计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200200081

基于聚类与特征融合的蛋白质亚细胞定位预测

王艺皓　丁洪伟　李　波　保利勇　张颖婕
云南大学信息学院　昆明 650500
(893885847@qq.com)

摘　要　蛋白质亚细胞的定位预测不仅是研究蛋白质结构和功能的重要基础,还对了解某些疾病的发病机理、药物设计与发现具有重要意义。然而,如何利用机器学习精准预测蛋白质亚细胞的位置一直是一项具有挑战性的科学难题。针对这一问题,提出了一种基于聚类与特征融合的蛋白质亚细胞定位方法。首先将自相关系数法和熵密度法引入蛋白质特征表达模型的构建,并在传统的 PseAAC(Pseudo-amino Acid Composition)的基础上提出了一种改进型 PseAAC 方法。为了更好地表达蛋白质序列信息,文中首先将自相关系数法、熵密度法和改进型 PseAAC 进行融合,构造了一种全新的蛋白质序列表征模型;然后利用主成分分析法对融合后的特征向量进行降维,将结果输入到 LibD3C 集成分类器,对蛋白质亚细胞进行分类预测,并采用留一法在 Gram-positive 和 Gram-negative 数据集上进行交叉检验;最后将取得的实验结果与其他现有算法进行比较。实验结果表明,所提方法在 Gram-positive 和 Gram-negative 数据集上分别取得了 99.24%和 95.33%的预测准确率,说明所提方法具有科学性和有效性。

关键词:特征融合;聚类;自相关系数;伪氨基酸组分法;主成分分析法
中图法分类号　TP391

Prediction of Protein Subcellular Localization Based on Clustering and Feature Fusion

WANG Yi-hao,DING Hong-wei,LI Bo,BAO Li-yong and ZHANG Ying-jie
School of Information Science and Engineering,Yunnan University,Kunming 650500,China

Abstract　The prediction of protein subcellular location is not only an important basis for the study of protein structure and function,but also of great significance for understanding the pathogenesis of some diseases,drug design and discovery. However,how to use machine learning to accurately predict the location of protein subcellular has always been a challenging scientific problem. To solve this problem,this paper proposes a protein subcellular localization method based on clustering and feature fusion. Firstly,autocorrelation coefficient method and entropy density method are introduced into the construction of protein feature expression model,and an improved PseAAC(Pseudo-amino acid composition) method is proposed on the basis of traditional PseAAC. In order to express protein sequence information better,this paper fuses autocorrelation coefficient method,entropy density method and the improved PseAAC to construct a new protein sequence representation model. Secondly,we use principal component analysis (PCA) to reduce the dimension of the fused feature vector. Thirdly,we adopt the LibD3C ensemble classifier to classify and predict protein subcellular,and the prediction accuracy is evaluated by leave-one-out cross validation on Gram-positive and Gram-negative datasets. Finally,the experimental results are compared with other existing algorithms. The results show that the new method achieves the prediction accuracy of 99.24% and 95.33% on Gram-positive and Gram-negative datasets respectively,and the new method is scientific and effective.

Keywords　Feature fusion,Clustering,Autocorrelation coefficient,Pseudo-amino acid composition,Principal component analysis

1　引言

随着破译生命密码的人类基因组计划(Human Genome Project,HGP)的不断推进和完善,生命科学进入了后基因组时代,大量的生物基因信息被不断发掘,蛋白质序列更是呈现指数式的增长。然而,通过传统的生物实验方法获取数据,不但过度消耗实验成本,造成资源浪费,效率低下,而且处理速度远远落后于数据的增长速度[1]。因此,需要通过生物信息学方法来提高相关数据分析的效率。其中,蛋白质亚细胞的定位预测是蛋白质组学和蛋白质功能研究的基础,更是生物信息学的关键环节,因此对蛋白质亚细胞定位的研究具有重要意义。

到稿日期:2020-02-16　返修日期:2020-05-21　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金项目(61461053,61461054)
This work was supported by the National Natural Science Foundation of China(61461053,61461054).
通信作者:丁洪伟(dhw1964@163.com)

一般来讲，蛋白质亚细胞的定位预测分为两个重要环节，即构建合适的特征表达模型和选取有效的分类预测方法。目前构建表征蛋白质序列信息的特征表达模型主要有以下几种：氨基酸组成成分法（Amino Acid Composition，AAC）[2]、伪氨基酸组分法（Pseudo-Amino Acid Composition，PseAAC）[3]、基因本体（Gene Ontology，GO）[4]、自相关系数法（Autocorrelation Coefficient Functions，ACF）[5] 等。Chen 等[2]利用 AAC 法提取蛋白质序列特征，并结合多层次稀疏编码对蛋白质亚细胞进行定位预测，取得了较好的预测结果。这种 AAC 法虽然充分考虑了氨基酸的组成信息，且简单易操作，但它忽略了各类氨基酸的顺序信息以及它们之间的相互作用[6] 对序列产生的影响。Chou 等[7] 提出了一种 PseAAC 法，在 AAC 的基础上将氨基酸的 3 种理化性质考虑在内，提升了定位预测的准确率，但是由于亚细胞种类的多样性和复杂性，该方法的研究工作还有极大的上升空间。Chou 等[8] 提出了功能域结合的方法，将 GO 和 PseAAC 相结合来构建特征表达模型。Li 等[9] 在此基础上融合了加权自相关函数等信息，进一步优化了特征提取算法。

对于蛋白质亚细胞定位预测的分类算法，目前比较常用的主要有以下几种：K 近邻算法（K-Nearest Neighbor，KNN）[10]、支持向量机（Support Vector Machine，SVM）[11]、随机森林（Rand Forest，RF）[12]、贝叶斯网络（Bayesian Network）[13]、集成学习[14] 和深度学习[15] 等。其中，集成学习方法具有强大的泛化能力和良好的鲁棒性，因此被广泛应用于多标签分类任务中。

