生 物 工 程 学 报
Chinese Journal of Biotechnology
http://journals.im.ac.cn/cjbcn
DOI: 10.13345/j.cjb.180403

Apr. 25, 2019, 35(4): 687−696

·生物技术与方法·

基于多层次稀疏编码预测蛋白质亚细胞定位

陈行健，胡雪娇，薛卫

南京农业大学 信息科学技术学院，江苏 南京 210095

陈行健, 胡雪娇, 薛卫. 基于多层次稀疏编码预测蛋白质亚细胞定位. 生物工程学报, 2019, 35(4): 687–696.
Chen XJ, Hu XJ, Xue W. Prediction of protein subcellular localization based on multilayer sparse coding. Chin J Biotech, 2019, 35(4): 687–696.

摘 要: 文中提出了一种简单有效的蛋白质亚细胞区间定位预测方法，为进一步了解蛋白质的功能和性质提供理论基础。运用稀疏编码，结合氨基酸组成信息提取蛋白质序列特征，基于不同字典大小对得到的特征进行多层次池化整合，并送入支持向量机进行分类。经 Jackknife 检验，在数据集 ZD98、CH317 和 Gram1253 上的预测成功率分别达到 95.9%、93.4%和 94.7%。实验证明基于多层次稀疏编码的分类预测算法能显著提高蛋白质亚细胞区间定位的预测精度。

关键词: 稀疏编码，氨基酸组成，多层次池化，支持向量机，亚细胞区间定位

Prediction of protein subcellular localization based on multilayer sparse coding

Xingjian Chen, Xuejiao Hu, and Wei Xue

School of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, *Jiangsu, China*

Abstract: In order to provide a theoretical basis for better understanding the function and properties of proteins, we proposed a simple and effective feature extraction method for protein sequences to determine the subcellular localization of proteins. First, we introduced sparse coding combined with the information of amino acid composition to extract the feature values of protein sequences. Then the multilayer pooling integration was performed according to different sizes of dictionaries. Finally, the extracted feature values were sent into the support vector machine to test the effectiveness of our model. The success rates in data set ZD98, CH317 and Gram1253 were 95.9%, 93.4% and 94.7%, respectively as verified by the Jackknife test. Experiments showed that our method based on multilayer sparse coding can remarkably improve the accuracy of the prediction of protein subcellular localization.

Keywords: sparse coding, amino acid composition, multilayer pooling, support vector machine, subcellular localization prediction

Received: September 30, 2018; **Accepted:** October 29, 2018
Supported by: National Key Technology R&D Program of China (No. 2017YFD0800204), the Fundamental Research Funds for the Central Universities (No. KYZ201600175).
Corresponding author: Wei Xue. Tel: +86-25-84396350; E-mail: xwsky@njau.edu.cn
国家重点研发计划 (No. 2017YFD0800204)，中央高校基本科研业务费专项资金 (No. KYZ201600175) 资助。
网络出版时间：2018-12-08 网络出版地址：http://kns.cnki.net/kcms/detail/11.1998.Q.20181207.0950.002.html

蛋白质作为生物体的基本组成物质，在生命活动中发挥着重要作用。蛋白质的功能与其亚细胞区间密切相关，不同蛋白质只有处于特定亚细胞区间才能发挥其功能，因而通过已有方法预测确定某种蛋白质所处的亚细胞区间，对明确蛋白质的功能和性质、认识蛋白质间的相互作用具有重要意义[1]。随着蛋白序列数据的不断增加，使用传统人工实验手段获取蛋白质亚细胞区间位置已远不能满足科研需要，这促使了机器学习在蛋白质亚细胞定位预测中的发展。

通过对目前研究现状的分析，可将近年来使用机器学习方法对蛋白质亚细胞区间进行预测的研究方向大致分为两类，分别为序列特征提取和分类模型构建[2]。目前用于蛋白质序列的特征提取算法主要有氨基酸组成 (Amino acid composition，AAC)、伪氨基酸组成 (Pseudo amino acid composition，PseAAC)、基因本体 (Gene ontology, GO)、位置特异性得分矩阵 (Position specific scoring matrix，PSSM) 和基于不同特征的融合等。如 Zhou 等基于 Mahalanobis 距离提取蛋白质序列的组分信息，使用斜变判别函数对蛋白质亚细胞区间进行预测，在 Jackknife 检验下 ZD98 数据集上的准确率约为 72.5%[3]；Wan 等提出了 GOASVM 算法，基于 GO 注释信息与蛋白非相邻区域的同源性来表示蛋白质序列，取得了较好的效果[4]；Chen 等采用混合增量的方式对蛋白质序列的 N 端、C 端以及疏水性 3 种特征进行融合，在 ZD98 和 CH317 数据集上的成功率分别为 90.8%和 82.7%[5]；Zhao 等提出蛋白质序列的词袋特征，将词袋模型与基于伪氨基酸组成的特征提取算法相结合，获得了较高的准确率[6]。同时，在分类预测模型方面，国内外研究者也开展了大量工作，如 Wan 等通过 GO 数据库的注释信息，提出自适应决策支持向量机，实现了对多功能膜蛋白序列的区间预测[7]；Ali 等提取蛋白质序列的伪氨基酸特征，采用区间投票、最邻近算法和概率神经网络等多种分类器进行对比预测，取得了较好的结果[8]；除此之外还有基于逻辑回归、贝叶斯集成和长短期记忆网络等多种分类模型的预测方法[9-11]。

