高通量测序与高性能计算理论和实践

陈禹保　黄劲松◎主编

北京科学技术出版社

高通量测序与高性能计算理论和实践

陈禹保　黄劲松　主编

北京科学技术出版社

图书在版编目（CIP）数据

高通量测序与高性能计算理论和实践/陈禹保，黄劲松主编. —北京：北京科学技术出版社，2017.5

ISBN 978-7-5304-8474-6

Ⅰ.①高… Ⅱ.①陈… ②黄… Ⅲ.①基因组-序列-测试-研究 Ⅳ.①Q343.1

中国版本图书馆 CIP 数据核字（2016）第 158306 号

高通量测序与高性能计算理论和实践

主　　编：陈禹保　黄劲松
责任编辑：李　鹏
责任印制：吕　越
封面设计：八度出版服务机构
出 版 人：曾庆宇
出版发行：北京科学技术出版社
社　　址：北京西直门南大街 16 号
邮政编码：100035
电话传真：0086-10-66135495（总编室）
0086-10-66113227（发行部）
0086-10-66161952（发行部传真）
电子信箱：bjkj@bjkjpress.com
网　　址：www.kjcbs.cn
经　　销：新华书店
印　　刷：北京宝隆世纪印刷有限公司
开　　本：787mm×1092mm　1/16
字　　数：448 千字
印　　张：19.5
版　　次：2017 年 5 月第 1 版
印　　次：2017 年 5 月第 1 次印刷
ISBN 978-7-5304-8474-6/Q·143

定　价：186.00 元

编 委 会

序

20 世纪 80 年代以来，我国的一些科学家与研究者已经意识到，生物信息学技术将在生命科学领域，以及医药、健康、农业等方面发挥重大的作用，因而，一批理论物理学家、数学家、计算机科学家投入生物信息领域，促进了我国生物信息学的发展。进入 21 世纪，随着人类基因组计划的实施与完成，生命科学家得到了大量的关于生物大分子的原始数据，借助于现代计算机技术，他们可以对这些原始数据进行收集、整理、分析、管理、检索与应用开发，如对数据进行比对、同源分析、建立计算模型、仿真与验证等，生物信息学进入蓬勃发展的时期。在数学、物理学的基础之上，信息技术和生物技术在学科层面进行更具深度和广度的交叉融合，再次推动了生物信息学这一新兴科学的发展。这一发展的速度之快、进步之巨、涉及之广，让许多研究者始料未及。尤其是近几年来，以新一代测序技术、单分子测序技术为代表的技术进步，使各类数据的产生速度陡然增加，而生物信息学的普及应用，总体看来，却还远落后于数据的增长。对这些海量数据的分析要求越来越高，计算量和复杂度也越来越大。

庆幸的是，高性能计算、云计算、大数据等技术的发展，为解决这一难题，提供了重要的技术支撑。目前在生命科学数据分析领域，高通量测序 + 高性能计算，已经逐渐成为研究重要生命科学问题的通用方法。然而要应用这些方法，不仅需要生命科学知识、生物实验设计方法，还涉及计算机科学领域的硬件配置、软件应用、数据库构建、算法开发、分析流程等一系列技术。换言之，不仅需要生物学、数理及计算理论的指导，还需要生物学、数理和计算信息技术的实践指导。一方面，生物医学研究者需要知道生物组学大数据的类型及产生的技术，是 WGS、Target Sequencing、RNA-seq、Chip-seq，还是 GWAS；另一方面，生物医学研究者还要了解通过什么样的计算平台、算法及软件来挖掘分析生物医学数据，解读这些数据的生物学内涵。

《高通量测序与高性能计算理论和实践》这本书，主要是由在一线进行数据分析的工作人员编撰的，是从解决实际问题的角度编写的一本专业参考书，不仅有丰

富的理论知识，也有大量的实际流程代码与结果示意图。因此，这本书在很大程度上，是一本方法与实践相结合的专业论著，对国内的生物信息学研究者，以及相关的从业人员，具有很好的指导作用和实际应用价值。

