生 物 工 程 学 报
Chinese Journal of Biotechnology
http://journals.im.ac.cn/cjbcn
DOI: 10.13345/j.cjb.160389

April 25, 2017, 33(4): 683−691

·生物技术与方法·

集成改进 KNN 算法预测蛋白质亚细胞定位

薛卫，王雄飞，赵南，杨荣丽，洪晓宇

南京农业大学 信息科学技术学院，江苏 南京 210095

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.
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.

摘 要：基于 Adaboost 算法对多个相似性比对 K 最近邻 (K-nearest neighbor，KNN) 分类器集成实现蛋白质的亚细胞定位预测。相似性比对 KNN 算法分别以氨基酸组成、二肽、伪氨基酸组成为蛋白序列特征，在 KNN 的决策阶段使用 Blast 比对决定蛋白质的亚细胞定位。在 Jackknife 检验下，Adaboost 集成分类算法提取 3 种蛋白序列特征，3 种特征在数据集 CH317 和 Gram1253 的最高预测成功率分别为 92.4%和 93.1%。结果表明 Adaboost 集成改进 KNN 分类预测方法是一种有效的蛋白质亚细胞定位预测方法。

关键词：亚细胞区间，蛋白序列特征，K-nearest neighbor，basic local alignment search tool，Adaboost

Prediction of protein subcellular locations by ensemble of improved K-nearest neighbor

Wei Xue, Xiongfei Wang, Nan Zhao, Rongli Yang, and Xiaoyu Hong

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

Abstract: Adaboost algorithm with improved K-nearest neighbor classifiers is proposed to predict protein subcellular locations. Improved K-nearest neighbor classifier uses three sequence feature vectors including amino acid composition, dipeptide and pseudo amino acid composition of protein sequence. K-nearest neighbor uses Blast in classification stage. The overall success rates by the jackknife test on two data sets of CH317 and Gram1253 are 92.4% and 93.1%. Adaboost

Received: October 18, 2016; **Accepted:** December 22, 2016
Supported by: Fundamental Research Funds for the Central Universities (No. KYZ201668), Natural Science Foundation of Jiangsu Province (No. BK2012363), National Science and Technology Support Program Project (No. 2015BAK36B05).
Corresponding author: Wei Xue. Tel: +86-21-84396350; E-mail: xwsky@njau.edu.cn
中央高校基本科研业务费专项资金 (No. KYZ201668)，江苏省自然科学基金 (No. BK2012363)，国家科技支撑计划 (No. 2015BAK36B05) 资助。
网络出版时间：2017-01-09 网络出版地址：http://www.cnki.net/kcms/detail/11.1998.Q.20170109.1240.002.html

algorithm with the novel K-nearest neighbor improved by Blast is an effective method for predicting subcellular locations of proteins.

Keywords: subcellular locations, protein sequence characteristics, K-nearest neighbor, basic local alignment search tool (Blast), Adaboost

蛋白质的功能与其所属的亚细胞定位有着紧密的联系，不同蛋白质只有处于特定的亚细胞定位才能发挥其功能，保障生命活动的正常进行，因此对蛋白序列的亚细胞定位预测研究有着重要意义[1]。

利用机器学习实现蛋白质亚细胞定位预测是目前获取定位信息的主要方法，并取得了一系列进展[2-6]。Zhou 等构建凋亡蛋白数据集，在氨基酸组成特征下，利用协变判别函数实现区间预测[7]。Huang 等用支持向量机对氨基酸组成特征实现了对凋亡蛋白的预测[8]。Bulashevska 等用贝叶斯分类器同样得到较好的分类预测效果[9]。Chen 等在使用凋亡蛋白数据集的同时，构建了数据集 CH317，将多种特征融合后使用混合增量的方式实现预测[10]。Ding 等在伪氨基酸特征下，将模糊 K 近邻 (Fuzzy K-nearest neighbor, FKNN) 分类器与遗传算法相结合，预测准确率有一定提高[11]。Lin 等采用伪氨基酸结合支持向量机方法对蛋白质亚细胞定位进行预测[12]。Zhang 等运用支持向量机融合距离频率实现蛋白序列的定位预测[13]。Liao 等将伪氨基酸、二肽等多种特征进行融合后通过支持向量机在 CH317 上取得较好的预测效果[14]。Hu 等提取序列之间的网状信息，对位于 19 个区间的酵母菌数据集进行预测，实现效果较好[15]。Yao 等基于序列之间的进化信息，通过位置特异性得分矩阵 (PSSM)，统计各氨基酸的突变率取得较好的预测效果[16]。Liu 等提取序列 PSSM 特征输入 SVM 进行预测，得到较好的预测效果[17]。Wang 等提取序列 GO 注释信息特征，在支持向量机中实现了革兰氏阴性菌的多区间分类问题[18]。Chen 等利用序列的物化属性、PSSM 和 GO 注释 3 种特征，对多个细菌数据集进行预测，得到较高的预测成功率[19]。总而言之，序列特征越来越丰富，提取过程更复杂，以上所述特征各有优缺点，结合适当的预测分类器可以取得一定的成功率，其中支持向量机、贝叶斯分类器、神经网络等训练较为复杂与耗时。故如何在一般低维或简单特征和低复杂度的预测模型前提下提高识别率是本文重点解决的问题。

