DOI: 10.13718/j.cnki.xdzk.2017.04.021

微孢子虫 PolyA 位点的预测①

孙　康[1]，　杨　明[1]，　马　立[1]，　李　田[2]，　赵玉芳[3]

1. 西南大学 计算机与信息科学学院，重庆 400715；
2. 西南大学 生物技术学院 家蚕基因组生物学国家重点实验室，重庆 400716；
3. 西南大学 教师教学发展中心，重庆 400715

摘要：多聚腺苷酸化是真核细胞内形成成熟 mRNA 的一个重要步骤，其位点的预测对基因组序列中编码基因的发掘具有重要的参考价值. 本研究以缺乏有效基因预测方法的微孢子虫基因组为对象，根据该物种的基因表达偏好设计了一个算法，对其 PolyA 位点进行预测分析. 首先，采用 k 阶核苷酸频率形式和位置权重矩阵形成初始的特征，然后用 PCA 降低特征空间的维数，得到的数据用机器学习方法进行分析，产生一个较好的分类结果. 其中基于支持向量机的实验得到的敏感度(S_p)和 ACC 分别达到了 87.33%和 85.14%，这在微孢子虫的 PolyA 位点预测上取得了较为理想的效果，并为以后机器学习算法在微孢子虫基因预测领域做了很好的尝试.

关　键　词：PolyA 信号；微孢子虫；位置权重矩阵；机器学习

中图分类号：TP399　　**文献标志码**：A　　**文章编号**：1673-9868(2017)04-0138-06

微孢子虫是一类专性细胞内寄生的单细胞真核生物，宿主范围广泛，能够寄生在几乎所有的无脊椎动物和脊椎动物上[1]. 微孢子虫的细胞核含有多条染色体，核糖体为原核型的 70S 核糖体. 微孢子虫的基因组高度减缩，绝大部分编码基因丢失了内含子，基因及基因间区的长度均变短，如感染哺乳动物的兔脑炎微孢子虫的基因平均长度为 1 080 bp，基因间区长度仅为 129 bp[2]. 微孢子虫基因组的特殊性导致其编码基因的准确预测是一个亟待解决的问题. 绝大多数真核基因具有多个 PolyA 位点，在形成成熟的 mRNA 过程中，外界环境的细微改变导致在 mRNA 的不同剪切位点进行选择性剪切和多聚腺苷酸化，这个现象叫做可选择性多聚腺苷酸化(APA). APA 能够影响胞外信号刺激、生长与发育、细胞增殖和多种疾病的发生发展[3-4]. 多聚腺苷酸化是真核细胞内 mRNA 转录后处理形成成熟 mRNA 的一个重要步骤(mRNA 转录处理的 3 个主要步骤分别是：5' 帽子结构的形成，内含子的剪切和 3' 端加尾巴[5])，它影响着基因的表达，对预测基因结构有着巨大的作用. 多聚腺苷酸化作用机制：切割及多聚腺苷酸化特异因子(Cleavage Polyadenylation Specific Factor，CPSF)绑定到 PolyA 信号序列，切割活化因子(Cleavage stimulation Factor，CstF)识别下游的 U-rich 和 G/U-rich 序列并相互作用，切割因子 CFI 在 PolyA 信号和下游作用元件之间的某个位置对前体 mRNA 进行分裂，最后在 PolyA 聚合酶的作用下添加多聚腺苷酸尾巴[6]. PolyA 位点结构特征如图 1 所示.

由于 PolyA 位点的预测对基因结构的分析和 mRNA 的形成机制有着重要的作用，近年来 PolyA 信号的预测引起了越来越多的关注. 起初，人们基于线性判别函数的原理设计出了 POLYAH[7]；1999 年，

① 收稿日期：2016-06-08
基金项目：国家自然科学基金面上项目(31371055)；中央高校基本业务费专项资助项目(XDJK2015A010).
作者简介：孙　康(1990-)，女，河南郑州人，硕士研究生，主要从事人工智能与机器学习方面的研究.

