91.7	95.7	0.95	

3.5. 与其他方法的比较

为了进一步评估我们方法的有效性，我们将我们提出的方法与几种凋亡蛋白亚细胞定位的方法进行了比较。表 7 和表 8 分别显示了不同方法对 ZW225 和 ZD98 两个数据集的 OA 和每类亚细胞位置的敏感性的预测结果，所有结果都是通过 Jackknife 检验得到的。

Table 7. Comparison from different methods on ZW225 dataset by Jackknife test
表 7. ZW225 数据集基于不同方法的预测结果

方法	cy	me	mi	nu	OA(%)
DF_SVM [30]	87.1	92.1	64.0	73.2	84.0
Liang et al. [31]	87.1	89.1	68.0	75.6	84.4
Zhang et al. [32]	93.5	92.1	96.0	93.5	92.2
ERT-ECT-PSSM-IS [33]	80.0	91.0	92.0	87.8	87.1
我们的方法	98.6	96.6	84.0	90.2	94.6

Table 8. Comparison from different methods on ZD98 dataset by Jackknife test
表 8. ZD98 数据集基于不同方法的预测结果

方法	cy	me	mi	oth	OA(%)
DF_SVM [30]	97.7	96.7	92.3	75.0	93.9
Liang et al. [31][32][33]	95.4	90.0	92.3	83.3	91.8
Zhang et al. [32]	95.3	88.9	97.4	91.7	93.2
ERT-ECT-PSSM-IS [33]	90.7	100	92.3	91.7	93.9
我们的方法	97.7	96.7	100	91.7	96.9

从表 7 和表 8 可以看出，我们提出的方法在两个数据集的 cy 类上都取得了较好的结果，分别为 98.6%和 97.7%，这是因为这两个数据集类中数量最多的都是 cy 类，从而导致了预测效果较好。而 nu 和 oth 类，分别在两个数据集中数量都最少，使得训练不足导致预测性能较差。但从 OA 值来看，我们的方法要高于其他方法的，这表明我们的方法具有较好的预测性能。

4. 结论

首先，基于矩阵分块的思想从 PSSM 中提取 SePSSM 特征，然后将 SePSSM 和 7 种理化性质得到的特征融合构建凋亡蛋白序列的特征表示方法，通过实验结果可知，对 PSSM 进行平均分割比不分割或其他的分割比例的预测效果更好。最后，ZW225 和 ZD98 两个数据集在 RBF 核的 SVM 分类器上分别进行预测分类，分别得到了 94.6%和 96.9%的总体准确率，这已高于大多数已有的凋亡蛋白亚细胞定位算法，这表明我们所提出的方法是可行的。鉴于我们使用的数据集为不平衡数据集，数据集类中数量存在较大差异，因此在下一步研究中，我们将考虑对数据集进行采样处理或构建一个平衡的数据集来对凋亡蛋白进行预测研究。

基金项目

国家自然科学基金(62062067)。

参考文献

[1] Reed, J.C. and Paternostro, G. (1999) Postmitochondrial Regulation of Apoptosis during Heart Failure. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 7614-7616. https://doi.org/10.1073/pnas.96.14.7614

[2] Schulz, J.B., Weller, M. and Moskowitz, M.A. (1999) Caspases as Treatment Targets in Stroke and Neurodegenerative Diseases. *Annals of Neurology*, **45**, 421-429. https://doi.org/10.1002/1531-8249(199904)45:4%3C421::AID-ANA2%3E3.0.CO;2-Q

[3] Kaufmann, S.H. and Hengartner, M.O. (2001) Programmed Cell Death: Alive and Well in the New Millennium. *Trends in Cell Biology*, **11**, 526-534. https://doi.org/10.1016/S0962-8924(01)02173-0

[4] Evan, G. and Littlewood, T. (1998) A Matter of Life and Cell Death. *Science*, **281**, 1317-1322. https://doi.org/10.1126/science.281.5381.1317

[5] Zhou, G.P. and Doctor, K. (2003) Subcellular Location Prediction of Apoptosis Proteins. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **50**, 44-48. https://doi.org/10.1002/prot.10251