KNN 是目前理论成熟、应用最为广泛的分类预测算法之一[20]，算法简单易优化，这一点对于近年来蛋白序列数量的激剧增长是有利的，但现有研究直接利用 KNN 进行定位预测效果并不理想，原因在于 KNN 受模式特征和决策机制影响较大。再考虑到 Blast 可用于推断结构和功能相似。本文尝试将两者结合起来，改进 KNN 算法，利用 KNN 过滤出与预测序列特征相似度较高的序列，再对这些序列进行更精细的 Blast 比对，作为最终预测依据。具体策略分别以序列的氨基酸组成、二肽和伪氨基酸作为 KNN 搜索阶段的特征，KNN 决策阶段用 Blast 比对确定蛋白所属定位，最后用 Adaboost 集成多个 KNN 子分类器进行定位预测，算法在多个数据集上取得较好的实验效果。文中预测算法通过网站 http://www.wsns.org/subloc/homepage_final.jsp 实现。

1 材料与方法

1.1 数据集

为了客观评价预测算法的有效性，本文将 Chen[10]等使用的 CH317 作为实验基准数据集。CH317 数据集中包含 317 条蛋白序列，分布在 6 个位置，其中细胞质蛋白 (Cytoplasmic proteins，cy) 112 条，膜蛋白 (Membrane proteins，me) 55 条，细胞核蛋白 (Nuclear proteins，nu) 52 条，线粒体蛋白 (Mitochondrial proteins，mi) 34 条，内质网蛋白 (Endoplasmic reticulum proteins，en) 47 条，分泌蛋白 (Secreted proteins，se) 17 条。CH317 中涉及到的所有蛋白序列均可在 uniprot 网站下载 (http://www.uniprot.org/)。

除了 CH317，为了对预测算法进行进一步评估，本文参考 Fan 的数据集构建方法[21]，具体参数本文不再复述。构建了革兰氏阴性菌数据集 (Gram1253)，Gram1253 共包含符合规则蛋白序列 1 253 条，分布于 5 个位置，如表 1 所示。

1.2 序列特征提取

对蛋白序列进行不同特征的提取从而实现蛋白质的亚细胞区间预测是目前研究所采用的主要方法。本文使用氨基酸组成、二肽和伪氨基酸特征。

表 1 革兰氏阴性菌数据集分布

Table1 Distribution of Gram1253

Subcellular locations	Number of protein sequences
Membrane	166
Cytoplasm	443
Periplasm(Pe)	423
Secreted	199
Nucleoid	22

1.2.1 氨基酸组成 (Amino acid composition，AAC)

不同亚细胞定位中的蛋白质在组成上有很大差别，基于这一特性提出了 AAC 特征提取方法[22]，Nakashima 等首次在 AAC 特征基础上实现了对亚细胞定位的预测[23]。AAC 的基本思想：对于任意的蛋白序列 P，统计构成序列的 20 种氨基酸各自出现的频率，那么序列 P 的 AAC 特征 $\vec{P}_{AAC}$ 可用公式 1 表示：

$$\vec{P}_{AAC}=\left[f_1,f_2,f_3,\cdots f_{20}\right]^T \tag{1}$$

上式中，f_i 表示第 i 种氨基酸在序列 P 中出现的频率。

1.2.2 二肽 (Dipeptide，Dipe)