通过传统的生物实验方法来获取信息存在耗时费力、效率低下的缺点，而且基于单一特征的提取方法存在一定的局限性，即会造成蛋白质序列部分信息表达不完整或丢失。本文针对这些问题，提出了一种基于聚类与特征融合的蛋白质亚细胞定位预测方法。首先将自相关系数法和熵密度法引入蛋白质序列特征表达模型的构建中，并在传统的 PseAAC 的基础上加入了 12 种氨基酸的理化性质，进而提出了改进型 PseAAC 法。为了更好地表达蛋白质的序列信息，本文融合了自相关系数、熵密度和改进型 PseAAC，从而构建了一种全新的蛋白质序列信息特征提取模型。接着通过 PCA 算法[16] 对融合后的特征向量进行降维，然后输入 LibD3C[17] 集成分类器进行定位预测，并采用留一法[18] 分别在 Gram-positive 和 Gram-negative 数据集上进行交叉验证，最后将本文提出的新方法与自相关系数法、改进型 PseAAC 法以及其他现有算法的实验结果进行比较。实验结果表明，本文方法可以有效提升蛋白质亚细胞定位预测的准确性。

2 蛋白质定位预测新方法

2.1 构建蛋白质序列特征提取模型

2.1.1 自相关系数

传统的特征表达模型以 AAC 为主，这种方法虽然考虑了各类氨基酸的组成信息和出现频率，但并未考虑到蛋白质序列中氨基酸的排列顺序以及耦合信息的影响。相比传统的蛋白质特征表达模型，ACF 加入了氨基酸的位置信息和不同距离氨基酸之间的相互影响，从而更加真实地表达出了蛋白质序列的特征信息。本文选取蛋白质的 15 种理化特征[19] 来表达蛋白质序列，分别用 $H_1(R_i)$，$H_2(R_i)$，…，$H_{15}(R_i)$ 表示，原始的特征值如表 1 所列，其缩略词的对应关系如表 2 所列。

表 1 原始特征值

Table 1 Original eigenvalue

Code	$H_1(R_i)$	$H_2(R_i)$	$H_3(R_i)$	$H_4(R_i)$	$H_5(R_i)$	$H_6(R_i)$	$H_7(R_i)$	$H_8(R_i)$	$H_9(R_i)$	$H_{10}(R_i)$	$H_{11}(R_i)$	$H_{12}(R_i)$	$H_{13}(R_i)$	$H_{14}(R_i)$	$H_{15}(R_i)$
A	−0.40	−0.5	15	8.1	0.046	0.67	1.28	0.3	0	0.687	115	0.28	27.5	1.181	0.0072
C	0.17	−1.0	47	5.5	0.128	0.38	1.77	0.9	2.75	0.263	135	0.28	44.6	1.461	−0.0370
D	−1.31	3.0	59	13.0	0.105	−1.20	1.60	−0.6	1.38	0.632	150	0.21	40.0	1.587	0.0238
E	−1.22	3.0	73	12.3	0.151	−0.76	1.56	−0.7	0.92	0.669	190	0.33	62.0	1.862	0.0068
F	1.92	−2.5	91	5.2	0.290	2.30	2.94	0.5	0	0.577	210	2.18	115.5	2.228	0.0376
G	−0.67	0	1	9.0	0	0	0	0.3	0.74	0.670	95	0.18	0	0.881	0.1791
H	−0.64	−0.5	82	10.4	0.230	0.64	2.99	−0.1	0.58	0.594	195	0.21	79.0	2.025	−0.0110
I	1.25	−1.8	57	5.2	0.186	1.90	4.19	0.7	0	0.564	175	0.82	93.5	1.810	0.0216
K	−0.67	3.0	73	11.3	0.219	−0.57	1.89	−1.8	0.33	0.407	200	0.09	100.0	2.258	0.0177
L	1.22	−1.8	57	4.9	0.186	1.90	2.59	0.5	0	0.541	170	1.00	93.5	1.931	0.0517
M	1.02	−1.3	75	5.7	0.221	2.40	2.35	0.4	0	0.328	185	0.74	94.1	2.034	0.0027
N	−0.92	0.2	58	11.6	0.134	−0.61	1.60	−0.5	1.33	0.489	160	0.25	58.7	1.655	0.0054
P	−0.49	0	42	8.0	0.131	1.20	2.67	−0.3	0.39	0.600	145	0.39	41.9	1.468	0.2395
Q	−0.91	0.2	72	10.5	0.180	−0.22	1.56	−0.7	0.90	0.527	183	0.35	80.7	1.932	0.0492
R	−0.59	3.0	101	10.5	0.291	−2.10	2.34	−1.4	0.64	0.591	225	0.10	105.0	2.560	0.0436
S	−0.55	0.3	31	9.2	0.062	0.01	1.31	−0.1	1.41	0.693	116	0.12	29.3	1.298	0.0043
T	−0.28	−0.4	45	8.6	0.108	0.52	3.03	−0.2	0.71	0.713	142	0.21	51.3	1.525	0.0034
V	0.91	−1.5	43	5.9	0.140	1.50	3.67	0.6	0	0.529	157	0.60	71.5	1.645	0.0570
W	0.50	−3.4	130	5.4	0.409	2.60	3.21	0.3	0.12	0.632	258	5.70	145.5	2.663	0.0380
Y	1.67	−2.3	107	6.2	0.298	1.60	2.94	−0.4	0.21	0.495	234	1.26	117.3	2.368	0.0236

由表 1 中的数据可以看出，各类特征的原始值差异较大，因此需要对这些原始数据进行归一化处理。首先根据最大最小规范化原则，采用式(1)将表 1 中的数据放缩到[−1,1]中：

$$H(R_i)=\frac{h^0(R_i)-[h^0(R_i)]_{\min}}{[h^0(R_i)]_{\max}-[h^0(R_i)]_{\min}} \tag{1}$$

其中，$H(R_i)$ 表示归一化后的特征值，$h^0(R_i)$ 表示第 i 种氨基酸对应的某一特征的原始特征值，$[h^0(R_i)]_{\max}$ 和 $[h^0(R_i)]_{\min}$ 分别对应该特征的最大和最小原始特征值。

