计算机科学与探索 2012 6 (1): 67-77 ISSN: 1673-9418 CN: 11-5602/TP

学术研究

植物抗性基因识别中的随机森林分类方法

郭颖婕，刘晓燕，郭茂祖，邹 权

1. 哈尔滨工业大学 计算机科学与技术学院, 哈尔滨 150001
2. 厦门大学 信息科学与技术学院, 福建 厦门 361005

收稿日期 修回日期 网络版发布日期 接受日期

摘要 为了解决传统基于同源序列比对的抗性基因识别方法中假阳性高、无法发现新的抗性基因的问题，提出了一种利用随机森林分类器和K-Means聚类降采样方法的抗性基因识别算法。针对目前研究工作中挖掘盲目性大的问题，进行两点改进：引入了随机森林分类器和188维组合特征来进行抗性基因识别，这种基于样本统计学习的方法能够有效地捕捉抗性基因内在特性；对于训练过程中存在的严重类别不平衡现象，使用基于聚类的降采样方法得到了更具代表性的训练集，进一步降低了识别误差。实验结果表明，该算法可以有效地进行抗性基因的识别工作，能够对现有实验验证数据进行准确的分类，并在反例集上也获得了较高的精度。

关键词 随机森林 分类器 抗性基因 聚类 降采样

分类号

Identification of Plant Resistance Gene with Random Forest

GUO Yingjie, LIU Xiaoyan, GUO Maozu, ZOU Quan

1. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China 2. School of Information Science and Technology, Xiamen University, Xiamen, Fujian 361005, China

Abstract

The traditional homology sequence alignment based approaches usually have high false positive rate and consequently new resistance genes are difficult to be identified. This paper presents a resistance gene identification approach by applying random forest classifier and K-Means under-sampling method. In order to solve the aimless problem in gene-mining research, two main contributions are provided. Firstly, it introduces random forest and 188 dimension features to identify resistance genes, accordingly the sample statistic learning approach can efficiently capture the internal characteristic of resistance genes. Secondly, it selects a more representative training subset and reduces the identification errors for solving the serious imbalanced classification during the training process. The experimental results indicate that the approach can efficiently identify the resistance genes, not only precisely clas-sifying the existing experimental verified data, but also obtaining high accuracy on the negative sample dataset.

Key words random forest classifier resistance gene cluster under-sampling

DOI:

通讯作者