141

国外生命科学优秀教材

基 因 组

〔英〕T.A.布朗 著

袁建刚 周 严 强伯勤 主译

科 学 出 版 社

2002

内 容 简 介

本书以清晰而简明的写作风格将基因组学的新观点与研究基因表达的传统方法相结合，为基因组作为生命蓝图所起的重要作用提供了最新理解。全书从三个方面论述了基因组学的核心主题："基因组的研究方法"涵盖了基因组的作图和测序技术，以及对功能基因组学的详细介绍；"基因组的功能"描述了基因组中基因的表达如何决定细胞的生化特征，以及基因的调控是如何与发育途径相统一的；"基因组的复制和进化"在分子和细胞水平解释了复制、突变和重组，以及它们如何决定基因组的进化。书中还包含所有最新的资料、基因组测序计划的最新信息和一些关键性的研究发现，使学生能熟悉真正的科学工作和数据处理。全书采用大量的图表，形象而简洁，是一本必备的现代分子生物学教材。

本书可作为生物类及相关专业本科生和研究生的课程教材，也可供专业科研人员阅读。

T. A. Brown
Genomes
Original edition published in English under the title of *Genomes*

图字 01-2000-0189 号
图书在版编目(CIP)数据

基因组/(英)布朗(Brown, T. A.)著；袁建刚等译. —北京：科学出版社，2002.9
(国外生命科学优秀教材)
ISBN 7-03-010114-6

Ⅰ. 基… Ⅱ. ①布… ②袁… Ⅲ. 基因组-教材 Ⅳ. Q343.2

中国版本图书馆 CIP 数据核字(2002)第 009345 号

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
源海印刷厂 印刷

科学出版社发行 各地新华书店经销

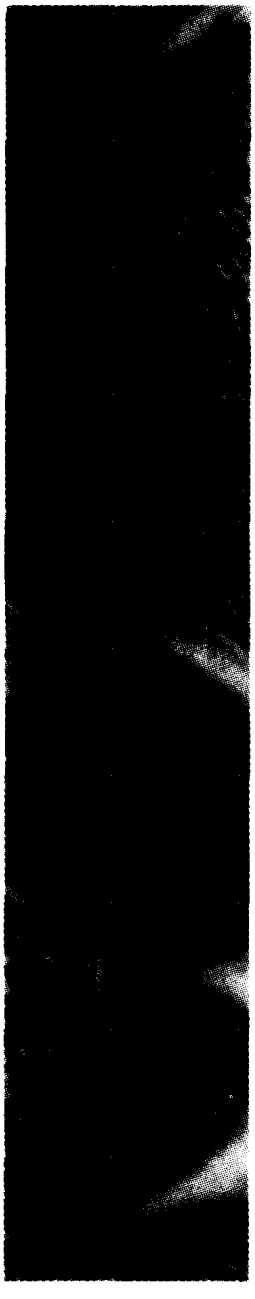
*

2002 年 8 月第 一 版 开本:850×1168 1/16
2002 年 8 月第一次印刷 印张:33 1/4 插页:1
印数:1—3 000 字数:748 000

定价:48.00 元

(如有印装质量问题，我社负责调换〈北燕〉)

前　言



本书旨在为讲授大学分子生物学提供一个全新的途径。其出发点是大学分子生物学的教学大纲应该反映新千年的主要研究方向，而不应再是20世纪70年代或80年代流行的论题。因此，本书以基因组，而非基因为中心内容。事实上今天的分子生物学主要是由基因组测序和功能分析推动的，而不再是单个基因水平上的研究。今天许多学习分子生物学的大学生在他们的研究生阶段都将涉及基因组研究的工作，那时，他们将发现他们的工作或多或少地被基因组计划所影响。如果对大学生授课的目的是为他们将来的研究生涯作准备的话，那么，就应该向他们教授基因组方面的知识!

认为基因不再重要的看法当然是愚昧的。我撰写《基因组》时面临的最大挑战是如何将传统分子生物学课程的精髓与有关基因组的新内容结合在一起。因为无法完全用“从基因组到蛋白质组”的笼统方式充分描述从DNA到蛋白质这一过程中的每一事件，因此，《基因组》中有相当一部分实质性的内容集中在单个基因的表达途径上。本书与其他书不同之处在于：尽量把基因组的活动和功能作为一个整体，在该背景下描述单基因的表达途径。与此相仿，对DNA的复制、突变和重组的叙述也是讲述它们对整个基因组、而不是简单对基因复制和变化过程的影响。