总结前人研究成果可发现，能否准确描述蛋白质序列特征直接影响了最终分类器的预测效果。由于蛋白质序列中包含的信息量较大，且分属同一亚细胞区间的序列长度不等，序列特征分布不均，导致单一使用传统蛋白质序列特征提取算法的分类结果不佳。而对于一些较为复杂的特征融合及其改进算法，虽然取得了较高的准确率，但特征提取过程复杂，且最终得到的特征向量维数较大，造成分类器的时间和空间复杂度过高。因此，本研究结合氨基酸组分信息，提出一种基于多层次稀疏编码的蛋白质序列特征提取算法，该算法能够基于简单的 AAC 方法对蛋白质序列进行稀疏表示，进而提取序列底层特征；根据不同字典大小对特征进行多层次池化整合，能有效增加序列特征的区分性；将得到的特征向量经主成分分析 (Principal component analysis, PCA) 降维，能在选取有效特征的同时降低算法的计算量。最后将得到的特征向量送入分类器进行分类。实验结果表明，本方法不仅能简化特征提取过程，降低分类器的时间及空间复杂度，也能更加全面地反映序列特征，提高分类性能。

1 材料与方法

1.1 数据集

为了对本文算法进行客观评价，方便与同类算法进行对比，采用近年来相关领域中使用最多且国际公认有效的 ZD98 和 CH317 作为实验基准数据集[12-19]，其中 ZD98 由 Zhou 和 Doctor[3]构建，共有 98 条蛋白质序列，分为 4 个亚细胞定位类别，分别是细胞质蛋白 (Cytoplasmic proteins，Cy) 43 条、

线粒体蛋白 (Mitochondrial proteins, Mi) 13 条、细胞膜蛋白 (Membrane proteins, Me) 30 条和其他类蛋白 (Other) 12 条。CH317 是由 Chen 和 Li[5] 构建，分为 6 个亚细胞定位类别，共有 317 条蛋白质序列，分别是分泌蛋白 (Secreted proteins，Se) 17 条、细胞核蛋白 (Nuclear proteins，Nu) 52 条、细胞质蛋白 (Cytoplasmic proteins，Cy) 112 条、内质网蛋白 (Endoplasmic reticulum proteins，En) 47 条、膜蛋白 (Membrane proteins，Me) 55 条和线粒体蛋白 (Mitochondrial proteins，Mi) 34 条。

考虑到上述数据集构建时间较长，参考 Wang 等的方法对 ZD98 和 CL317 数据集进行了更新[12]，删除了部分重复及错误序列，其具体方法不再复述。经处理后 ZD98 数据集剩余 96 条蛋白质序列，CH317 数据集剩余 314 条蛋白质序列。此外，为了对算法进行进一步评估，除了上述两个数据集外，本研究还采用了 Xue 等按照同样标准构建的 Gram1253 数据集进行测试[20]。Gram1253 数据集共包含 1 253 条蛋白质序列，分为 Me、Cy、Nu、Se 及细胞周质 (Periplasm，Pe) 等 5 个亚细胞定位类别。以上 3 种数据集中的所有蛋白质序列均来自最新版本的 UniProt 数据库 (Release 2018_08)，其具体区间分布如表 1 所示。

1.2 序列特征编码

将稀疏编码引入蛋白质亚细胞区间定位预测中，目的是在每条蛋白序列与相应的数值向量间建立一种能够更为准确表达此条蛋白序列特征的映射关系。基于多层次稀疏编码的特征提取算法主要包括局部特征提取、稀疏编码和多层次池化等 3 个步骤。首先对蛋白质序列进行分割处理得到若干个序列单词，使用传统蛋白质特征提取算法对序列单词进行特征编码得到特征单词，然后选取部分特征单词作为局部特征块学习字典，用字典对原始序列特征进行稀疏表示。采用平均池化的方法对稀疏矩阵降维，将基于不同字典大小得到的特征向量进行组合，即为蛋白质序列的最终特征表示。其提取流程如图 1 所示。

表 1 三种数据集中不同区间的蛋白质序列条数

Table 1 Numbers of protein sequences in different class of 3 datasets

Dataset	Number of sequences in each class						Total
ZD98	Cy	Me	Mi	Other			96
	43	28	13	12			
CL317	Cy	En	Me	Mi	Nu	Se	314
	110	47	55	34	51	17	
Gram1253	Me	Cy	Pe	Se	Nu		1 253
	166	443	423	199	22		