[6] Zhou, H., Yang, Y. and Shen, H.B. (2017) Hum-mPLoc 3.0: Prediction Enhancement of Human Protein Subcellular Localization through Modeling the Hidden Correlations of Gene Ontology and Functional Domain Features. *Bioinformatics*, **33**, 843-853. https://doi.org/10.1093/bioinformatics/btw723

[7] Chen, Y.L. and Li, Q.Z. (2007) Prediction of Apoptosis Protein Subcellular Location Using Improved Hybrid Approach and Pseudo-Amino Acid Composition. *Journal of Theoretical Biology*, **248**, 377-381. https://doi.org/10.1016/j.jtbi.2007.05.019

[8] Jones, D.T. (1999) Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices. *Journal of molecular biology*, **292**, 195-202. https://doi.org/10.1006/jmbi.1999.3091

[9] Yu, X.Q., Zheng, X.Q., Liu, T.G., Dou, Y.C. and Wang, J. (2012) Predicting Subcellular Location of Apoptosis Proteins with Pseudo Amino Acid Composition: Approach from Amino Acid Substitution Matrix and Auto Covariance Transformation. *Amino acids*, **42**, 1619-1625. https://doi.org/10.1007/s00726-011-0848-8

[10] Li, B., Cai, L., Liao, B., Bing, P. and Yang, J. (2019) Prediction of Protein Subcellular Localization Based on Fusion of Multi-View Features. *Molecules*, **24**, Article No. 919. https://doi.org/10.3390/molecules24050919

[11] Wang, S.F., Cao, Z.C., Li, M.Y. and Yue, Y.T. (2019) G-DipC: An Improved Feature Representation Method for Short Sequences to Predict the Type of Cargo in Cell-Penetrating Peptides. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **17**, 739-747. https://doi.org/10.1109/TCBB.2019.2930993

[12] Zhang, S. and Liang, Y. (2018) Predicting Apoptosis Protein Subcellular Localization by Integrating Auto-Cross Correlation and PSSM into Chou's PseAAC. *Journal of Theoretical Biology*, **457**, 163-169. https://doi.org/10.1016/j.jtbi.2018.08.042

[13] 刘太岗, 王春华. 基于 SVM-RFE 算法的凋亡蛋白亚细胞定位预测[J]. 计算机工程与应用, 2017(10): 155-159.

[14] Ding, Y.S. and Zhang, T.L. (2008) Using Chou's Pseudo Amino Acid Composition to Predict Subcellular Localization of Apoptosis Proteins: An Approach with Immune Genetic Algorithm-Based Ensemble Classifier. *Pattern Recognition Letters*, **29**, 1887-1892. https://doi.org/10.1016/j.patrec.2008.06.007

[15] Zhang, Z.H., Wang, Z.H., Zhang, Z.R. and Wang, Y.X. (2006) A Novel Method for Apoptosis Protein Subcellular Localization Prediction Combining Encoding Based on Grouped Weight and Support Vector Machine. *FEBS Letters*, **580**, 6169-6174. https://doi.org/10.1016/j.febslet.2006.10.017

[16] Xiang, Q., Liao, B., Li, X., Xu, H., Chen, J., Shi, Z., *et al*. (2017) Subcellular Localization Prediction of Apoptosis Proteins Based on Evolutionary Information and Support Vector Machine. *Artificial Intelligence in Medicine*, **78**, 41-46. https://doi.org/10.1016/j.artmed.2017.05.007

[17] Fu, H.Y., Cao, Z.C., Li, M.Y. and Wang, S.F. (2020) ACEP: Improving Antimicrobial Peptides Recognition through Automatic Feature Fusion and Amino Acid Embedding. *BMC Genomics*, **21**, Article No. 597. https://doi.org/10.1186/s12864-020-06978-0

[18] Chou, K.C. and Shen, H.B. (2007) Recent Progress in Protein Subcellular Location Prediction. *Analytical Biochemistry*, **370**, 1-16. https://doi.org/10.1016/j.ab.2007.07.006

[19] Chou, K.C. and Maggiora, G.M. (1998) Domain Structural Class Prediction. *Protein Engineering*, **11**, 523-538. https://doi.org/10.1093/protein/11.7.523

