微生物学通报
Microbiology China
tongbao@im.ac.cn

Oct. 20, 2016, 43(10): 2298−2305
http://journals.im.ac.cn/wswxtbcn
DOI: 10.13344/j.microbiol.china.150862

生物实验室

相似性比对预测蛋白质亚细胞区间

王雄飞[1] 张梁[2] 薛卫[1,3*] 赵南[1] 徐焕良[1]

(1. 南京农业大学信息科学技术学院 江苏 南京 210095)
(2. 江南大学粮食发酵工艺与技术国家工程实验室 江苏 无锡 214122)
(3. 苏州市康绿农产品发展有限公司 江苏 苏州 215155)

摘 要:【目的】对蛋白质所属的亚细胞区间进行预测，为进一步研究蛋白质的生物学功能提供基础。【方法】以蛋白质序列的氨基酸组成、二肽、伪氨基酸组成作为序列特征，用 BLAST 比对改进 K 最近邻分类算法(K-nearest neighbor，KNN)实现蛋白序列所属亚细胞区间预测。【结果】在 Jackknife 检验下，数据集 CH317 三种特征的成功率分别为 91.5%、91.5%和 89.3%，数据集 ZD98 成功率分别为 93.9%、92.9%和 89.8%。【结论】BLAST 比对改进 KNN 算法是预测蛋白质亚细胞区间的一种有效方法。

关键词: 亚细胞区间，KNN，BLAST，蛋白序列特征

Prediction of protein subcellular locations by similarity comparison

WANG Xiong-Fei[1] ZHANG Liang[2] XUE Wei[1,3*] ZHAO Nan[1] XU Huan-Liang[1]

(1. School of Information Science and Technology, Nanjing Agricultural University, Nanjing, Jiangsu 210095, China)
(2. National Engineering Laboratory for Cereal Fermentations Technology, Jiangnan University, Wuxi, Jiangsu 214122, China)
(3. Suzhou Kanglü Agricultural Products Development Co., Ltd., Suzhou, Jiangsu 215155, China)

Abstract: [Objective] A new subcellular location prediction algorithm is proposed that provides basis for further experimental study of protein biological function. **[Methods]** Nearest neighbor classification algorithm improved by Blast comparison is used to predict the protein subcellular locations by three sequence features including amino acid composition, two peptides and pseudo amino acid composition of protein sequence. **[Results]** Through Jackknife test, on data set CH317 the success rates of 3 characteristics were 91.5%, 91.5% and 89.3%, on data set ZD98 success rates were 93.9%, 92.9% and 89.8%. **[Conclusion]** K-Nearest Neighbor algorithm improved by Blast comparison is an effective method for predicting subcellular locations of proteins.

Keywords: Subcellular locations, K-Nearest Neighbor, Blast, Protein sequence characteristics

Foundation item: The Fundamental Research Funds for the Central Universities (No. KYZ201668); Natural Science Foundation of Jiangsu Province (No. BK2012363, BK20140002); Jiangsu Postdoctoral Scientific Research Foundation of China (No. 1302038B)

***Corresponding author:** E-mail: xwsky@njau.edu.cn

Received: November 02, 2015; **Accepted:** January 26, 2016; **Published online** (www.cnki.net)**:** February 23, 2016
基金项目：中央高校基本科研业务费专项资金项目(No. KYZ201668)；江苏省自然科学基金项目(No. BK2012363，BK20140002)；江苏省博士后科研项目(No. 1302038B)

*通讯作者：E-mail：xwsky@njau.edu.cn

收稿日期：2015-11-02；接受日期：2016-01-26；优先数字出版日期(www.cnki.net)：2016-02-23

蛋白质是生命活动中的重要参与者，了解蛋白质的功能对于疾病的控制、新药的研发等有着重大意义。蛋白序列的功能与其所属的亚细胞区间有着紧密联系，因此对蛋白序列的亚细胞区间预测研究有着重要意义[1-6]。

