新生物学丛书

第二代测序信息处理

Next-Generation DNA Sequencing Informatics

〔美〕Stuart M. Brown 等编著
吴佳妍 肖景发 于 军 主译

科学出版社

新生物学丛书

第二代测序信息处理

Next-Generation DNA Sequencing Informatics

〔美〕Stuart M. Brown 等　编著

吴佳妍　肖景发　于　军　主译

科学出版社

北　京

图字: 01-2013-7494

内 容 简 介

本书几乎涵盖了 NGS 技术在生命科学领域的全部应用，包括从头测序（含基因组注释）、针对稀有变异检测和元基因组研究的扩增子测序、染色质免疫共沉淀测序（ChIP-seq）、RNA 测序（RNA-seq）和肿瘤体细胞变异检测（包括单碱基替换、插入、缺失和易位）等。通过广泛使用的一线软件充分讨论数据分析方法，详述最优工作流程（包括部分学习指南），实用性强、可靠性强、专业指导性强。

本书非常适用于从事生命科学研究的研究生和青年学者。他们不仅可以在这里了解到不同软件的详细使用方法和参数设置，还可以在作者提供的软件评估和优化流程的基础上找到自身研究项目所需的第二代测序信息处理的最佳解决方案。

图书在版编目(CIP)数据

第二代测序信息处理 / (美)布朗(Brown, S. M.)等编著; 吴佳妍等主译. —北京：科学出版社, 2014.6

(新生物学丛书)

书名原文：Next-Generation DNA Sequencing Informatics

ISBN 978-7-03-040673-6

Ⅰ. ①第… Ⅱ. ①布… ②吴… Ⅲ. ①人类基因–基因组–序列–测试–数据处理 Ⅳ. ①Q78

中国版本图书馆 CIP 数据核字(2014)第 107676 号

责任编辑：罗 静 白 雪 / 责任校对：鲁 素
责任印制：赵德静 / 封面设计：美光制版

科学出版社 出版
北京东黄城根北街 16 号
邮政编码: 100717
http://www.sciencep.com

骏杰印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2014 年 6 月第 一 版 开本：787×1092 1/16
2014 年 6 月第一次印刷 印张：12 3/4 插页：4
字数：303 000
定价：80.00 元

(如有印装质量问题, 我社负责调换)

《新生物学丛书》专家委员会成员名单

译 者 名 单

主　　译　吴佳妍　肖景发　于　军

翻译人员（按姓氏汉语拼音排序）

李茹姣　刘晶星　苏明明

孙世翔　张若思　张玉玉

《新生物学丛书》丛书序

当前，一场新的生物学革命正在展开。为此，美国国家科学院研究理事会于2009年发布了一份战略研究报告，提出一个“新生物学”（New Biology）时代即将来临。这个“新生物学”，一方面是生物学内部各种分支学科的重组与融合，另一方面是化学、物理、信息科学、材料科学等众多非生命学科与生物学的紧密交叉与整合。

在这样一个全球生命科学发展变革的时代，我国的生命科学研究也正在高速发展，并进入了一个充满机遇和挑战的黄金期。在这个时期，将会产生许多具有影响力、推动力的科研成果。因此，有必要通过系统性集成和出版相关主题的国内外优秀图书，为后人留下一笔宝贵的“新生物学”时代精神财富。

科学出版社联合国内一批有志于推进生命科学发展的专家与学者，联合打造了一个21世纪中国生命科学的传播平台——《新生物学丛书》。希望通过这套丛书的出版，记录生命科学的进步，传递对生物技术发展的梦想。

《新生物学丛书》下设三个子系列：科学风向标，着重收集科学发展战略和态势分析报告，为科学管理者和科研人员展示科学的最新动向；科学百家园，重点收录国内外专家与学者的科研专著，为专业工作者提供新思想和新方法；科学新视窗，主要发表高级科普著作，为不同领域的研究人员和科学爱好者普及生命科学的前沿知识。

如果说科学出版社是一个“支点”，这套丛书就像一根“杠杆”，那么读者就能够借助这根“杠杆”成为撬动“地球”的人。编委会相信，不同类型的读者都能够从这套丛书中得到新的知识信息，获得思考与启迪。

