遗传学报　2004 31 (5): 432-443　ISSN: 0379-4172　CN: 11-5450/R

通过新基因计算机识别与实验确认对NCBI人类基因数据库一些模式参考序列错误的分析与纠正

张德礼,1,2 季　梁,1 李衍达1

1.清华大学生物信息学研究所生物信息学教育部重点实验室;清华大学信息科学技术学院自动化系智能技术与系统国家重点实验室;北京100084;2.北京大学医学部;北京100083

收稿日期　修回日期　网络版发布日期　接受日期

摘要　采用生物信息学分析与实验确认相结合的技术路线，通过所识别的基因在非冗余数据库比对发现了网上公布的计算机注释人类基因组编码序列存在各种类型的多处错误，包括cDNA水平的一个或一段碱基插入、缺失或突变，或是这些错误的不同排列组合，其中以错误插入为多，往往导致编码氨基酸的移码突变。最先举证了NCBI GENOME Annotation Project 预测人类新基因的下列错误类型：（1）开放读码框架（ORF）中错误插入一个碱基造成编码氨基酸移码；（2）错误拼接；（3）开放读框中错误插入一个或一段碱基造成该读框提前终止。只编码N端氨基酸的cDNA序列而不完整；（4）只有编码C端氨基酸序列的cDNA而不完整；（5）只是正确基因ORF中间的一段编码蛋白cDNA序列而不完整，缺N端与C端氨基酸序列，并且将不完整蛋白氨基酸序列的第一个非起始码氨基酸错误地预测为起始码氨基酸，如将L错误地预测为M；（6）开放读框中错误插入一个或一段碱基造成前面出现不该有的终止码，因而编码蛋白缺开头部分氨基酸；（7）可能将污染基因组序列当作完整基因cDNA序列对待而预测出所谓单一外显子基因。即便真是基因，也只是较长单一外显子mRNA中有一小ORF，而ORF起始码上游同一相位确实存在终止码，无其它特点符合基因条件；（8）所预测基因只有ORF，而ORF两端没有任何EST证据,可据此ORF拼接出受EST和人类基因组双重支持的完整基因cDNA（开放读框上游同一相位有终止码），预示所预测ORF参考序列可能不正确；（9）有EST实验证据支持存在基因的人类基因组序列范围内又被预测出一条相似但更小的蛋白编码基因，因而新预测基因有可能是错误的。

关键词　人类基因组　表达序列标签　计算机克隆　基因纠正　模式参考序列　生物信息学

分类号

1. MOE Key Laboratory of Bioinformatics; State Key Laboratory of Intelligent Technology and Systems; Department of Automation; School of Information Science and Technology; Tsinghua University; Beijing 100084; China;2. Peking University Health Science Center; Beijing 100083; China

Abstract

Key words human genome　EST　in silico cloning　gene identification　REFSEQs　bioinformatics

DOI:

通讯作者