接着将一条长度为 L 的蛋白质序列中的各氨基酸符号进行数值化处理，即将各氨基酸编码符号用其对应的离散数值序列来表示：

$$H=h_1,h_2,\cdots,h_L \tag{2}$$

其中，h_i 表示第 i（$i=1,2,\cdots,20$）个氨基酸对应的数值。然后，通过式(3)来计算离散数值序列 H 的自相关函数：

$$r_n=\frac{1}{L-n}\sum_{i=1}^{L-n}h_i+h_{i+n},n=1,2,\cdots,\phi \tag{3}$$

其中，ϕ 为自相关系数的阶数，满足 $\phi<L$，表示相隔距离为 ϕ 的两个氨基酸之间的相关特性，反映了蛋白质序列的局部物化特性。$\phi=1$ 代表相邻氨基酸之间的特性；$\phi=2$ 代表间隔距离为 2 的氨基酸之间的特性。因此，每类氨基酸的各种理化特征都可以用 ϕ 维自相关系数的特征向量来表示，即：

$$R=(k_1,k_2,\cdots,k_\phi)^{\mathrm{T}} \tag{4}$$

本文选取了 15 种氨基酸的理化特性，则可以构建 $15*\phi$ 维自相关系数的特征向量，即：

$$W_k=\begin{pmatrix} k_{1,1} & k_{1,2} & \cdots & k_{1,\varphi} \\ k_{2,1} & k_{2,2} & \cdots & k_{2,\varphi} \\ \vdots & \vdots & \vdots & \vdots \\ k_{15,1} & k_{15,2} & \cdots & k_{15,\varphi} \end{pmatrix} \tag{5}$$

实验中取 $\phi=45$，通过 ACF，每条蛋白质序列可以转化成一个 $675(15*45)$ 维的特征向量。

表 2 特征值符号对应的内容

Table 2 Contents corresponding to eigenvalue symbols

Abbreviation	Original word
$H_1(R_i)$	Hydrophobicity
$H_2(R_i)$	Hydrophilicity
$H_3(R_i)$	Side chain molecular weight
$H_4(R_i)$	polarity
$H_5(R_i)$	Polarizability
$H_6(R_i)$	Solvation free energy
$H_7(R_i)$	Curve shape index
$H_8(R_i)$	Transfer free energy
$H_9(R_i)$	Amino acid composition
$H_{10}(R_i)$	Regression analysis correlation coefficient
$H_{11}(R_i)$	Residue accessible surface
$H_{12}(R_i)$	Partition coefficient
$H_{13}(R_i)$	Amino acid side chain volume
$H_{14}(R_i)$	Surface area dissolving ability
$H_{15}(R_i)$	Network Load Index

2.1.2 熵密度

生物信息学中存在着许多结构复杂的组合体系，由于信息熵在提取有效特征信息方面有着良好的表现，因此本文引入熵密度来进一步丰富蛋白质序列信息的表达。

若想利用熵密度法进行特征提取，则首先通过计算得出蛋白质序列中 20 种氨基酸残基的信息熵，即：

$$H(P)=-\sum_{i=1}^{20}F_i\log F_i \tag{6}$$

其中，F_i 表示第 i 种氨基酸出现在蛋白质序列 P 中的频率。由信息熵计算公式可得，第 i 种氨基酸对应的熵密度可以定义为：

$$E_i(P)=-\frac{1}{H(P)}F_i\log F_i \tag{7}$$

由此，每个蛋白质序列 P 可以由一个 20 维的熵密度特征向量来表示，即：

$$W_E=[E_1(P),E_2(P),\cdots,E_{20}(P)]^{\mathrm{T}} \tag{8}$$

2.1.3 改进型伪氨基酸组成模型

Chou 等提出了 PseAAC 算法[20]，即结合氨基酸组成和 λ 阶相关因子来共同表达序列信息。在此方法中，每条序列都用 $20+\lambda$ 维向量来表征：前 20 维是各类氨基酸出现在蛋白质序列中的频率，λ 维为加入的序列次序效应的相关因子。其中，对于 PseAAC 的紧邻相关因子 $J_{i,i+k}$，仅考虑了疏水性、亲水性和侧链分子量 3 种理化特性对蛋白质的影响。本文在此基础上又加入了 12 种理化特性，即极性、极化率、溶剂化自由能、曲线形状指数、转移自由能、氨基酸组分、回归分析相关系数、残基可及表面、分配系数、氨基酸边链体积、表面区域溶解能力和网络负荷指数，进而提出了一种改进型 PseAAC 算法。其理化特性对应的特征值如表 1 所列。

由改进型 PseAAC 算法可得，每条蛋白质序列可用式(9)来表示，其中每个元素 p_u 可由式(10)求出：

$$W_{PseAAC}=[p_1,p_2,p_3,\cdots,p_{20},p_{20+1},\cdots,p_{20+\lambda}]^{\mathrm{T}} \tag{9}$$

$$p_u=\begin{cases} \dfrac{f_u}{\sum\limits_{i=1}^{20}f_i+\omega\sum\limits_{j=1}^{\lambda}\gamma_j}, & 1\leqslant u\leqslant 20 \\ \dfrac{\omega\gamma_{u-20}}{\sum\limits_{i=1}^{20}f_i+\omega\sum\limits_{j=1}^{\lambda}\gamma_j}, & 20+1\leqslant u\leqslant 20+\lambda \end{cases} \tag{10}$$

其中，f_u 表示每种氨基酸在蛋白质序列中出现的概率，ω 是权重因子，本文默认取值为 0.05，γ_j 可由式(11)求得：

$$\gamma_k=\frac{1}{L-k}\sum_{i=1}^{L-k}J_{i,i+k},\ 1\leqslant k\leqslant 15 \tag{11}$$

$J_{i,i+k}$ 为相关函数，其定义为：

$$J_{i,i+k}=\frac{1}{15}\{[H_1(R_{i+k})-H_1(R_i)]^2+\cdots+[H_{15}(R_{i+k})-H_{15}(R_i)]^2\} \tag{12}$$

其中，$H_1(R_i)$，$H_2(R_i)$，…，$H_{15}(R_i)$ 分别表示氨基酸残基 R_i 的 15 种理化特性的索引值。$H(R_i)$ 可由式(13)求得：

$$H(R_i)=\frac{h^0(R_i)-average(h^0)}{\nu(h^0)} \tag{13}$$

其中，$h^0(R_i)$ 表示氨基酸 R_i 对应物化性质的原始特征值，$average(h^0)$ 表示对应物化性质特征下 20 种氨基酸原始特征值的平均值，$\nu(h^0)$ 表示其对应的方差。