1.2.1 局部特征提取

在进行稀疏编码过程之前，首先需要提取蛋白质序列的局部特征作为特征块，组成训练样本构造字典。由于每条蛋白质序列长度不等，其主要特征可能分布在序列的不同位置，因此参考 Zhao 等的方法[6]采用滑动窗口分割法对原始蛋白质序列进行切分得到序列单词。滑动窗口分割法即按照一定长度对每条蛋白质序列进行切片，通过设定窗口大小和滑动间距得到若干个序列单

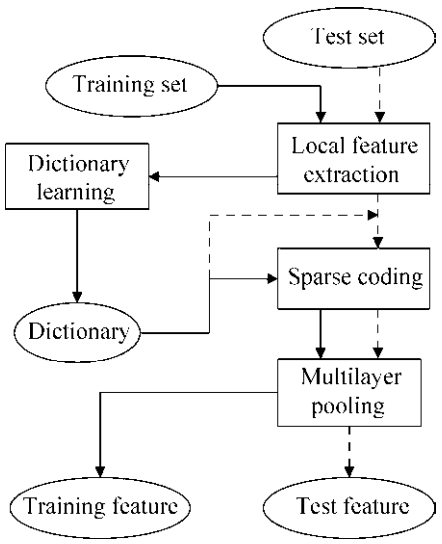

图 1 稀疏编码特征提取流程

Fig. 1 The feature extraction of sparse coding. The solid black line represents the progress of the training set, and the dotted black line represents the flow of the test set.

词，经特征提取后得到特征单词集合形成构建字典的基础。这种方法能完整地保留蛋白质序列的全部信息。本研究取滑动间隔为 1，滑动窗口大小决定序列单词长度，需满足以下条件：

$$L_{min}=Min\{L_1,L_2,\ldots,L_{num}\},\frac{L_{min}}{2}\le s\le L_{min},m\in Z \quad (1)$$

式中 $L_1,L_2,\cdots,L_{num}$ 分别表示蛋白质序列集合中每一条蛋白质序列的长度，取 L_{min} 为数据集中最短的蛋白质序列长度，s 为滑动窗口大小，即序列单词长度在 $L_{min}/2$ 到 L_{min} 之间选取，具体值根据实验经验选取。

切分后每条蛋白质序列被表示为若干个长度相等的序列单词，运用已有的蛋白质序列特征提取算法统计序列单词的组分信息，即可得到相应的特征单词。Nakashima 和 Nishikawa[21]最早将氨基酸组成和蛋白质亚细胞区间定位预测联系起来，提出 AAC 编码方式，统计每个氨基酸残基在蛋白质序列中出现的频率，其定义如下：

$$P=\left[f_1f_2f_3\cdots f_{20}\right]^T \quad (2)$$

其中 $f_1,f_2,f_3,\ldots,f_{20}$ 分别代表蛋白质序列中每种氨基酸残基出现的次数，具体求解方式如下：

$$f_u=\frac{1}{L}\sum_{i=1}^{L}F_i,F_i=\begin{Bmatrix}1,if\ R_i=A(u)\\0,if\ R_i\ne A(u)\end{Bmatrix} \quad (3)$$

L 表示每条蛋白质序列的长度，即包含所有氨基酸残基的总数目。首先对 20 种氨基酸从 1 到 20 进行编号。$f_u(u=1,2,3,\cdots,20)$ 为相应编号在蛋白质序列中出现的频率，$A(u)$表示与编号 u 对应的氨基酸残基。

通过 AAC 算法计算蛋白质序列 P 的序列单词特征，将每条蛋白质序列的所有特征单词进行组合，则每条序列都被表示为一个片段特征矩阵，如公式 (4) 所示：

$$\begin{bmatrix}v_{11}&v_{1n}&\cdots&v_{1n}\\v_{21}&\cdots&\cdots&v_{2n}\\\vdots&\cdots&\cdots&\vdots\\v_{m1}&v_{m2}&\cdots&v_{mn}\end{bmatrix} \quad (4)$$

其中 m 表示一条蛋白质序列经切割后得到的片段条数，n 为经过特征提取算法处理后的序列单词的特征维度，此时 n 为 20。矩阵的每一行均代表蛋白质序列不同序列单词的特征，在选取局部特征作为训练样本学习字典时，由于原始数据集的序列数据量较小，因此本研究选取数据集中所有蛋白质序列经切割得到的特征单词进行组合，构成稀疏编码字典学习过程中的训练样本矩阵 $X=\left[x_1,x_2,x_3,\cdots,x_N\right]$，其中 $x_i\in R^n(i=1,2,3,\cdots,N)$，$x_i$ 分别代表不同蛋白质序列的特征单词，N 为数据集中所有蛋白质序列经分割后得到的片段数之和。