传统的用实验方式确定蛋白质亚细胞区间位置信息耗时久、代价高，而且随着蛋白序列数量的增长已无法满足需求[7]，因此利用机器学习的思想对蛋白质亚细胞区间进行预测成为了获取区间信息的主要研究方法。1991 年，Nakai 等首次将机器学习方法应用于对革兰氏阴性细菌的亚细胞区间预测[8-9]，随后的二十多年，对蛋白质亚细胞区间预测的研究取得了一系列进展。2003 年 Zhou 等采用基于马氏距离的协变判别函数，结合降维的氨基酸组成特征首次对凋亡蛋白进行区间预测，构建数据集 ZD98 在 Jackknife 检验下成功率为 72.5%[10]。2005 年，Huang 等在 Zhou 的基础上对氨基酸组成特征开方后采用支持向量机对 ZD98 进行预测，成功率达到 90.8%[11]。2006 年，Bulashevska 和 Eils 提出了基于贝叶斯的集成分类思想，对 ZD98 数据集进行预测，成功率达到 89.8%[12]。2007 年 Chen 等构建 CH317 数据集，将蛋白质的 N 端、C 端以及疏水性 3 种特征进行融合后，采用混合增量的方式分别对 CH317 和 ZD98 进行预测，Jackknife 检验下成功率为 82.7%和 90.8%[13]。2008 年 Ding 等提取蛋白序列伪氨基酸特征，采用模糊 K 近邻(Fuzzy K-nearest neighbor，FKNN)分类器并利用遗传算法寻优，对 CH317 进行预测，成功率为 90.9% (Jackknife)[14]。2009 年 Lin 等基于伪氨基酸特征结合支持向量机对 CH317 和 ZD98 分别预测，成功率为 91.1%和 92.9% (Jackknife)[15]。2009 年，Zhang 等在氨基酸位置分布信息特征基础上融合距离频率，运用支持向量机对 CH317，ZD98 进行预测，成功率为 88.0%和 93.9%[16]。2011 年 Liao 等将伪氨基酸、三肽组成以及氨基酸位置分布信息 3 种特征进行融合，在支持向量机中对 CH317 数据集预测的成功率为 91.2%[17]。2012 年 Hu 等提取蛋白序列之间相互作用的网状信息，对包含 4 683 条酵母菌分布于 19 个不同区间的序列进行预测，取得较好的实验效果[18]。2014 年 Yao 等考虑蛋白序列的进化信息，根据位置特异性得分矩阵(PSSM)统计序列中氨基酸的突变率，在支持向量机中对 CH317 和 ZD98 预测的成功率分别为 90.5%和 92.9%[19]。

上述预测方法中，基于氨基酸组成的预测方法只提取了氨基酸的出现频率，忽略了序列之间的位置信息。伪氨基酸特征预测在氨基酸组成基础上融合了序列位置信息，包含了更多的序列特征，但没有考虑到各序列之间的关系。疏水性等物化属性的预测方法只是考虑到蛋白序列本身的特性，特征相对单一，预测的准确率较低。基于支持向量机和贝叶斯分类的预测方法相对复杂，并且训练时间较长，对于大量序列的预测不利。近年来理论成熟、算法简单高效的 KNN 算法在机器学习领域得到大量运用，但直接用它来预测蛋白区间成功率较低。实际上，除了序列本身的特征之外，各序列之间的结构关系也是确定区间的有效依据，序列之间的结构越相似，所属同一区间的可能性就越大[20-23]。本文以蛋白序列的氨基酸组成、二肽、伪氨基酸作为序列特征，采用 BLAST 比对改进的 KNN 算法，对蛋白质亚细胞区间信息进行预测，取得较好的实验结果。

1 数据集

为了便于将实验结果与传统的预测方法进行比较，本文采用 Chen[13]、Zhou[10]等使用的 CH317、ZD98 数据集，数据集中所有序列在 Uniprot 网站下载(http://www.uniprot.org/)。CH317 数据集中包含 317 条蛋白序列，分布在 6 个区间，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。ZD98 数据集中包含细胞质蛋白(Cytoplasmic proteins，Cy) 43 条，膜蛋白(Membrane proteins，Me) 30 条，线粒体蛋白(Mitochondrial proteins，Mi) 13 条以及 12 条其它蛋