随着本书的撰写，我认为分子生物学授课应该集中在基因组上的信念越来越坚定。我还发现相对于传统教学大纲来说，这种方式更令人满意，也能提供更多的信息。许多过去对我来说属于边缘性的主题现在逐渐凸现出来，并显出其新的重要性。希望我能将在撰写本书过程中所感受到的兴奋之情传递给读者。

T. A. Brown

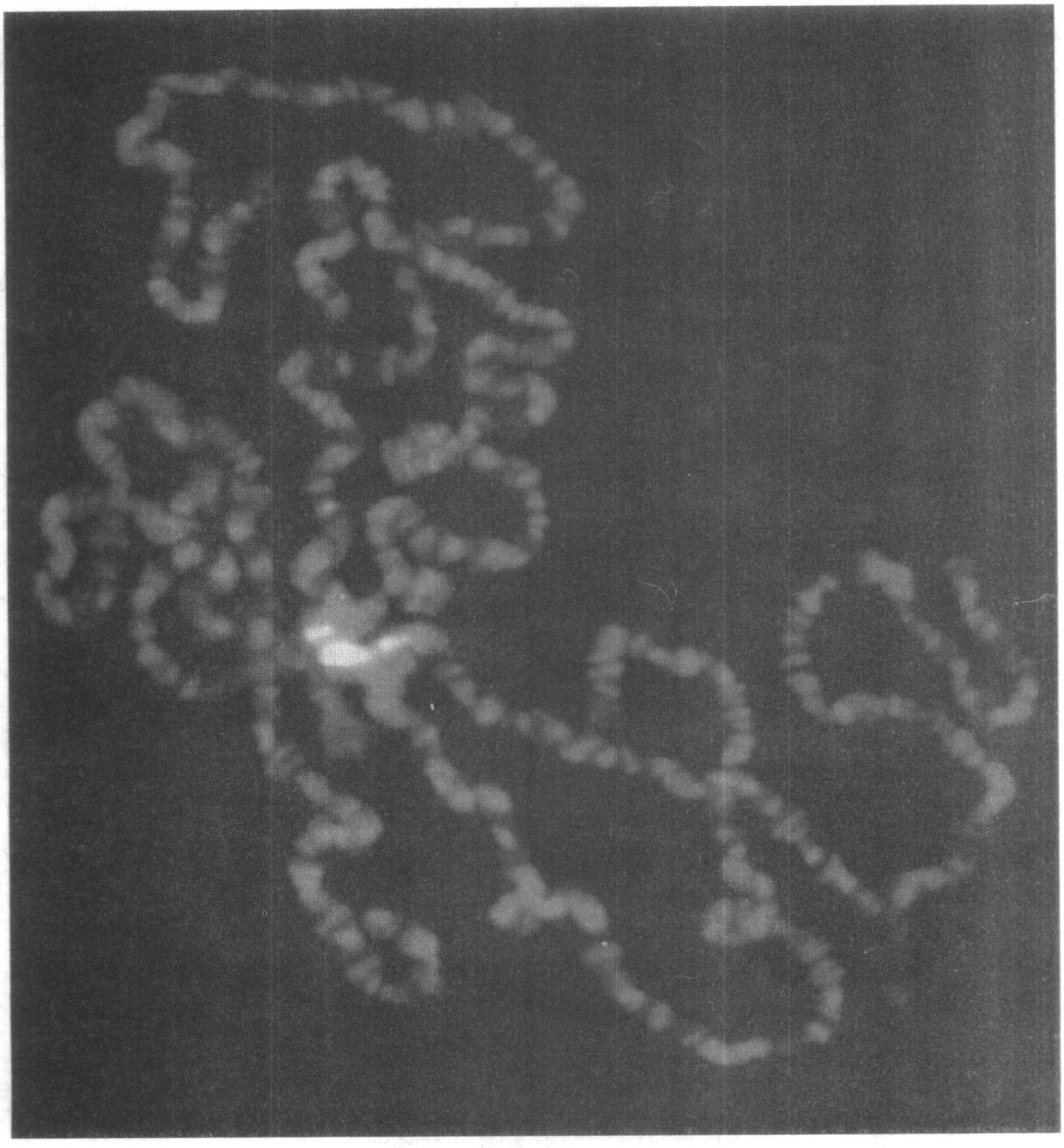
曼彻斯特

目　录

第 3 篇 基因组的复制和进化

PART

第1篇 基因组的研究方法



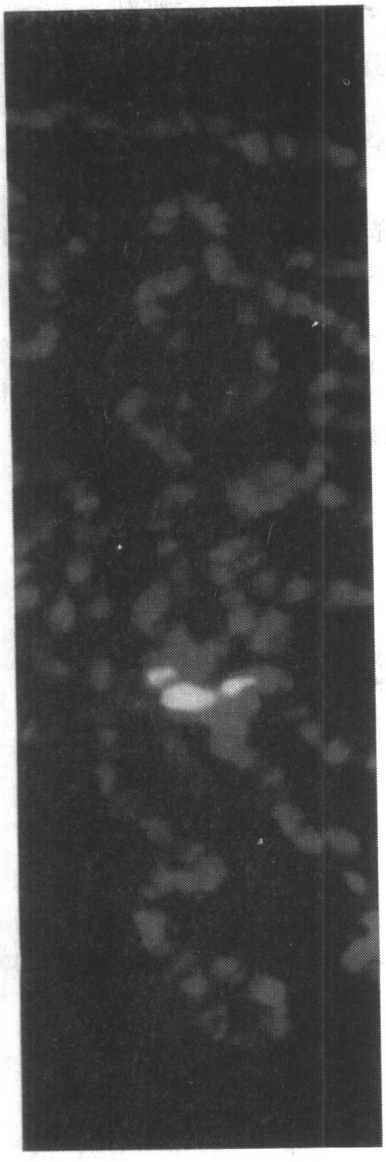
上页：黑腹果蝇（*Drosophila melanogaster*）的一个染色体，利用着丝粒重复序列特异的 DNA 探针进行荧光原位杂交。染色单体是灰白色的，明亮的白色位于着丝粒。有关原位杂交技术见 3.2，着丝粒的信息见 6.1.1 和 6.3.1。

致谢：在 P. A. Coelho 和 C. E. Sunkel（葡萄牙 Universidade do Porto 分子和细胞生物学研究所）的授权下复制该图。

该图由 Bio-Rad Microscience Ltd. 提供。

CHAPTER

第1章 什么是基因组



内 容

概　念

- 基因组是细胞中所有的 DNA，包括所有的基因和基因间区域
- 人类基因组大约有 80 000 个基因，这些基因的编码区仅占整个基因组的 3%
- 酵母基因组有 6000 个基因，并且组织形式更紧密
- DNA 重复序列在一些植物基因组中占优势
- 原核生物基因组小，基因间几乎没有间隔
- 充分理解基因组序列中包含的信息将是 21 世纪早期面临的主要挑战