译 者 前 言

第二代 DNA 测序（next-generation DNA sequencing，NGS）技术的应用普及极大地刺激了生物学新假说的提出和已有假说的验证，并驱动了当代基础生物医学和临床转化科学的迅猛发展。各种专为 NGS 技术开发或改进的新颖精致的生物信息学工具层出不穷，一方面使得 NGS 技术被广泛应用，每种主流 NGS 技术都对应很多不同的软件包；另一方面各种软件多以命令行用户界面和简约文档形式发布，却少有标志性研究指导用户选择最佳解决方案。作为国内最早一批接触使用 NGS 技术的生物信息学研究者，我们在科研实践中走过弯路但也积累了经验，深知现在迫切需要一部科学严谨、前沿、实用的专著，在生物信息学科的所有主流方向上指导研究人员，使其能够成功操作并充分应用 NGS 技术。《第二代测序信息处理》一书几乎涵盖了 NGS 技术在生命科学领域的全部应用，包括从头测序（含基因组注释）、针对稀有变异检测和元基因组研究的扩增子测序、染色质免疫共沉淀测序（ChIP-seq）、RNA 测序（RNA-seq）和肿瘤体细胞变异检测（包括单碱基替换、插入、缺失和易位）等，并通过广泛使用的软件讨论数据分析方法，详述最优工作流程（包括部分学习指南）。

本书非常适用于从事生命科学研究的研究生和青年学者。他们不仅可以在这里了解到不同软件的详细使用方法和参数设置，还可以掌握软件背后的相关算法和原理，在作者提供的软件评估和优化流程的基础上找到自身研究项目所需的第二代测序信息处理的最佳解决方案。

衷心感谢科技部 973 计划项目“重要热带作物木薯品种改良的基础研究”的“木薯基因组注释和信息整合”（课题号：2010CB126604）课题对于本书翻译工作的支持。

在译校过程中，虽力求忠于原文、通顺信达，但限于水平，谬误之处在所难免，敬希读者批评指正。

吴佳妍　肖景发　于　军

2014 年 5 月 13 日

前　言

第二代 DNA 测序（next-generation DNA sequencing, NGS）技术极大地刺激了生物学新假说的提出和验证，也提供了创新且广阔的视角去重新审视已有假说。毫不夸张地说，NGS 技术驱动了当代基础生物医学和临床转化科学的迅猛发展。

专为 NGS 技术开发或改进的各种新颖精致的生物信息学工具使得 NGS 技术得以广泛应用。为各种类型的数据处理和创新应用而开发的新软件陆续诞生，为解决序列比对和从头组装等原有问题的新算法也层出不穷，所有这一切都是为了应对新测序仪所产生的海量数据。

软件开发过程的持续加速主要是由于供应商不断升级测序仪器，不同研究组争相发表新方法以满足科研人员的需要。如此紧锣密鼓的开发流程导致 NGS 数据分析软件多数以命令行用户界面和简约文档形式发布。更为复杂的是，每种主流 NGS 技术都对应很多不同的软件包，却少有标志性研究指导用户选择最佳解决方案。所以，现在迫切需要一部科学严谨、前沿、实用的专著，在生物信息学科的所有主流方向上指导研究人员，使其能够成功操作并充分应用 NGS 技术。

作为纽约大学 Langone 医学中心的成员，我们感到十分荣幸。该中心很早就开始在 NGS 实验技术和信息处理及人力资源建设上进行了大量投入。特别是 2008 年，Langone 医学中心成立了基因组技术中心，让基础科研人员和临床转化科学家在微阵列芯片和实时定量聚合酶链反应（qPCR）等早期技术的基础上接触最新的 DNA 测序技术。与此同时，Langone 医学中心的信息学中心组建了测序信息学小组，为 Langone 医学中心和其他所有测序用户提供研究方案及上游数据处理、数据管理和数据分析服务。

我们研究小组在实践中持续积累经验，评估过很多不同的软件包，为很多不同类型的 NGS 项目建立了最优工作流程，涵盖了从头测序（包括基因组注释）、针对稀有变异检测和元基因组研究的扩增子测序、染色质免疫共沉淀测序（ChIP-seq）、RNA 测序（RNA-seq）和肿瘤体细胞变异检测（包括单碱基替换、插入、缺失和易位）。

在本书中，我们以 30 多个美国国立卫生研究院（National Institutes of Health, NIH）资助项目的大量经验为基础，综合本领域著作并去粗取精，为读者提供了多种 NGS 研究的全景展示，通过广泛使用的软件讨论数据分析方法，详述最优工作流程（包括部分学习指南）。我们也提出一些建议，希望能帮助生物信息学家更好地实施他们自己的数据分析方法，也希望能帮助实验室和临床研究人员利用 NGS 方法来落实他们自己的研究课题。