白(Other)。

2 特征提取和区间预测

2.1 特征提取

由于蛋白序列的长度较大且排列较为复杂，直接使用序列本身来预测是不现实的，因此需要对序列进行简化，提取不同的特征来代替复杂的序列本身，实现对区间的预测[24]。本文提取序列的 3 种不同特征来实现对蛋白质亚细胞的区间预测。

2.1.1 氨基酸组成(Amino acid composition，AAC)：处于不同亚细胞位置的蛋白质序列的组成有很大区别，基于 AAC 的特征提取正是利用了这一特性[25]，随后 Nakashima 等首次将该特征应用于对亚细胞区间的预测[26]。AAC 的基本思想：20 种氨基酸排列组合形成一条蛋白序列，对于给定的任意一条预测序列 P，统计这 20 种氨基酸在序列 P 中出现的频率，那么序列 P 可用公式(1)表示：

$$P_{\mathrm{AAC}}=[f_1,f_2,f_3,\quad f_{20}]^T \tag{1}$$

上式中，f_i 表示第 i 种氨基酸在序列 P 中出现的频率。

2.1.2 二肽(Dipeptide，Dip)：相较于 AAC 特征而言，二肽模型不再考虑单个氨基酸出现的频率，而是一个氨基酸对(称为二肽)在序列中出现的频率[27]。常见氨基酸有 20 种，所以二肽共有 400 种，对于任意的预测序列 P 可用公式(2)表示：

$$P_{\mathrm{Dip}}=[f_1,f_2,f_3,\quad f_{400}]^T \tag{2}$$

上式中，f_i 表示第 i 种氨基酸对在序列 P 中出现的频率。

2.1.3 伪氨基酸(Pseudo amino acid composition，PseAAC)：AAC 特征虽然能够表示一条序列，但是它忽略了序列中氨基酸的位置信息，而二肽特征的维数较多计算相对复杂。为了能够包含蛋白质序列中氨基酸的位置信息，同时减少特征向量的维数，Chou 等提出了伪氨基酸模型[28-29]，对于任意预测序列 P，它的伪氨基酸特征向量可用公式(3)表示：

$$P_{\mathrm{PseAAC}}=[f_1,f_2,\cdots f_{20},f_{20+1}\cdots f_{20+\lambda}]^T \tag{3}$$

上式中，前 20 维表示 20 种氨基酸在蛋白质序列中出现的频率，后面的 λ 维用来表示氨基酸之间的位置信息。

2.2 蛋白序列相似性搜索算法

序列相似性常被用来推断结构和功能相似性[21]，因此，序列比对技术出现在一些区间预测算法中，如将序列比对作为集成分类器的一个子分类器[20]，从 Needleman-Wunsch 算法的得分矩阵提取特征用于预测[22-23]。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区间。

BLAST 算法是目前较为常用的一种序列局部比对搜索算法，其基本思想可简单描述为：

(1) 从两个序列中找出一些长度相等且无空位完全匹配的子序列，即序列片段对；

(2) 筛选出满足一定匹配值的序列片段对；

(3) 将得到的序列片段对根据给定的相似性阈值延伸，得到一定长度的相似性片段，称为高分值片段对。

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。

本文中使用的 BLAST 搜索比对程序版本为 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+/)。程序共包含 5 个子程序：BLASTp、BLASTn、BLASTx、TBBLASTn 和 TBBLASTx，提供对蛋白序列和核苷酸等的搜索比对。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-inDB.fasta-parse_seqids-hash_index-dbtype prot

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。

2.3 基于 BLAST 比对改进 KNN 算法的序列预测

经典 KNN 分类算法分为两阶段：搜索得到 K 个相邻元素集；通过投票统计最终类别。本文修改 KNN 算法的决策阶段，使用 BLAST 比对决定最终区间。基本思想如下：对于一条预测序列 P，计算该序列的特征向量与训练集中所有序列特征向量的欧氏距离，取训练集中距离最短的前 K 个蛋白序列同预测序列做 BLAST 比对，找到同预测序列结构最为相似的序列，该序列所在的区间就是预测蛋白序列的发生区间，具体步骤如下：

