生物物理学报(2006) 22(2), 110-116

PDF文档

基于k-tuple组合的酵母ncRNA与mRNA的比较研究

李华、应晓敏、查磊、李伍举*
军事医学科学院基础医学研究所

ncRNA和mRNA一样，都是重要的功能分子。以k-tuple（k字）含量为特征，对酵母ncRNA成熟序列和mRNA的编码区、上游序列与下游序列进行了分类与比较研究，结果显示：基于ncRNA成熟序列与mRNA编码区的3-tuple的含量，ncRNA和mRNA的交叉有效性分类精度（leave-one out cross-validation, LOOCV）平均值达到93.93%；基于上游序列4-tuple和5-tuple的含量，分类精度分别为92.49%和92.76%；基于下游序列4-tuple和5-tuple的含量，分类精度分别为91.58%和90.60%；利用上游序列和下游序列的4-tuple与5-tuple的含量，其平均分类精度分别为94.68%和94.83%；通过t检验，得到了在ncRNA和mRNA上、下游序列中具有显著统计学差异的k-tuple。上述结果表明，基于ncRNA成熟序列与mRNA编码区的3-tuple含量和基于ncRNA与mRNA上、下游序列的4或5-tuple含量可以有效地区分ncRNA与mRNA。此研究结果不仅有助于准确识别ncRNA与mRNA，还有助于发现ncRNA特异的转录因子结合位点。

The comparison of the yeast ncRNA and mRNA based on k-tuple combinations

Both ncRNAs and mRNAs are important functional molecules. In this report, the Na?鍛ve Baye's classification method has been used to classify ncRNA and mRNA based on their k-tuple content, composition of coding regions, upstream sequences and downstream sequences, and the vector concatenation of both upstream and downstream sequences. The results show that the average leave-one-out cross-validation (LOOCV) classification accuracy is 93.93% for 3-tuple content in ncRNA sequences and mRNA coding regions. In upstream 1 kb sequences of the ncRNAs and mRNAs for 4-tuple and 5-tuple content, the classification accuracies are 92.49% and 92.76%, respectively. For the downstream 1 kb sequences of the ncRNAs and mRNAs, the classification accuracy are 91.58% and 90.60% for 4-tuple and 5-tuple content, respectively. For the vector concatenation of upstream and downstream sequences, the average classification accuracies are 94.68% and 94.83% for 4-tuple and 5-tuple content respectively. Finally, the t-test has been used to verify if k-tuples are statistically different between the upstream and downstream sequences of the ncRNAs and mRNAs. The approach may be used to identify ncRNAs from other genomes and to identify binding motifs within ncRNA sequences.

关键词