众所周知，生命是由**基因组**（genome）决定的。每个生物都具有基因组，携带着构成和

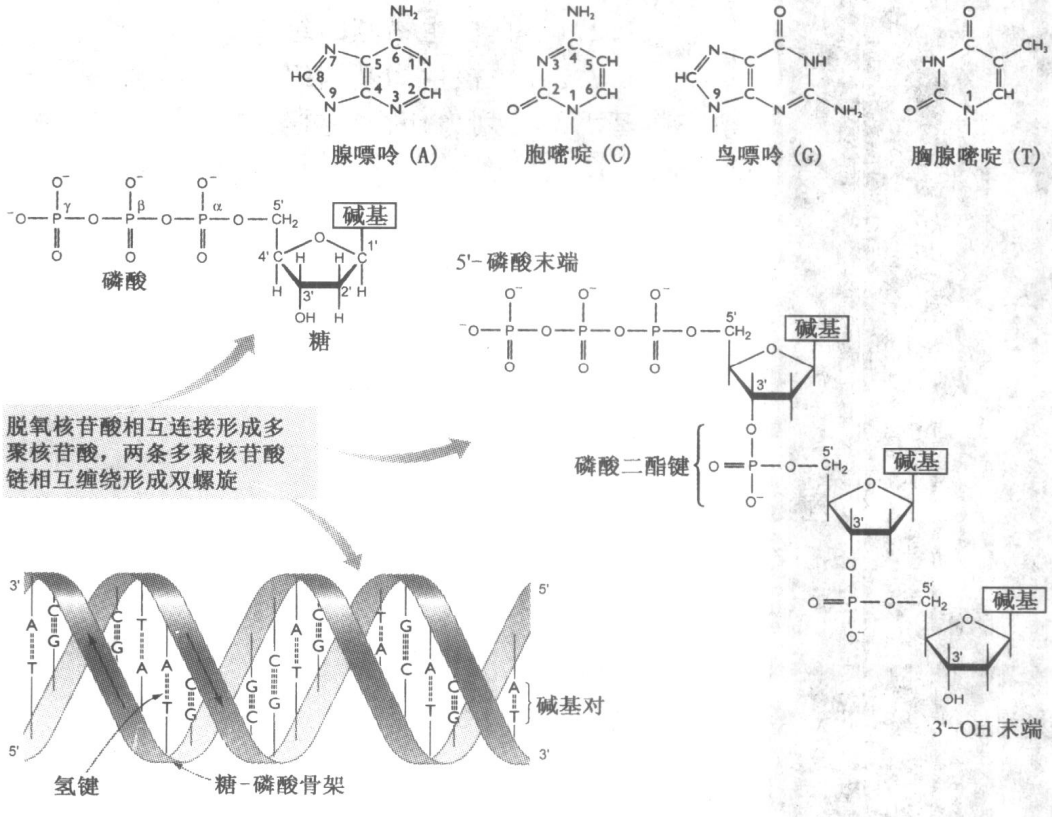


图 1.1　DNA 的结构

每一个脱氧核糖核苷酸由 2′-脱氧核糖与 1 个磷酸基团和 4 种碱基（碱基是腺嘌呤、胞嘧啶、鸟嘌呤和胸腺嘧啶）中的一个组成。核糖核苷酸通过磷脂键结合在一起形成多聚核糖核苷酸，并产生两个化学特性不同的末端。双螺旋中，两个多聚核糖核苷酸以不同的方向通过碱基对之间的氢键相互缠绕在一起。7.1 节对 DNA 和 RNA 的结构有更详细的描述。

维持该生物体生命形式所必需的所有**生物信息**（biological information）。绝大部分基因组，包括所有的细胞生命形式的基因组，由 **DNA**（脱氧核糖核酸）组成，但有一些病毒具有 **RNA**（核糖核酸）基因组。DNA 和 RNA 是由**核苷酸**（nucleotide）单体构成的线性、无分支的多聚分子。每个核苷酸包含三部分：一个单糖、一个磷酸基团和一个碱基（图 1.1）。DNA 中，糖是 2′-脱氧核糖，碱基分别是腺嘌呤（A）、胞嘧啶（C）、鸟嘌呤（G）和胸腺嘧啶（T）。核苷酸之间通过**磷酸二酯键**形成包含几百万个核苷酸的 DNA 多聚体或称**多聚核苷酸**（polynu-