由于 λ 的取值会对最终的分类结果产生影响，因此在本文实验中，分别取 $\lambda=1,\cdots,\lambda=30$，将其代入模型输入支持向量机在数据集上进行实验，并采用留一法进行检验，通过比较，最终选取 $\lambda=14$。因此，通过改进型 PseAAC 算法，每条蛋白质序列可以转化为一个 34 维的特征向量。

2.1.4 多信息融合特征表达模型

本文基于多特征融合的思想，将前文提到的 ACF、熵密度法和改进型 PseAAC 相结合，构成了一种全新的蛋白质序列信息特征提取模型。该新模型的特征向量中既包含了氨基酸组成信息，又充分考虑了氨基酸的排列顺序和耦合信息对序列的影响。因此，每条蛋白质序列可用式(14)来表达：

$$Z=(Z_1,Z_2,\cdots,Z_{20},\cdots,Z_{20+\lambda},\cdots,Z_{20+20+\lambda},\cdots,Z_{20+20+\lambda+15*\varphi}) \tag{14}$$

其中，前 $20+\lambda$ 维是改进型 PseAAC 模型提取的特征向量，第 21 维到第 40 维是利用熵密度法提取的特征向量，最后 $15*\varphi$ 维是通过 ACF 提取的特征向量。将 $\lambda=14$，$\varphi=45$ 代入式(14)，则每条蛋白质序列可以用 729 维的特征向量表示。

2.2 PCA 降维

本文实验中采用了多信息融合特征表达模型，虽然这一做法丰富了蛋白质序列的特征信息，但同时也带来了较多的冗余信息和噪声。为了避免维度过高对分类预测准确性的影响，本文引入了 PCA 降维算法，以获取融合特征中更有效的特征向量，进而达到提高分类预测准确性的目的。主成分分析法[16](PCA)是目前应用得较为广泛的线性降维算法之一，其基本思想是将可能具有相关性的高维变量合成线性无关的

低维向量，新得到的低维向量会尽可能地保留原始数据的变量信息。经过 PCA 降维，最大程度地剔除了冗余信息并保留了主要信息，从而有效降低了算法的复杂度。

2.3 LibD3C

LibD3C 是 Lin 等[17]提出的一种基于 k-means 聚类和动态选择与循环集成的混合模型的集成分类器算法框架，但是由于 k-means 算法的优化结果严重依赖于随机初始化的结果和聚类个数 k 的先验知识，而且在实验过程中发现，k-means 算法并不适用于弱分类器的筛选，因此本文进一步借鉴了文献[21]来改进该算法，并采用基于图模型的聚类算法 Affinity Propagation[22]替换 k-means 算法。与 k-means 算法不同的是，Affinity Propagation 算法不需要先验性地指定聚类个数，即所有样本点均可以视为潜在的聚类中心，同时这一做法也避免了由于随机初始化聚类中心所带来的影响。图 1 为改进后的 LibD3C 算法的流程图。

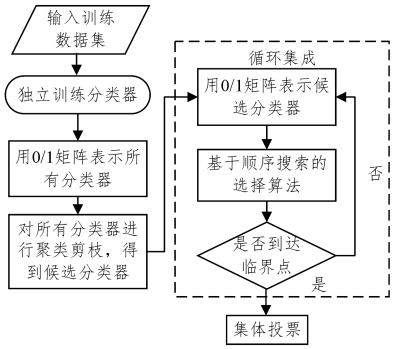

图 1 改进后的 LibD3C 算法的流程图

Fig. 1 Flowchart of improved LibD3C algorithm

改进后的 LibD3C 算法主要由两层模型组成。第一层模型是基于 Affinity Propagation 聚类进行预筛选，并通过并行的方式训练出多种弱分类器。聚类筛选的目的是利用 Affinity Propagation 算法过滤掉冗余的弱分类器，在保证分类准确性的前提下进一步提高集成分类器的预测速度。假设 $C=\{c_1,\cdots,c_n\}$ 为 n 个样本，$d(i,j)$ 表示 c_i 和 c_j 之间的距离或相似度，$\vartheta(i)$ 表示样本 c_i 聚类中心点的下标，则 Affinity Propagation 算法的优化目标为：

$$C[\vartheta]=\sum_{i=1}^{n}\alpha(c_1,c_{\vartheta(i)})-\sum_{i=1}^{n}\varphi_i[\delta] \tag{15}$$

其中，$\alpha(c_1,c_{\vartheta(i)})=\begin{cases}-d(i,j), & i\neq j\\ -\alpha^*, & i=j\end{cases}$ 且偏好系数 $\alpha^*\geqslant 0$，$\varphi_i[\delta]=\begin{cases}\infty, & \delta(\delta(i))\neq\delta(i)\\ 0, & \delta(\delta(i))=\delta(i)\end{cases}$。

第二层模型是动态选择与循环集成，将相互一致性度量 k 作为多样性度量方式。其关键步骤在于确定分类器个数阈值并记录局部最优分类器组合。假设训练集为 $T=\{(\vec{x}_1,y_1),\cdots,(\vec{x}_m,y_m)\}$，则非成对多样性度量 k 可用式(16)表示：

$$k=1-\frac{1}{2\,\vec{p}(1-\vec{p})}Dis_{average} \tag{16}$$

其中，$\vec{p}$ 为 t 个分类器的平均精度，$Dis_{average}$ 表示平均差异性度量，$\vec{p}$ 和 $Dis_{average}$ 分别由式(17)和式(18)计算得出：

$$\vec{p}=\frac{1}{mt}\sum_{j=1}^{m}\sum_{i=1}^{t}h_i(\vec{x}_j,y_j) \tag{17}$$

$$Dis_{average}=\frac{t}{t(t-1)}\sum_{i=1}^{t}\sum_{\substack{k=1\\ i\neq k}}^{t}Dis_{i,k} \tag{18}$$

其中，$h_i(\vec{x}_j,y_j)$ 表示正确或者错误的决定。

由于本文选用的两个数据集中均包含多个位点蛋白，因此此次蛋白质亚细胞定位预测任务在本质上也属于多标签分类问题。图 2 为 LibD3C 算法用于多标签分类的过程图。