Graber等利用马尔科夫模型来预测PolyA位点[8]；2003年，通过Erpin统计PolyA位点上下游的序列中一些位置特异的二核苷酸对所出现的频率进行预测[9]；随后开始采用机器学习的方法来预测PolyA位点，例如基于SVM的PROBE[10]；现在人们最常用的是POLYAR方法，它把PolyA位点分成3类(PAS-strong，PAS-weak，PAS-less)进行预测[11]，还有研究者利用神经网络进行预测[12]. 然而这些方法却很少在微孢子虫基因组上尝试，主要原因在于人类基因与植物基因的研究颇为成熟，许多软件是针对研究较为成熟的人类基因或者植物基因表达偏好而设计的. 目前已报道的微孢子虫有1 400多种，不同种属微孢子虫的基因组大小差异较大(2.3 Mbp～24 Mbp)[13-15]. 人类基因和植物基因相较于微孢子虫基因更为复杂，然而许多方法即使能够用于病原体的研究，结果也会存在较大误差.

本文以Encephalitozoon cuniculi的基因组为数据材料[16-17]，对其进行分析，设计了微孢子虫PolyA信号的特异性预测算法. 文中提出了一种新的特征提取方法，然后基于SVM机器学习算法来对PolyA信号进行预测. 具体过程是通过运用一些特征表达方法，例如位置权重矩阵(PWM)[18]、k阶核苷酸出现频率来进行特征提取，运用主成分分析(PCA)法进行冗余特征的筛选，最后使用SVM分类器进行分类训练，从而建立微孢子虫PolyA信号的预测算法.

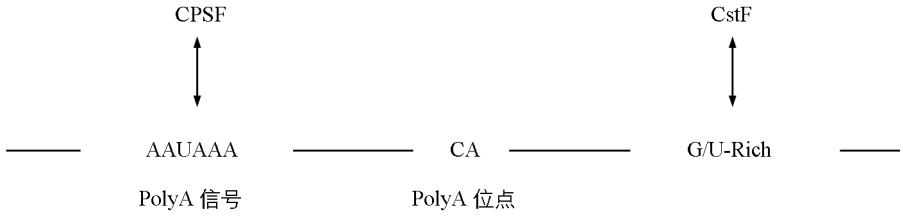


图1 PolyA位点结构特征

1 方 法

1.1 数据集

本研究从NCBI数据库中下载Encephalitozoon cuniculi的基因组序列作为数据集. 数据分为训练集和测试集. 扫描全部基因序列，得到348条含有AATAAA的片段，训练集和测试集的正集各取174条序列，余下的序列分别作为训练集和测试集的负集，经过实验取证，本文用174条序列作为训练集的负集，821条序列作为测试集的负集. 并且对数据集中的每一条序列都进行处理，剪切形成以PolyA信号AATAAA为中心，前后各100 bp的核苷酸，共206个核苷酸长的序列.

1.2 支持向量机(SVM)

现在比较流行的生物数据特征分析是机器学习领域的支持向量机(SVM)[19]，SVM经常用于数据分类和回归问题. 它的原理是将所有待分类的点映射到高维空间，然后在高维空间中找到一个能将这些点分开的“超平面”. SVM提供了一种避开高维空间的复杂性，直接用此空间的内积函数(核函数)，再利用线性可分情况下的求解方法直接求解对应的高维空间的决策问题. 当核函数已知时，可以简化高维空间问题的求解难度. 支持向量机具有很好的泛化推广能力，它在生物信息领域的应用越来越广. SVM的核函数有3种：线性核函数、多项式核函数、高斯核函数[20]. 考虑到SVM具有良好的分类效果和在生物学中的广泛应用，文中将通过特征提取和空间降维所得到的数据，运用多项式核函数来进行分类.

