DOI:10.12113/202206004
:

基于机器学习的蛋白质编码区识别

包晓娜,何黎黎*,崔景安
(北京建筑大学 理学院 北京 102616)

摘 要:针对DNA序列编码区的识别问题,本研究提出一个特征向量和逻辑回归的组合模型。首先对DNA序列进行数值处理转化为特征向量,并结合k字符相对频率技术提取特征向量的元素特征,之后利用二分类逻辑回归算法,对编码区和非编码区进行准确区分。选取了HMR195和BG570两个基准数据集进行五折交叉验证,结果表明,平均AUC(Area Under Curve)值分别为0.981 3和0.987 4,明显优于传统的贝叶斯判别法和VOSSDFT等方法。此外,本文提出的特征向量的维度很低,提高了运算效率。因此,本文组合模型能够较为高效准确地识别蛋白质编码区。

关键词:编码区;特征向量;逻辑回归;机器学习
中图分类号:TP181 **文献标志码**:A **文章编号**:1672-5565(2023)04-270-07

Identification of protein coding region based on machine learning

BAO Xiaona, HE Lili*, CUI Jingan
(*School of Science*, *Beijing University of Civil Engineering and Architecture*, *Beijing* 102616, *China*)

Abstract: In order to identify the coding region of DNA sequence, a combined model of eigenvector and logistic regression is proposed in this article. Firstly, the DNA sequence is transformed into a feature vector by numerical processing, and the element features of the feature vector are extracted by combining the k-character relative frequency technology. Then, the binary classification logistic regression algorithm is used to accurately distinguish the coding region from the non-coding region. Two benchmark data sets, HMR195 and BG570, were selected for five-fold cross-validation. The results showed that the average AUC (Area Under Curve) values were 0.981 3 and 0.987 4 respectively, which are significantly better than the traditional Bayesian discriminant method and VOSSDFT. In addition, the dimension of the feature vector in this article is very low, which improves the operation efficiency. Therefore, the combined model in this article can identify protein coding regions more efficiently and accurately.

Keywords: Protein coding region; Feature vector; Logistic regression; Machine learning

大多数真核生物的编码区是不连续的,编码蛋白质的序列在基因序列中被非编码序列隔开(见图1)。编码的序列又称为外显子(Exon),携带着遗传信息,能够决定和指导生物的性状;非编码序列又称为内含子(Intron)[1]。如果一个基因有n个内含子,一般总是把基因的外显子分隔成n+1个部分。且内含子的核苷酸数量比外显子多许多倍[1-2]。因此,外显子和内含子的准确识别是一个具有挑战性的研究。外显子和内含子区分也有助于研究基因功能、基因表达、基因注释、基因转录调控,对于内含子功能的研究也具有一定的辅助作用[3-4],故外显子和内含子的分类具有重要的意义。

多年来,学者们已经提出了基因编码区(外显子)预测的多种方法。一般可以分为基于同源比对的方法和不依赖同源比对的方法。基于序列同源性的方法是以现有的基因数据库为标准,对待检测DNA序列进行相似性识别,从而根据已有经验判断未知序列的外显子和内含子区域。BLAST[5]、MUSCLE[6]是常见的比对工具,近年来也有诸如GeMoMa[7]的基因预测程序被提出。基于序列同源

收稿日期:2022-06-09;**修回日期**:2022-10-27.**网络首发日期**:2022-12-15.
网络首发地址:https://kns.cnki.net/kcms/detail//23.1513.q.20221214.1057.001.html
基金项目:国家自然科学基金项目(No.11871093);北京建筑大学青年教师科研能力提升计划项目(No.X21026).
*通信作者:何黎黎,女,讲师,研究方向:生物信息学.E-mail: helili@bucea.edu.cn.
引用格式:包晓娜,何黎黎,崔景安.基于机器学习的蛋白质编码区识别[J].生物信息学,2023,21(4):270-276.
BAO Xiaona, HE Lili, CUI Jingan.Identification of protein coding region based on machine learning[J].Chinese Journal of Bioinformatics, 2023,21(4):270-276.

