

# 四色荧光标记 DNA 自动测序中一例典型错读码的校正

## THE CORRECTION OF A TYPICALLY FALSE READ CODE IN DNA AUTO- SEQUENCE LABELLED BY 4 COLOR FLUORESCENT DYES

马洪明<sup>1</sup>      路新枝<sup>1</sup>      王 勇<sup>2</sup>

(<sup>1</sup> 青岛海洋大学食品工程系 266003)

(<sup>2</sup> 中国科学院海洋研究所 青岛 266071)

**关键词**      DNA, 自动测序, 荧光标记

 DNA 测序是现代分子生物学研究中的重要技术。随着人类基因组计划的实施, DNA 测序技术虽然在原理上没有新突破, 仍然依赖于 Sanger 的酶法, 但在不断提高自动化程度、降低成本方面取得了很大的进展, 测序效率目前已有很大提高。大规模测序一个碱基的成本从 1986 年的 0.5 美元降到目前的约 7 美分。

目前各种 DNA 自动测序技术都是在荧光标记技

术的基础上发展起来的。采用荧光标记的自动测序仪已在各个领域被广泛应用。四色荧光标记自动测序采用 4 种具有不同发射波长的荧光染料标记由聚合酶

---

第一作者: 马洪明, 出生于 1970 年, 硕士, 讲师。

E mail: mahongm@ouqd.edu.cn

收稿日期: 2000-10-23; 修回日期: 2001-03-16



链终止反应产生的一系列终止片段,经过聚丙烯酰胺或毛细管电泳分离后,采用激光对分离的 DNA 片段携带的荧光信号进行激发,检测并记录不同发射波长的信号,经计算机处理,最后得到序列信息。因此,对不同荧光激发信号的精确区分是影响 DNA 自动测序数据准确性的关键因素。

目前采用的荧光染料是荧光素 FAM, JOE 和罗丹明 TAMRA, ROX。它们的优点是激发光谱窄,信号强度高,但不同染料的发射波长有部分重叠,而且电泳行为有差异,使被标记的 DNA 片段迁移率发生不同的变化,导致荧光激发信号一定程度的混杂。仪器须对检测到的信号通过软件进行计算校正才能得到最终结果。目前 ABI 公司的分析软件 ABI Prism Sequencing Ver. 2.1.1, 保证对测出的前 450 bp 不低于 98% 的准确率。这样的准确率对于大规模测序是可以接受的,但在某些工作中,如杂合性分析、点突变检测、关键调控序列的克隆等,是远远不够的。产生分析错误的原因,一方面可能是测序反应或电泳分离不够理想,而另一方面则是由于分析软件对数据的处理能力的缺陷所导致。这些软件处理错误通过优化操作和改良试剂是无法消除的,必须通过人工分析测序图谱进行矫正。这些错误往往带有一定的规律性,总结这些规律进行错误校正对于获得更高准确率的数据是非常必要的。

作者在测序工作中发现了一例有规律的分析错误,且未见报道。应用这一规律指导对测序数据的处理,获得较好的效果。

## 1 材料和方法

仪器: ABI 310 Genetic Analyzer 软件: ABI Prism 310 Collection Ver 1.0, ABI Prism Sequencing Ver 2.1.1 (PE 公司), Gene Amp PCR System 9600 (PE 公司)。试剂: Dye Terminator cycle sequencing kit (Ampli Taq, FS) (PE 公司)。材料: 质粒 pGEM, 购自 PE 公司, 质粒 pBluescript 及青岛文昌鱼 Bb1m 基因由中科院海洋所发育室提供, 人乙型肝炎病毒 HBV 由青岛大学医学院遗传学教研室提供。方法: 测序反应参照试剂盒说明书进行。电泳分离数据收集及处理参照有关说明书进行。

## 2 结果

图 1 所示序列和图谱截取自 pGEM 质粒的自动测序结果。图中所示两行序列, 上行为仪器自动读出的序列结果, 下行为该质粒的标准序列。其中误读之处分别是在 53 位后多读了一个 G 和一个 A; 在 324 位后多读了一个 N。