1.3 特征提取

PolyA的序列与其他序列(例如启动子)保守性较弱，使得很难分析PolyA位点的位置保守性. 近几年位置权重矩阵(PWM)已经被广泛应用到分子生物学中，用来描述PolyA位点附近的碱基保守水平. 构建位置权重矩阵时，横坐标代表4个核苷酸碱基(A、T、C、G)，纵坐标代表在这条基序中对应的位置信息，矩阵中的值代表每一个可能出现的核苷酸在对应位置的频率. 计算公式如下：

$$f_{i,b}=n_{i,b}/N \tag{1}$$

$n_{i,b}(i=1, 2, \cdots, N; b=A, T, C, G)$代表的是序列的第 b 个碱基在第 i 个位置上出现的次数，其中 N 是所对应的序列的碱基总数.

PWM 被定义为

$$S_i = \ln(f_{i,b}/\theta_{0,b}) \tag{2}$$

$\theta_{0,b}(b=A, T, C, G)$ 是序列上碱基 b 出现的随机频率，为了避免分母为 0 的情形，令 $\theta_{0,b}=0.25^3$. $f_{i,b}$ 是碱基 b 在位置 i 出现的频率. 在这个算法中，用 3-mer 频率作为位置权重矩阵的参数.

k 阶是长度为 $k(k=1, 2, 3, \cdots)$的核苷酸的低聚物. 例如 A 为 1 阶核苷酸类型，CG 为 2 阶核苷酸类型，ACG 为 3 阶核苷酸类型. 借鉴蛋白质序列的特征组织方式来形成特征，本算法中的特征提取只考虑 1～3 阶的核苷酸出现的频率. 对 PolyA 信号前后的 200 bp 序列综合考虑，共得到 84 个特征，在此基础上，本文又提取了 T 在上游和下游序列出现的频率和 G 在下游出现的频率共得到 87 个特征数据.

提取得到的原始特征空间的维度很大，变量太多会增加计算量和问题的复杂性，为了更加全面系统地分析问题，在这里我们采用 PCA 的方法来进行空间的降维. PCA 的计算步骤如下：① 计算相关系数矩阵；② 计算特征值和特征向量；③ 计算主成分载荷；④ 得出综合信息并进行递减排序，从而降低原始空间的维度.

对于一个给定的微孢子虫 PolyA 位点，该 PolyA 位点预测基本流程图如图 2 所示.