NGS 技术和生物信息学的蓬勃发展非常鼓舞人心，我们为本书能对这个领域的发展做出贡献而感到欣慰。

Stuart M. Brown

致　　谢

全体作者向纽约大学 Langone 医学中心的高层领导、院长和首席执行官 Robert Grossman 博士及整个行政和科研领导团队表达诚挚的谢意。感谢他们提供了舒适的环境与和谐的氛围，鼓励我们坚持不懈地进行 NGS 信息学的科学研究。

我们由衷感谢纽约大学 Langone 医学中心所有同仁，他们将 NGS 项目的成功经验与我们分享，支持我们通过广泛的基础科学和临床转化研究去研发、测试、总结大量高质量而具创新性的信息学问题解决方案。

我们深深感谢冷泉港实验室出版社（Cold Spring Harbor Laboratory Press, CSHLP）的 John Inglis 发现本书的价值。感谢 CSHLP 员工极大的耐心，感谢你们对我们不规律日程的迁就，以及在出版过程中始终如一的卓越品质。我们特别感谢 Inez Sialiano 在本书写作各阶段给予的编辑指导，感谢 Kathleen Bubbeo 对高质量图表的坚持及对全书的勘误。

纽约大学测序信息学小组成员：

Alexander Alekseyenko
Constantin Aliferis
Silvia Argimón （附属机构）
Stuart M. Brown
Efstratios Efstathiadis
Frank Hsu（附属机构）
Kranti Konganti
Eric R. Peskin
Christina Schweikert（附属机构）
Steven Shen
Phillip Ross Smith
Alexander Statnikov
Zuojian Tang
Jinhua Wang

关 于 作 者

Alexander Alekseyenko 是纽约大学医学院医学系助理教授、卫生信息学和生物信息学中心生物信息学咨询小组营运副总监。Alekseyenko 博士在洛杉矶的加利福尼亚大学获得生物数学博士学位。他先后在英国剑桥的欧洲生物信息研究所和斯坦福大学完成博士后培训。Alekseyenko 博士是纽约大学第一批信息学教职人员，研究领域是元基因组学，主要通过第二代测序技术，利用进化和生态统计模型研究人体内微生物多样性。

Silvia Argimón 是纽约大学牙医学院龋病学和口腔综合治疗系副研究科学家。她的研究内容包括口腔细菌多样性和毒性。Argimón 博士在苏格兰阿伯丁大学获得分子生物学博士学位。

Stuart M. Brown 是纽约大学医学院细胞生物学系助理教授、卫生信息学和生物信息学中心高级教职人员，同时是生物信息学咨询小组营运总监，也是序列信息学小组组长。他在纽约大学教授了 12 年生物信息学研究生课程，是生物信息学和医学基因组学教材的作者。Brown 博士在康奈尔大学获得分子生物学博士学位。

Efstratios Efstathiadis 是纽约大学 Langone 医学中心助理教授和高性能计算设施技术总监。他曾作为技术架构师供职于布鲁克海文国家实验室的计算机科学中心。Efstathiadis 博士于 1996 年在纽约城市大学获得核物理博士学位。

Jeremy Goecks 是埃默里大学生物学和数学计算机科学系的博士后。他是被广泛使用的基于网络的计算生物医学研究平台 Galaxy 开发团队的核心成员。Goecks 博士在佐治亚理工学院获得计算机科学博士学位。

D. Frank Hsu 是佛罕大学克拉维斯特聘科学教授和计算与信息科学教授。他是佛罕大学计算机科学系前任主任，《互联网络杂志》前任主编。Hsu 博士在密歇根大学获得博士学位。

Kranti Konganti 是纽约大学医学院卫生信息学和生物信息学中心的程序员/生物信息学研究人员，主要负责 Roche 454 测序数据的分析和 GBrowse 系统的基因组数据可视化。他在美国东北大学获得生物信息学硕士学位。

Eric R. Peskin 是纽约大学医学院卫生信息学和生物信息学中心高性能计算设施的技术副总监。Peskin 博士在犹他大学获得计算机科学博士学位。他曾作为逻辑技术开发的高级软件工程师供职于英特尔，也曾在罗彻斯特理工学院担任电气工程助理教授。