(1) 对于预测蛋白序列 P，提取特征向量 $\vec{P}$ ；

(2) 遍历训练集中 n 条蛋白序列，设 i 表示第 i 条序列：

For i from 1 to n

{

提取序列 i 的特征向量 $\vec{P}_i$

计算 $\vec{P}_i$ 同 $\vec{P}$ 的欧氏距离 Dis 并保存到集合 d 中；

}

(3) 获取(2)中集合 d 最短的前 K 个距离代表的蛋白序列并保存到集合 S；

(4) 将预测序列 P 同集合 S 进行 BLAST 比对，找到 S 中同 P 结构最为相似的序列 S_p；

(5) 序列 S_p 所在的区间就是预测序列的发生区间

对于两个 n 维特征向量 $\{f_1, f_2, f_3, \cdots f_n\}$ 和 $\{t_1, t_2, t_3, \cdots t_n\}$，他们的欧氏距离的计算用公式(4)求出：

$$Dis = \sqrt{\sum_{j=1}^{n}(f_j - t_j)^2} \tag{4}$$

3 结果与分析

3.1 测试方法

自身一致性测试(Re-substitution)、独立数据集测试(Independent)和刀切法测试(Jackknife)是最为常用的 3 种测试验证方法。在自身一致性测试中采用相同数据集作为测试和训练样本，因此该方法容易对类别产生偏见，可靠性不高。独立数据集测试中将整个样本分为测试集和训练集两部分(两个集合互不相容)，但由于样本划分过程中存在人为因素，也会对算法的评估造成影响。刀切法测试是目前最为常用一种测试方法，在该测试中，从数据集中取出一条蛋白序列作为测试序列，其余序列作为训练集，测试完毕后将测试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。

由刀切法测试的过程得知，该测试方法屏蔽了人为的干扰因素，是目前认可度较高的测试方法[29-30]。

3.2 评价指标

为了便于实验结果的比较，同时对预测方法进行有效评估，引入敏感性、特异性和相关系数 3 个评价指标，敏感性(S_n)、特异性(S_p)、相关系数(MMC_i)以及总得准确率($Total$)的定义如下：

$$S_n = TP_i/(TP_i + FN_i) \tag{5}$$

$$S_p = TP_i/(TP_i + FP_i) \tag{6}$$

$$MMC_i = \frac{(TP_i \times TN_i) - (FP_i \times FN_i)}{\sqrt{(TP_i + FP_i) \times (TN_i + FN_i) \times (TP_i + FN_i) \times (TN_i + FP_i)}} \tag{7}$$

$$Total = \frac{1}{M} \frac{\sum_{i=1}^{M} TP_i}{\sum_{i=1}^{M}(TP_i + FN_i)} \tag{8}$$

上式中，TP_i 是第 i 类亚细胞区间正确预测的序列个数，FN_i 是第 i 类亚细胞区间中没有正确预测的序列个数，FP_i 是非第 i 类亚细胞区间但被预测为第 i 类区间的序列个数，TN_i 是被正确预测的非第 i 类亚细胞区间的序列数。评价指标的引入从三个方面对预测方法进行客观、有效的评估：敏感性(S_n)体现了预测算法的准确性，特异性(S_p)是对算法置信度的评价，而相关系数(MMC_i)则体现了预测算法整体的有效性，$Total$ 是对 M 个区间总的预测准确率的统计。

3.3 参数选择(K)

由序列预测算法的过程我们知道，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。因此找到一个最优的 *K* 值显得尤为重要。图 1 显示了在 Jackknife 检验下，两个数据集 CH317 和 ZD98 分别取不同 *K* 值对应的预测准确率的分布情况。

如图所示预测的准确率随着 *K* 值的增大而逐渐增加，当 *K* 达到 20 时准确率达到最大且趋于稳定，当前 *K* 的取值为最优值。在本文中所有数据集的序列预测过程 *K* 值均取 20。

3.4 结果分析

为了检验预测算法的性能，对数据集 CH317、ZD98 进行 AAC、PseAAC、Dip 的特征提取后，预测算法在 Jackknife 方式下检验，实验结果列于表 1 和表 2。