二肽特征是基于 AAC 特征的改进，所谓二肽是指任意 2 个氨基酸构成的氨基酸对，组成蛋白序列的氨基酸共有 20 种，因此二肽共有 400 种，通过统计二肽的频率来描述一条蛋白序列的特征是二肽特征的基本思想[24]。对于任意的序列 P，其二肽特征 $\vec{P}_{Dipe}$ 可用公式 2 表示：

$$\vec{P}_{Dipe}=\left[d_1,d_2,d_3,\cdots d_{400}\right]^T \tag{2}$$

上式中，d_i 表示第 i 种二肽在序列 P 中出现的频率。

1.2.3 伪氨基酸 (Pseudo amino acid composition，PseAAC)

伪氨基酸特征同样是基于 AAC 特征的改进，在统计氨基酸频率的基础上，利用 λ 维来表示氨基酸之间的位置信息[25]。同 AAC 特征相比，伪氨基酸特征对序列的刻画更加全面。对于任意的序列 P，其伪氨基酸特征 $\vec{P}_{PseAAC}$ 可用公式 3 表示：

$$\vec{P}_{PseAAC}=\left[p_1,p_2,\cdots,p_{20},p_{20+1},\cdots p_{20+\lambda}\right]^T \tag{3}$$

上式中，前 20 维表示 20 种氨基酸的频率，

后面的 λ 维表示氨基酸之间的位置信息。

1.3 预测算法

1.3.1 基于相似性改进 KNN 分类算法

分类器的设计是对传统 KNN 算法做改进，修改 KNN 决策阶段，利用 Blast 比对取代投票机制确定所属区间位置。

序列相似性常被用来推断结构和功能相似[26]，因此，序列比对技术出现在一些区间预测算法中，如将 Blast 比对作为集成分类器的一个子分类器[27]，从 Needleman-Wunsch 算法的得分矩阵提取特征用于预测[28]。本文采用 Blast 序列局部比对搜索算法计算蛋白序列之间氨基酸残基的相似比率，从而确定蛋白序列所属位置。通过 Blast 序列局部比对搜索算法计算得分后，得分最高的蛋白序列便是与检索序列相似度最高的序列。基于改进 KNN 分类器算法流程见图 1。

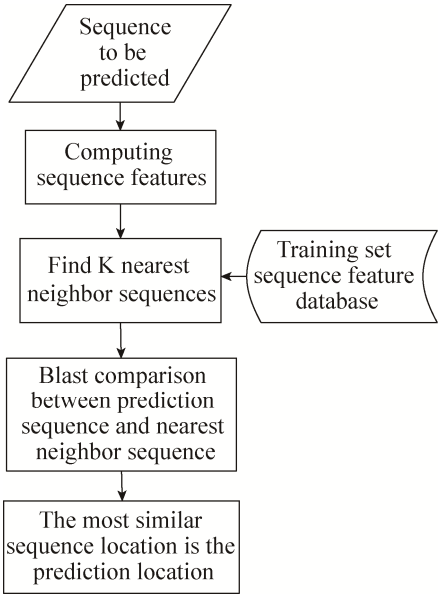

图 1 基于 Blast 改进的 KNN 分类算法

Fig. 1 Improved KNN classification algorithm based on Blast.

文中使用的 Blast 程序版本为 2.2.30，在 National Center for Biotechnology Information (NCBI) 官方网站下载 (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/)。这里采用 blastp 子程序对蛋白序列的亲缘性进行比对，具体用到的命令及主要参数如下：

1) 数据库格式化

makeblastdb.exe –in DB.fasta-parse_seqids-hash_index–dbtypeprot

其中 makeblastdb.exe 为格式化数据库命令，–in 指定数据库文件，-parse_seqids-hash_index 为子序列比对的参数，-dbtype 指定比对类型，prot 为蛋白序列。

2) 序列比对命令

blastp.exe-task blastp-query que-db DB–out out

使用 blastp.exe 命令实现蛋白序列比对，-query 指定要比对的序列文件，-db 为格式化后的数据库文件，–out 指定结果输出文件。

1.3.2 Adaboost 集成分类预测算法

Adaboost 集成分类算法对多个基于 Blast 改进的 KNN 分类器进行集成，得到一个较强的分类器[29]。在分类器训练过程中，由于每个分类器的权重都基于前一个分类器的分类效果，因此最后得到的集成分类器效果较好。