1.2.2 稀疏编码

得到由蛋白质序列的局部特征组成的训练样本之后，下一步即是对这些训练样本进行稀疏编码。稀疏编码是一种无监督的机器学习算法，通过在高维数据中寻找一组超完备的基向量来对样本进行稀疏表示，主要分为字典学习和稀疏重构两个过程。其公式表示如下：

$$X=UD \quad (5)$$

其中 X 即为特征单词组成的训练样本矩阵；$D=\left[d_1,d_2,d_3,\cdots,d_K\right]\in R^{K*n}$ 是由特征单词的基元向量组成的矩阵，也称为字典；$U=\left[u_1,u_2,u_3,\cdots,u_N\right]\in R^{N*K}$ 即训练样本的稀疏矩阵，u_i 表示第 i 个特征单词在稀疏特征空间中的表示系数。

字典学习是指通过学习局部特征，从中找到最能够为有效重构训练样本的元素组合。在字典学习过程中，为了确保能够按照约束条件自适应地训练出超完备字典，一般要求字典中的原子个数较大，即若字典 D 的大小为 $K*n$，则需使 $K>n$，这样得到的字典更容易学习到训练样本深层次的特征。同时训练样本的稀疏系数 u_i 需满足以下条件：

$$\min_{D,U}\sum_{i=1}^{N}x_i-Du_i\ {}_2^2$$
$$s.t\|u_i\|_0\ T_0,i=1,2,3,\cdots,N. \quad (6)$$

其中$\|\cdot\|_2$表示L_2范数，$\|\cdot\|_0$表示L_0范数。公式 (6) 对应字典训练过程中的两个条件规约：1) 重构样本具有可稀疏性，即u_i中的非零元素个数小于等于 T_0，T_0 即稀疏相关系数；2) 字典 D 与重构样本 U 的不相关性，即重构误差项需小于某一固定值。

本研究采用 K-SVD 算法训练字典。K-SVD 算法是由 Aharon 等提出的一种基于 K-means 算法扩展而来的字典学习算法[22]，其实质是迭代交替学习字典原子并优化其相应的稀疏系数。该算法要经过 K 次迭代，每次迭代时都需要对误差项进行奇异值分解，采用逐列更新的方式对字典进行优化，每次只更新其中的一个原子和其对应的稀疏系数，选择使重构误差最小的分解项作为新的元素值，经过不断迭代得到最优化的解。K-SVD 算法主要分为以下几个步骤：(1) 随机初始化字典 D，设置迭代终止条件；(2) 固定字典 D，求解稀疏矩阵 U；(3) 固定稀疏矩阵 U，求解字典 D；(4) 交替执行步骤 (2) 和 (3)，直至迭代结束。

得到字典后，通常使用正交匹配追踪 (Orthogonal matching pursuit，OMP) 算法，求得原始样本的稀疏矩阵。OMP 算法的核心思想是在每次迭代过程中使用最小二乘法对原始样本进行稀疏逼近，选择字典中最匹配的基元对其进行稀疏重构，求出残差并继续选择下一个最匹配的基元。这种更新方式能保证在下一次迭代过程中不会重复选择相同基元，在一定程度上加快了算法的收敛速度，克服了传统匹配追踪 (Matching pursuit，MP) 算法容易陷入局部最优解的问题。

使用 OMP 算法对原始蛋白质序列的片段特征矩阵进行稀疏表示，则每一条的蛋白质序列可被表示为一个 $m*K$ 稀疏矩阵 Z，m 即为每一条蛋白质序列切割得到的片段条数，K 为字典的原子数目。此过程即为蛋白质序列的稀疏编码。

1.2.3 多层次池化

经稀疏编码后所得到的稀疏矩阵维度较高，如果直接展开进行串接表示数据量过大，训练分类器时的内存和时间消耗代价过高。所以需要对特征矩阵进行降维，通常使用池化方法。池化是指把特征向量集映射为单个向量的过程，对不同位置的特征进行聚合统计，能提取有效特征，减少计算量。常用的池化算法有最大池化 (Max-pooling)[23]和平均池化 (Mean-pooling)[24]。Max-pooling 即对邻域内的特征点取最大值，能更多地保留矩阵的边缘信息。而 Mean-pooling 则是对邻域内特征点求平均值，能更多地保留矩阵的背景信息。考虑到序列数据的特殊性，本研究选择 Mean-pooling 作为最终的池化方法。公式表示如下：

$$Z=\left[z_1,z_2,z_3,\cdots,z_K\right]^T$$
$$z_i=mean\left\{z_{i1},z_{i2},z_{i3},\cdots z_{im}\right\} \quad (7)$$

式中，$i=1,2,3,\cdots K$，z_i 是由稀疏矩阵 Z 中第 i 行的 m 个元素取平均值求得。经平均池化后每条蛋白质序列被表示成一个 K 维的特征向量。