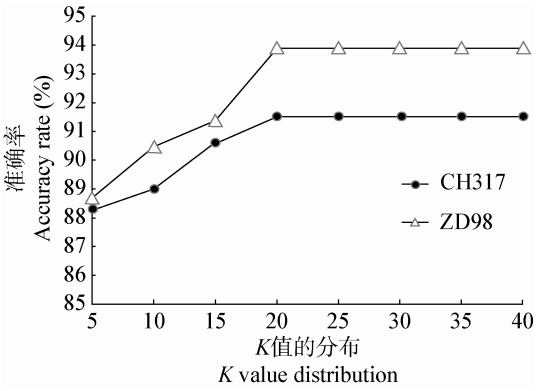

图 1 *K* 值准确率分布

Figure 1 *K* value accuracy rate distribution

由表 1 结果可知，在 Jackknife 检验模式下，BLAST 比对改进 KNN 算法对 3 种特征的预测都取得了较好的效果。总的准确率 AAC、PseAAC 特征高于特征 Dip，提高了 2.2%。

如表 2 中对于 ZD98 在 Jackknife 检验下，BLAST 比对改进 KNN 算法实现蛋白质区间预测，在 3 种特征下都取得了较好的实验效果。其中基于 AAC、PseAAC 的特征预测在 3 个区间 Cy、Me 和 Mi 上的准确率都高于 Dip 特征，而 AAC、Dip 特征的预测效果在区间 Other 上与 PseAAC 特征相比，准确率提高了 8.4%。整体预测效果上 AAC、PseAAC 特征的预测效果优于 Dip。

3.5 其他方法比较

为了对预测算法进行进一步的评估，将 BLAST 比对改进 KNN 算法的预测结果与其他预测方法进行比较，采用各个区间预测的准确率和整体准确率来作为算法的评价指标，这里将 3 种特征中准确率较高的作为整体预测准确率列于表 3、表 4 中，预测方法采用缩写并在表注中详细说明。

表 1 Jackknife 检验下 CH317 的预测结果

Table 1 The predictive results by Jackknife test on data set CH317

区间 Class	敏感性 Sensitivity (S_n) (%)			特异性 Specificity (S_p) (%)			相关系数 Correlation coefficient (*MMC*) (%)		
	AAC	PseAAC	Dip	AAC	PseAAC	Dip	AAC	PseAAC	Dip
细胞质 Cytoplasmic	96.4	93.8	94.6	90.0	92.9	89.1	89.0	89.4	86.8
膜 Membrane	94.5	92.7	89.1	96.3	92.7	96.1	94.4	91.1	90.9
线粒体 Mitochondrial	88.2	91.2	79.4	93.8	91.2	100	89.8	90.0	87.9
分泌 Secreted	82.4	70.6	76.5	100	92.3	100	90.3	79.7	86.8
细胞核 Nuclear	78.8	88.5	80.8	91.1	92.0	100	81.9	88.2	88.1
内质网 Endoplasmic reticulum	95.7	95.7	97.9	86.5	88.2	70.8	89.3	90.3	79.8
总体 Total	91.5	91.5	89.3						

表 2 Jackknife 检验下 ZD98 的预测结果
Table 2 The predictive results by Jackknife test on data set ZD98

区间 Class	敏感性 Sensitivity (S_n) (%)			特异性 Specificity (S_p) (%)			相关系数 Correlation coefficient (*MMC*) (%)		
	AAC	PseAAC	Dip	AAC	PseAAC	Dip	AAC	PseAAC	Dip
细胞质 Cytoplasmic	97.7	97.7	95.3	95.5	95.5	89.1	93.7	93.6	85.4
膜 Membrane	86.7	86.7	83.3	100	100	96.2	90.4	90.4	85.2
线粒体 Mitochondrial	100	100	84.6	81.3	76.5	84.6	88.5	85.3	82.1
其它 Other	91.7	83.3	91.7	91.7	90.9	84.6	90.4	85.2	86.2
总体 Total	93.9	92.9	89.8						

表 3 Jackknife 检验下 CH317 其它方法预测结果比较(敏感性，%)
Table 3 The predictive results of different methods by Jackknife test for data set CH317 (sensitivity, S_n, %)