中国科学院院士，中国科学院生物物理所研究员

陈润生

2017 年 4 月 26 日

前 言

下一代测序技术（next generation sequencing technology，NGS）的应用始于2005年Roche公司的454测序仪，可以说开启了基因组学研究的新篇章。随着NGS及设备不断更新换代，测序成本快速下降，高通量测序技术不仅已经广泛应用于生物医学基础研究及应用研究，而且推进了农林分子育种领域的快速发展。随着NGS应用的深入，不仅需要样品制备及测序技术，而且需要计算技术和数据挖掘分析技术的支持；需要一本较为系统的把测序生物学实验应用和计算技术、数据分析相结合的著作。

2010年10月，北京市计算中心生物计算事业部基于北京市计算中心的计算资源优势，开始构建生物医学大数据整体解决方案平台，首先以NGS生物信息数据分析高性能计算整体解决方案为起点，建立了NGS数据分析计算环境及数据分析流程，构建了*de novo*测序数据分析流程、RNA－Seq数据分析流程、宏基因组数据分析流程、表观组数据分析流程及重测序数据的分析流程等；随着NGS数据分析业务的扩大，开展了组学数据与表型数据整合分析，启动了生物信息技术培训；2011年开启了NGS生物信息云计算平台构建，历经各种挫折后终于在2015年初开始服务于NGS数据分析。2014年，与北京市理化分析测试中心、中国人民解放军总医院联合成立了“北京市基因测序与功能分析工程技术中心”，中心经过NGS测序实验与数据分析计算资源一体运行的经验，结合几年的生物信息技术培训中学员反映的问题，把实践经验整理成这本《高通量测序与高性能计算理论和实践》。

本书整体设计为两大部分，即高通量测序实验篇和NGS数据分析篇。全书共四章，前两章为实验篇，包括第一章：测序技术进展（陈禹保、黄劲松、李春瑞、童贻刚），第二章：高通量测序实验（陈禹保、武会娟、卜祥霞、邢玉华、安云鹤、钱嘉林、李俊博、徐媛）。后两章为NGS数据分析篇，包括第三章：生物信息分析环境构建（潘勇、吴一雷、黄劲松、陈超、赵屹、闫鹏程、裴智勇、孙亮、吴杨、桂小柯、张丽杰、贾利鹃、陈禹保），第四章：高通量测序数据生物信息分析（裴智勇、闫鹏程、陈超、徐媛、刘满姣、陈禹保、童贻刚、朱怀球、赵屹）。最后由陈禹保、黄劲松统一审稿。

本书编写过程中得到了北京市科学技术研究院科技处、人事处等领导的大力支持，得到了北京市计算中心、北京市理化分析测试中心及北京市科学技术出版社等单位领导的大力支持。得到了中国科学院生物物理所陈润生院士、军事医学科学院二所童贻刚教授、北京大学生物医学工程系朱怀球教授、中国科学院基因组所曾长青研究员、方向东研究员，生命科学研究院赵方庆研究员，计算所赵屹研究员的学术支持。

感谢中国科学院生物物理所陈润生院士为本书作序，感谢北京市计算中心裴智勇博士做的大量的校对工作，感谢北京市计算中心生物计算事业部全体同仁的大力支持和辛勤工作。

本书编写过程中，笔者尽管努力发挥最好水平，但是技术发展迭代迅速，书中难免存在不足和缺点，敬请批评指正！笔者的邮箱：allexchen@126.com。

陈禹保　黄劲松

目 录

1 测序技术进展

1.1 绪论

基因组包含了生物全部的遗传信息，获得生物体基因组的全序列对于生物学研究和破解生命密码具有重要的意义。

1990 年启动的由 6 个国家 16 个中心以及全球众多科学家参加的“人类基因组计划”（HGP），于 2003 年圆满完成，与“曼哈顿计划”和“阿波罗计划”并称为人类自然科学史上的“三大计划”，这是基因组学研究领域的一个里程碑式的工作，它首度解读了人体 DNA 中所隐藏的、完整描述造就和维持人体生命的密码。HGP 的核心内容是测定人基因组的全部 DNA 序列，是人类获得自身最重要的生物学信息、实现对自身认识的一次重大的飞跃。