性的方法准确率较高,但测序成本高、比对效率等因素制约了该项技术的发展。基于此,许多的学者将研究重点转向不依赖比对技术的模型。数字信号处理技术在该领域发挥着关键的作用[8]。且数字信号处理前通常需对 DNA 序列进行数值映射[9]。VOSS[10]是一种广泛使用的固定映射技术,它将 DNA 序列转化为 4 个二进制指示符序列 $X_A[n]$, $X_C[n]$,$X_G[n]$,$X_T[n]$。核苷酸在特定碱基位置出现用 1 表示,未出现用 0 表示。Z 曲线理论[11]是基于物理化学性质的映射方式。利用传统四面体的对称性开发,它将 DNA 或 RNA 序列映射到折叠曲线中。Z 曲线表示出 DNA 序列携带的所有信息[8],可用于基因鉴定和 DNA 或 RNA 序列分析[12]、识别细菌和古细菌基因组中蛋白质编码基因[13]等。此外,在众多序列编码方法中,k 字符相对频率技术(k-mer)[14]是较常见和简便的方法。图 2 展示了当 k 为 4 步长为 1 时的短序列的 k-mer 生成过程。机器学习的迅猛发展也为蛋白质编码区的识别带来了许多新的算法。如 CNN-MGP[15]、GeneMark EP+[16]、DBN[17]。CNN-MGP[15]是用于宏基因组学基因预测的卷积神经网络,能够提取编码区和非编码区的特征。GeneMark EP+[16]是用于真核基因预测的算法和工具。深度置信网络 DBN[17]通过多层玻尔兹曼机对 DNA 序列进行数值转换,用深度置信网络模型对外显子和内含子分类判别。尽管已经有许多的外显子与内含子分类方法被提出,但是准确率、敏感度、特异度、AUC 值等评价参数还有待提升。

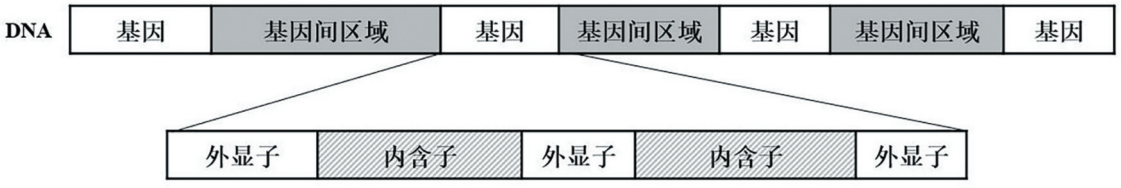


图 1　真核生物外显子与内含子交替示意图

Fig.1　Schematic diagram of exon intron alternation in eukaryotic coding region

图 2　k 字符相对频率技术提取 k-mer 示意图(k=4)

Fig.2　Schematic diagram of k-mer extraction by k-character relative frequency technology(k=4)

将数值映射和机器学习分类器相结合,提出了一个组合算法(具体流程见图 3)。首先,给定一个外显子或内含子,将其通过密码子与氨基酸的对应转换为特定的氨基酸序列,此处的转换不同于标准的翻译过程。然后,利用经典的 k-mer 技术获取序列的特征向量。最后,将外显子与内含子的特征向量输入逻辑回归分类器中,训练模型并识别蛋白质编码区(外显子)。利用真核生物基准数据集 HMR195 和 BG570 对模型进行了五折交叉验证,AUC 值分别达到了 0.981 3 和 0.987 4。将两个数据集合并计算时,敏感度和特异度分别为 0.954 1、0.942 8。通过对比发现,新算法的识别结果明显优于 VOSSDFT、传统的贝叶斯判别等方法。新算法识别 HMR195 和 BG570 数据集的时间为 1.46 s、3.58 s,表明组合模型能够高效又准确地鉴定真核生物的外显子和内含子。

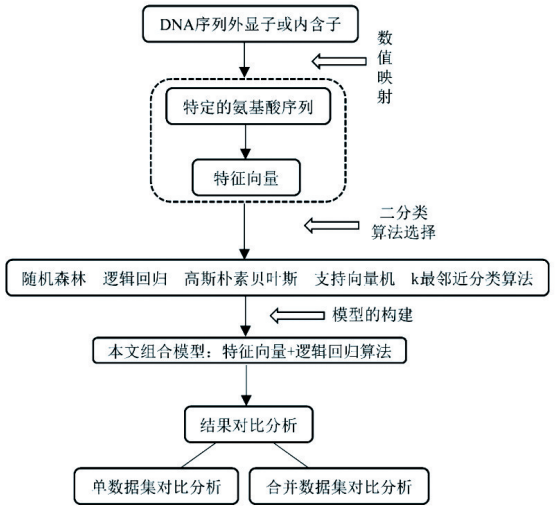


图 3　本文算法的框图

Fig.3　Block diagram of the algorithm in this article

1　数据

1.1　数据的获取

本文对真核生物的 DNA 序列进行编码区判别分析,实验中用到 2 个基准数据集,分别是 HMR195[18]和 BG570[19]数据。HMR195 数据由 195 个哺乳动物 DNA 序列组成,包括人类、小鼠和大鼠,