给定一个训练数据集 $T=\{(x_1,y_1),\ (x_2,y_2)\cdots(x_n,y_n)\}$，其中实例 $x\in X$，而实例空间 $X\subset R^n$，y_i 属于标记集合 $\{-1,+1\}$，Adaboost 的算法流程如下：

步骤 1：初始化训练数据的权值分布。每一个训练样本最开始时都被赋予相同的权重：1/N。

$$D_1=\left(w_{11},w_{12}\cdots w_{1i}\cdots,w_{1N}\right),\quad w_{1i}=\frac{1}{N},i=1,2,\cdots N \tag{4}$$

步骤 2：进行多轮迭代，用 m=1,2, …, M 表示迭代的第 M 轮。

使用具有权值分布 D_m 的训练数据集学习，得到基本分类器：

$$G_m(x):\chi \to \{-1,+1\} \tag{5}$$

计算 $G_m(x)$在训练数据集上的分类误差率

$$\begin{aligned} e_m &= P\left(G_m(x_i) \neq y_i\right) \\ &= \sum_{i=1}^{N} w_{mi} I\left(G_m(x_i) \neq y_i\right) \end{aligned} \tag{6}$$

$G_m(x)$在训练数据集上的误差率 e_m 即被 $G_m(x)$误分类样本的权值之和。

计算 $G_m(x)$的系数， m 表示 $G_m(x)$在最终分类器中的比重

$$\alpha_m = \frac{1}{2}\log\log\frac{1-e_m}{e_m} \tag{7}$$

更新训练数据集的权值分布，用于下一轮迭代。

$$D_{m+1} = (w_{m+1,1}, w_{m+1,2}\cdots w_{m+1,i}\cdots, w_{m+1,N}) \tag{8}$$

$$\begin{aligned} D_{m+1,i} &= \frac{w_{mi}}{Z_m}\exp\left(-\alpha_m y_i G_m(x_i)\right), \\ & i = 1,2,\cdots,N \end{aligned} \tag{9}$$

更新权重使得被基本分类器 $G_m(x)$误分类样本的权值增大，而被正确分类样本的权值减小。通过这样的方式，Adaboost 方法能“聚焦于”那些较难分的样本上。

其中，Z_m 是规范化因子，使得 D_{m+1} 成为一个概率分布：

$$Z_m = \sum_{i=1}^{N} w_{mi}\exp\left(-\alpha_m y_i G_m(x_i)\right) \tag{10}$$

步骤 3：组合各个弱分类器

$$f(x) = \sum_{m=1}^{M}\alpha_m G_m(x) \tag{11}$$

从而得到最终分类器，如下：

$$\begin{aligned} G(x) &= \mathrm{sign}\left(f(x)\right) \\ &= \mathrm{sign}\left(\sum_{m=1}^{M}\alpha_m G_m(x)\right) \end{aligned} \tag{12}$$

Adaboost 分类通过对多个相似性比对改进 KNN 分类器进行集成，实现亚细胞定位预测。一个 Adaboost 分类器只能完成二分类，所以需要训练多个分类器达到多区间预测，分类器构建过程如下：

1) 对于位于 N 个区间位置的数据集，随机取其中一个区间作为类别 1，其余位置作为类别 2，转化成一个二分类问题；2) 初始化数据集中蛋白序列的权重；3) 随机生成 k 值，得到对应的相似性比对改进 KNN 分类器；4) 使用该分类器对数据集进行预测，由预测误差确定分类器系数；5) 重复步骤 3−4 M 次，得到一个 Adaboost 分类器；6) 根据预测效果更新数据集中样本的权重，用于下一个分类器的训练；7) 对于类别 2，重复步骤 1–6，获取对应的分类器，直到区间无法再细分为止；8) 对于 N 个区间的样本，进行 N–1 次训练，得到 N–1 个 Adaboost 分类器；9) 对 N–1 个 Adaboost 分类器进行集成，得到最终的集成分类器。

1.3.3 评价指标

Jackknife 检验是蛋白质亚细胞定位预测中较为常用的检验方法，基本原理为：从数据集中取出一条蛋白序列作为测试序列，剩余序列作为训练集，测试完毕后将该序列放入数据集并取出下一条序列作为测试序列，以此类推直至所有序列预测完毕。本文中的预测算法在 Jackknife 检验下完成。