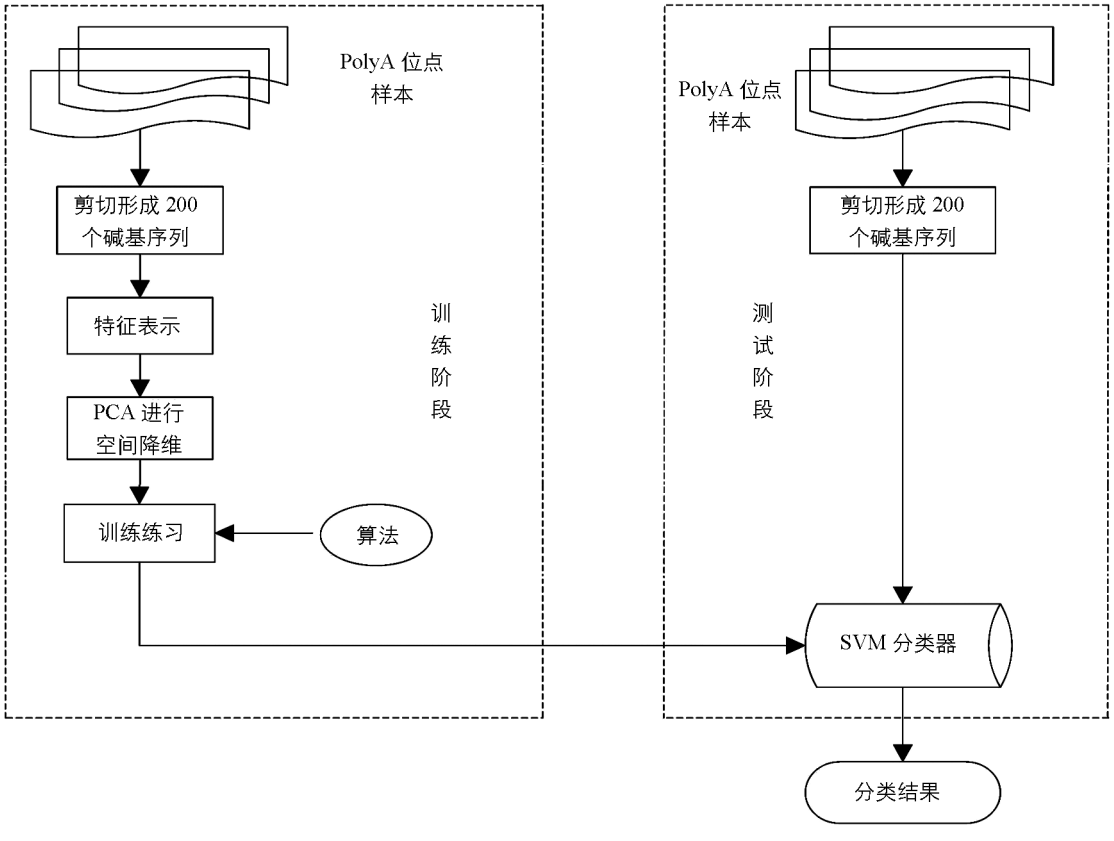


图 2　PolyA 位点预测基本流程图

2　结果与分析

2.1　评价指标

本文通过敏感度(Sensitivity，S_n)、特异度(Specificity，S_p)[21]、准确度(Accuracy，ACC)、假阳性率(False Positive Rate，FPR)[22]和真阳性率(True Positive Rate，TPR)来评价模型的好坏. 它们的定义如下：

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

$$S_p = \frac{TN}{TN + FP}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

$$TPR = \frac{TP}{TP + FN}$$

其中：TP 表示真阳性，FP 表示假阳性，TN 表示真阴性，FN 表示假阴性.

ROC 曲线的横纵坐标分别为 FPR 和 TPR，它代表横纵坐标之间的协同变化关系，是一种分类模型的评判标准. ROC 曲线被广泛应用在分类模型的评价指标中，它的精准率一般用曲线下的面积表示，曲线下的面积越小，表示模型越不精准.

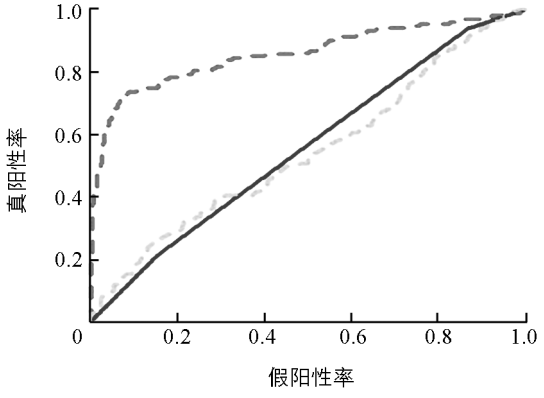
2.2 微孢子虫 PolyA 位点的识别算法比较

经过特征提取得到原始空间的特征，然后用 PCA 进行降维，再用机器学习算法进行分类分析. 分别对原始空间的特征和 PCA 降维后空间的特征进行分类，得出的特异度、敏感度、精确度大致相同. 由于家蚕病原体的 Encephalitozoon cuniculi 的基因组序列的实验数据集的限制，本文使用原始空间的特征进行实验，采用 SVM、决策树算法和 KNN 算法对本实验的数据进行分析比较，但是决策树算法和 KNN 算法分类出来的效果并没有 SVM 的好，SVM 算法的精确度能达到 85.14%，因此 SVM 算法有明显的算法优势. 表 1 为微孢子虫 PolyA 位点识别算法的性能表.

表 1 微孢子虫 PolyA 位点的识别算法的性能

	Sn/%	Sp/%	ACC/%
SVM 算法	74.86	87.33	85.14
决策树算法	53.71	49.09	50
KNN 算法	61.7	45	48.2

在相同数据的情况下，用 ROC 曲线来对 SVM 算法、决策树算法和 KNN 算法作分析，得到的结果如图 3 所示，图中的红色虚线表示 SVM 算法，黄色虚线表示决策树算法，蓝色实线代表 KNN 算法. 在微孢子虫 PolyA 位点的特异性预测算法中，SVM 算法表现出了较高的性能.