预测方法 Method	细胞质 Cy	膜 Me	线粒体 Mi	分泌 Se	细胞核 Nu	内质网 En	总体 Total
混合增量 ID[a]	81.3	81.8	85.3	88.2	82.7	83.0	82.7
模糊 K 近邻 FKNN[b]	93.8	92.7	82.4	76.5	90.4	93.6	90.9
伪氨基酸结合支持向量机 PseAAC_SVM[c]	93.8	90.9	85.3	76.5	90.4	95.7	91.1
距离频率结合支持向量机 DF_SVM[d]	92.9	85.5	76.5	76.5	86.5	93.6	88.0
多特征融合结合支持向量机 Mul_SVM[e]	94.6	90.9	93.8	70.6	88.5	95.7	91.2
位置特异性得分矩阵结合支持向量机 PSSM_SVM[f]	92.0	92.7	82.4	76.5	90.4	93.6	90.5
单一 BLAST 预测 BLAST Prediction	89.3	87.3	79.4	76.5	76.9	89.4	85.2
本文方法 Our	96.4	94.5	88.2	82.4	78.8	95.7	91.5

注：[a]：Chen 使用混合增量的预测方法(2007)；[b]：Ding 使用模糊 K 近邻预测方法(2008)；[c]：Lin 使用伪氨基酸结合支持向量机预测方法(2009)；[d]：Zhang 使用距离频率结合支持向量机预测方法(2009)；[e]：Liao 使用多特征融合结合支持向量机预测方法(2011)；[f]：Yao 使用位置特异性得分矩阵(PSSM)结合支持向量机预测方法(2014). ID：Increment of diversity 混合增量预测方法；FKNN：Fuzzy K nearest neighbor 模糊 K 近邻预测方法；PseAAC_SVM：Pseudo amino acid composition and SVM 伪氨基酸组成结合支持向量机预测方法；DF_SVM：Distance frequency and SVM 距离频率与支持向量机结合预测方法；Mul_SVM：Multi-feature fusion and SVM 多特征融合结合支持向量机预测方法；PSSM_SVM：PSSM and SVM 位置特异性得分矩阵结合支持向量机预测方法.

Note: [a] comes from Chen, by using increment of diversity method (2007); [b] comes from Ding, by using fuzzy K nearest neighbor method (2008); [c] comes from Lin, using SVM by pseudo amino acid composition (2009); [d] comes from Zhang, by using distance frequency method and SVM (2009); [e] comes from Liao, by using multi-feature fusion method and SVM (2011); [f] comes from Yao, by using PSSM and SVM (2014). ID: Increment of diversity method; FKNN: Fuzzy k nearest neighbor method; PseAAC_SVM: Pseudo amino acid composition and SVM method; DF_SVM: Distance frequency and SVM method; Mul_SVM: Multi-feature fusion and SVM method; PSSM_SVM: PSSM and SVM method.

表 4 Jackknife 检验下 ZD98 其它方法预测结果比较(敏感性，%)
Table 4 The predictive results of different methods by Jackknife test for data set ZD98 (sensitivity, S_n, %)

预测方法 Method	细胞质 Cy	膜 Me	线粒体 Mi	其它 Other	总体 Total
协变判别函数 Covariant[a]	97.7	73.3	30.8	25.0	72.5
氨基酸组成结合支持向量机 AAC_SVM[b]	86.0	90.0	100	100	90.8
贝叶斯分类 BC[c]	95.3	90.0	92.3	66.7	89.9
混合增量 ID[d]	90.7	90.0	92.3	91.7	90.8
伪氨基酸结合支持向量机 PseAAC_SVM[e]	95.3	93.3	92.3	83.3	92.9
距离频率结合支持向量机 DF_SVM[f]	97.7	96.7	92.3	75.0	93.9
位置特异性得分矩阵结合支持向量机 PSSM_SVM[g]	95.3	93.3	84.6	91.7	92.9
单一 BLAST 预测 BLAST Prediction	93.0	73.3	84.6	83.3	84.7
本文方法 Our	97.7	86.7	100	91.7	93.9