这里根据不同字典大小对特征矩阵进行多层次池化，可以分别帮助提取蛋白质序列的整体和局部特征。将几种池化的结果融合在一起作为蛋白质序列的最终特征，可使其表述更具有代表性。具体方法如下：在稀疏编码的字典学习过程中，将字典原子数目分别设置为 K_1、K_2 和 K_3，经 K-SVD 算法后得到 3 种不同层次大小的字典，然后针对不同字典使用 OMP 算法对原始蛋白质序列的片段特征矩阵进行稀疏表示，将得到的稀疏矩阵分别进行平均池化得到不同层次的特征向量。将每个池化块内的特征向量进行串接融合在一起，得到一个 $K_1+K_2+K_3$ 维的向量作为最终特征向量。K 值分别设置为 30、50 和 70，最后生成一个 150 维的向量，经 PCA 进行特征选择后送入分类器完成分类。

1.3 支持向量机

为了方便与其他算法进行对比，选择支持向量机 (Support vector machine, SVM) 建立分类模型。SVM 是 Vapnik 领导的 AT&T Bell 实验室在 1995 年提出的一种基于统计学习理论的分类方法，通过核函数将输入样本从原空间非线性映射到高维特征空间，利用线性方法解决非线性问题，在高维特征空间中构造最优分类超平面，在解决小样本、非线性及高维模式识别问题中表现出强大的泛化能力[25]。蛋白质序列经特征编码后，使用 LIBSVM 通用软件包，基于一对一算法 (One-versus-one) 构造 SVM 多类分类器，在训练阶段为任意两类样本设计一个 SVM，则 k 个类别的数据集就需要设计 $k(k–1)/2$ 个 SVM。当对一个未知样本进行分类时，最后取得票最多的类别为该未知样本的类别。运用 SVM 进行蛋白质亚细胞区间定位预测的流程图如图 2 所示。

将数据集中的样本分为训练样本和预测样本，先送入训练样本的特征向量，设定输出为相应的亚细胞定位 y_i，训练 SVM，确定模型参数，再送入预测样本的特征向量，SVM 分类器会给出一个预测结果，用 x_i 表示，若 $x_i=y_i$，则预测正确，若 $x_i \neq y_i$，则预测错误，最后统计整个数据集的预测准确率作为蛋白质亚细胞区间定位的评价指标。

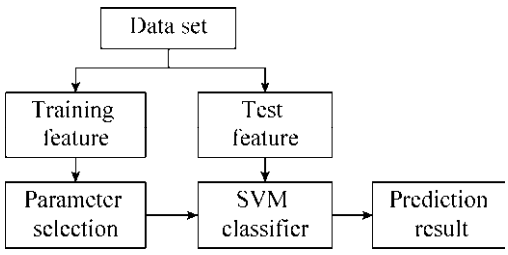

图 2 基于 SVM 的亚细胞定位预测流程

Fig. 2 Subcellular localization prediction process based on SVM. In the training process of SVM, it is necessary to manually set parameters according to different kinds of kernel functions, that is the parameter selection.

2 结果与分析

2.1 测试方法

为了验证方法的有效性，采用 Jackknife 进行假设检验。Jackknife 是蛋白质亚细胞定位预测研究中公认且使用最多的一种测试方法[12-20]，每次仅用一条序列作为测试集进行验证，其余全部序列作为训练集送入分类器进行训练，以此类推直至所有序列均预测完毕，是一种客观有效的假设检验方法[26]。为了便于比较实验结果，同时对预测方法进行有效评估，引入敏感性 (Sensitivity, Se)、特异性 (Specificity, Sp) 和相关系数 (Matthews correlation coefficient, MCC) 等 3 个评价指标，并统计总的预测准确率 (Overall accuracy, OA)，定义如下[27]：

$$S_e = \frac{TP_i}{TP_i + FN_i} \tag{8}$$

$$S_p = \frac{TP_i}{TP_i + FP_i} \tag{9}$$

$$MCC = \frac{(TP_i \times TN_i) - (FP_i \times FN_i)}{\sqrt{(TP_i + FP_i) \times (TN_i + FN_i) \times (TP_i + FN_i) \times (TN_i + FP_i)}} \tag{10}$$

$$OA = \frac{\sum_{i=1}^{M} TP_i}{\sum_{i=1}^{M} (TP_i + FN_i)} \tag{11}$$

其中，TP_i 是第 i 类亚细胞区间预测正确的序列条数，FN_i 是第 i 类亚细胞区间预测错误的序列条数，FP_i 是非第 i 类亚细胞区间但被预测为第 i 类区间的序列条数，TN_i 是被正确预测的非第 i 类亚细胞区间的序列条数，M 为亚细胞类别总数。

2.2 参数选择

在使用 PCA 对最终的特征向量进行选择时，维数 D 的设置对于整个算法的准确度存在一定影响。选取的维数越多，包含的特征就越多，但可能造成分类器的训练时间过长；维数越小，则越有可能丢失一些真正有意义的特征，影响分类效