红色虚线代表 SVM 算法，黄色虚线代表决策树算法，蓝色实线代表 KNN 算法.

图 3 决策树，KNN 算法和 SVM 算法的 ROC 曲线图

3 结 论

机器学习在选择性剪切位点预测、启动子的预测等应用中已经十分普遍，本文采用 SVM 机器学习的算法来对微孢子虫 PolyA 位点进行预测分析. 为了使结果更加精确，根据 PolyA 位点附近的特征(下游是 U,U/G 丰富的区域)，在提取 1～3 阶的核苷酸出现的频率的基础上又提取了 T 在上游和下游出现的频率以及 G 在下游出现的频率来作为特征数据. 使用目前应用非常广泛的特征提取方法—位置权重矩阵(PWM)，并运用核苷酸的三阶频率来计算位置权重矩阵. 然后用 PCA 的方法进行数据降维，最后用机器学习中的 SVM 算法来进行分类，通过对模型的评估与实验，

得出了一个较好的结果. 家蚕病原体的 Encephalitozoon cuniculi 的基因组序列的数据集不像人类基因组序列的实验数据集那样的庞大、丰富，虽然通过 PCA 降维去除了一些冗余特征，但由于研究对象的实验数据的限制，本文采用 PCA 降维之前的特征进行实验，与决策树和 KNN 算法相比，SVM 有较大的优势. 由于微孢子虫的基因组并没有一个完整注释 PolyA 位点的数据库，虽然算法产生出一个良好的结果，但在这一点还是有限制的.

由于微孢子虫的基因库还不够完善，在这里用 SVM 算法对其进行分类预测，得出了较高的 S_n，S_p 和 ACC. 相较于在水稻基因中的研究，本算法的精确度还是比较高的. 在接下来的研究中，我们将更加深入地了解微孢子虫的基因结构特征，进而对算法进行改进，使之更加精确和实用. 我们希望通过运用计算机知识对微孢子虫进行 PolyA 位点预测，为生物学微孢子虫的研究者提供一个更好的思路.

参考文献：

[1] 罗　洁，林立鹏，潘国庆，等. 家蚕微孢子虫 NbTom40 的原核表达及定位 [J]. 西南大学学报(自然科学版)，2013，35(5)：30－36.

[2] KATINKA M D，DUPRAT S，CORNILLOT E，et al. Genome Sequence and Gene Compaction of the Eukaryote Parasite Encephalitozoon Cuniculi [J]. Nature，2001，414(6862)：450－453.

[3] 杨　亮，张红星，崔　英，等. 可选择性多聚腺苷酸化的生物学功能 [J]. 军事医学，2015，39(5)：393－397.

[4] EARLY P，ROGERS J，DAVIS M，et al. Two mRNAs Can be Produced from a Single Immunoglobulin Mu Gene by Alternative Rna Processing Pathways [J]. Cell，1980，20(2)：313－319.

[5] HAFEZ D，NI T，MUKHERJEE S，et al. Genome-Wide Identification and Predictive Modeling of Tissue-Specificalter Native Polyadenylation [J]. Bioinformatics，2013，29(13)：i108－i116.

[6] ZHAO J，HYMAN L，MOORE C. Formation of mRNA 3' Ends in Eukaryotes：Mechanism，Regulation，and Interrelationships with Other Steps in mRNA Synthesis [J]. Microbiol Mol Biol Rev，1999，63(2)：405－445.

[7] SALAMOV A A，SOLOVYEV V V. Recognition of 3'-Processing Sites of Human mRNA Precursors [J]. Comput Appl Biosci，1997，13(1)：23－28.

[8] GRABER J H，CANTOR C R，MOHR S C，et al. Genomic Detection of New Yeast Pre-mRNA 3'-End-Processing Signals [J]. Nucleic Acids Research，1999，27(3)：888－894.

[9] LEGENDRE M，GAUTHERET D. Sequence Determinants in Human Polyadenylation Site Selection [J]. BMC Genomics，2003，4(1)：7.