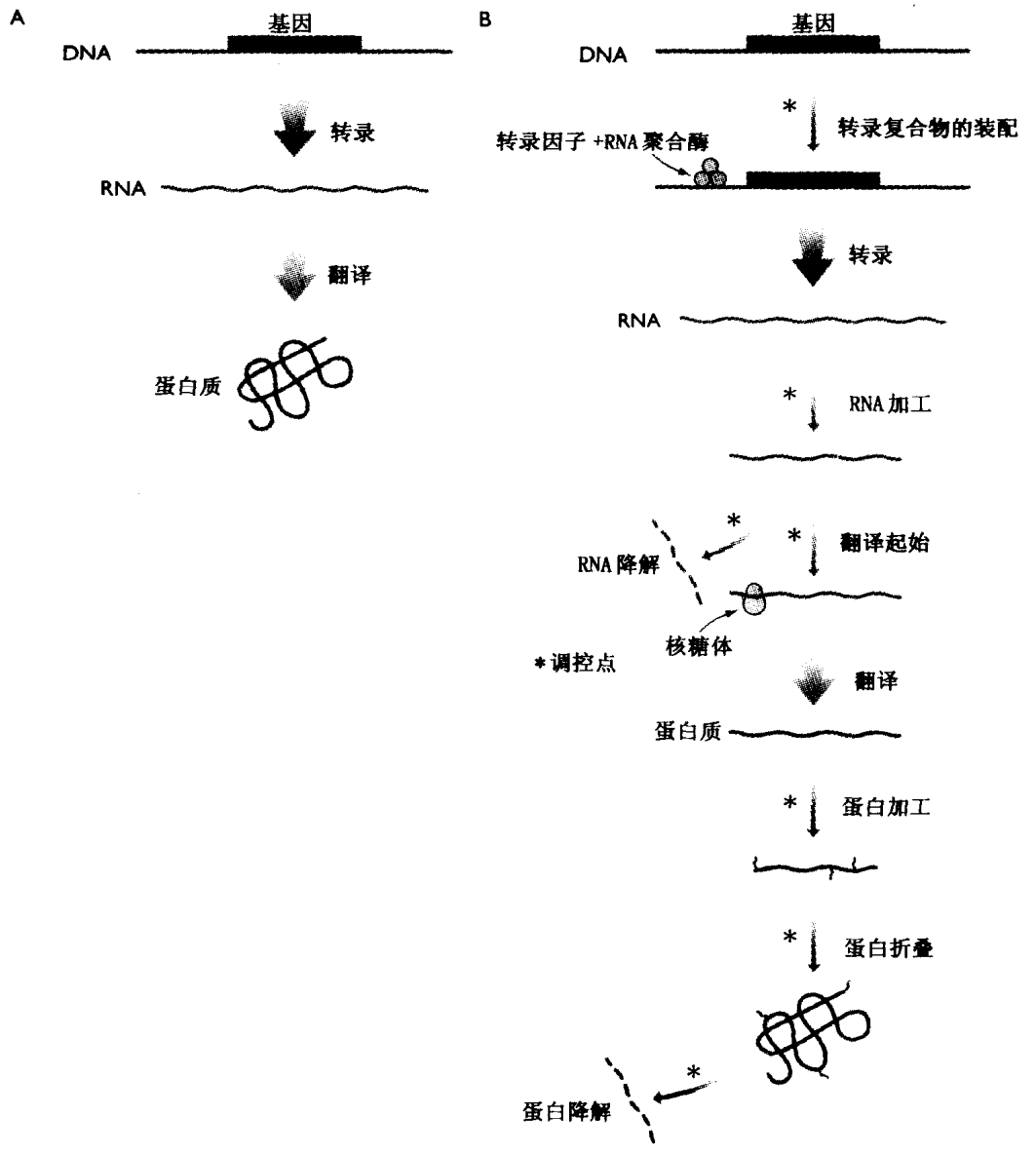


图 1.2　基因表达的两个见解

A. 传统基因表达的描述，总结为 DNA 产生 RNA，RNA 产生蛋白质，这两个步骤分别叫做转录和翻译。B. 高等生物基因表达更正确的略图。关键的调控点是突出的。注意，这个设计仅适用于编码蛋白的基因。一些基因产生非编码的 RNA，如核糖体 RNA 和转运 RNA：这些基因由所示那样转录和加工，但这些 RNA 并不翻译（9.1.1）。基因表达所有步骤的细节见本书的第二部分。

cleotide)。活细胞中的 DNA 是**双链**（double-strand）的，两链之间相互缠绕形成**双螺旋**（doublehelix)。双链通过碱基对间的**氢键**（hydrogenbond）维持着双螺旋结构。碱基配对原则是 A 与 T 配对，G 与 C 配对。因此，双螺旋中的两条 DNA 链是**互补**（complementary）序列。

RNA 核苷酸中的糖是核糖而不是 2′-脱氧核糖，胸腺嘧啶（T）被尿嘧啶（U）所取代。RNA 很少超过几千个核苷酸，细胞中的 RNA 通常是**单链**（single-strand）的，但单链分子的不同区域间也可形成碱基对。

基因组中的生物信息由 DNA 或 RNA 的核苷酸序列所编码，并分别称作**基因**（gene）的不连续单元。基因信息由一些蛋白质来解读，这些蛋白质结合在基因组的适当位置，引发称为**"基因表达"**（gene expression）的一系列生物化学反应。最初认为，对于含 DNA 基因组的生物，这一过程包括两个阶段，即**转录**（transcription）和**翻译**（translation），前者产生基因的 RNA 转录物，后者合成蛋白质，蛋白质的氨基酸序列是由 RNA 转录物（图 1.2A）中核苷酸序列代表的**遗传密码**（genetic code）决定的。在简单生物如细菌中，这种描述较为精确。但对高等生物中，基因组信息向功能蛋白的转换过程而言，该描述尚不完美（图 1.2B)。这种对于转录-翻译过程阐述的一个重要缺点在于，它转移了我们对基因表达通路中信息流的调节关键点的注意力。

每一次细胞分裂都一定会产生一个完整的基因组拷贝。**DNA 复制**（DNA replication）必须极其精确，以避免将**突变**（mutation）导入基因组拷贝中。然而，某些突变的确发生了，它们可能是由于复制过程中产生了错误或是化学或物理诱变剂直接改变了 DNA 的化学结构所造成的。DNA 修复酶可以更正许多这样的错误；那些逃避了修复的突变就作为已突变的基因组的永久性特征而保留下来。这些基因突变事件以及由重组导致的基因组重排是分子进化的原因，而分子进化又是生物进化的推动力量。

1.1 人类基因组

在现有的基因组中，我们自然对自己的基因组最感兴趣。因此，我们首先介绍人类基因组的结构和组织概况。在本章的第 2 节，我们将讨论人类和其他生物基因组之间的相似和不同之处。