共 948 个外显子。BG570 是指由 570 个脊椎动物序列组成的基因组测试数据集,共2 649个外显子。两个数据集可从网址 http://www.imtech.res.in/raghava/genebench 中获取。基准数据集的长度范围、外显子数目和内含子数目如表 1 所示。为了保证对外显子和内含子分类的全面性,将短的(长度低于 20 bp)外显子和内含子序列也加入了实验中。

表 1 基准数据的外显子和内含子分布表

Table 1 Exon and intron distribution table of benchmark data

数据	序列数	长度范围	外显子	内含子
HMR195	195	795~56 500	948	1 143
BG570	570	398~36 845	2 649	3 211
总计	765	398~56 500	3 597	4 354

1.2 数据的预处理

1.2.1 DNA 序列的数值转化

在实现外显子和内含子的精准分类与预测前,通常需要对 DNA 序列进行数值映射,即将 DNA 序列转化为一个数值形式的表示[17]。本文提出了一个全新的 DNA 序列数值化映射方法,结合 k-mer 技术[14],将 DNA 序列中的外显子和内含子分别转化为一个特征向量。下面介绍特征向量的提取过程:

给定一个外显子 ACAGCGACC:第 1 步,从第一个核苷酸 A 处开始,通过每次仅移动一个核苷酸,将外显子转化为一段特定氨基酸序列,具体为,'ACA'对应氨基酸 T,'CAG'对应氨基酸 Q,'AGC'对应氨基酸 S,'GCG'对应氨基酸 A,'CGA'对应氨基酸 R,'GAC'对应氨基酸 D,'ACC'对应氨基酸 T,由此得到一段特定氨基酸序列为 TQSARDT;第 2 步,结合经典的 k 字符相对频率技术,规定 k 值从 1 至 5 变化,将 TQSARDT 转化为特征向量。假设 $k=2$,则 2-mer 种类包括 TQ、QS、SA、AR、RD、DT。特征向量由 2-mer 频数构成,即 $(f_{TQ}, f_{QS}, f_{SA}, f_{AR}, f_{RD}, f_{DT}) = (1,1,1,1,1,1)$,其中 f_{TQ} 表示 TQ 的频数。

1.2.2 DNA 序列的特征提取

DNA 序列特征提取源于特定氨基酸序列 k-mer 的种类和数值频率。具体来说,特征向量的元素(即所有的 k-mer 种类)是 DNA 序列的特征,即 1.2.1 节提到的TQ、QS等。通常来讲,一段氨基酸序列中的 k-mer 种类数为 20^k。但是,由于特定氨基酸序列的转化不同于生物学中标准的翻译过程,且存在不同的密码子对应同一种氨基酸,所以本算法的 k-mer 种类数远远少于 20^k,这大大节约了计算时的内存消耗。以脯氨酸 P 为例,如图 4,它由 4 个密码子编译 CCT、CCC、CCA、CCG,由本文的转化过程,P 后的下一个氨基酸共 5 种,分别为亮氨酸 L、组氨酸 H、谷氨酰胺 Q、脯氨酸 P、精氨酸 R,远少于 20 种。因此,每个氨基酸后可能出现的氨基酸种类少于 20 种。最终 k-mer 组合种类数随之大大减少,也就是说本文的转化过程大大降低了特征向量的维度。表 2 列出了 $k=2$ 时的全部 95 种特征向量的元素特征。

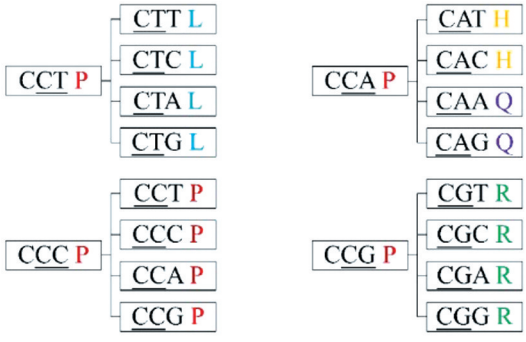


图 4 脯氨酸 P 后面会出现的氨基酸种类示意图

Fig.4 Schematic diagram of amino acid types that will appear after P

表 2 $k=2$ 时,特征向量的 95 种元素

Table 2 When k = 2, 95 elements of eigenvector

2-mer 种类	数目
AH AL AP AQ AR CA CV DI DM DT	10
EK EN ER ES FF FL FS GA GD GE GG GV	12
HI HM HT IF IK IL IN IR IS IY KK KN KR KS	14
LC LD LE LF LK LL LN LR LS LW LY	11
MC MD ME MW NI NM NT	7
PH PL PP PQ PR QK QN QR QS	9
RA RE RD RG RV SA SH SL SP SQ SR SV	12
TH TL TP TQ TR WG	6
VC VD VE VF VK VL VN VR VS VW VY	11
YI YM YT	3
总计	95

