**电 子 科 技 大 学**

**2016级本科毕业设计（论文）开题报告表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **学号：2016090202016** | | | **姓名：高鑫辰** | **学院：生命科学与技术学院** | **专业：生物技术** |
| **论文题目** | | **基于间隔序列寡核苷酸的必需基因预测算法研究** | | | |
| **题目来源：科研** | | | | | |
| **成果形式：软件** | | | | | |
| **学位论文　　研究内容** | **（1500-2000字，参考毕业论文正文格式：宋体小4号，英文字体为Times New Rome，行间距固定20磅，可另加页）：**  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，必需基因的总和构成了维持一个细胞生命活动的最小基因集合，简单地说，去除一个必需基因会使生物体无法继续存活，虽然在不同的物种之间具有不同的基因组大小和基因组成，但是这些差别很大的基因组中都包含各自一套的必需基因用以维持关键的细胞功能。通过对必需基因的研究，有助于理解生命的起源以及在进化过程中不同物种之间的联系，有利于全面理解生命的本质，人为地重建最小的基因集，并筛选出治疗病原体疾病的潜在药物靶标。所以，识别某一基因序列是否是必需基因意义重大。  近年来，在原核生物中，许多研究小组利用了实验方法，例如转座子诱变和RNA干扰来识别必需基因；也有一些鉴定人类必需基因的研究相继发表，这些基因是通过实验技术确定的，例如RNA干扰、CRISPR-Cas9或逆转录病毒基因陷阱。然而，由于实验技术的限制，确定必需基因的实验方法耗费巨大且无法对一个基因组中所有基因进行判断。此外，因为必需基因的进化保守性较强，可以基于序列比对利用同源性搜索进行必需基因预测，但是在漫长的物种进化过程中，直向同源的基因经常会在各自的生物体内部进行复制，经常会出现一个基因组中多个基因和另外一个基因组的同一个基因同源，这样的情况导致了利用同源性搜索预测必需基因时，经常预测出假的必需基因。为了更高效准确的识别必需基因，许多研究小组采用了更具成本效益和时间效率的计算生物学方法来预测必需基因，目前，计算方法已经在预测必需基因领域具有广泛的研究应用。再利用机器学习技术预测必需基因中，有很多不同类型的特征信息可以利用，例如进化保守性、域信息、网络拓扑结构、序列组成和表达水平等，而选择合适的特征对于预测必需基因是至关重要的。  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。随机过程理论中的马尔科夫链模型假设下一个字符的状态由前面邻接的连续几个字符决定。本课题假设间隔一段距离(λ个字符)的核苷酸间也存在相关性，这种相关不是Markov链的实际关联。而是在一级序列中虚拟的，只有三级结构中才真实存在的相关。这种关联的实验证据就是现在广为研究的三维基因组学。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。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。  本课题的特色在于由于DNA序列也可以理解为一个普通的字符序列，因此本项目发展的特征表述方法可以推广到其他具有周期性的字符序列，有望成为一种普适性的字符序列特征表述形式。 | | | | |
| **指导教师**  **审查意见** | 背景清楚，课题方法理解准确。  **签名：**  **日期：**2020年4月15日 | | | | |