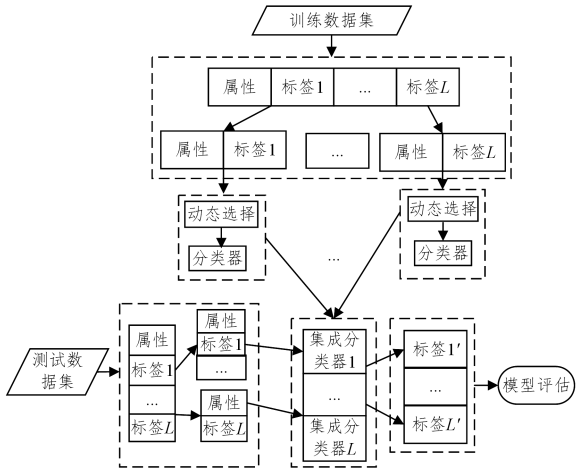

图 2 LibD3C 算法用于多标签分类的过程图

Fig. 2 Process diagram of LibD3C algorithm for multi-label classification

从图 2 可以看出，LibD3C 算法在本质上就是将多标签分类问题转化为多个单标签分类问题。本文所用的 LibD3C2.0 来自 Weka 平台，可以通过 URL 在线获取 LibD3C.zip 进行集成[1)]。

2.4 性能检验和评估指标

目前模型性能检测的方法主要有：K 折交叉验证[23]、自相容检验[24]、独立性检验[25]、留一法(leave-one-out cross validation)[18]等。留一法也称为 n 折交叉验证(n 是数据集中样本的数目)，其主要思想是从数据集中依次选择每一条蛋白质序列作为测试集，其余序列均作为训练集，用于训练分类器，直到完成对所有蛋白质序列的测试。由于留一法在每次迭代中都使用了最大可能数目的样本来进行训练，因此该方法得出的结果与训练整个测试集的期望值最接近。留一法[26]被认为是最客观、最严格的检测方法之一，因此留一法已被广泛应用于检测各种蛋白质亚细胞定位预测模型的性能[27]。本文将采用留一法对模型分类预测性能进行验证。

对于模型性能评估，本文采用 4 个指标，分别如下。

(1)特异性(Specificity)：

$$Spec=\frac{TP_i}{TP_i+FP_i} \tag{19}$$

(2)灵敏性(Sensitivity)：

$$Sens=\frac{TP_i}{TP_i+FN_i} \tag{20}$$

(3)整体准确率(Overall Accuracy，OA)：

$$OA=\frac{TP_i+TN_i}{N} \tag{21}$$

1) http://datamining.xmu.edu.cn/main/~chenwq/downloads/LibD3C.zip

2) http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc/

(4)马修斯相关系数(Matthews Correlation Coefficient, MCC):

$$MCC=\frac{TP_i\times TN_i-FP_i\times FN_i}{\sqrt{(TP_i+FP_i)\times(TP_i+FN_i)\times(TN_i+FN_i)\times(TN_i+FP_i)}} \tag{22}$$

其中,TP_i 表示第 i 类亚细胞位点中预测正确的样本数,TN_i 表示除了第 i 类亚细胞位点以外的其他类别预测正确的样本数,FP_i 表示其他类别被错误预测为第 i 类亚细胞位点的样本数,FN_i 表示第 i 类亚细胞位点被错误预测为其他类别的样本数。

3 实验结果与分析

3.1 数据集

本文选用两个革兰氏细菌数据集[2)]。其中,数据集 Gram-positive 共包含 523 条蛋白质序列,有 4 个亚细胞位点标签;数据集 Gram-negative 共包含 1 456 条蛋白质序列,有 8 个亚细胞位点标签。数据集 Gram-positive 和 Gram-negative 的详细信息如表 3 所列。

表 3 实验中用到的数据集 Gram-positive 和 Gram-negative

Table 3 Gram-positive and Gram-negative datasets used in experiments

Subcellular locus tag	Gram-positive		Gram-negative	
	Protein subcellular class	Number of protein sequences	Protein subcellular class	Number of protein sequences
1	Cytoplasmic proteins	208	Extracellular proteins	133
2	Extracellular proteins	123	Flagellum proteins	12
3	Cell wall proteins	18	Outer membrane proteins	124
4	Cell membrane proteins	174	Fimbrium proteins	32
5			Cytoplasm proteins	410
6			Periplasm proteins	180
7			Nucleoid proteins	8
8			Inner membrane proteins	557
	total	523		1 456

3.2 实验结果与分析

3.2.1 特征融合算法性能分析

本文采用来自 Weka 平台集成的 LibD3C2.0 软件包进行实验。首先在 Gram-positive 和 Gram-negative 数据集上分别用 ACF、改进型 PseAAC 法和多特征融合法进行实验,并使用留一法进行验证,实验结果如图 3 所示。

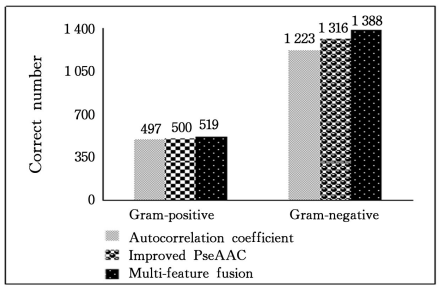

图 3 3 种方法在 Gram-positive 和 Gram-negative 数据集上的预测结果比较

Fig. 3 Prediction results comparison of 3 methods on Gram-positive and Gram-negative datasets

由图 3 可知,本文在传统的 PseAAC 算法的基础上又加入了 12 种氨基酸理化特性,这一做法极大地丰富了所提取的蛋白质序列中包含的氨基酸信息,由此提出的改进型 PseAAC 算法相比传统 PseAAC 算法而言,在两个数据集上均取得了更高的预测准确率,这充分验证了该方法的有效性。改进型 PseAAC 法在数据集 Gram-positive 上的预测准确数为 500,在数据集 Gram-negative 上的预测准确数为 1 316;而本文的多特征融合法在改进型 PseAAC 的基础上又结合了熵密度法和 ACF,因此取得了更好的分类预测结果,即在数据集 Gram-positive 上的预测准确数达到了 519,准确率接近 99.24%,而在数据集 Gram-negative 上的预测准确数达到了 1 388,准确率达到 95.33%,均高于 ACF 和改进型 PseAAC 的预测结果。同时,这也证明了多特征融合方法的科学性和先进性。