表 3 列出了部分短外显子或内含子的特征向量(以 $k=2$ 为例)详细求解过程。规定外显子类别为 1,内含子类别为 0。

表 3 当 $k=2$ 时,部分外显子和内含子序列的特征向量

Table 3 When $k=2$, eigenvectors of some exon and intron sequences

类别	序列	特定氨基酸序列	2-mer 特征							特征向量
			TQ	QS	SA	SQ	SP	PR	RA	
1	ACAGCG	TQSA	1	1	1	0	0	0	0	(1,1,1,0,0,0,0)
0	TCAGCG	SQSA	0	1	1	1	0	0	0	(0,1,1,1,0,0,0)
0	TCAGCA	SPRA	0	0	0	0	1	1	1	(0,0,0,0,1,1,1)
…	…	…	…	…	…	…	…	…	…	…

2　模型的构建

2.1　二分类算法的选择

在完成 DNA 序列的数值转换后,为了找到最适合特征向量的二分类模型,本文对五种分类器进行了尝试和验证,分别是随机森林(Random forest)[20]、逻辑回归(Logistic regression)[21]、高斯朴素贝叶斯(Gaussian naive bayes)[22]、支持向量机(SVM)[23]、k 最邻近分类算法(KNN)[24]。计算时,采用五折交叉验证[25]。五折交叉验证是判断分类器性能的一种统计分析方法。它将原始数据分为 5 组,不重复地抽取其中 4 组作为训练集,剩余的 1 组作为测试集,共得到 5 种测试结果,最终取用平均数。

为了对 5 种不同的算法进行有效的对比和测度,此处使用三个评价指标 ROC(Receiver operating characteristic)曲线、AUC 值和近似相关系数 AC 值。ROC 曲线[26]是以假阳率(False positive rate)作为横轴线(成本),以真阳率(True positive rate)作为纵轴线(收益),来说明在各种阈值条件下的假阳率和真阳率的关系曲线。ROC 曲线与对角线的距离愈接近,表明试验中识别编码区与非编码区的能力愈弱,亦即该方法的分类预测能力愈弱。为了更准确地描述算法的判别能力,通常将 ROC 曲线下的区域面积用 AUC[26]进行定量和比较,AUC 数值愈接近 1,说明分类的有效性越好。近似相关系数 AC[26]是一种得到普遍认可的综合评估指标,TP(True positive)为外显子被正确预测的个数,FP(False positive)为预测为外显子但实际为内含子的个数,TN(True negative)为内含子被正确预测的个数,FN(False negative)为预测为内含子但实际为外显子的个数。此外,为了检验结果的统计学显著性,采用 Delong 检验[27]对 ROC-AUC 进行成对比较,$p<0.05$ 被认为具有统计学意义。

$$AC=\frac{1}{2}\times\left[\left(\frac{TP}{TP+FN}+\frac{TP}{TP+FP}+\frac{TN}{TN+FP}+\frac{TN}{TN+FN}\right)-2\right] \tag{1}$$

具体实验结果如图 5、图 6 和表 4 所示。图 5 中,$k=2$ 时,在 HMR195 数据集,逻辑回归的 AUC 平均数分别为 0.981 3,明显高于其他模型的结果。如图 6,BG570 数据集也得到类似的结果,逻辑回归算法在所有 k 值优于其他算法。

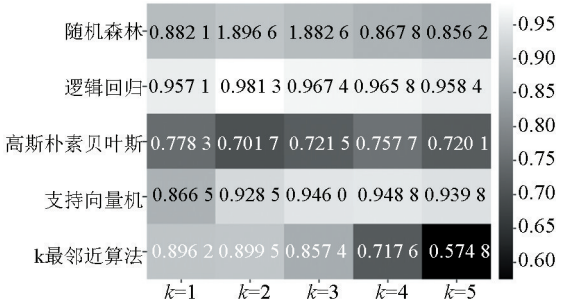


图 5　HMR195 数据集中,五种算法 AUC 值的热图

Fig.5　In HMR195 data set, heatmaps of AUC values of five algorithms

AC 值对比结果如表 4 所示,$k=2$ 时,逻辑回归取得了最大的 AC 值。AC 值兼顾了 TP、TN、FP、FN 四个参数的值。AC 值越大,表明分类效果越好。同时可以发现,当 k 取其他值时,逻辑回归算法相较其余四种方法也具有明显的优势。因此,由特征向量与逻辑回归组合的分类模型较准确。