参考 Chen 使用的评价指标，引入敏感性 (S_n)、特异性 (S_p)、相关系数 (MMC_i) 以及总体准确率 (A)[10]。

2 结果与分析

KNN 分类器 K 值的选取对于整个算法的准

确度有很大影响。*K* 值越大，包含的蛋白序列数量越多，算法的时间复杂度越高。*K* 值越小，则越有可能丢弃掉一些真正有意义的蛋白序列，影响算法的准确度。故文中算法在各数据集的序列预测过程 *K* 值均取 20。

2.1 本文算法在多数据集及多特征下预测结果与分析

基于相似性比对改进 KNN 的 Adaboost 集成分类预测算法，对数据集 CH317 和 Gram1253 提取 3 种特征进行预测，CH317 的实验结果如表 2 所示。

由表 2 可知，除了位置 cy 和 en 外，数据集 CH317 的 AAC、PseAAC 特征准确率都高于 Dipe 特征并且总的预测准确率也较高，说明在 KNN 搜索阶段 Dipe 特征会误排除一些同模式序列。

基于相似性比对改进 KNN 的 Adaboost 集成分类预测算法在数据集 Gram1253 的实验结果如表 3 所示。

由表 3 结果可知，基于相似性比对改进 KNN 的 Adaboost 集成分类预测算法在数据集 Gram1253 上，AAC、PseAAC 特征各位置的准确率都高于 Dipe 特征并且总的预测准确率也较高。总之，与表 2 的结果一致的是，AAC、PseAAC 特征预测准确率都较高，而维数更高的 Dipe 效果均要差些，带有相邻位置信息的氨基酸对频率不能较准确地代表序列特征。

表 2 通过 Jackknife 检验在数据集 CH317 上的预测结果

Table 2 The predictive results by Jackknife test on data set CH317

Class	S_n (%)			S_p (%)			*MMC* (%)		
	AAC	PseAAC	Dipe	AAC	PseAAC	Dipe	AAC	PseAAC	Dipe
cy	96.4	93.8	94.6	91.5	93.8	90.6	90.3	90.2	88.1
me	94.5	92.7	89.1	96.3	94.4	96.1	94.4	92.2	90.9
mi	88.2	91.2	85.3	93.8	91.2	100	89.8	90.0	91.5
se	82.4	88.2	76.5	100	93.8	100	90.3	90.4	86.8
nu	82.7	88.5	80.8	91.5	92.0	100	84.5	88.2	88.1
en	95.7	95.7	97.9	86.5	88.2	70.8	89.3	90.4	79.8
A	92.1	92.4	90.0						

表 3 通过 Jackknife 检验在数据集 Gram1253 上的预测结果

Table 3 The predictive results by Jackknife test on data set Gram1253

Class	S_n (%)			S_p (%)			*MMC* (%)		
	AAC	PseAAC	Dipe	AAC	PseAAC	Dipe	AAC	PseAAC	Dipe
me	94.6	94.6	96.4	88.2	86.7	69.3	89.9	89.0	78.3
cy	98.6	98.6	97.1	96.0	96.0	93.7	95.7	95.7	92.3
pe	98.3	98.1	92.9	91.4	90.8	89.1	91.9	91.2	85.7
se	68.8	66.3	46.7	96.5	96.4	91.2	78.8	77.1	61.1
nu	86.4	86.4	63.6	82.6	82.6	70.0	84.2	84.2	66.1
A	93.1	92.6	87.0						

2.2 本文算法与其他算法预测结果比较

在数据集 CH317 上，将基于相似性比对改进 KNN 的 Adaboost 集成分类预测算法的预测结果同其他方法进行比较，并将结果列于表 4 中，由于一些其他预测方法不涉及特异性和相关系数，所以这里只对敏感性(S_n)进行比较。

ID 算法使用混合增量 (Increment of diversity, ID) 作为特征进行预测；FKNN 使用 PseAAC 作为特征，结合模糊 K 近邻算法预测；PseAAC_SVM 使用 PseAAC 特征结合支持向量机预测；DF_SVM 使用距离频率 (Distance frequency, DF) 结合支持向量机预测；Mix_SVM 提出新的 PseAAC 计算方法结合支持向量机预测；PSSM_SVM 使用位置特异性得分矩阵 (PSSM) 结合支持向量机预测。