果。因此需要通过实验寻求一个最优的 *D* 值。图 3 显示了数据集 ZD98、CH317 和 Gram1253 在 PCA 进行特征选择过程中分别取不同的 *D* 值所对应的预测准确率。在特征向量的维数较低时，3 个数据集的预测准确率也相对较低，在维数高于某一确定值时，预测准确率也随之降低。在维数取 60 到 70 之间时，在 ZD98、CH317 和 Gram1253 数据集上的预测准确率均达到最大且趋于稳定，当前的 *D* 值即为最优值。本研究使用的 3 种数据集的最优 *D* 值分别为 60、65 和 65。

2.3 结果分析

将本方法在 ZD98、CH317 和 Gram1253 数据集上采用 Jackknife 进行实验的预测结果列于表 1 中，为了进一步说明本文方法的有效性，表中分别列出了 3 个数据集在各个亚细胞区间进行预测得到的不同实验结果。

由表 1 可知，本方法在 3 个数据集上均获得了较好的实验结果，总的准确率分别达到了 95.9%、93.4%和 94.7%，实验证明本方法能有效增加蛋白质亚细胞区间定位预测的准确率。同时为了方便进行对比，将部分同领域内基于蛋白质序列特征提取的改进算法得到的实验结果也一并列出。

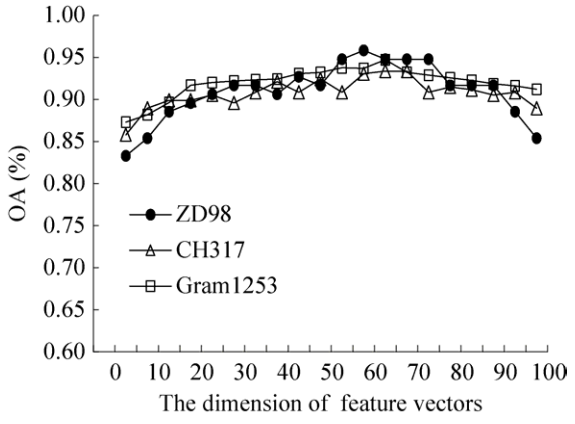

图 3 基于不同维度的预测准确率

Fig. 3 The prediction accuracies based on different dimensions. The abscissa represents the dimensions of feature vectors and the ordinate represents the total accuracy.

表 1 数据集实验结果

Table 1 The experimental results of data sets

Dataset	Jackknife test (%)				
	Location	Se (%)	Sp (%)	MCC (%)	OA (%)
ZD98	Cy	100	95.6	95.9	95.9
	Me	96.7	96.7	95.2	
	Mi	84.6	91.7	86.4	
	Other	91.7	91.7	90.5	
CL317	Cy	94.6	93.8	90.9	93.4
	Me	92.7	92.7	91.1	
	Mi	97.1	97.1	96.7	
	Se	76.5	92.9	83.4	
	Nu	92.3	90.1	89.6	
	En	95.7	90.0	91.5	
Gram1253	Cy	99.3	97.1	96.7	94.7
	Me	94.0	90.7	90.2	
	Pe	98.9	93.2	93.4	
	Se	85.6	98.6	80.5	
	Nu	86.4	83.2	85.7	

从表 2 可以看出，在 ZD98 数据集上本文算法相比 DCC、OF 和 DE 等复杂的特征融合算法在总体预测精度上最大提升了约 7 个百分点，在 Cyto 这一亚细胞类上的预测准确率达到了 100%，预测全部正确，且整体预测准确率方面也均优于其他方法。将本方法与 BOW、GA 和 OA 等改进算法的实验结果进行对比，在相同数据集上的准确率也都提高了约 2 到 5 个百分点，实验表明本文算法较基于传统蛋白质序列特征提取的改进算法仍具有显著优势。通过表 3 的比较可以看出，在 CH317 数据集上，本文算法在 Mito 这一亚细胞类上的预测准确率最高达到了 97.1%，相比其他算法最大提升了约 14.7 个百分点，在 Nucl 这一亚细胞类上的准确率最高也提升了 12.3 个百分点，这一实验结果也充分说明了本文算法对少数类别序列进行特征提取的有效性，使得序列底层特征更加具有区分性。对比 BOW、IAC 和 CF 等改进算法，在总预测准确率上均提升了 2–4 个

表 2 ZD98 数据集预测结果比较

Table 2 Comparison of the accuracy of ZD98 data set

Methods	Jackknife test (%)					References
	Cyto	Memb	Mito	Other	OA (%)	
DCC_SVM	93.0	86.7	92.3	75.0	88.9	[28]
OF_SVM	97.7	86.3	92.3	66.7	90.8	[16]
DE_SVM	95.4	93.3	76.9	83.3	90.8	[17]
BOW_SVM	97.7	92.9	76.9	83.3	91.7	[6]
GA_SVM	95.4	90.0	92.3	83.3	91.8	[19]
OA_SVM	95.3	88.9	97.4	91.7	93.2	[13]
This study	100	96.7	84.6	91.7	95.9	–