Christina Schweikert 是佛罕大学计算机与信息科学系助理教授。Schweikert 博士在纽约城市大学获得计算机科学博士学位。

Steven Shen 是纽约大学医学院卫生信息学和生物信息学中心与生物化学系副教授。他的工作重点是开发用来探索蚂蚁类物种基因组表观遗传变化的第二代测序相关技术和计算方法。Shen 博士曾经是波士顿大学医学院助理教授、麻省理工学院研究科学家。他还曾供职于 Helicos 生物科学公司，参与开发单分子测序技术。

Phillip Ross Smith 是纽约大学医学院细胞生物学系副教授、卫生信息学和生物信息学中心高级教职人员。他是纽约大学医学院前任首席信息官、《结构生物学杂志》前任编辑。Smith 博士在英国剑桥大学获得高能物理博士学位，在纽约大学医学院获得医学博士学位。

Zuojian Tang 是纽约大学医学院卫生信息学和生物信息学中心的副研究科学家，负责 Illumina 第二代测序的计算支持。她在麦吉尔大学获得计算机科学和生物信息学硕士学位。

James Taylor 是埃默里大学生物学和数学计算机科学系的助理教授，是被广泛使用的基于网络的计算生物医学研究平台 Galaxy 最初的开发者之一。Taylor 博士在宾夕法尼亚州立大学获得计算机科学博士学位，并在那里参与了几个脊椎动物基因组项目和 ENCODE 项目。

Jinhua Wang 是纽约大学医学院助理教授、纽约大学肿瘤研究所成员。Wang 博士在中国科学院获得计算生物学和基因组学博士学位。他曾作为生物信息学研究经理供职于中国国家人类基因组南方研究中心。他曾在冷泉港实验室进行博士后研究，主要开发识别真核生物基因组功能元件的数学和统计方法，特别是针对调控基因转录和 mRNA 前体剪切的序列元件。他还曾作为生物信息学科学家供职于圣犹达儿童研究医院。

目　录

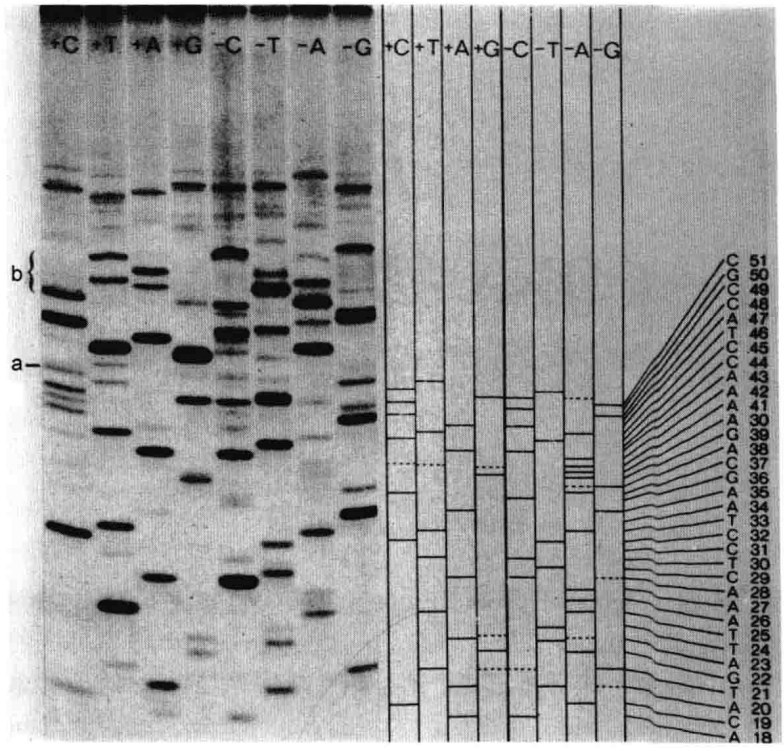

图 1.1 “DNA 聚合酶通过引物合成”的 DNA 测序法所产生的放射自显影图像
来自 1975 年 Sanger 和 Coulson 在 *Journal of Molecular Biology*（94:441-448）发表的文章

从各个角度来看，**Sanger 测序法**都是具有革命性意义的。其中最重要的是，它能对任何 DNA 分子进行测序，它可以用来检测长 DNA 序列。然而，这个系统在首次提出时有两个严重局限，导致它并未立刻被广泛采用。第一，对于寡聚核苷酸引物的需求意味着必须有一段与待测 DNA 序列区域直接相邻的 DNA 序列是已知的；第二，引物的“随机延伸”未必能生成平均分布的所有理论长度片段。