中国作为参加人类基因组计划的 6 个成员国中唯一的发展中国家，承担了 1% 的测序任务。2001 年 8 月 26 日，国际人类基因组计划中国部分“完成图”提前两年高质量地绘制完成。虽然只做了 1%，但这表明中国把握住了 21 世纪生物产业发展的机遇，因而意义重大。

图 1-1 人类基因组计划完成

HGP 的完成有力地推动了测序技术的发展，HGP 主要是通过 Sanger 测序法完成的，通过对并行度、自动化和微量化的改进，Sanger 测序法的成本已经降低到发明之初的 1% 以下，效率也提高了 100 倍以上。第二代测序技术（next - generation sequencing technology）的出现，不仅大大地降低了测序成本（人的基因组测序仅需 1000 美元左右），而且大大地提高了测序通量，随着技术的不断推进，测序已经成为基因组学强有力的研究工具。本章将就测序技术的发展、原理、流程及应用等做一概述。

1.2 测序技术发展历程

1.2.1 发展中的 DNA 测序技术

1.2.1.1 自动测序仪

自动测序仪是20世纪80年代中期，应用双脱氧终止法的原理，用非放射性荧光标记代替同位素标记，在电泳过程中通过激光激发荧光，然后用探测器收集荧光信号，再通过计算机进行图像识别与分析的测序方法。这种方法实现了 DNA 测序的全自动化，并大大节约了人力与物力。

1.2.1.2 毛细管凝胶电泳测序技术

1990 年 Zagursky 和 McCornick 建立了毛细管凝胶电泳质谱测序技术。毛细管凝胶电泳技术将凝胶电泳对大分子的高分离率与 cap 电泳的快速、微量相结合。电泳中凝胶的抗对流性大大提高了分辨率。

1.2.1.3 杂交测序技术

杂交测序是根据 DNA 分子中碱基互补配对的特性，通过标记的单链 DNA 模板，与一系列短链寡核苷酸探针分子杂交，来实现 DNA 测序的策略。杂交测序检测速度快，采用标准化的高密度寡核苷酸芯片能够大幅度降低检测的成本。

1.2.1.4 基因芯片测序技术

早在 1980 年，Bains 等人就采用将探针固定于载体上再杂交的方法进行 DNA 测序，这就是基因芯片测序技术的最初模型。基因芯片测序技术是建立在杂交测序基础之上的一种 DNA 测序方法。

1.2.1.5 PCR 直接测序技术

PCR 直接测序技术是以 PCR 扩增引物作为测序引物，这极大地提高了 DNA 测序分析的效率。

1.2.1.6 cDNA 微阵列技术

cDNA 微阵列技术是以荧光标记的 DNA 探针，与 cDNA 微阵列进行杂交，从而进行扩大规模基因表达分析的一种新方法。

1.2.2 经典的 DNA 测序方法

1977 年 Maxam 和 Gilbert 等发明的化学降解法和 Sanger 等发明的双脱氧核苷酸末端终止法，标志着第一代测序技术的诞生。二者都是先得到随机长度的 DNA 链，再通过电泳方法读出序列。不同之处在于，Gilbert 法是先用特定的化学试剂标记碱基，再用化学方法打断待测序列，而 Sanger 法是通过 ddNTP 随机中断所合成的待测序列。

1.2.2.1 化学降解法

化学降解法是将模板 DNA 的一端标记之后，在四组或五组互为独立的化学反应中

分别得到部分降解，其中每一组反应特异地针对某一种或某一类碱基。在这几组反应中，通过化学裂解形成具有共同起点而终点不同的放射性标记的分子。经过电泳及放射后自显影可以读出距离标记位点 250 个核苷酸以内的 DNA 序列，不仅适用于单链，也可用于双链测序。

