生物物理学报(2007) 23(5), 389-396

PDF文档

拟南芥基因组中新的microRNA预测及分析

金伟波[1,2]、孔栋[2]、应晓敏[1]、郭蔼光*[2]、李伍举*[1]
1 军事医学科学院基础医学研究所
2 西北农林科技大学

MicroRNA（miRNA）是一类存在于动植物体内、长度为21~25 nt的内源性小RNA，对生物体的转录后基因调控起着关键作用，但一些低丰度的miRNA 和组织特异性miRNA往往很难发现。为了系统识别拟南芥基因组中新的非同源miRNA，首先基于已报道的拟南芥miRNA的特征，从全基因组范围中筛选出453条可能的miRNA前体；其次，为了进一步对上述miRNA前体进行筛选，利用人的miRNA前体数据构建了支持向量机模型GenomicSVM，该模型对人测试集的敏感性和特异性分别为86.3%和 98.1%（30个人miRNA前体和1 000个阴性miRNA前体），对拟南芥测试集的正确率为93.6%（78个miRNA前体）；最后，利用GenomicSVM预测上述453条miRNA前体序列，得到了37条候选的新的拟南芥miRNA前体，为进一步的miRNA实验发现研究提供了指导。

Prediction and aanalysis of novel miRNA in Arabidopsis thaliana

MicroRNAs (miRNAs), ranging in size from 20~25 nt, are a growing family of noncoding RNAs. They play an important role in the regulation of gene expression. The low abundance of some miRNAs and their time- and tissue-specific expression patterns make them difficult to be identified. To identify the novel miRNA systematically in A. thaliana, the authors firstly found 453 pre-miRNA candidates from the genome using the characteristics of the known A. thaliana miRNAs and comparative genomics methods. Then, in order to reduce the number of putative pre-miRNA candidates, the authors developed a SVM (support vector machine) model, GenomicSVM, using the human miRNA dataset as the training dataset. The model had the sensitivity 86.3% and specificity 98.1% respectively on the human test dataset, which contained 30 positive human pre-miRNAs and 1000 negative pre-miRNAs. For the 78 positive pre-miRNAs in A. thaliana, the model could pick up 73 pre-miRNAs and therefore the correct rate was 93.6%. Finally, the GenomicSVM was used to discriminate whether each 453 pre-miRNA-like sequence was pre-miRNA or not. The results indicated that there were 37 novel miRNA candidates. Therefore, the study in this report provides bioinformatics help for the experimental identification of miRNAs in A. thaliana.

关键词

拟南芥(A.thaliana); 基因组(genome); microRNA; 预测(prediction)