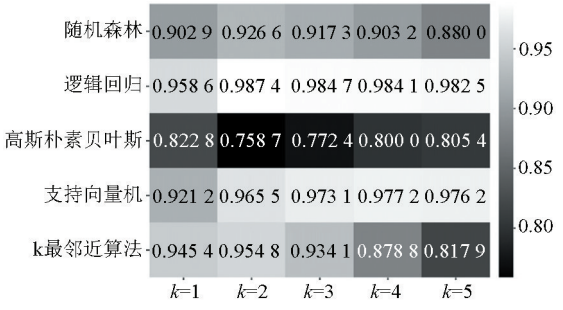


图 6　BG570 数据集中,五种算法 AUC 值的热图

Fig.6　In BG570 data set, heatmaps of AUC values of five algorithms

2.2　组合模型的确定

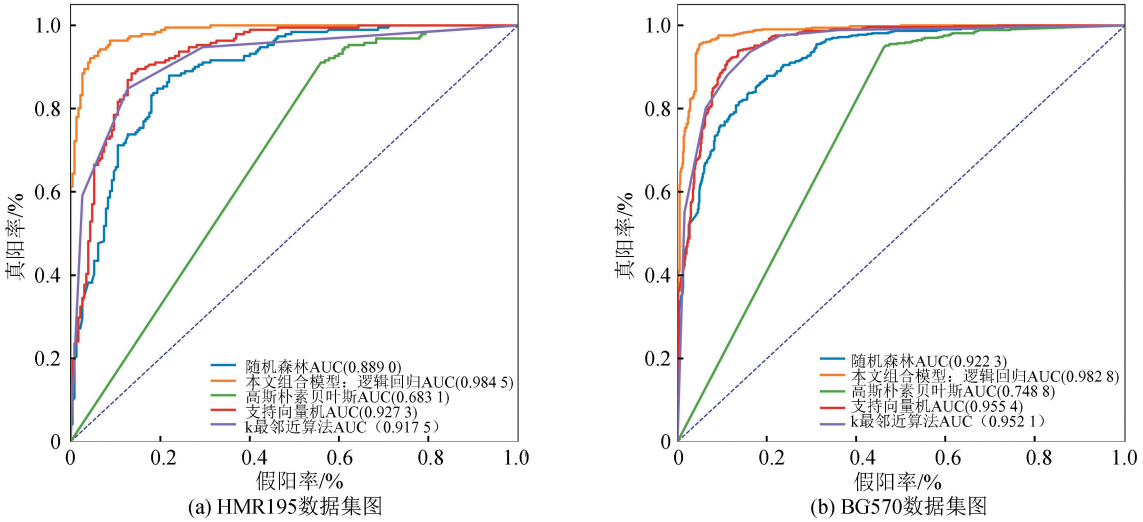
最终,组合模型确定为特征向量与逻辑回归分类器的结合。首先,将 DNA 序列转化为特定的氨基酸序列;其次,由特定氨基酸序列得到特征向量。最后,将特征向量放入逻辑回归分类器中,获得外显子和内含子的预测结果。

如图 7,选取五折交叉验证中的一次实验结果,画出 ROC 曲线图($k=2$)。可以明显看出,组合模型最贴近面积为 1 的四边形线,分类效果较好。并且,HMR195 的结果具有统计学显著性(逻辑回归 VS 随机森林:$p=5.07\times10^{-8}$;逻辑回归 VS 朴素贝叶斯:$p=4.99\times10^{-16}$;逻辑回归 VS 支持向量机:$p=7.74\times10^{-10}$;逻辑回归 VS k 最邻近算法:$p=8.91\times10^{-7}$)。BG570 数据的试验结果也显著(逻辑回归 VS 随机森林:$p=8.05\times10^{-16}$;逻辑回归 VS 朴素贝叶斯:$p=3.70\times10^{-54}$;逻辑回归 VS 支持向量机:$p=4.67\times10^{-9}$;逻辑回归 VS k 最邻近算法:$p=1.24\times10^{-7}$)。

表 4 k 从 1 至 5,5 种算法的 AC 平均值

Table 4 K from 1 to 5, mean AC value of 5 algorithms

数据集	k 值	逻辑回归	随机森林	朴素贝叶斯	支持向量机	k 最邻近算法
HMR195	1	0.771 0	0.588 9	0.313 0	0.599 7	0.680 0
	2	0.873 4	0.662 6	0.437 1	0.671 5	0.682 9
	3	0.817 9	0.633 0	0.522 8	0.706 0	0.532 8
	4	0.814 8	0.583 0	0.587 5	0.725 6	0.230 7
	5	0.809 9	0.563 4	0.632 2	0.692 6	0.105 6
BG570	1	0.792 2	0.664 7	0.327 9	0.697 7	0.774 7
	2	0.896 5	0.691 1	0.463 2	0.789 8	0.805 4
	3	0.884 9	0.680 0	0.539 5	0.831 3	0.754 9
	4	0.873 1	0.661 3	0.594 1	0.848 8	0.633 7
	5	0.873 8	0.624 2	0.639 5	0.829 4	0.516 6