1.1.1 人类基因组的物理结构

人类基因组由两类不同的组分构成（图 1.3)：

- **核基因组**（nuclear genome） 由大约 30 亿个**碱基对**（bp）组成。这个数字也可以说成 300 万**千碱基对**（kb）或 3000 **兆碱基对**（Mb)。核基因组分为 24 个线性 DNA 分子，最短的 55Mb，最长的 250Mb，每一个分子包含在不同的**染色体**中。24 条染色体中有 22 条常染色体，两条性染色体——X 和 Y。
- **线粒体基因组**（mitochondrion genome） 是一个长为 16 569bp 的环状 DNA 分子，它有许多拷贝，位于产生能量的细胞器——**线粒体**（mitochondron）中。

成人体内的约 10^{13}个细胞中均含有单拷贝或多拷贝基因组，只有少数细胞类型例外，如血红细胞，在完全分化阶段缺少细胞核。绝大多数的细胞是**二倍体**（diploid），含两个拷贝的常染色体和两条性染色体，雌性个体中是 XX，雄性个体中是 XY，总共 46 条染色体。这些细胞

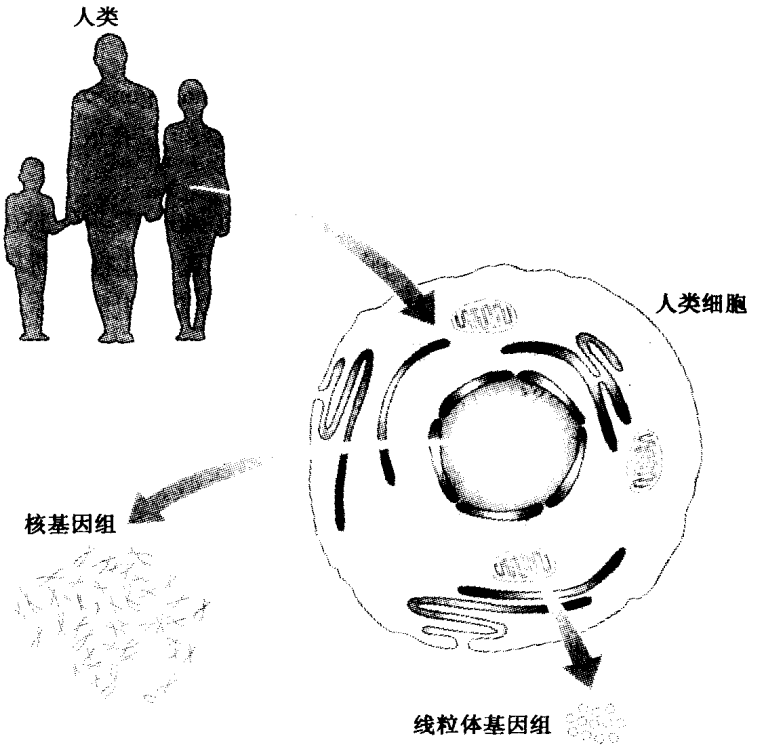


图 1.3　人类基因组的核与线粒体成分
要更详细的了解人类基因组结构请见 6.1。

叫做**体细胞**（somatic cell）。而**性细胞**（sex cell）或**配子**（gamete）是**单倍体**（haploid），仅有 23 条染色体，包括一套常染色体和一条性染色体。这两类细胞中大约有 8000 个拷贝的线粒体基因组，每个线粒体中有 10 个拷贝左右。

在我们进行深入讨论之前，先要对人类基因组之巨大有一个感性认识。30 亿是一个如此大的数目，我们很难理解它所代表的规模，因此，类比的方法是很有帮助的。如果以本书的字体大小打印，DNA 序列中每 60 个核苷酸排列长度约为 10cm。如果按这种格式打印，人类基因组序列长度可为 5000km，相当于从蒙特利尔到伦敦、洛杉矶到巴拿马、东京到加尔各答、开普敦到亚的斯亚贝巴或奥克兰到珀思的距离（图 1.4），可写满 3000 本与本书大小一致的书。即使是最简单的细菌基因组按这种格式打印也可排出千米的距离。如果我们想要理解基因组的构成及功能行使方式，那么我们正面临着无比巨大的任务。

