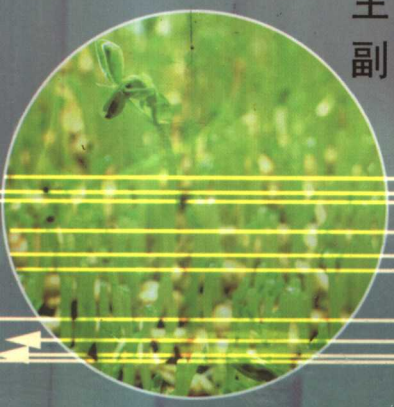
分子生物学理论与常用实验技术

主　编　张　涛　王　昭　江旭东
副主编　王明富　赵守琪　李　健



哈尔滨地图出版社

高等学校教材

分子生物学理论与常用实验技术

FENZI SHENGWUXUE LILUN YU CHANGYONG SHIYAN JISHU

主　编　张　涛　王　昭　江旭东

副主编　王明富　赵守琪　李　健

参　编　王树清　吕秀芳　田丽华

刘志伟　邢立娟　杨　波

张　杰　张　军　张　红

张义英　姚嵩坡　袁晓环

崔荣军　徐　辉　谢　红

哈尔滨地图出版社

·哈尔滨·

图书在版编目(CIP)数据

分子生物学理论与常用实验技术/张涛，王昭，江旭东主编. —哈尔滨：哈尔滨地图出版社，2007.6
ISBN 978-7-80717-663-3

Ⅰ.分… Ⅱ.①张…②王…③江… Ⅲ.分子生物学
Ⅳ.Q7

中国版本图书馆 CIP 数据核字(2007)第 101006 号

哈尔滨地图出版社出版发行
(地址：哈尔滨市南岗区测绘路 2 号　邮政编码：150086)
哈尔滨市动力区哈平印刷厂印刷
开本：787 mm×1 092 mm　1/16　印张：11.25　字数：295 千字
ISBN 978-7-80717-663-3
2007 年月 6 第 1 版　2007 年 6 月第 1 次印刷
印数：1～1 000　定价：25.60 元

前　言

分子生物学是20世纪迅速发展起来的从分子水平研究生命本质的一门新兴学科，是21世纪生命科学的带头学科。分子生物学的理论和实验技术已渗透到生命科学的各个领域，分子生物学技术已经成为医学各个学科的关键技术，是教学、科研、医疗、诊断和治疗中最常用的技术手段。本书从分子生物学的基本原理和分子生物学常用的实验技术二方面对现代分子生物学进行了论述，系统介绍了基因克隆、基因表达、蛋白质分析等经典理论和实验技术，既简明扼要地阐述了分子生物学理论又介绍了分子生物学最常见和实用的技术方法和实验方案。本书适合高等院校生命科学专业的师生和从事生命科学的研究工作者的参考和使用。

本书分二大部分：第一部分是分子生物学基础理论，包括基因与基因组，基因表达与调控，DNA重组技术，分子杂交，聚合酶链反应，基因诊断与基因治疗和生物芯片技术等。第二部分是分子生物学常用实验技术，包括质粒DNA的提取、酶切与电泳鉴定，重组DNA分子的构建、筛选与鉴定，大肠杆菌感受态细胞的制备及转化，PCR基因扩增，哺乳动物基因组DNA的提取，酵母tRNA的制备(苯酚法)，DNA序列测定，DNA的分子杂交，RNA的分子杂交，蛋白质的分子杂交，外源基因转染哺乳动物细胞，真核基因在原核细胞的表达等十二个实验，并附有分子生物学实验中常用数据及换算关系，分子生物学常用试剂，溶液和缓冲液的配制和常用培养基和抗生素的配制等。书中基本概念论述准确，深度适中，紧扣分子生物学的基本内容，又力求反映分子生物学研究的新成果、新进展、新的研究手段和方法，以达到拓宽基础、开拓视野、加强对学生科学素质和能力的培养之目的。

本书由张涛(佳木斯大学)、王昭(佳木斯大学)、江旭东(佳木斯大学)担任主编，由王明富(佳木斯大学)、赵守琪(齐齐哈尔医学院)、李健(燕山大学环境与化学院生物工程系)担任副主编，参加编写的还有王树清(佳木斯市中心医院)、吕秀芳(牡丹江医学院)、田丽华(佳木斯大学)、刘志伟(燕山大学环境与化学院生物工程系)、邢立娟(佳木斯市中心医院)、杨波(佳木斯市中心医院)、张杰(牡丹江医学院)、张军(佳木斯市中心医院)、张红(佳木斯市中心医院)、张义英(佳木斯大学)、姚嵩坡(佳木斯大学)、袁晓环(牡丹江医学院)、崔荣军(牡丹江医学院)、徐辉(佳木斯大学)、谢红(天津武警医学院)。

本书编写过程中，受到兄弟院校同行、同事以及哈尔滨地图出版社的鼓励和支持，在此表示由衷的感谢。由于分子生物学理论和技术的飞速发展，新的技术和研究成果的大量涌现，加上编者水平和经验有限，书中难免有不当之处，敬请读者批评指正。