表 3 CH317 数据集预测结果比较

Table 3 Comparison of the accuracy of CH317 data set

Methods	Jackknife test (%)							References
	Cyto	Memb	Mito	Secr	Nucl	Endo	OA (%)	
DCC_SVM	91.1	92.7	82.4	76.5	80.0	93.6	88.3	[28]
GA_SVM	92.9	89.1	82.4	76.5	84.6	93.6	89.0	[19]
BOW_SVM	94.6	87.3	82.4	82.4	84.3	91.5	89.2	[6]
IAC_SVM	96.4	94.5	82.4	76.5	80.8	93.6	90.5	[18]
EI_SVM	94.6	95.7	92.7	82.4	90.4	70.6	91.1	[14]
CF_SVM	96.4	90.9	92.3	95.7	82.4	64.7	91.5	[29]
This study	94.6	92.7	97.1	76.5	92.3	95.7	93.4	–

百分点，进一步表明通过多层次池化分别提取序列的整体和局部信息，能有效提高蛋白质亚细胞定位预测精度。对于较大数据集 Gram1453 而言，本文引用了文献[20]中基于不同蛋白质序列特征提取算法的实验结果进行对比，如 AAC、Dipe 和 PseAAC 等，同时也基于 PSSM 特征进行了相关的对比实验，如 PSSM_SVM 等，表 4 结果表明，本方法在各个区间类别的预测率上均有一定提高，且相较于传统算法，如 PSSM_SVM 和 BLAST_KNN 等，本文方法不需要依靠复杂工具实现，在算法的可移植性上也具有明显优势。

与传统蛋白质序列特征提取及其改进方法相比，本文算法时间及空间复杂度低，在较简单的 AAC 特征下也能取得较好的效果，且通过平均池化提取特征序列特征矩阵的背景信息，将不同层次特征进行整合后经 PCA 降维，得到一种低维向量的形式反映序列特征的分布规律，能显著提高大数据处理的效率。

表 4 Gram1253 数据集预测结果比较

Table 4 Comparison of the accuracy of Gram1253 data set

Methods	Jackknife test (%)					
	Cyto	Memb	Peri	Secr	Nucl	OA (%)
AAC_SVM	98.4	94.6	97.6	70.0	40.9	92.1
Dipe_SVM	97.5	93.4	96.7	70.0	40.9	91.3
PseAAC_SVM	98.4	94.6	97.6	70.9	59.1	92.6
PSSM_SVM	98.4	93.4	95.6	77.2	65.3	92.8
BLAST_KNN	98.6	94.6	98.3	68.8	86.4	93.1
Our	99.3	94.0	98.9	85.6	86.4	94.7

3 讨论

蛋白质亚细胞定位预测一直是国内外生物信息学专家研究的热点方向。本研究在传统蛋白质序列特征提取算法 AAC 的基础上,提出了一种基于多层次稀疏编码的蛋白质序列特征提取算法对序列特征进行优化整合。相比其他算法，本方法提取过程简单，不需要经过复杂的特征融合步骤也能得到较高的预测准确率,且最后使用 PCA 对特征向量进行降维，在提高准确率的同时也降低了分类器的时间及空间复杂度。算法的主要流程包括：首先使用滑动窗口分割法对蛋白质序列进行切分提取序列单词，结合传统蛋白质特征提取算法对序列单词进行特征编码;采用 K-SVD 算法对序列单词特征进行字典学习，再通过 OMP 算法对序列特征矩阵进行稀疏表示；基于不同字典大小对特征矩阵进行多层次平均池化，分别帮助提取稀疏矩阵的整体信息和局部信息;使用 SVM 多类分类器对蛋白的亚细胞区间位置进行分类预测。实验表明，本文算法能在绝大部分亚细胞区间的预测成功率上获得较好的效果，对提升传统蛋白质序列特征提取算法的特征表达能力方面具有重要指导意义，是一种较为有效的蛋白质亚细胞区间预测方法。算法的源代码和所用数据集均可从 https://github.com/Multisc/Multi_sc_subloc/tree/master 获取。

REFERENCES

[1] Xu YY, Yang F, Shen HB. Incorporating organelle correlations into semi-supervised learning for protein subcellular localization prediction. Bioinformatics, 2016, 32(14): 2184–2192.

[2] Wei L, Ding Y, Su R, et al. Prediction of human protein subcellular localization using deep learning. Journal of Parallel & Distributed Computing, 2018, 117: 212–217.

[3] Zhou GP, Doctor K. Subcellular location prediction of apoptosis proteins. Proteins, 2003, 50(1): 44–48.

[4] Wan SB, Mak MW, Kung SY. GOASVM: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition. J Theor Biol, 2013, 323: 40–48.

[5] Chen YL, Li QZ. Prediction of the subcellular location of apoptosis proteins. J Theor Biol, 2007, 245(4): 775–783.