在 Sanger 的“引物延伸”测序法发表短短两年后，Allan Maxam 和 Walter Gilbert 发明了一种基于 DNA 化学裂解的测序方法（Maxam and Gilbert 1977）。Maxam-Gilbert 测序与 Sanger 法类似，将 DNA 模板分成 4 个反应。在每个反应中，先在模板的 5′端进行放射性标记，再加入能特异性在其中一种碱基处切开 DNA 的化学试剂。反应进行时，平均一个 DNA 分子只在随机位点产生一次裂解。接着，和 Sanger 法一样将 4 个反应的产物加在丙烯酰胺凝胶的相邻泳道，通过电泳根据片段大小进行分离。最后，通过对丙烯酰胺凝胶的放射自显影图像读取 DNA 序列。起初，Maxam 和 Gilbert 测序相对于 Sanger 测序更受欢迎，因为它能直接通过纯化后的 DNA 片段进行测序，而不需要单链模板和互补寡聚核苷酸引物。

Sanger 随即改进了他在 1975 年提出的方法，他在引物延伸反应中使用双脱氧核苷酸作为“链终止子”代替复杂的双阶段反应（Sanger et al. 1977）。改进后的测序法仍以单链 DNA 模板和短互补寡聚核苷酸引物的杂交为起始。将杂交后的模板分成 4 组反应混合物，每组包括 DNA 聚合酶、4 种三磷酸脱氧核苷酸（其中一种用放射性同位素标记）和一种

双脱氧核苷酸。当引物在 DNA 聚合酶作用下延伸时，一旦连接上双脱氧核苷酸反应就会停止，随即生成不同长度的、以同样引物为起始、以同一碱基终止的短片段混合物。最后，将 4 个反应生成物加在丙烯酰胺凝胶上，通过电泳分离大小不同的片段，通过放射自显影图像读取 DNA 序列。Sanger 表示在一次电泳中能读取最多 300 个碱基的长度。

很多年来，Sanger 的双脱氧核苷酸末端终止测序法和 Maxam-Gilbert 的化学降解测序法一直被互相比较。也许是由于 Maxam-Gilbert 法步骤烦琐并使用了有毒试剂，Sanger 法越来越受欢迎。很多细化完善 Sanger 法的研究方法被开发出来，包括多种跨待测基因（或整个基因组）的单链模板克隆方法和进行试剂准备流水作业的商业试剂盒。一项对 Sanger 技术非常重要的改进是荧光染料取代了新合成 DNA 片段上的放射性同位素标记（Smith et al. 1986）。由此促使 Leroy Hood、Michael Hunkapiler 等开发出半自动 DNA 测序仪，并且由美国应用生物系统公司（Applied Biosystems Inc.，ABI）投入商业化生产。ABI 测序仪的重要创新包括将 4 种不同颜色的荧光标记连接在 4 种双脱氧核苷酸链终止子上，使 4 个碱基终止的片段都在同一个反应管中生成，并且在同一块丙烯酰胺凝胶上进行电泳，通过电脑监控进行实时荧光检测，这样在凝胶电泳进行时序列数据就能被自动收集。人类基因组计划数据基本上都是由这些 ABI 测序仪获得的。ABI 自动荧光测序仪的另一处改进是用毛细管取代两个玻璃盘之间又大又薄的平板来盛放丙烯酰胺凝胶。这为测序实验室的技术人员节省了准备工作，保证了电泳结果的持续稳定并且提高了电泳速度，也使测序仪能够同时处理更多的样本（图 1.2）。

测 序 克 隆

Sanger 测序反应需要单链 DNA 模板、与模板互补的短单链寡聚核苷酸引物、DNA 聚合酶及链延伸和链终止的核苷酸混合物。准备测序 DNA 的常规策略是将 DNA 的目标片段**克隆**到质粒载体上，质粒载体在可以被单链 DNA 聚合酶 II 使用的标准测序引物结合位点之间提供克隆位点。由此任何 DNA 目标片段都可以通过标准寡聚核苷酸引物从两个方向被测序，因此不需要提前知道目标 DNA 分子的序列（图 1.3）。