注：[a]：Zhou 使用协变判别函数的预测方法(2003)；[b]：Huang 使用氨基酸组成结合支持向量机预测方法(2005)；[c]：Bulashenvska 使用贝叶斯分类预测方法(2006)；[d]：Chen 使用混合增量的预测方法(2007)；[e]：Lin 使用伪氨基酸结合支持向量机预测方法(2009)；[f]：Zhang 使用距离频率结合支持向量机预测方法(2009)；[g]：Yao 使用位置特异性得分矩阵(PSSM)结合支持向量机预测方法(2014). Covariant：Covariant discriminant function 协变判别函数预测方法；AAC_SVM：Amino acid composition and SVM 氨基酸组成结合支持向量机预测方法；BC：Bayesian classifier 贝叶斯分类预测方法；ID：Increment of diversity 混合增量预测方法；PseAAC_SVM：Pseudo amino acid composition and SVM 伪氨基酸组成结合支持向量机预测方法；DF_SVM：Distance frequency and SVM 距离频率与支持向量机结合预测方法；PSSM_SVM：PSSM and SVM 位置特异性得分矩阵结合支持向量机预测方法.

Note: [a] comes from Zhou, by using covariant discriminant function (2003); [b] comes from Huang, using SVM by amino acid composition (2005); [c] comes from Bulashenvska, by using bayesian classifier (2006); [d] comes from Chen, by using increment of diversity method (2007); [e] comes from Lin, using SVM by pseudo amino acid composition (2009); [f] comes from Zhang, by using distance frequency method and SVM (2009); [g] comes from Yao, by using PSSM and SVM (2014). Covariant: Covariant discriminant function method; AAC_SVM: Amino acid composition and SVM method; BC: Bayesian classifier method; ID: Increment of diversity method; PseAAC_SVM: Pseudo amino acid composition and SVM method; DF_SVM: Distance frequency and SVM method; PSSM_SVM: PSSM and SVM method.

本文算法与 BLAST 比对，Chen[13]、Ding[14]、Lin[15]、Zhang[16]、Liao[17]、Yao[19]等预测算法在数据集 CH317 上的预测结果列于表 3。在 BLAST 的比对过程中，为了测试 evalue、word_size、matrix、best_hit_overhang、best_hit_score_edge 等参数对预测结果的影响，通过循环遍历，优化选择后观察预测结果，实验结果表明 evalue=10、word_size=2、matrix=BLOSUM62 、 best_hit_overhang=0.2 、best_hit_score_edge=0.15 时预测准确率较高，因此本文算法使用这些参数值。

由表 3 可以看出，BLAST 比对改进 KNN 算法对 6 个区间的预测，在区间 Cy、Me、En 上的预测成功率均高于其它预测方法，且总体预测效果优于其它预测方法。

本文算法与 BLAST 比对，Zhou[10]、Huang[11]、Bulashevska[12]、Chen[13]、Lin[15]、Zhang[16]、Yao[19]等预测算法在数据集 ZD98 上的预测结果列于表 4。由于 ZD98 中其它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。

由表 4 比较结果看出，BLAST 比对改进 KNN 算法对 4 个区间的预测与其它方法相比，除 Me 区间预测结果略低，其余 3 个区间都取得了较好的预测效果且整体预测准确率优于其它方法。

4 结论

本文提取序列 3 种不同特征 AAC、PseAAC、

Dip，采用 BLAST 比对改进的 KNN 算法对数据集 CH317 和 ZD98 在 Jackknife 检验模式下进行预测，CH317 的预测成功率分别为 91.5%、91.5%和 89.3%，ZD98 的成功率分别为 93.9%、92.9%和 89.8%。与传统的神经网络、SVM 等有监督学习方法相比，BLAST 比对改进的 KNN 算法不需要复杂费时的网络训练，属于无监督学习，因此对训练集的元素增减具有自适应能力，并且在较简单的 AAC 特征向量下也能取得较好的效果，在预测准确率上有了一定程度提高，是一种较为有效的蛋白质亚细胞区间预测方法。

参 考 文 献

[1] Cai YD, Liu XJ, Xu XB, et al. Support vector machines for prediction of protein subcellular location[J]. Molecular Cell Biology Research Communications, 2000, 4(4): 230-233