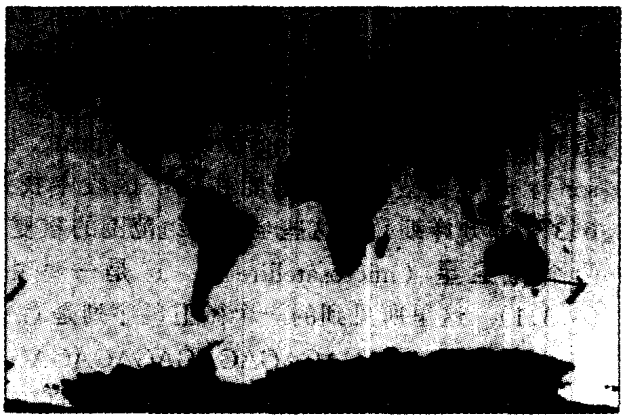


图 1.4　人类基因组的长度
该图显示的是如果按本书的字体打印人类基因组序列所覆盖的距离。

1.1.2 人类基因组的遗传内容

认识人类基因组是如何构成的是**人类基因组计划**（human genome project）的目标之一。于1984年酝酿、并始于1990年的人类基因组计划旨在2003年前完成人类基因组的测序工作。序列将揭示什么呢?

尽管有科学家提出过基因的数目（Cohen 1997）少至50 000或多至150 000个，但现有的共识是人类基因组大约有80 000个基因。我们将在第6章讨论这些推测所依据的基础（解释框6.2）。同时我们也知道这80 000个基因中包括的信息仅占核基因组的3%。图1.5显示了7号染色体中一长为50kb的区段的遗传组成，此区段形成人类βT细胞受体基因座的一部分，该基因座是一很大的基因组区域（685kb），编码与免疫应答相关的蛋白（Rowen *et al*. 1996）。此50kb的区段包含下列遗传特征：

- 一个基因　该基因称为 *TRY4*，编码胰蛋白酶原，是消化酶胰蛋白酶的无活性前体。*TRY4* 是位于βT细胞受体座位两端的胰蛋白酶原基因家族成员之一。这些基因与免疫应答毫无关系，它们只不过是与βT细胞受体基因座共享该部分染色体而已。
- 两个基因片段　即 *V28* 和 *V29 - 1*，编码βT细胞受体蛋白的一部分，该基因座由此而得名。*V28* 和 *V29 - 1* 的特殊性在于它们不是完整的基因，而是一个基因的片段。表达之前，它们必须与该座位的其他基因片段相连。该过程发生在T淋巴细胞中，是细胞分化过程中基因组活性永久改变的一个例子（11.2.1）。注意 *TRY4*、*V28* 和 *V29 - 1*，像其他的人类基因一样，是**不连续**（discontinuous）的，由包含蛋白质编码信息的**外显子**（exon）组成，这些外显子被非编码的**内含子**（intron）所间断。外显子加在一起是1414bp，占该50kb片段的2.8%。该片段的编码能力是基因组整体编码能力的一个相当典型的例子。
- 一个**拟基因**（pseudogene）　拟基因是基因的非功能性拷贝，通常是因为发生了突变，其生物信息变得不可解读（6.1.1）。该拟基因叫做 *TRY5*，它与胰蛋白酶原家族的功能成员有很近的关系。
- 全基因组范围分布的52个重复序列　这是一些在基因组的多处都能发现的序列。**全基因组范围的重复序列**（genome-wide repeat sequence）有4个主要类型，它们分别是 **LINE**（long interspersed nuclear element，长散布核元件）、**SINE**（**short interspersed nuclear element**，短散布核元件）、**LTR**（**long terminal repeat**，长末端重复）和 **DNA 转座子**（DNA transposon）。每一类型的例子都能在这一区段中找到，加起来共占全序列的39.1%。我们将在6.3.2部分详细介绍这些全基因组范围的重复序列。
- 两个**微卫星**（microsatellite）　这是一些由短元件首尾相接、线性排列的重复序列（6.3.1）。这里所见到的一个微卫星序列是GA重复16次所组成的，其序列是：

  5′-GAGAGAGAGAGAGAGAGAGAGAGAGAGAGA-3′

  3′-CTCTCTCTCTCTCTCTCTCTCTCTCTCTCT-5′

  第二个微卫星是由TATT重复6次所成。许多微卫星都是多态的，重复次数随个体而异。微卫星是基因组中很有用的标记位点，在第2章和第3章我们将看到它们在构建基因组图谱中的作用。
- 最后，此50kb区段的50%是一些既非基因、也非重复，功能和意义未知的单拷贝DNA序列。

任何一个短的区段都不可能真正代表基因组整体。就某些方面来说，图 1.5 所示的 50kb 的区段又是相当不典型的，但它阐明了人类基因组中的遗传特征概貌。我们将讨论的下一个问题是：人类基因组与其他生物基因组的相似性。