用 Sanger 法进行 DNA 测序一次能读取 500~800 个碱基，这个界限受两个因素影响，一是 Sanger 引物延伸/链终止反应，二是通过单碱基分辨率用电泳准确区分 DNA 片段的能力。由于多数研究人员感兴趣的生物学核酸分子（如基因、mRNA 转录物、质粒和基因组）都远比 800 bp 长，DNA 测序项目一般先将 DNA 分子打成短片段，对短片段进行测序，再用生物信息学工具将数据组装成目标分子的完整序列。

对于较小的测序目标，基于限制性消化片段的策略颇为有效，但是却难以追踪所有序列组成片段的大小和方向。Henikoff 在 1984 年提出的策略包括通过核酸外切酶III的有向消化生成逐级变小的 DNA 片段，再将这些巢式序列通过重叠 read 组装被构建成为**重叠群**（**contig**）的邻接序列。由于所有 DNA 测序法都会产生少量错误，逐渐形成的标准流程是将全部目标区域中重叠的 **read** 结合起来，在理想状态下 read 来自 DNA 分子的两个方向。将全部来自两个方向的重叠 read 组装成为共有序列的方法成为 20 世纪八九十年代软件开发的焦点。一篇发表于 1994 年的综述（Miller and Powell 1994）比较了 11 个不同 DNA **序列组装**软件的性能。