在数据集 Gram-positive 上进行进一步的实验,3 种方法对各位点蛋白的预测结果如图 4 所示。

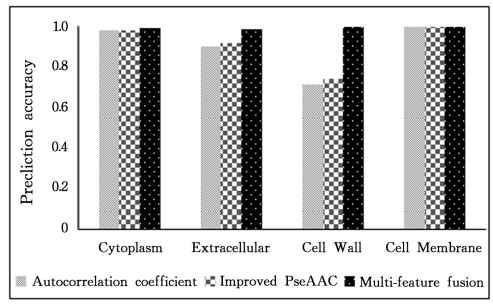

图 4 3 种方法在 Gram-positive 数据集上的预测准确率比较

Fig. 4 Prediction accuracy comparison of 3 methods on Gram-positive dataset

由图 4 可知,对于 Gram-positive 数据集,除了在 Cytoplasm 和 Cell membrane 两个位点上 ACF 和改进型 PseAAC 法的预测准确率基本持平外,在其他位点上改进型 PseAAC 法均取得了比自相关系数法更高的准确率。而本文提出的多特征融合法在任何位点上的预测准确率都是三者中最高的,尤其在 Cell wall 位点上,其预测准确率相比另外两种方法有明显的提升,甚至在 Cell wall 和 Cell membrane 位点上取得了高达 100%的预测准确率。

在数据集 Gram-negative 中,3 种方法在各位点蛋白上的预测结果如图 5 所示。

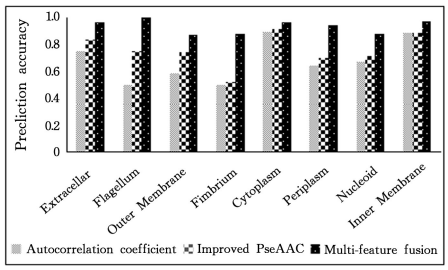

图 5 3 种方法在 Gram-negative 数据集上的预测准确率比较

Fig. 5 Prediction accuracy comparison of 3 methods on Gram-negative dataset

由图 5 可知，对于 Gram-negative 数据集，除了在 Inner Membrane 位点上改进型 PseAAC 法相比 ACF 的预测准确率提升幅度较小外，在其他位点上改进型 PseAAC 法的预测准确率相比 ACF 均有明显的提高。从图 5 中可以明显看出，本文提出的多特征融合法在各位点的预测准确率相比另外两种方法均有明显的提高，甚至在 Flagellum 位点取得了 100% 的预测准确率。

为了更加客观全面地对本文方法的预测效果进行评价，引入了灵敏性(Sens)、特异性(Spec)、整体准确率(OA)和马修斯相关系数(MCC)4 个指标，并将 3 种方法分别在 Gram-positive 和 Gram-negative 数据集上取得的实验结果进行比较。实验结果的详细对比如表 4、表 5 所列。

表 4 自相关系数法、改进型 PseAAC 和本文方法在 Gram-positive 数据集上的实验结果对比

Table 4 Experimental results comparison of autocorrelation coefficient method, improved PseAAC and the proposed method on Gram-positive dataset

(单位：%)

Index	Autocorrelation coefficient				Improved PseAAC				Proposed Method			
	Sens	Spec	MCC	OA	Sens	Spec	MCC	OA	Sens	Spec	MCC	OA
Cytoplasm	97.95	93.88	94.52	95.13	97.86	94.75	93.21	95.58	99.34	99.87	99.65	99.24
Extracellular	89.72	97.18	90.56		91.59	97.98	92.13		100	99.59	95.73	
Cell wall	71.42	95.20	88.72		73.78	96.08	87.36		100	99.24	99.47	
Cell Membrane	100	93.02	94.03		100	93.76	94.05		100	98.86	97.92	

表 5 自相关系数法、改进型 PseAAC 和本文方法在 Gram-negative 数据集上的实验结果对比

Table 5 Experimental results comparison of autocorrelation coefficient method, improved PseAAC and the proposed method on Gram-negative dataset

(单位：%)

Index	Autocorrelation coefficient				Improved PseAAC				Proposed Method			
	Sens	Spec	MCC	OA	Sens	Spec	MCC	OA	Sens	Spec	MCC	OA
Extracellular	75.00	84.14	79.83	84.72	83.33	86.72	84.79	90.43	96.62	91.75	94.43	95.33
Flagellum	50.00	83.97	75.38		74.84	85.48	83.27		100	91.52	97.78	
Outer Membrane	58.62	85.74	73.82		74.14	87.45	80.28		87.92	90.27	88.64	
Fimbrium	50.00	84.04	76.24		51.79	86.25	80.72		77.45	90.51	84.81	
Cytoplasm	88.95	83.26	83.81		91.16	85.75	88.69		95.31	89.75	92.15	
Periplasm	64.02	85.63	78.42		69.57	86.86	80.93		87.02	91.17	88.12	
Nucleoid	66.67	83.91	73.95		71.49	86.52	80.82		99.68	92.56	94.97	
Inner Membrane	88.02	80.52	85.46		88.64	86.52	87.42		90.78	92.97	90.72	

由表 4 可知，在 Gram-positive 数据集中，本文方法的整体预测准确率达到了 99.24%，相比 ACF 的 95.13%和改进型 PseAAC 法的 95.58%，本文方法提升了 3%～4%。而且新方法的灵敏性、特异性和马修斯相关系数的指标也优于另外两种方法，尤其是在 Extracellular，Cell wall 和 Cell Membrane 位点的灵敏性指标达到了 100%。

由表 5 可知，在 Gram-negative 数据集中，本文方法的整体预测准确率由 ACF 的 84.72%提升到了 95.33%，提升了 10%。特别地，其在 Flagellum 位点的灵敏性指标由自相关系数法的 50%上升到了 100%，在其他位点上的各项指标也是碾压 ACF 和改进型 PseAAC 法。综上，对表 4、表 5 的分析结果充分说明了构建合理的多特征融合模型可以明显提升亚细胞定位预测的准确性。