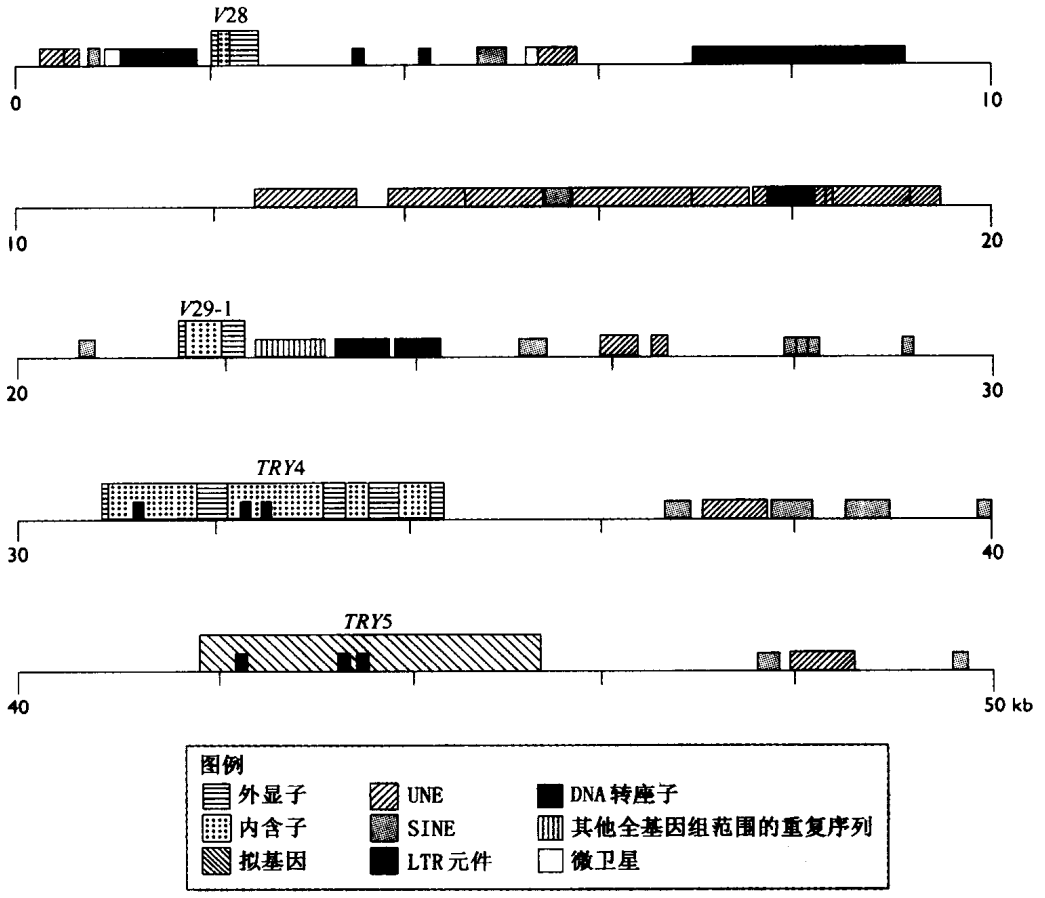


图 1.5　人类基因组的一个区段

该图显示了位于人 7 号染色体上的 βT 细胞受体座位上的基因、基因片段、拟基因、全基因组范围的重复序列和微卫星的位置。摘自 Rowen *et al*. 所编书（1996）。

解释框 1.1　关于基因的一些关键概念

基因是可以转录成 RNA 的基因组片段。如果 RNA 是蛋白编码基因的转录物，那么这种 RNA 称为**信使 RNA**（mRNA），它能翻译成蛋白质。如果 RNA 是非编码的**核糖体 RNA**（rRNA）和**转运 RNA**（tRNA），那么它不能翻译成蛋白质。非编码 RNA 在细胞中具多种功能（9.1.1）。

编码蛋白的基因中翻译成蛋白质的那部分称**可读框**(open reading frame, ORF)。中的每一个核苷酸三联体是一个**密码子**(codon)，根据**遗传密码子**(genetic code)原则每一个密码子决定一个**氨基酸**(10.1.2)。ORF 沿 mRNA 从 5′到 3′的方向阅读。ORF 始于一个**起始密码子**(initiation codon)，终止于一个**终止子**(termination codon)。ORF 前面的 mRNA 部分称**前导**(leader)片段，紧随其后的部分叫**后随**(trailer)片段。

真核生物中，许多基因是断裂基因，间断成**外显子**和**内含子**。在**剪接**（splice）过程中，内含子从最初的转录本中被切除，产生功能性的 RNA 分子（9.2.3）。