由表 4 可以看出，与其他预测算法相比，基于相似性比对改进 KNN 的 Adaboost 集成分类预测算法的最高准确率高于其他算法，尤其超过其他以 PseAAC 为特征的预测算法；在各位置上的预测准确率也较高，特别是 CH317 的 me、se、en 位置获得最高准确率，且总体准确率也有一定提高。

为了便于对 Gram1253 的预测结果进行比较，采用支持向量机作为分类器，分别在 3 种特征下统计准确率，并与 KNN 分类器预测结果进行比较，结果列于表 5 中。

表 4 通过 **Jackknife** 检验在数据集 **CH317** 上不同方法的预测结果

Table 4 The predictive results of different methods by Jackknife test for data set CH317

Method	S_n (%)						
	cy	me	mi	se	nu	en	A (%)
ID[10]	81.3	81.8	85.3	88.2	82.7	83.0	82.7
FKNN[11]	93.8	92.7	82.4	76.5	90.4	93.6	90.9
PseAAC_SVM[12]	93.8	90.9	85.3	76.5	90.4	95.7	91.1
DF_SVM[13]	92.9	85.5	76.5	76.5	86.5	93.6	88.0
Mix_SVM[14]	94.6	90.9	93.8	70.6	88.5	95.7	91.2
PSSM_SVM[16]	92.0	92.7	82.4	76.5	90.4	93.6	90.5
Our method	93.8	92.7	91.2	88.2	88.5	95.7	92.4

表 5 通过 **Jackknife** 检验在数据集 **Gram1253** 上不同方法的预测结果

Table 5 The predictive results of different methods by Jackknife test for data set Gram1253

Method	S_n (%)					
	cy	me	pe	se	nu	A(%)
AAC_SVM	98.4	94.6	97.6	70.0	40.9	92.1
PseAAC_SVM	98.4	94.6	97.6	70.9	59.1	92.6
Dipe_SVM	97.5	93.4	96.7	70.0	40.9	91.3
AAC_KNN	98.4	87.3	86.8	39.7	0.0	82.0
PseAAC_KNN	98.4	88.0	87.5	38.6	4.5	82.1
Dipe_KNN	95.7	97.0	59.6	13.1	0.0	68.9
Our method	98.6	94.6	98.3	68.8	86.4	93.1

由表 5 可以看出，与支持向量机以及 KNN 算法相比，当使用 AAC 特征时 Adaboost 集成分类预测算法在 4 个区间 cy、me、pe、nu 位置的预测效果较好，总体预测准确率有了一定提高。表 4、5 中，与文中算法预测率接近的是基于支持向量机的预测技术，与它相比，本文算法更适合大数据的处理，算法简单易实现，而支持向量机处理大数据效率低。

3 讨论

蛋白质亚细胞定位预测是生物信息学领域

较复杂的研究内容，研究者在序列特征提取与预测算法设计上做了大量工作。在此基础上，不失一般性，本文以常见的 AAC、Dipe、PseAAC 作为蛋白序列特征，基于相似度高的蛋白序列出现在同一个亚细胞位置中可能性较高的思想构建改进 KNN 分类器，进而集成改进 KNN 分类器，实现蛋白质亚细胞定位预测。算法架构可满足大数据处理的要求，对于大数据集，改进 KNN 分类器便于实现 Hadoop 等分布式处理架构，缩短算法运行时间。

基于通用性考虑，选用国际公认有效的数据集 CH317，并按通用标准构建一个较大数据集用于测试。通过严格的 Jackknife 检验，数据集 CH317 和 Gram1253 在 3 种特征下最高预测成功率分别为 92.4%和 93.1%。与一些报道的预测算法相比，集成改进 KNN 预测算法在 3 种特征下都取得较好的实验效果，且总体成功率有一定提高，优于直接使用 Blast 比对预测，说明同源性比对不适合直接用于蛋白质亚细胞定位预测。其中，AAC、PseAAC 特征的准确率最为稳定，AAC 总体更优，表明在 KNN 的搜索阶段无需考虑复杂的理化特性。总之，通过在 3 种特征及多个数据集下的验证测试，集成改进 KNN 预测算法均取得较好的效果，该算法是一种较为有效的蛋白质亚细胞定位预测算法。

REFERENCES

[1] Cai YD, Liu XJ, Xu XB, et al. Support vector machines for prediction of protein subcellular location. Mol Cell Biol Res Commun, 2000, 4(4): 230–233.