[2] Chou KC, Cai YD. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology[J]. Biochemical and Biophysical Research Communications, 2003, 311(3): 743-747

[3] Chou KC, Elrod DW. Prediction of membrane protein types and subcellular locations[J]. Proteins, 1999, 34(1): 137-153

[4] Chou KC, Elrod DW. Protein subcellular location prediction[J]. Protein Engineering, 1999, 12(2): 107-118

[5] Reed JC, Paternostro G. Postmitochondrial regulation of apoptosis during heart failure[J]. Proceedings of the National Academy of Sciences of the United States of America, 1999, 96(14): 7614-7616

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

[7] Murphy RF, Boland MV, Velliste M. Towards a systematics for protein subcellular location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images[J]. Proceedings / ... International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology, 2000, 8: 251-259

[8] Nakai K, Kanehisa M. Expert system for predicting protein localization sites in gram-negative bacteria[J]. Proteins: Structure, Function, and Bioinformatics, 1991, 11(2): 95-110

[9] Nakai K, Kanehisa M. A knowledge base for predicting protein localization sites in eukaryotic cells[J]. Genomics, 1992, 14(4): 897-911

[10] Zhou GP, Doctor K. Subcellular location prediction of apoptosis proteins[J]. Proteins: Structure, Function, and Bioinformatics, 2003, 50(1): 44-48

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

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

[13] Chen YL, Li QZ. Prediction of the subcellular location of apoptosis proteins[J]. Journal of Theoretical Biology, 2007, 245(4): 775-783

[14] 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[J]. Pattern Recognition Letters, 2008, 29(13): 1887-1892

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

[16] Zhang L, Liao B, Li DC, et al. A novel representation for apoptosis protein subcellular localization prediction using support vector machine[J]. Journal of Theoretical Biology, 2009, 259(2): 361-365

[17] Liao B, Jiang JB, Zeng QG, et al. Predicting apoptosis protein subcellular location with PseAAC by incorporating tripeptide composition[J]. Protein & Peptide Letters, 2011, 18(11): 1086-1092

[18] Hu LL, Feng KY, Cai YD, et al. Using Protein-protein Interaction Network Information to Predict the Subcellular Locations of Proteins in Budding Yeast[J]. Protein & Peptide Letters, 2012, 19(6): 644-651

[19] Yao YH, Shi ZX, Dai Q. Apoptosis protein subcellular location prediction based on position-specific scoring matrix[J]. Journal of Computational and Theoretical Nanoscience, 2014, 11(10): 2073-2078

[20] Cherian BS, Nair AS. Protein location prediction using atomic composition and global features of the amino acid sequence[J]. Biochemical and Biophysical Research Communications, 2010, 391(4): 1670-1674

[21] Nair R, Rost B, Sequence conserved for subcellular localization[J]. Protein Science, 2002, 11(12): 2836-2847

[22] Kim JK, Bang SY, Choi S. Sequence-driven features for prediction of subcellular localization of proteins[J]. Pattern Recognition, 2006, 39(12): 2301-2311

[23] Kim JK, Raghava GPS, Bang SY, et al. Prediction of subcellular localization of proteins using pairwise sequence alignment and support vector machine[J]. Pattern Recognition Letters, 2006, 27(9): 996-1001

[24] Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition[J]. Proteins: Structure, Function, and Bioinformatics, 2001, 43(3): 246-255

[25] Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition[J]. Journal of Biochemistry, 1986, 99(1): 153-162

[26] Nakashima H, Nishikawa K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies[J]. Journal of Molecular Biology, 1994, 238(1): 54-61

[27] Wu C, Whitson G, McLarty J, et al. Protein classification artificial neural system[J]. Protein Science, 1992, 1(5): 667-677

[28] Chou KC, Shen HB. Recent progress in protein subcellular location prediction[J]. Analytical Biochemistry, 2007, 370(1): 1-16

[29] Chou KC, Shen HB. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms[J]. Nature Protocols, 2008, 3(2): 153-162

[30] Chou KC, Zhang CT. Prediction of protein structural classes[J]. Critical Reviews in Biochemistry and Molecular Biology, 1995, 30(4): 275-349