**电　子　科　技　大　学**

**2016级本科毕业设计（论文）中期报告**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **学院：生命科学与技术学院** | | | | | | **填表日期：2020/4/8 21:51:52** | | |
| **姓名** | **高鑫辰** | | **学号** | **2016090202016** | | | **指导教师** | **郭锋彪** |
| **题目名称** | | **基于间隔序列寡核苷酸的必需基因预测算法研究** | | | | | | |
| **1. 课题研究概述（不少于300字）**  本课题的主要任务是理解并采用计算机编写出间隔序列寡核苷酸组成的代码，基于特征描述代码构建必需基因预测模型，获得85%（严格交叉验证）以上的AUC或准确率。必需基因是指在生物体整个基因组中负责基本生命活动（例如生存或繁殖）的对一个物种的生存和发展起决定性作用的基因，必需基因所编码的蛋白质的功能被认为是生命活动的基础，简单地说，去除一个必需基因会使生物体无法继续存活。本课题拟利用序列组成，基于间隔序列寡核苷酸进行必需基因的识别，序列组成是指可以表示为一组不同的寡核苷酸组成形式的DNA序列（例如K-mer），这种特征提取方法可以将短或长碱基序列转化为目标序列中K-mer的出现频率。SVM（支持向量机）是目前预测必需基因最流行的机器学习方法，是一类按监督学习方式对数据进行二元分类的广义线性分类器。课题计划以大肠杆菌的基因序列为对象，以上文介绍的频率作为特征，用基于SVM的分类器构建必需基因预测模型，对数据集进行多重的交叉验证，获得85%（严格交叉验证）以上的AUC。  参考文献：  [1] Mobegi FM, Zomer A, de Jonge MI, van Hijum SA. Advancesand perspectives in computational prediction of microbialgene essentiality. Brief Funct Genomics 2017;16:70–9.  [2]Chuan Dong, Yan-Ting Jin, Hong-Li Hua, Qing-Feng Wen, Sen Luo, Wen-Xin Zheng, Feng-Biao Guo, Comprehensive review of the identification of essential genes using computational methods: focusing on feature implementation and assessment,Briefings in Bioinformatics, Volume 21, Issue 1, January 2020, Pages 171–181. | | | | | | | | |
| **2. 目前已完成学位论文工作的内容及进展（不少于800字）**  目前学位论文工作的进展基本按照进度计划表进行，目前所完成的工作如下：  1、论文的阅读和翻译：阅读了相关文献，对近年来许多科研工作者对必需基因预测的研究工作有了了解，从论文《Accurate prediction of human essential genes using only nucleotidecomposition and association information》获得启发，将基因的序列特征与支持向量机（SVM）算法结合起来，用机器学习打分的机制自动地去学习每个特征对区分必需基因和非必需基因的重要性度量。  2、数据的收集和整理：NCBI是美国国家生物技术信息中心，其保管了GeneBank的基因测序数据，本课题采用大肠杆菌作为研究对象，从NCBI上下载了大肠杆菌的基因信息文件，将其整理成两个文件，一个存放大肠杆菌基因序列的fasta格式文件，一个存放标签信息。数据集总共包含4144条基因序列，其中287条必需基因，3857条非必需基因。  3、特征提取：本课题假设间隔一段距离(λ个字符)的核苷酸间存在相关性，这种相关不是Markov链的实际关联。对于单核苷酸，我们提取的特征为3个密码子（相位）位上每一位的4种核苷酸字符的频率，一共12个变量。对于邻接双核苷酸，我们提取三个相位每一相位的16种相邻双核苷酸的频率作为描述变量。同时，我们准备多组变量用以描述具有间隔距离的两个核苷酸间的虚拟关联。间隔距离用λ表示，λ的最大取值则用L表示。由此我们一共可以得到16×3+16×3×L个变量来描述双核苷酸（邻接的和具有间隔距离虚拟的双核苷酸）的关联组成频率。对于三核苷酸、四核苷酸、五核苷酸…进行类似处理。类似于马尔科夫模型，对于k核苷酸频率，间隔前面的核苷酸片段取k-1个字符，间隔后面的核苷酸只取1个字符，然后计算这个长度为k的虚拟核苷酸片段的频率。目前特征提取进行到6核苷酸，间隔为6。  4、特征筛选：过多的特征中包含了太多的无用信息，需要对上一步获得的大量特征进行筛选，采取的是t检验的方法，具体做法是把数据集的必需基因和非必需基因分开，然后根据每一特征在不同类别的基因中是否存在显著的差异，选出差异性较大的特征。  5、模型构建：采取线性核的svm作为预测模型，对数据集进行五折交叉验证计算得到auc，目前的平均auc大约在82%，在后续的工作中，尝试增加特征的数目以及改变特征筛选的方法以期提高准确率。 | | | | | | | | |
| **3. 存在的问题及解决办法（若无可以不填或填“无”）**  目前的准确率尚未达到预期，在后期工作中将和指导老师商讨对相关工作进行完善。 | | | | | | | | |
| **指导教师**  **审查意见** | | 目前课题进展良好。 | | | | | | |
|  | | | 签名：郭锋彪  日期：2020/4/21 20:17:37 | | | |

**2016级本科毕业设计（论文）中期检查表**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **学院名称：生命科学与技术学院** | | | | | | **填表日期：2020年4月21日** |
| **以下内容由学生本人填写** | | | | | | |
| **学生姓名** | | **高鑫辰** | | **题目名称** | **基于间隔序列寡核苷酸的必需基因预测算法研究** | |
| **学号** | | **2016090202016** | | **题目性质** | **科研** | |
| **指导教师** | | **郭锋彪** | | **工作地点** | **⦿校内 🌕校外 详细地点校内：电子科技大学清水河校区创新中心D422** | |
| **设计时间** | | **2019年11月 1日至 2020年6月 1日** | | | | |
| **以下内容由检查教师填写** | | | | | | |
| **课题核心** | **课题采用支持向量机（SVM）这一机器学习方法，基于大肠杆菌基因序列，构建必须基因预测模型。** | | | | | |
| **课题进展情况** | **课题进展良好。** | | | | | |
| **存在困难** | **暂无困难。** | | | | | |
| **解决办法**  **或**  **建议** | **建议多和指导老师沟通，确保课题的开展。**    **检查教师签名：** | | | | | |
| **当期完成情况成绩** | | | **⦿优秀 🌕良好 🌕中等 🌕合格 🌕不合格** | | | |

**说明：**1、本表内容应如实填写；

2、本表应妥善保管，以便装订在毕业论文中；

3、学院教务科对检查情况分类汇总后，报送教务处实践教学科备案。