[2] Chou KC, Cai YD. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. Biochem Biophys Res Commun, 2003, 311(3): 743–747.

[3] Chou KC, Elrod DW. Prediction of membrane protein types and subcellular locations. Prot: Struct, Funct, Bioinform, 1999, 34(1): 137–153.

[4] Chou KC, Elrod DW. Protein subcellular location prediction. Protein Eng, Des Select, 1999, 12(2): 107–118.

[5] Reed JC, Paternostro G. Postmitochondrial regulation of apoptosis during heart failure. Proc Natl Acad Sci USA, 1999, 96(14): 7614–7616.

[6] Suzuki M, Youle RJ, Tjandra N. Structure of bax: coregulation of dimer formation and intracellular localization. Cell, 2000, 103(4): 645–654.

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

[8] Huang J, Shi F. Support vector machines for predicting apoptosis proteins types. Acta Biotheor, 2005, 53(1): 39–47.

[9] Bulashevska A, Eils R. Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. BMC Bioinformatics, 2006, 7(1): 298.

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

[11] Ding YS, Zhang TL. Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. Pattern Recognit Lett, 2008, 29(13): 1887–1892.

[12] Lin H, Wang H, Ding H, et al. Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition. Acta Biotheor, 2009, 57(3): 321–330.

[13] Zhang L, Liao B, Li DC, et al. A novel representation for apoptosis protein subcellular localization prediction using support vector machine. J Theoret Biol, 2009, 259(2): 361–365.

[14] Liao B, Jiang JB, Zeng QG, et al. Predicting apoptosis protein subcellular location with PseAAC by incorporating tripeptide composition. Protein

Pept Lett, 2011, 18(11): 1086–1092.
[15] Hu LL, Feng KY, Cai YD, et al. Using protein-protein interaction network information to predict the subcellular locations of proteins in budding yeast. Protein Pept Lett, 2012, 19(6): 644–651.
[16] Yao YH, Shi ZX, Dai Q. Apoptosis protein subcellular location prediction based on position-specific scoring matrix. J Computat Theoret Nanosci, 2014, 11(10): 2073–2078.
[17] Liu TG, Tao PY, Li XW, et al. Prediction of subcellular location of apoptosis proteins combining tri-gram encoding based on PSSM and recursive feature elimination. J Theoret Biol, 2015, 366: 8–12.
[18] Wang X, Zhang J, Li GZ. Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble. BMC Bioinformatics, 2015, 16(S12): S1.
[19] Chen J, Xu H, He PA, et al. A multiple information fusion method for predicting subcellular locations of two different types of bacterial protein simultaneously. Biosystems, 2016, 139: 37–45.
[20] Jiang Y, Zhou ZH. Editing training data for kNN classifiers with neural network ensemble [M]//Yin FL, Wang J, GuoCG, Eds. Advances in Neural Networks–ISNN 2004. Berlin Heidelberg: Springer, 2004: 356–361.
[21] Fan GL, Li QZ. Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition. J Theoret Biol, 2012, 304: 88–95.
[22] Nakashima H, Nishikawa K, Tatsuo O. The folding type of a protein is relevant to the amino acid composition. J Biochem, 1986, 99(1): 153–162.
[23] 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.
[24] Wu C, Whitson G, Mclarty J, et al. Protein classification artificial neural system. Protein Sci, 1992, 1(5): 667–677.
[25] Chou KC, Shen HB. Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms. Nat Protoc, 2008, 3(2): 153–162.
[26] Nair R, Rost B. Sequence conserved for subcellular localization. Protein Sci, 2002, 11(12): 2836–2847.
[27] Cherian BS, Nair AS. Protein location prediction using atomic composition and global features of the amino acid sequence. Biochem Biophys Res Commun, 2010, 391(4): 1670–1674.
[28] Kim JK, Bang SY, Choi S. Sequence-driven features for prediction of subcellular localization of proteins. Pattern Recognit, 2006, 39(12): 2301–2311.
[29] Lin J, Wang Y. Using a novel Adaboost algorithm and Chou's pseudo amino acid composition for predicting protein subcellular localization. Protein Pept Lett, 2011, 18(12): 1219–1225.

(本文责编　陈宏宇)