由图谱可见, 两处误读均出现在连续多个 (大于等于 3) 鸟嘌呤 G 后紧接腺嘌呤 A 的情况 (以下记作

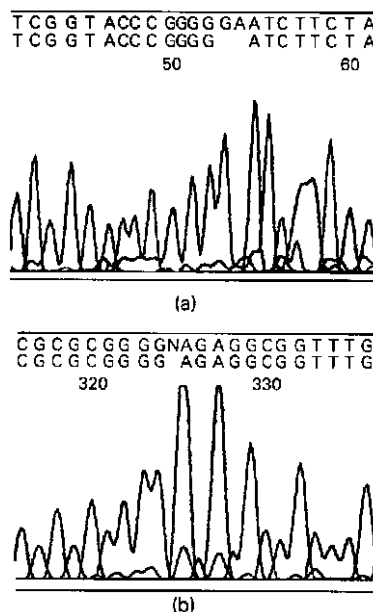


图 1 pGEM 质粒的自动测序结果  
上排序列为原始序列, 下排序列为标准序列。

GsA)。从图谱来看, 这种情况下峰谱表现为: (1) 连续的 G 峰峰宽压缩变小; (2) G 与 A 峰之间的间距增大; (3) 相邻接的 G 和 A 的峰高和峰宽均显著增大。

对多份不同样品所做的统计表明在满足连续大于等于 3 个 G 后接 A 的条件下, 软件发生误读的频率为 70%。将 GsA 引起的错误校正后, 准确率提高。见表 1。

利用这一规律指导未知序列的测定, 获得了较好效果。青岛文昌鱼 Bb1m 基因是一段富含 GC 的序列。图 2 是文昌鱼 Bb1m 基因经 BamHI, EcoRI 双酶切, 克隆入 pBluescript 质粒后, 用 M13 正向引物进行荧光标记测序的结果。图中 80~374 之间的顺序为目的片段的顺序, 余为载体顺序。在这段顺序中共有两处被判读为 N, 分别为 81 位和 85 位。其中, 85 位的误读符合 GsA 规律, 由此判定此处不是 G 而是无信号, 将 G 删除。另外 170 位出现 GsA, 对照图谱可以确认此处软件多读了一个 G, 予以删除。经人工修正后的序列与其他研究者发表的该基因序列完全一致。

## 3 讨论

采用 4 种不同的荧光染料对测序反应得到的随机终止片段进行标记的技术是目前最先进的自动测序技术。由于 4 种不同的染料在同一电场中的电泳行为有所差异, 导致经这些染料标记的随机终止单链 DNA 片段电泳迁移率发生改变, 经电泳分离后, 其前后位置并不完全对应于其分子大小, 而是有倒错和重叠。

表 1 不同样品中 GsA 出现和出错的情况

| 样品             | 序列前 450bp 中<br>GsA 出现次数 | GsA 判读错误数 | 矫正前的准确率<br>(%) | 校正 GsA 错误后<br>的准确率(%) |
|----------------|-------------------------|-----------|----------------|-----------------------|
| pGE M 质粒       | 4                       | 3         | 98.7           | 99.3                  |
| pbluescript 质粒 | 1                       | 1         | 94.17          | 96.67                 |
| 青岛文昌鱼 Bblim 基因 | 4                       | 2         | 98.98          | 100                   |
| 人乙型肝炎病毒( HBV)  | 1                       | 1         | 95             | 96.25                 |
| 总计             | 10                      | 7         | GsA 出错率: 70 %  |                       |