(a) HMR195数据集图 (b) BG570数据集图

图 7 五个算法模型的 ROC 曲线图

Fig.7 ROC curves of 5 algorithm models

3 实验结果

3.1 单独数据集对比分析

为了说明本文新方法与其余方法的优劣,将其与经典的 VOSSDFT[10,28]、EIIPDFT[28-29]、SPDFT[28,30]和 Code13-Marple[28]进行了比较。VOSSDFT、EIIPDFT、SPDFT 均是基于离散傅里叶变换的技术(Discrete Fourier Transform, DFT)来区分真核生物外显子和内含子[10,29-30]。Code13-Marple 是一种基于自回归谱分析和小波变换的集成算法。由表 5,以 HMR195 为例,新方法(k=2)的 AUC 值达到了0.981 3,比其余四种方法分别高出了 0.418 7、0.470 0、0.385 1、0.263 4;在 BG570 数据集上,AUC 和 AC 值也远远超过其余四种模型中的最大值。新算法明显优于其他三种传统的基于 DFT 的方法和 Code13-Marple。

表 5 组合模型与其他方法的比较

Table 5 Comparison of eigenvector method with other methods

数据集	评估指标	本文组合模型	VOSSDFT	EIIPDFT	SPDFT	Code13-Marple
HMR195	AUC	0.981 3	0.562 6	0.511 3	0.596 2	0.717 9
	AC	0.873 4	0.104 5	0.057 2	0.130 0	0.250 8
BG570	AUC	0.987 4	0.532 9	0.486 7	0.547 0	0.652 2
	AC	0.896 5	0.109 3	0.059 9	0.126 3	0.126 3

3.2 合并数据集对比分析

为验证算法在较大数据集上的分类效果,将 HMR195 和 BG570 两组数据合并得到合并数据集,共 3 597 个外显子、4 354 个内含子。此外,为了更加全面的评估组合模型的性能,增加了准确率、敏感度、特异度以及运行时间这四个对比维度,并与经典的贝叶斯判别法[31]进行比较。贝叶斯判别法是进行判别分析的一种多元统计分析方法。合并数据集后,k 值取 3 时本文算法得到最好的预测结果。

$$acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$S_n = \frac{TP}{TP + FN} \tag{3}$$

$$S_p = \frac{TN}{TN + FP} \quad (4)$$

表 6 是两种方法的对比分析表,其中准确率 acc[26]为全部序列中被正确预测的序列的比例;敏感度 S_n [26]为所有实际外显子中被正确预测为外显子的比例;特异度 S_p [26]为所有真实的内含子被正确预测为内含子的比例。在合并后的较大数据集上,组合模型的敏感度 S_n 为 0.954 1 远远大于贝叶斯判别法的 0.787 2。在运行时间方面,组合模型只需要 8.91 s,而贝叶斯判别法需要 27.28 s。因此,本文方法不仅适用于小数据集,在较大数据集上同样表现优异,并且运行速度快于贝叶斯判别法。本文组合模型以及贝叶斯判别法的计算基于处理器为 Intel(R) Core(TM) i7-8550U CPU@ 1.80 GHz 和 16.0 GB RAM 的设备,使用 Python3.8 编程获得。

表 6　二种模型的比对结果分析表

Table 6　Analysis table of comparison results of two models

数据集	方法	AUC	AC	准确率	敏感度	特异度	运行时间/s
HMR195	贝叶斯判别法	0.720 1	0.626 7	0.815 4	0.771 7	0.851 3	6.94
	本文组合模型	0.981 3	0.878 4	0.939 2	0.944 2	0.934 9	1.46
BG570	贝叶斯判别法	0.805 4	0.641 5	0.822 6	0.795 0	0.845 5	19.89
	本文组合模型	0.987 4	0.898 5	0.949 5	0.957 6	0.942 6	3.58
合并数据集	贝叶斯判别法	0.796 3	0.528 7	0.795 7	0.787 2	0.829 8	27.18
	本文组合模型	0.986 8	0.895 4	0.947 9	0.954 1	0.942 8	8.91