[10] CHENG Y，MIURA R M，TIAN B，Prediction of mRNA Polyadenylation Sites by Support Vector Machine [J]. Bioinformatics，2006，22(19)：2320－2325.

[11] AKHTAR M N，BUKHARI S A，FAZAL Z，et al. POLYAR，a New Computer Program for Prediction of Poly(A) Sites in Human Sequences [J]. BMC Genomics，2010，11：646.

[12] HAN J，LIU Z，ZHONG D，et al. A Hybrid Model for the Prediction of mRNA Polyadenylation Signals [J]. Conf Proc IEEE Eng Med Biol Soc，2013，2013：3511－3514.

[13] VAVRA J，LUKES J. Microsporidia and 'the Art of Living Together' [J]. Adv Parasitol，2013，82：253－319.

[14] PEYRETAILLADE E，ELALAOUI H，DIOGON M，et al. Extreme Reduction and Compaction of Microsporidian Genomes [J]. Res Microbiol，2011，162(6)：598－606.

[15] KEELING P J，CORRADI N. Shrink it or Lose it：Balancing Loss of Function with Shrinking Genomes in the Microsporidia [J]. Virulence，2011，2(1)：67－70.

[16] 董战旗，张　军，胡　楠，等. 家蚕核型多角体病毒 IE1 的多克隆抗体制备及鉴定 [J]. 西南大学学报(自然科学版)，2014，36(10)：76－81.

[17] BELKORCHIA A，GASC C，POLONAIS V，et al. The Prediction and Validation of Small CDSs Expand the Gene Repertoire of the Smallest Known Eukaryotic Genomes [J]. PLoS ONE，2015，10(9)：e0139075.

[18] STORMO G D，SCHNEIDER T D，GOLD L，et al. Use of the 'Perceptron' Algorithm to Distinguish Translational Ini-

tiation Sites in E. coli [J]. Nucleic Acids Res, 1982, 10(9): 2997－3011.
[19] VAPNIK V N. Statistical Learning Theory [M]. New York: Wiley-Interscience, 1998.
[20] 晏 勇. 基于SKLLE和SVM的人脸表情识别 [J]. 西南师范大学学报(自然科学版), 2014, 39(1): 55－60.
[21] 李 琴, 张 瑾, 骈 聪, 等. 基于位置关联权重矩阵及序列组分的多样性增量识别剪接位点 [J]. 生物物理学报, 2014, 30(5): 391－400.
[22] 廖 堃, 段江波, 周艳红. 人类基因PolyA位点预测 [J]. 计算机学报, 2008, 31(6): 927－933.

Prediction of Polyadenylation Sites in Microsporidian Genome

SUN Kang[1], YANG Ming[1], MA Li[1],
LI Tian[2], ZHAO Yu-fang[3]

1. *School of Computer and Information Science, Southwest University, Chongqing 400715, China;*
2. *School of Biotechnology, State Key Laboratory of Silkworm Genome Biology, Southwest University, Chongqing 400716, China;*
3. *Teacher Education Development Center, Southwestern University, Chongqing 400715, China*

Abstract: Polyadenylation is a critical cellular process that forms mature mRNAs in eukaryotic cells. The prediction of its sites is of an important reference value for the discovery of encoding genes in the genome sequence. At present, no effective gene prediction methods for microsporidian genomes are available. Here, we studied microsporidia genomes and, according to the preference of gene expression of the species, proposed a method to predict and analyze poly(A) sites of microsporidium. First, we employed the K-gram nucleotide acid pattern, position weight matrix and increment of diversity to form the initial features. Then we used PCA to reduce the dimension of the initial feature space. Finally, a classification model integrating SVM classifiers was built to predict poly(A) sites. By the proposed algorithm, we achieved a specificity (S_p) of 87.33% and an accuracy (ACC) of 85.14% in the specific dataset. This method also gave an ideal result in the prediction of the poly(A) sites in the microsporidium genome.

Key words: poly(A) signal (polyadenylation signal); Nosema bombycis; positional weight matrix; machine learning

责任编辑 崔玉洁