3.2.2 分类器性能分析

本文主要对目前用于蛋白质亚细胞定位任务的 4 种表现较好的分类器进行了实验和对比。这 4 种分类器分别为：LibD3C、朴素贝叶斯(NB)、支持向量机(SVM)以及随机森林(RF)。首先将融合后的特征向量分别输入 LibD3C，NB，SVM 和 RF 分类器中，采用留一法在 Gram-positive 和 Gram-negative 两个数据集上进行交叉验证，输出灵敏性、特异性、马修斯相关系数和整体准确率 4 个指标，来评估 4 种分类算法的性能。其中，LibD3C，NB 和 RF 均采用默认参数；SVM 中的核函数选择高斯核函数。在不同分类器下，Gram-positive 和 Gram-negative 两个数据集得到的预测结果如表 6 和表 7 所列。

表 6 4 种分类算法在 Gram-positive 数据集上的预测结果对比

Table 6 Prediction results comparison of 4 classification algorithms on Gram-positive dataset

(单位：%)

Classifiers	jackknife cross-validation test			
	Sens	Spec	MCC	OA
NB	93.31	95.52	88.34	92.93
SVM	99.83	98.87	97.83	98.89
RF	97.98	98.89	97.21	99.04
LibD3C	97.85	98.95	98.02	99.24

表 7 4 种分类算法在 Gram-negative 数据集上的预测结果对比

Table 7 Prediction results comparison of 4 classification algorithms on Gram-negative dataset

(单位:%)

Classifiers	jackknife cross-validation test			
	Sens	Spec	MCC	OA
NB	86.13	87.92	82.27	85.78
SVM	91.83	93.26	90.52	94.37
RF	93.86	94.12	93.51	94.92
LibD3C	93.75	95.42	94.12	95.33

由表 6 可以看出,在 Gram-positive 数据集上 LibD3C 的分类性能最好,整体准确率达到了 99.24%,相比 NB 提高了 6%左右,而且除了灵敏性指标略低于 RF 以外,其特异性和马修斯相关系数的指标均高于其他分类器。

由表 7 可以看出,在 Gram-negative 数据集上,通过对比灵敏性、特异性、马修斯相关系数以及整体准确率 4 个评估指标可以发现,LibD3C 的整体表现优于另外 3 种算法。其中,整体准确率达到了 95.33%,比 NB 高了将近 10%。综上,本文选择 LibD3C 可以有效提高蛋白质定位预测的准确性。

3.2.3 与其他算法比较

为了进一步验证本文方法的先进性,将在各位点标签取得的实验结果与其他现有算法的实验结果进行比较,其详细对比结果如表 8、表 9 所列。

表 8 Gram-positive 数据集上不同算法性能的比较结果

Table 8 Performance comparison of different algorithms on dataset Gram-positive

(单位:%)

Subcellular locus tag	iLoc-Gpos [28]	Gneg-ECC-mPLoc[29]	Gram-LocEN[4]	Proposed Method
Cytoplasm	95.2	96.2	97.1	99.0
Extracellular	89.4	92.7	97.1	98.4
Cell wall	66.7	66.7	94.4	100
Cell membrane	96.0	96.5	97.7	100
Overall	93.1	94.4	96.8	99.2

表 9 Gram-negative 数据集上不同算法性能的比较结果

Table 9 Performance comparison of different algorithms on Gram-negative dataset

(单位:%)

Subcellular locus tag	Gneg-PLoc[30]	Gneg-mPLoc[31]	iLoc-Gneg [32]	Proposed Method
Extracellular	44.4	59.4	86.5	96.2
Flagellum	0.0	0.0	100	100
Outer Membrane	54.8	84.7	83.1	87.1
Fimbrium	34.4	87.5	93.8	87.5
Cytoplasm	88.3	87.1	89.5	96.1
Periplasm	48.3	85.6	89.4	94.4
Nucleoid	0.0	0.0	50	87.5
Inner Membrane	81.5	94.3	96.8	96.9
Overall	71.5	85.7	91.4	95.3

由表 8 可知,对于 Gram-positive 数据集,本文方法与现有的 iLoc-Gpos 算法、Gneg-ECC-mPLoc 算法和 Gram-LocEN 算法相比,在各位点的预测准确率均有明显的提升,新方法的总体预测准确率相比现有算法中表现较好的 Gram-LocEN 算法提升了 2.4%;由表 9 可得,对于 Gram-negative 数据集,本文方法与现有的 Gneg-Ploc 算法和 Gneg-mPLoc 算法相比,在各位点的预测准确率都有显著的提高,而与 iLoc-Gneg 算法相比,除了在 Fimbrium 位点的预测准确率有小幅度下降外,在其他位点的预测准确率均有提升,新方法的总体预测准确率比 iLoc-Gneg 算法提高了将近 4%。

综上所述,本文提出的基于聚类与特征融合的蛋白质亚细胞定位预测新方法在 Gram-positive 和 Gram-negative 数据集上均取得了良好的预测结果,这证明了新方法的科学性和有效性,同时也说明了构建合理的序列特征提取模型和选取高性能的分类方法对于提高亚细胞定位预测的准确性是非常重要且有意义的。

结束语 提高蛋白质亚细胞定位预测的准确性一直是蛋白质组学和生物信息学研究的热点问题之一。本文基于多特征融合和集成学习的思想,提出了一种基于聚类和特征融合的蛋白质亚细胞定位方法。首先利用自相关系数法和熵密度法提取特征向量,并基于传统的 PseAAC 提出了一种改进型 PseAAC;接着将自相关系数法、熵密度法和改进型 PseAAC 进行融合,构造了一种全新的蛋白质序列表征模型;然后利用 PCA 算法对融合后的特征向量进行降维,并输入 LibD3C 集成分类器中进行分类预测;最后采用留一法在 Gram-positive 和 Gram-negative 两个数据集上进行交叉检验。通过将本文方法与其他现有算法的实验结果进行比较可知,本文方法有效地提高了蛋白质亚细胞定位预测的准确性。虽然本文方法在实验中取得了良好的预测结果,但是由于本文所用的 LibD3C 分类算法相比其他分类算法耗时较长,导致整个模型的运行时间也有所增加,而且还可以考虑许多其他的方法对蛋白质序列信息的特征进行提取,如进化信息等,因此下一步的研究工作将进一步丰富蛋白质序列表征模型,并对分类器进行优化,保证在提高预测精度的同时也能够提高算法效率。

参 考 文 献

[1] Q1AO S P, YAN B Q. Review of protein subcellular localization prediction[J]. Application Research of Computers, 2014, 31(2): 321-327.