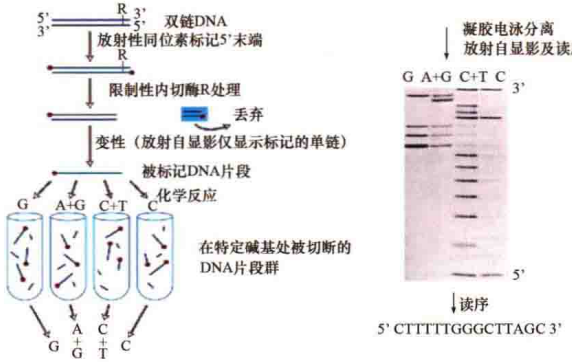

图 1－2　Maxam-Gilbert 化学降解法测序原理

1.2.2.2　双脱氧链终止法

在 Maxam 和 Gilbert 提出通过化学降解测定 DNA 序列的方法的同一年，Sanger 推出了双脱氧链终止法，因是 Sanger 发明，故又称 Sanger 测序法。双脱氧法不仅具有化学法的优点，而且适用于大规模测序。目前，最主要的测序技术都是以 Sanger 法为基础的。本章第三节将对 Sanger 测序技术的原理、流程、影响因素分别进行介绍。

1.3　Sanger 测序技术的原理和流程

1.3.1　Sanger 测序技术的原理

核酸模板在核酸聚合酶、引物、四种单脱氧核苷酸存在条件下复制或转录时，在四管反应系统中分别按比例引入四种双脱氧核苷酸，只要双脱氧核苷酸掺入链端，该链就停止延长，链端掺入单脱氧碱基的片段可继续延长。如此每管反应体系中便合成以共同引物为 5′端，以双脱氧碱基为 3′端的一系列长度不等的核酸片段。反应终止后，分四个泳道进行电泳。以分离长短不一的核酸片段（长度相邻者仅差一个碱基），根据片段 3′端的双脱氧核苷酸的碱基种类，便可依次阅读合成片段的碱基排列顺序。双脱氧核苷酸（ddNTP）分子结构及 DNA 链合成终止反应见图 1－3。

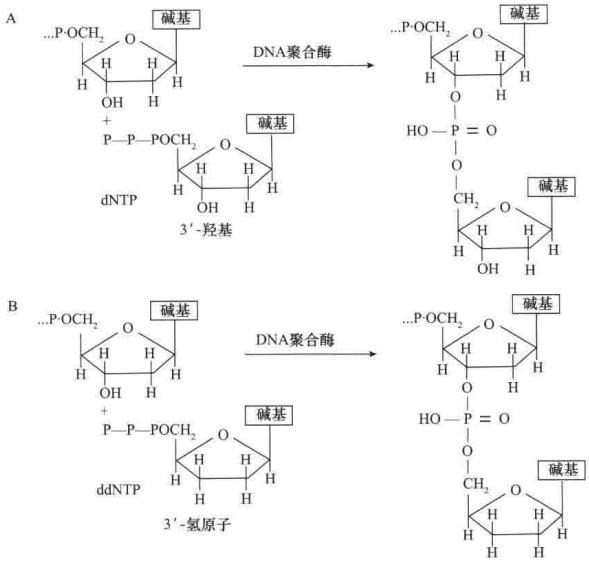

图 1－3　双脱氧核苷酸（ddNTP）分子的结构及 DNA 链合成终止反应

A. 正常的 DNA 合成反应；B. ddNTP 掺入 DNA 合成反应后导致反应终止

1.3.2　Sanger 测序技术流程

（1）文库制备：将随机 DNA 片段的大量拷贝克隆到质粒并转化大肠杆菌（随机从头测序），或使用引物对目标片段进行 PCR 扩增（目标片段重测序）。

（2）测序反应：即 DNA 体外合成的变性、退火、延伸循环，引物延伸片段在每个循环都可能因末端加入荧光双脱氧核苷酸而终止延伸，最后可得到包含所有长度的末端标记片段的混合物。

（3）测序示踪：毛细管电泳过程中，用激光检测各种单链 DNA 走出凝胶的时间和荧光类型。

（4）计算机分析：软件将示踪信号翻译成序列，并计算出误差概率。

1.3.3　影响 DNA 测序的因素

DNA 测序成功与否，与构建文库所用的模板的浓度、纯度，模板自身的序列结构及测序过程有直接的关系。DNA 模板的类型主要有 ssDNA（如 M13）、dsDNA（如各类质粒）、黏粒、BAC、细菌基因组 DNA 等（表 1－1）。其中 PCR 产物及质粒 DNA 模板是最常见的科研用测序 DNA 模板。DNA 模板的浓度与纯度，以及在循环测序反应中的终浓度是其测序成功与否的前提条件，也是最易出现问题的环节。