编　者

2007年6月

目　录

第一部分　分子生物学基础理论

第二部分　分子生物学常用实验技术

第一部分　分子生物学基础理论

第一章　基因与基因组

遗传因子(hereditary factor)的概念，最初是由奥地利学者 Mendel 提出的。Mendel 从 1857 年到 1864 年根据豌豆为材料进行的杂交实验，提出遗传因子的概念，并总结出 Mendel 遗传定律。令人遗憾的是，Mendel 的理论直到 1900 年才被荷兰的 H. De Vries、德国的 C. Correns 和奥地利 E. Tschermak 等植物学家重新发现。1909 年，丹麦植物学家和遗传学家 W. Johannsen 首次提出“基因”这一名词，用以表达 Mendel 的遗传物质的分子。1910 年，美国著名遗传学家 T. H. Morgan 和他的研究小组对基因学说的建立做出了杰出的贡献，使基因学说得到了普遍的承认。1953 年，Watson—Crick 通过实验提出 DNA 分子的双螺旋模型，使遗传学家能够从分子水平分析遗传与变异现象。基因以一种真正的分子物质呈现在我们面前，科学家们能够像研究其他大分子一样，客观地探索基因的结构及功能。这样，人们便开始从分子的层次上研究基因的遗传现象。1969 年，科学家成功分离出第一个基因。1990 年 10 月国际人类基因组计划启动。1998 年一批科学家在美国罗克威尔组建赛莱拉遗传公司，与国际人类基因组计划展开竞争。1998 年 12 月一种小线虫完整基因组序列的测定工作宣告完成，这是科学家第一次绘出多细胞动物的基因组图谱。1999 年 9 月中国获准加入人类基因组计划，负责测定人类基因组全部序列的 1%。中国是继美、英、日、德、法之后第 6 个国际人类基因组计划参与国，也是参与这一计划的唯一发展中国家。1999 年 12 月 1 日国际人类基因组计划联合研究小组宣布，完整破译出人体第 22 对染色体的遗传密码，这是人类首次成功地完成人体染色体完整基因序列的测定。2001 年 2 月 12 日中、美、日、德、法、英等 6 国科学家和美国塞莱拉公司联合公布人类基因组图谱及初步分析结果。

基因(gene，Mendelian factor)，也称为遗传因子。是指携带有遗传信息的 DNA 或 RNA 序列，是控制性状的基本遗传单位。基因通过指导蛋白质的合成来表达自己所携带的遗传信息，从而控制生物个体的性状表现。除某些病毒的基因由核糖核酸(RNA)构成以外，多数生物的基因由脱氧核糖核酸(DNA)构成，并在染色体上作线状排列。基因一词通常指染色体基因。在真核生物中，由于染色体都在细胞核内，所以又称为核基因。位于线粒体和叶绿体等细胞器中的基因则称为染色体外基因、核外基因或细胞质基因，也可以分别称为线粒体基因、质粒和叶绿体基因。在通常的二倍体的细胞或个体中，能维持配子或配子体正常功能的最低数目的一套染色体称为染色体组或基因组，一个基因组中包含一整套基因。相应的全部细胞质基因构成一个细胞质基因组，其中包括线粒体基因组和叶绿体基因组等。原核生物的基因组是一个单纯的 DNA 或 RNA 分子，因此又称为基因带，通常也称为它的染色体。基因在染色体上的位置称为座位，每个基因都有自己特定的座位。凡是在同源染色体上占据相同座位的基因都称为等位基因。在自然群体中往往有一种占多数的(因此常被视为正常的)等位基因，称为野生型基因；同一座位上的其他等位基因一般都直接或间接地由野生型基因突变产生，相对于野生型基因，称它们为突变型基因。在二倍体的细胞或个体内有两个同源染色体，所以每一个座位上有两个等位基因。如果这两个等位基因是相同的，那么就这个基因座位来讲，这种细胞或个体称为纯合体；如果这两个等位基因是不同的，就称为杂合体。在杂合体中，两个不同的等位基因往往只表现一个基因的性

状，这个基因称为显性基因，另一个基因则称为隐性基因。在二倍体的生物群体中等位基因往往不止两个，两个以上的等位基因称为复等位基因。不过有一部分早期认为是属于复等位基因的基因，实际上并不是真正的等位，而是在功能上密切相关、在位置上又邻接的几个基因，所以把它们另称为拟等位基因。某些表型效应差异极少的复等位基因的存在很容易被忽视，通过特殊的遗传学分析可以分辨出存在于野生群体中的几个等位基因。这种从性状上难以区分的复等位基因称为同等位基因。许多编码同工酶的基因也是同等位基因。

基因有两个特点，一是能忠实地复制自己，以保持生物的基本特征；二是基因能够"突变"，突变绝大多数会导致疾病，另外的一小部分是非致病突变。非致病突变给自然选择带来了原始材料，使生物可以在自然选择中被选择出最适合自然的个体

