生物物理学报(2004) 20(2), 125-131

PDF文档

拟南芥和线虫基因序列及剪切位点的理论预测

陈翠霞、李前忠*、林昊
内蒙古大学理工学院物理系

将拟南芥（A. thaliana）和线虫（C. elegans）基因组按外显子、内含子及基因间序列区分为3类。分别选取64、40、20 种三联体的概率作为信号参数构建离散源，根据离散增量预测序列所属类型。结果表明：拟南芥各条染色体标准集总预测成功率达到82.19%，检验集为87.95%；线虫各条染色体标准集总预测成功率达到79.67%，检验集达到81.93%。另外，将两种基因序列中的外显子分别划分成3类，用外显子剪切位点、翻译起始和结束位点附近的三联体的3个位点作为3条子链，以各条子链的12个参数构建离散源，用离散增量对3种序列类型进行预测，预测成功率都达80%以上。

A STUDY ON THE SEQUENCES AND SPLICE SITES OF A.thaliana AND C.elegans GENES

The complete sequences of A.thaliana and C.elegans genome are divided into three kinds: exons, introns and intergenicDNA. The 64, 40 and 20 trimers' probabilities of the three kinds of sequences are respectively selected as parameters of the sources of diversity. The classes of these sequences are predicted by the increments of diversity the minimum of the three increments. The results shown that the overall prediction accuracies of A.thaliana's every chromosome are 82.19% and 87.95% for the standard-sets and test-sets; the overall prediction accuracies of C.elegans' every chromosome are 79.67% and 81.93% for the standard-sets and test-sets, respectively. In addition, the exons in A.thaliana and C.elegans genome are divided into three types. Based on the frequencies of 4 kinds of bases in regions near intron/exon boundary, initiation and termination site for translation, the diversity source is composed of 12 sequence parameters. The three kinds of exons are predicted by using of an algorithm based on the increment of diversity. The rates of correct prediction higher than 80% are obtained.

关键词