4　结论及展望

本研究提出了一个基于特征向量的数值映射方法,之后结合逻辑回归算法,对基因外显子和内含子实现了精确的分类。将组合模型运用到编码区识别,给出了一个全新的研究视角。为了证明组合模型的可行性,利用 HMR195 和 BG570 两个真核生物数据集,将其与现有的成熟方法进行了对比(见表 5 和表 6),均证明了它的有效性。此外,为证实模型在更大数据集上的效果,本文新收集了 462 条人类 DNA 序列[32]进行试验,共包含2 843个外显子,2 381 个内含子。全部数据可从网址 https://www.fruitfly.org/sequence/human-datasets.html 获取。当全部数据共同训练时,共 6 440 个外显子,6 735 个内含子。本文方法实验结果:acc、S_n 、S_p 、AC、AUC 的值分别为 0.957 7、0.966 6、0.949 0、0.915 5、0.989 4(k=2)。当扩大数据集后,组合模型对于外显子和内含子依然能起到很好的识别效果。其次,1.2.2 节中特征向量的提取过程充分利用了密码子的简并性,降低了特征向量的维度。然而本文还未将外显子和内含子的结构信息作为特征的重要因素,之后的研究中会考虑加入结构信息,从而进一步提升模型的性能。并且,本文后续研究仍将扩大样本量,尝试更多更全面物种的蛋白质编码区分类,争取构建快速便捷的外显子与内含子识别工具。

参考文献(References)

[1]KORALEWSKI T E, KRUTOVSKY K V. Evolution of exon-intron structure and alternative splicing[J]. PLoS One, 2011, 6(3): e18055. DOI: 10.1371/journal.pone.0018055.

[2]SABERKARI H, SHAMSI M, HERAVI H, et al. A fast algorithm for exonic regions prediction in DNA sequences[J]. Journal of Medical Signals & Sensors, 2013, 3: 139-49. DOI: 10.4103/2228-7477.120977.

[3]ABELSON J, TROTTA C R, LI H. tRNA splicing[J]. Journal of Biological Chemistry, 1998, 273(21): 12685-12688. DOI: 10.1074/jbc.273.21.12685.

[4]KAR S, GANGULY M. Study of effectiveness of FIR and IIR filters in exon identification: A comparative approach[J]. Materials Today: Proceedings, 2022, 58(1): 437-444. DOI: 10.1016/j.matpr.2022.02.394.

[5]ALTSCHUL S F. Basic local alignment search tool (BLAST)[J]. Journal of Molecular Biology, 1990, 215(3): 403-410. DOI: 10.1016/S0022-2836(05)80360-2.

[6]EDGAR R C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput[J]. Nucleic Acids Research, 2004, 5: 1792-1797. DOI: 10.2460/ajvr.69.1.82.

[7]JENS K, MICHAEL W, ERICKSON J L, et al. Using intron position conservation for homology-based gene prediction[J]. Nucleic Acids Research, 2016, 44(9): e89. DOI: 10.1093/nar/gkw092.

[8]ABO-ZAHHAD M, AHMED S M, ABD-ELRAHMAN S A. Genomic analysis and classification of exon and intron sequences using dna numerical mapping techniques[J]. International Journal of Information Technology & Computer Science, 2012, 4(8): 22-36. DOI: 10.5815/ijitcs.2012.08.03.

[9]SABERKARI H, SHAMSI M, SEDAAGHI M, et al. Prediction of protein coding regions in DNA sequences using signal processing methods[C]// 2012 IEEE Symposium on Industrial Electronics and Applications, 2012, 355-360. DOI: 10.1109/ISIEA.2012.6496660.

[10]VOSS R F. Evolution of long-range fractal correlations and

1/f noise in DNA base sequences[J]. Physical Review Letters, 1992, 68(25): 3805-3808. DOI: 10.1103/PhysRevLett.68.3805.

[11] GUO F, ZHANG C T. ZCURVE_V: A new self-training system for recognizing protein-coding genes in viral and phage genomes[J]. BMC Bioinformatics, 2006, 7(1): 9. DOI: 10.1186/1471-2105-7-9.

[12] CHEN Jiahai, LIU Yongmin, LIAO Qing, et al. iEsGene-ZCPseKNC: Identify essential genes based on z curve pseudo k-Tuple nucleotide composition [J]. IEEE Access, 2019, 7: 165242. DOI: 10.1109/ACCESS.2019.2952237.

[13] GUO Fengbiao, OU Hongyu, ZHANG Chunting. ZCURVE: A new system for recognizing protein-coding genes in bacterial and archaeal genomes[J]. Nucleic Acids Research, 2003, 31(6): 1780-1789. DOI: 10.1093/nar/gkg254.

[14] LÜ Hao, ZHANG Zimei, LI Shihao, et al. Evaluation of different computational methods on 5-methylcytosine sites identification[J]. Briefings in Bioinformatics, 2019, 21(3): 982 - 995. DOI: 10.1093/bib/bbz048.