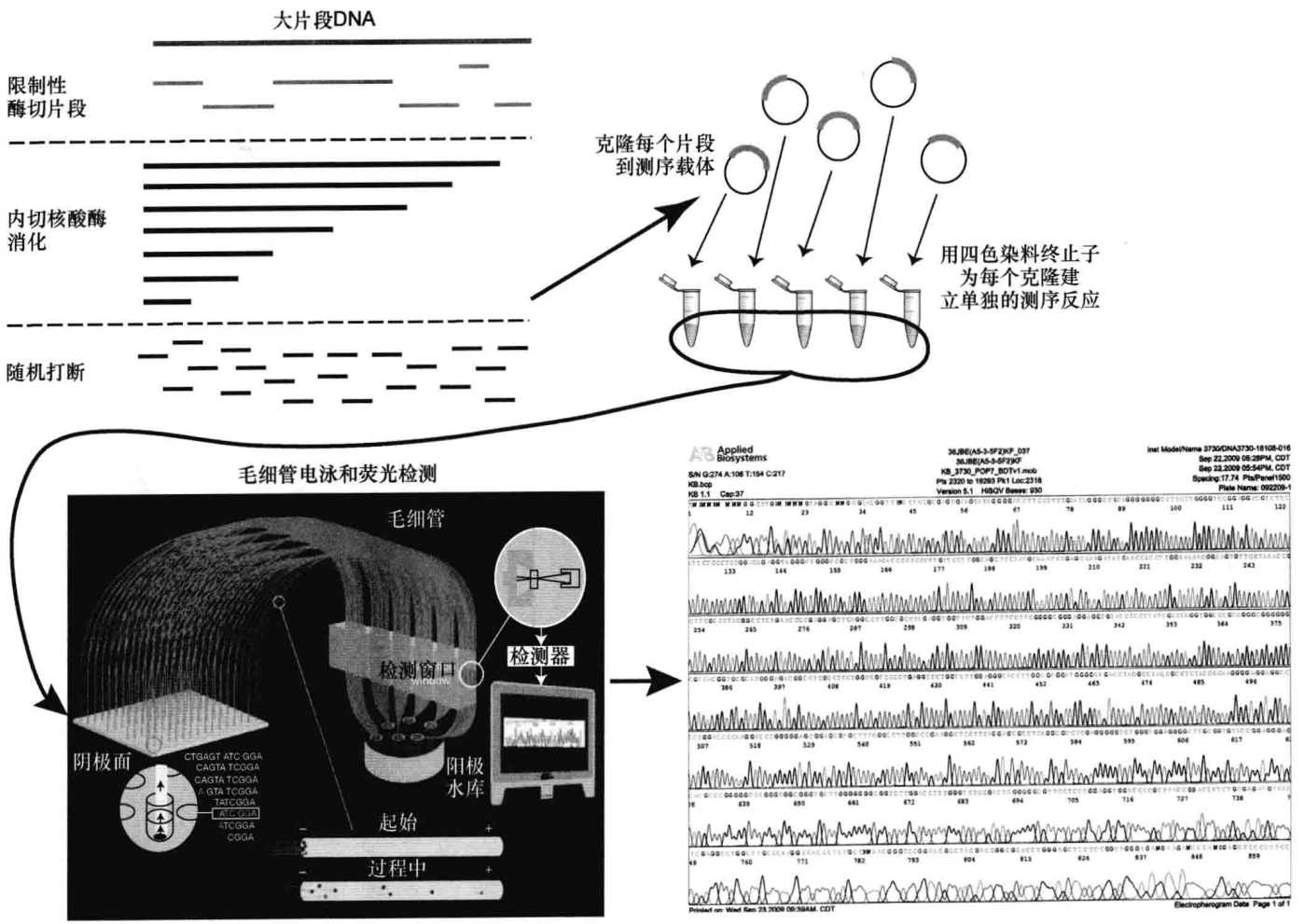

图1.2　毛细管DNA测序流程包括DNA片段化、子片段克隆、为每个片段建立单独的测序反应、把每个反应装载到单个毛细管纤维并对每个染料标记的终止碱基进行电泳和荧光检测

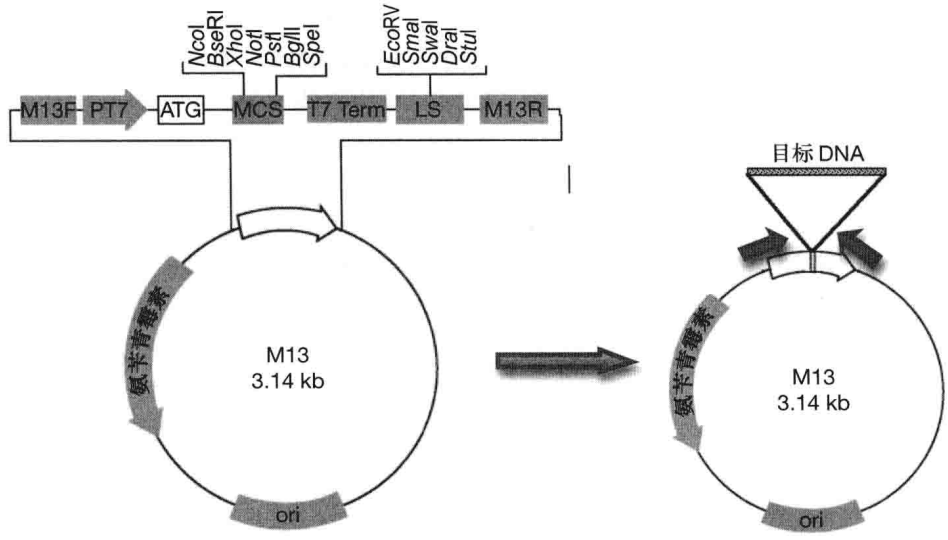

图 1.3　把 DNA 片段克隆到 M13 测序载体的多克隆位点上

测序项目的**片段组装**在算法上和**序列比对**相似，但也有其独特之处。首先，由于每一个 read 都由克隆到质粒（或病毒）载体上的 DNA 片段产生，读到的前几个碱基经常包含测序载体的序列。当克隆片段的长度比 read 短时，read 结尾也常包含载体。而当整个测序区域只包括载体 DNA 时，有可能会造成克隆干扰。所以，在将 read 组装成 contig 之前，从原始序列数据中识别和移除所有载体序列是很有必要的。

其次，通过 Sanger 法获得的 DNA 序列质量不稳定。由于电泳造成的不均匀分离、游离引物所引起的噪声，以及引物二聚体干扰，read 的前 50 个碱基质量较低。read 的末端（超过 500 个碱基）质量也较低，这是因为长片段 read 数量减少会导致信号减弱和扩散，以及电泳时微小的迁移率差异所导致的片段分离不明显。在理想状态下，序列组装软件应当具有识别低质量区域的能力，并且提供将低质量区域从高质量序列中分离出来的工具。

随着序列组装软件的发展，为大型测序项目开发的新策略应运而生。研究人员认识到 DNA 可以被随机打断成一系列无序片段，而无需小心翼翼地克隆限制性片段或使用核酸酶消化成巢式缺失片段。然后将这些无序片段克隆并测序，再通过软件组装，寻找其中重叠的部分（见第 5 章）。这整个过程被称为**鸟枪法测序**（Anderson et al. 1982）。用鸟枪法测得的 **DNA 片段**在目标 DNA 分子上呈**泊松分布**，所以需要对足够多的 DNA 片段进行测序，才能使原分子的每个碱基都有足够的**覆盖度**，进而拼出完整的 contig。例如，一个 10 000 个碱基的 DNA 目标片段（10 kb），需要对相当于全长 8~10 倍数量的片段进行测序。鸟枪法策略与需要大量人力进行限制性片段和巢式缺失克隆的测序方法相比，每一轮测序费用更加低廉，这使得鸟枪法策略越来越受欢迎。