[2] CHEN X J, HU X J, XUE W. Prediction of protein subcellular localization based on multilayer sparse coding[J]. Chinese Journal of Biotechnology, 2019, 35(4): 687-696.

[3] CHOU K C, XIANG C, XUAN X. PLoc_bal-mHum: Predict subcellular localization of human proteins by PseAAC and quasi-balancing training dataset[J]. Genomics, 2019, 111: 1274-1282.

[4] WAN S, MAK M W, KUNG S Y. Gram-LocEN: Interpretable prediction of subcellular multi-localization of Gram-positive and Gram-negative bacterial proteins[J]. Chemometrics and Intelligent Laboratory Systems, 2017, 162: 1-9.

[5] LIU Q H, LAI Y P, DING H W, et al. Protein subcellular localization prediction based on SVM[J]. Computer Engineering and Applications, 2019, 55(11): 136-141.

[6] ZHANG H C, GAO Y J, DENG M H, et al. A survey on algorithms for protein contact prediction[J]. Journal of Computer Research and Development, 2017, 51(1): 1-19.

[7] CHOU K C. Some remarks on protein attribute prediction and pseudo amino acid composition[J]. Journal of theoretical biolo-

gy,2011,273(1):236-247.
[8] CHOU K C,CAI Y D. Predicting protein localization in budding Yeast[J]. Bioinformatics,2005,21(7):944-950.
[9] LI L Z,DONG Z M. Using pseudo amino acid composition to predict protein subcellular localization:approached by incorporating evolutionary conservation information[J]. Acta Biophysica Sinica,2009,25:125-132.
[10] WANG M H,GONG Y,WANG Q,et al. Prediction of protein subcellular localization by incorporating sequence and protein-protein interaction features[J]. Journal of University of Electronic Science and Technology of China,2015,44(3):467-470.
[11] RAHMAN J,MONDAL N I,ISLAM K B,et al. Feature Fusion Based SVM Classifier for Protein Subcellular Localization Prediction[J]. Journal of Integrative Bioinformatics,2016,13(1):23-33.
[12] LI Z C,LAI Y H,CHEN L L,et al. Identifying subcellular localizations of mammalian protein complexes based on graph theory with a random forest algorithm[J]. Mol. Biosyst,2013,9(4):658-667.
[13] HE B,MORTUZA S M,WANG Y,et al. NeBcon:protein contact map prediction using neural network training coupled with naive Bayes classifiers[J]. Bioinformatics,2017,33(15):2296-2306.
[14] CHOU K C,SHEN H B. Hum-PLoc:a novel ensemble classifier for predicting human protein subcellular localization[J]. Biochemical and Biophysical Research Communications,2006,347(1):150-157.
[15] WEI L Y,DING Y J,SU R,et al. Prediction of human protein subcellular localization using deep learning[J]. Journal of Parallel and Distributed Computing,2018,117:212-217.
[16] ZHAO Q. A review of principal component analysis[J]. Software Engineering,2016,19(6):1-3.
[17] LIN C,CHEN W Q,QIU C,et al. LibD3C:Ensemble classifiers with a clustering and dynamic selection strategy[J]. Neurocomputing,2014,123:424-435.
[18] MAO W,MU X,ZHENG Y,et al. Leave-one-out cross-validation-based model selection for multi-input multi-output support vector machine[J]. Neural Computing and Applications,2014,24(2):441-451.
[19] ZHANG Y P,ZHA Y L,ZHAO S,et al. Protein structure class prediction based on autocorrelation coefficient and PseAAC[J]. Journal of Frontiers of Computer Science and Technology,2014,8(1):103-108.
[20] CHOU K C. Prediction of protein cellular attributes using pseudo-amino acid composition[J]. Proteins,2001,43(3):246-255.
[21] CHEN W Q. LibD3C2.0:An Ensemble Classifier Based on Clustering and Its Parallel Implementation[D]. Xiamen:Xiamen University,2014.
[22] FREY B J,DUECK D. Clustering by passing messages between data points[J]. Science,2007,315(5814):972-976.
[23] WONG T T. Parametric methods for comparing the performance of two classification algorithms evaluated by k-fold cross validation on multiple data sets[J]. Pattern Recognition:The Journal of the Pattern Recognition Society,2017,65:97-107.
[24] KROOPNICK M H,CHEN J,CHOI J,et al. Assessing Classification Bias in Latent Class Analysis:Comparing Resubstitution and Leave-One-Out Methods[J]. Journal of Modern Applied Statistical Methods,2010,9(1):52-63.
[25] NEI S Y,LI M H. Construction and comparative analysis of several conditional independence test statistics[J]. The Journal of Quantitative of Quantitative & Technical Economics,2014,31(2):137-147.
[26] CHOU K C,SHEN H B. Cell-PLoc:a package of Web servers for predicting subcellular localization of proteins in various organisms[J]. Nature Protocols,2008,3(2):153-162.
[27] JAVED F,HAYAT M. Predicting subcellular localization of multi-label proteins by incorporating the sequence features into Chou's PseAAC [J]. Genomics,2019,111:1325-1332.
[28] WU Z C,XIAO X,CHOU K C. iLoc-Gpos:a multi-layer classifier for predicting the subcellular localization of single plex and multiplex Gram-positive bacterial proteins [J]. Protein and Peptide Letters,2012,19(1):4-14.
[29] XIAO W,ZHANG J,LI G Z. Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble [J]. BMC Bioinformatics,2015,16(S12):S1.
[30] CHOU K C,SHEN H B. Large-scale predictions of gram-negative bacterial protein subcellular locations[J]. Journal of Proteome Research,2006,5:3420-3428.
[31] SHEN H B,CHOU K C. Gneg-mPLoc:a top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins [J]. Journal of Theoretical Biology,2010,264(2):326-333.
[32] XIAO X,WU Z C,CHOU K C. A multi-label classifier for predicting the subcellular localization of Gram-negative bacterial proteins with both single and multiple sites[J]. PLoS ONE,2011,6(6):e20592.

WANG Yi-hao, born in 1995, postgraduate, is a member of China Computer Federatio. His main research interests include machine lear-ning and computer vision.

DING Hong-wei, born in 1964, Ph. D, professor, Ph. D supervisor. His main research interests include multiple access communication and machine learning.