[15] AMANI A A, ACHRAF E A. CNN-MGP: Convolutional neural networks for metagenomics gene prediction[J]. Interdiplinary Ences Computational Life Ences, 2018, 11: 628 - 635. DOI: 10.1007/s12539-018-0313-4.

[16] BRŮNA T, LOMSADZE A, BORODOVSKY M. GeneMark-EP+: Eukaryotic gene prediction with self-raining in the space of genes and proteins[J]. NAR Genomics and Bioinformatics, 2020, 2(2): 1-14. DOI: 10.1093/nargab/lqaa026.

[17] 胡青渝, 刘广臣. DBN 在蛋白质编码区识别问题中的应用研究[J]. 计算机工程与应用, 2020, 56(4): 247-255. DOI: 10.3778/j.issn.1002-8331.1811-0045.
HU Qingyu, LIU Guangchen. Application of deep belief network in recognition of protein coding regions [J]. Computer Engineering and Applications, 2020, 56(4): 9: 247-255. DOI: 10.3778/j.issn.1002-8331.1811-0045.

[18] ROGIC S, MACKWORTH A K, OUELLETTE F B F. Evaluation of gene-finding programs on mammalian sequences[J]. Genome Research, 2001, 11: 817-832. DOI: 10.1101/gr.147901.

[19] BURSET M, GUIGÓ R. Evaluation of gene structure prediction programs[J]. Genomics, 1996, 34(3): 353-367. DOI: 10.1006/geno.1996.0298.

[20] MANAVALAN B, SHIN T H, KIM M O, et al. AIPpred: Sequence-based prediction of anti-inflammatory peptides using random forest[J]. Front Pharmacol, 2018, 9: 276. DOI: 10.3389/fphar.2018.00276.

[21] COURONNÉ R, PROBST P, BOULESTEIX A. Random forest versus logistic regression: A large-scale benchmark experiment[J]. BMC Bioinformatics, 2018, 19(1): 270. DOI: 10.1186/s12859-018-2264-5.

[22] LOU Wangchao, WANG Xiaoqing, CHEN Fan, et al. Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naïve Bayes[J]. PLoS One, 2014, 9(1): e86703. DOI: 10.1371/journal.pone.0086703.

[23] MANAVALAN B, SHIN T H, LEE G. PVP-SVM: Sequence-based prediction of phage virion proteins using a support vector machine[J]. Front Microbiology, 2018, 9: 476. DOI: 10.3389/fmicb.2018.00476.

[24] DENG Zhenyun, ZHU Xiaoshu, CHENG Debo, et al. Efficient kNN classification algorithm for big data[J]. Neurocomputing, 2016, 195: 143-148. DOI: 10.1016/j.neucom.2015.08.112.

[25] CHEN Xing, HUANG Yuan, YOU Zhuhong, et al. A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases[J]. Bioinformatics, 2016, 33(5): 733-739. DOI: 10.1093/bioinformatics/btw715.

[26] 马玉韬. 基于滤波理论和特征统计的蛋白质编码区预测算法研究[D]. 天津:天津大学,2013. DOI: 10.7666/d.D439439.
MA Yutao. Research on protein coding regions prediction algorithms based on filtering theories and statistics of characteristics of DNA sequences[D]. Tianjin: Tianjin University,2013. DOI: 10.7666/d.D439439.

[27] DELONG E R, DELONG D M, CLARKE-PEARSON D L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach [J]. Biometrics, 1988, 44(3): 837-845. DOI: 10.2307/2531595.

[28] LIU Guangchen C, LUAN Yihui. Identification of protein coding regions in the eukaryotic DNA sequences based on marple algorithm and wavelet packets transform[J]. Abstract and Applied Analysis, 2014, 2014: 402567. DOI: 10.1155/2014/402567.

[29] MAI S M. A study of the potential of EIIP mapping method in exon prediction using the frequency domain Tech-niques [J]. American Journal of Biomedical Engineering, 2012, 2(2): 17-22. DOI: 10.5923/j.ajbe.20120202.04.

[30] ZHANG Weifeng, YAN Hong. Exon prediction using empirical mode decomposition and Fourier transform of structural profiles of DNA sequences[J]. Pattern Recognition, 2012, 45(3): 947-955. DOI: 10.1016/j.patcog.2011.08.016.

[31] MORAES R M, FERREIRA J A, MACHADO L S. A new bayesian network based on gaussian naive bayes with fuzzy pa-rameters for training assessment in virtual simulators [J]. International Journal of Fuzzy Systems, 2021, 23: 849-861. DOI: 10.1007/s40815-020-00936-4.

[32] KULP D, HAUSSLER D, REESE M G, et al. A generalized hidden Markov model for the recognition of human genes in DNA[J]. Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology, 1996, 4: 134-142. DOI: 10.5555/645631.662887.