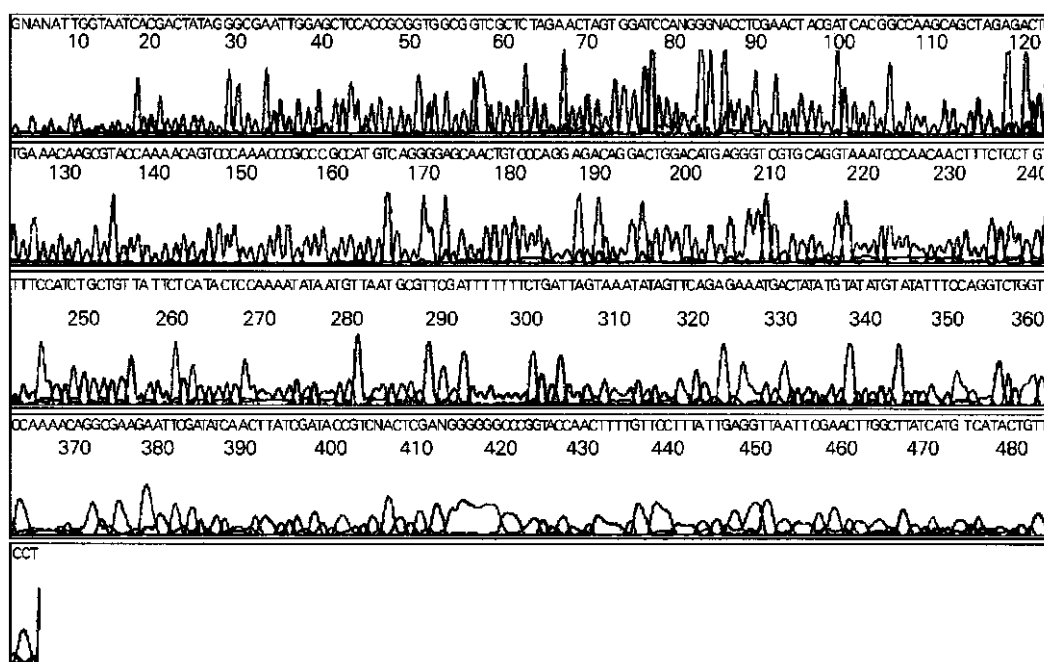


图 2 青岛文昌鱼 Bblim 基因测序结果

因此,原始的激发信号是混杂的。分析软件必须对原始信号进行计算修正才能得到最终的序列信息。软件对原始信号进行分析时首先根据片段的电泳分离情况计算出相邻信号之间的平均间距,以此间距值为标准,容许一定的波动范围,每过这样一个平均间距读取一个信号。由此推断若两信号之间的间距超过了平均间距的容许波动范围,则软件不能正确判读,往往读作 N 或其他碱基。

从图谱中不难发现,G 峰与 A 峰之间的间距远远大于图谱的平均间距,导致在此处不能正确判读,常读为 N 如图 1b);同时,G 和 A 峰高和峰宽的加大则导致在此处多读一个 G 或 A 如图 1a)。GsA 中有连续 3 个以上 G 时一般即可产生错读,当有小于 3 个 G 时虽然一般并未发生错读但从图谱中已可以看出明显的 G 和 A 之间的间距加大,相邻的 A 或 G 的信号强度增加等现象。

在常规的以同位素进行末端标记,聚丙烯酰胺凝胶电泳分离分析的 DNA 测序中,当有连续多个 G 信号时,电泳条带的间距也会出现压缩的现象。这反映了聚丙烯酰胺凝胶电泳与 ABI 310 自动测序仪所采用的毛细管电泳在对连续以 G 结尾的单链 DNA 片段分离上的一些共性。另外值得注意的是上述现象只在末端荧光标记测序法中出现,在对测序引物进行荧光标记的测序中则未发现上述现象,这提示间距压缩的发生可能与末端标记物有关。

标记染料的特性对电泳结果的影响至关重要。激发光强度均一,激发光波长范围窄而且互相之间没有重叠是从光学特性方面对标记染料提出的要求。另一方面标记染料必须尽可能小地影响 DNA 片段的电泳行为。据 Metzker 等 1996 年报道,目前国外正在开发的 BODIPY 荧光团(fluorophores),比前述传统染料除了在光学特性方面有所改善外,其电泳迁移率基本达到一

致。这将大大提高 DNA 测序的准确性并降低成本。

在末端荧光标记自动测序中,连续的 G (大于等于 3) 后跟 A 时, G 与 A 的间距增大造成误读的现象十分常见,有时甚至是造成数据准确率较低的主要因

素。而功能基因及其上游调控序列多是 GC 富含区。因此上述规律的利用对于在 DNA 序列分析中得到尽可能准确的基因序列信息显得尤为重要。

( 本文编辑 :刘珊珊)