[6] Zhao N, Zhang L, Xue W, et al. Application of bag of words model in the prediction of protein subcellular location. J Food Sci Biotechnol, 2017, 36(3): 296–301 (in Chinese).
赵南, 张梁, 薛卫, 等. 词袋模型在蛋白质亚细胞定位预测中的应用. 食品与生物技术学报, 2017, 36(3): 296–301.

[7] Wan SB, Mak MW, Kung SY. Mem-ADSVM: a two-layer multi-label predictor for identifying multi-functional types of membrane proteins. J Theor Biol, 2016, 398: 32–42.

[8] Ali F, Hayat M. Classification of membrane protein types using Voting feature interval in combination with Chou's pseudo amino acid composition. J Theor Biol, 2015, 384: 78–83.

[9] Wan SB, Mak MW, Kung SY. mPLR-Loc: an adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction. Anal Biochem, 2015, 473: 14–27.

[10] Sáez-Atienzar S, Martínez-Gómez J, Alonso-Barba JI, et al. Automatic quantification of the subcellular localization of chimeric GFP protein supported by a two-level Naive Bayes classifier. Expert Syst Appl, 2015, 42(3): 1531–1537.

[11] Sønderby SK, Sønderby CK, Nielsen H, et al. Convolutional LSTM networks for subcellular localization of proteins//2nd International Conference on Algorithms for Computational Biology. Mexico City, Mexico: Springer, 2015: 68–80.

[12] Wang X, Li H, Zhang QW, et al. Predicting subcellular localization of apoptosis proteins combining go features of homologous proteins and distance weighted KNN classifier. BioMed Res Int, 2016, 2016: 1793272.

[13] Zhang SL, Duan X. Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC. J Theor Biol, 2018, 437: 239–250.

[14] Xiang QL, Bo L, Li XH, et al. Subcellular localization prediction of apoptosis proteins based on evolutionary information and support vector machine. Artif Intell Med, 2017, 78: 41–46.

[15] Dai Q, Ma S, Hai YB, et al. A segmentation based model for subcellular location prediction of apoptosis protein. Chemom Intell Lab Syst, 2016, 158: 146–154.

[16] Zhang SL, Jin J. Prediction of protein subcellular localization by using λ-order factor and principal component analysis. Lett Org Chem, 2017, 14(9): 717–724.

[17] Liang YY, Zhang SL. Prediction of apoptosis protein's subcellular localization by fusing two different descriptors based on evolutionary information. Acta Biotheor, 2018, 66(1): 61–78.

[18] Zhang SL, Liang YY. Predicting apoptosis protein subcellular localization by integrating auto-cross correlation and PSSM into Chou's PseAAC. J Theor Biol, 2018, 457: 163–169.

[19] Liang YY, Liu SY, Zhang SL. Geary autocorrelation and DCCA coefficient: application to predict apoptosis protein subcellular localization via PSSM. Phys A, 2017, 467: 296–306.

[20] Xue W, Wang XF, Zhao N, et al. Prediction of protein subcellular locations by ensemble of improved K-nearest neighbor. Chin J Biotech, 2017, 33(4): 683–691 (in Chinese).
薛卫, 王雄飞, 赵南, 等. 集成改进 KNN 算法预测蛋白质亚细胞定位. 生物工程学报, 2017, 33(4): 683–691.

[21] Nakashima H, Nishikawa K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol, 1994, 238(1): 54–61.

[22] Aharon M, Elad M, Bruckstein A. *rmK-SVD*: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans Signal Process, 2006, 54(11): 4311–4322.

[23] Liu YH, Cheng JY, Ma YM, et al. Protein secondary structure prediction based on two dimensional deep convolutional neural networks//2017 3rd IEEE International Conference on Computer and Communications. Chengdu, China: IEEE, 2017: 1995–1999.

[24] Chen YH. Long sequence feature extraction based on deep learning neural network for protein secondary structure prediction//2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference. Chongqing, China: IEEE, 2017: 843–847.

[25] Silva MFM, Leijoto LF, Nobre CN. Algorithms analysis in adjusting the SVM parameters: an approach in the prediction of protein function. Appl Artif Intell, 2017, 31(4): 316–331.

[26] Ding H, Liang ZY, Guo FB, et al. Predicting bacteriophage proteins located in host cell with feature selection technique. Comput Biol Med, 2016, 71: 156–161.

[27] Xu YY, Yao LX, Shen HB. Bioimage-based protein subcellular location prediction: a comprehensive review. Front Comput Sci, 2018, 12(1): 26–39.

[28] Liang YY, Liu SY, Zhang SL. Detrended cross-correlation coefficient: application to predict apoptosis protein subcellular localization. Math Biosci, 2016, 282: 61–67.

[29] Chen HW, Chen X, Hu QM, et al. Predicting protein subcellular location based on a novel sequence numerical model. J Comput Theor Nanosci, 2015, 12(1): 82–87.

(本文责编 陈宏宇)