[20] Chou, K.C., Liu, W.M., Maggiora, G.M. and Zhang, C.T. (1998) Prediction and Classification of Domain Structural Classes. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **31**, 97-103. https://doi.org/10.1002/(SICI)1097-0134(19980401)31:1%3C97::AID-PROT8%3E3.0.CO;2-E

[21] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., *et al*. (2003) The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL in 2003. *Nucleic Acids Research*, **31**, 365-370. https://doi.org/10.1093/nar/gkg095

[22] Wang, J., Yang, B., Revote, J., Leier, A., Marquez-Lago, T.T., Webb, G., *et al*. (2017) POSSUM: A Bioinformatics Toolkit for Generating Numerical Sequence Feature Descriptors Based on PSSM Profiles. *Bioinformatics*, **33**, 2756-2758. https://doi.org/10.1093/nar/gkg095

[23] Wei, L., Liao, M., Gao, X., Wang, J. and Lin, W. (2016) mGOF-Loc: A Novel Ensemble Learning Method for Human Protein Subcellular Localization Prediction. *Neurocomputing*, **217**, 73-82. https://doi.org/10.1016/j.neucom.2015.09.137

[24] Shen, H.B. and Chou, K.C. (2007) Nuc-PLoc: A New Web-Server for Predicting Protein Subnuclear Localization by Fusing PseAA Composition and PsePSSM. *Protein Engineering, Design & Selection*, **20**, 561-567. https://doi.org/10.1093/protein/gzm057

[25] Lin, C., Zou, Y., Qin, J., Liu, X., Jiang, Y., Ke, C., *et al*. (2013) Hierarchical Classification of Protein Folds Using a Novel Ensemble Classifier. *PLoS ONE*, **8**, e56499. https://doi.org/10.1371/journal.pone.0056499

[26] Qiu, J.D., Luo, S.H., Huang, J.H., Sun, X.Y. and Liang, R.P. (2010) Predicting Subcellular Location of Apoptosis Proteins Based on Wavelet Transform and Support Vector Machine. *Amino Acids*, **38**, 1201-1208. https://doi.org/10.1007/s00726-009-0331-y

[27] Burges, C.J.C. (1998) A Tutorial on Support Vector Machines for Pattern recognition. *Data Mining and Knowledge Discovery*, **2**, 121-167. https://doi.org/10.1023/A:1009715923555

[28] Chou, K.C. (2011) Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **273**, 236-247. https://doi.org/10.1016/j.jtbi.2010.12.024

[29] Mei, S. (2012) Multi-Kernel Transfer Learning Based on Chou's PseAAC Formulation for Protein Submitochondria Localization. *Journal of Theoretical Biology*, **293**, 121-130. https://doi.org/10.1016/j.jtbi.2011.10.015

[30] Zhang, L., Liao, B., Li, D. and Zhu, W. (2009) A Novel Representation for Apoptosis Protein Subcellular Localization Prediction Using Support Vector Machine. *Journal of Theoretical Biology*, **259**, 361-365. https://doi.org/10.1016/j.jtbi.2009.03.025

[31] Liang, Y., Liu, S. and Zhang, S. (2017) Geary Autocorrelation and DCCA Coefficient: Application to Predict Apoptosis Protein Subcellular Localization via PSSM. *Physica A*: *Statistical Mechanics and Its Applications*, **467**, 296-306. https://doi.org/10.1016/j.physa.2016.10.038

[32] Zhang, S. and Duan, X. (2018) Prediction of Protein Subcellular Localization with Oversampling Approach and Chou's General PseAAC. *Journal of Theoretical Biology*, **437**, 239-250. https://doi.org/10.1016/j.jtbi.2017.10.030

[33] Ruan, X., Zhou, D., Nie, R., Hou, R. and Cao, Z. (2019) Prediction of Apoptosis Protein Subcellular Location Based on Position-Specific Scoring Matrix and Isometric Mapping Algorithm. *Medical & Biological Engineering & Computing*, **57**, 2553-2565. https://doi.org/10.1007/s11517-019-02045-3