对于非常大型的测序项目（全基因组），一种分而治之（divide-and-conquer）的策略经常被采用。把从 10 万~100 万个碱基不等的大片段 DNA 克隆到被称为细菌人工染色体（bacterial artificial chromosome，BAC）的载体上，再用鸟枪法对这些片段逐个测序。

即使具有较高的覆盖度（8×~10×），用鸟枪法组装的重叠的 **read** 也会在目标基因区域的序列中留下一些空位。可能是片段的泊松分布导致的随机低覆盖度区域，也可能是序列特异效应在克隆或测序过程中影响了目标区域的某些部位。空位可以用“引物步移策略”（primer walking strategy）来填补，该策略首先需要设计特异性测序引物，通过这些引物使测序反应从 contig 末端开始进行，并将序列延伸直至覆盖空位区域。随着每条 read 被添加到 contig 上，继续设计新的引物，直到和另一 contig 相遇。在相反方向进行延伸的引物能提供双链覆盖。引物步移策略在覆盖 DNA 片段的测序中所需反应数量很少，但是十分耗时，因为只有在得到上一个引物的测序数据时，才能设计和合成下一个引物（图 1.4）。

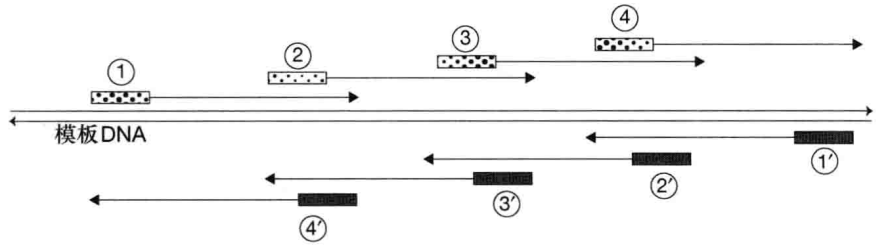

图 1.4　测序的引物步移策略

第二代测序

一些历史学家（Goldstein 1978；Kuhn 1996；Gladwell 2008）发现，当现有知识或新兴技术积累到足够充裕的程度时，会有很多不同的研究者同时钻研一个科学问题并同时产生新发现。其中比较成功的理论或方法会继续相互竞争，直到其中一个占据优胜地位，成为标准方法或主导范例。很明显 20 世纪 70 年代就是一个 DNA 测序的革命性时代。而另一场围绕**第二代 DNA 测序**（next-generation DNA sequcencing，NGS）的革命则正在进行。2004~2012 年，由于测序仪器的通量每年都会加倍，而平均到每个碱基的测序费用每年都会减半，新的 DNA 测序标准显然尚未确立。NGS 技术通常具有几个特征，即高数据通量、短序列读长及低于 Sanger 测序法的准确度。NGS 数据为生物信息学带来很多关键性挑战，包括将大量 read 定位到**参考基因组**上、从头组装新基因组、海量 read 的**多序列比对**、**扩增子**测序项目的稀有**变异检测**，以及高效存储的文件格式和运算工具、多千兆字节序列数据文件的操作等。

另一个有趣的方面是，NGS 影响了测序技术在科学界的地位。在 20 世纪 80 年代，多数 DNA 测序是在小型实验室完成的，全靠科研人员手工向大玻璃盘灌注聚丙烯酰胺凝胶，再耐心地将 X 射线胶片上的每个碱基逐个读出。在昂贵的自动化高通量 DNA 测序仪诞生后，大型测序计划都交由大型专业测序实验室、核心设备中心和专门进行 DNA 测序的承包商来完成。NGS 可能会通过降低仪器成本使 DNA 测序重新回归小型实验室，也可能利用昂贵的机器建立类似医药诊断实验室的永久性测序外包服务。即使 NGS 仪器的价格变得非常低廉，或者小型实验室可以签约测序公司以获得 NGS 数据，分析测序所得的大量复杂数据集仍然需要计算基础设施和生物信息技能，而这些条件是多数小