第一节 真核基因组

真核生物的基因组一般比较庞大，例如人类基因组由 3×10^9 个碱基对组成，按 1 000 个碱基编码一种蛋白质计，理论上可有 300 万个基因。但实际上，人细胞中所含基因总数不超过 4 万个。这就说明在人细胞基因组中有许多 DNA 序列并不转录成 mRNA 用于指导蛋白质的合成。DNA 的复性动力学研究发现这些非编码区往往都是一些大量的重复序列，这些重复序列或集中成簇，或分散在基因之间。在基因内部也有许多能转录但不翻译的间隔序列(内含子)。因此，在人细胞的整个基因组当中只有很少一部分(占 1%～2%)的 DNA 序列用以编码蛋白质。

一、真核生物基因组的特点：

真核生物与原核生物基因组有很大的差异，在基因结构、表达方式和过程都远比原核生物复杂。它的主要特点有：

1. 真核生物基因一般比较庞大，远远大于原核生物的基因组，约含 10 万个基因。且具有许多复制起点，而每个复制子的长度较小。

2. 真核生物基因组 DNA 与蛋白质结合形成染色体，储存于细胞核内。体细胞基因组是双倍体，即有两份同源的基因组。

3. 绝大多数真核生物编码蛋白质的基因为断裂基因(splitgene)，即结构基因是不连续排列的，由中间不编码的插入序列所隔开。编码序列称为外显子(exon)，编码序列中间的插入序列称为内含子(intron)，也称为间隔序列(intervening sequence)。

4. 真核生物基因组存在着许多重复序列，重复次数达几百万以上。

5. 真核生物基因组中不编码的区域多于编码区域，基因组中只有很小一部分是编码蛋白质的。如编码卵白蛋白的基因，内含子比外显子长得多，内含子占基因总量的 85%，不同生物所含内含子的数目、占基因总长度的比例和所处位置均可不同。

二、C 值矛盾

在真核生物中，一个物种单倍体基因组的 DNA 含量(bp)称为 C 值(C—Value)。C 值是相对恒定的，是每种生物的一个特性。不同物种的 C 值差异极大，最小的 C 值是支原体(mycoplasma)，小于 10^6 bp；最大的是某些显花植物和两栖动物，可达 10^{11} bp。随着生物的进化，生物体的结构和功能越复杂，其 C 值就越大，例如，真菌和高等植物同属于真核生物，而后者的 C 值就大得多。这一点是人们容易理解的，因为结构和功能越复杂，所需的蛋白质种类和基因就越多，因而 C 值就越大。但在结构和功能相似的同一类生物中，甚至在亲缘关系很近的物种之间，他们的 C 值差异仍可达 10 倍乃至上百倍。例如，在两栖类、被子植物的不同物种之间，其 C 值小的低于 10^9 bp，大的可达

10^{11} bp。特别是人类的C值只有 3.2×10^{9} bp，而肺鱼的C值则为 10^{11} bp，居然比人类高100倍。人们很难相信两栖类、肺鱼的结构和功能会比哺乳动物包括人类更复杂。另外，就哺乳动物来说，由于基因具有内含子，因而基因长达5 000～8 000 bp，少数的可达10 000 bp。按这样大小的基因进行推算，哺乳动物的基因组相当于 4×10^{5}～6×10^{6}个基因，但目前按各种方法估计的结果表明，哺乳动物编码蛋白质的基因总数在 3×10^{4}～1.2×10^{5}个之间，远远低于基因组应有的基因数。由此表明C值的大小并不能完全说明生物进化的程度和遗传复杂性的高低，即物种C值的大小与生物的进化程度之间不完全呈相关关系，这就是C值矛盾(C—value paradox)。

C值矛盾使人们认识到在真核生物基因组中存在着许多不编码蛋白质的DNA序列，而这些"额外"DNA具有什么功能？如果没有功能，为什么还会一代一代地遗传下来？这些问题都有待进一步的研究。

三、真核生物DNA序列的类型

真核生物染色体DNA中存在着许多重复序列(repeated sequence)，根据DNA序列出现频率的不同，可分为不同的类型。

(一)单拷贝序列(低度重复序列)

单拷贝序列也称单一拷贝序列(unique sequence)，在单倍体基因组中只出现一次或数次。单拷贝序列在基因组中占50%～80%，如人基因组中，大约有60%～65%的序列属于这一类。单拷贝序列中储存了巨大的遗传信息，编码各种不同功能的蛋白质。目前尚不清楚单拷贝基因的确切数字，但是在单拷贝序列中只有一小部分用来编码各种蛋白质，其他部分的功能尚不清楚。

在基因组中，单拷贝序列的两侧往往为散在分布的重复序列。由于某些单拷贝序列编码蛋白质，体现了生物的各种功能，因此对这些序列的研究对医学实践有特别重要的意义，由于结构基因的突变很容易造成遗传性状的改变或产生遗传性疾病。但由于其拷贝数少，在DNA重组技术出现以前，要分离和分析其结构和顺序几乎是不可能的，现在人们通过基因重组技术可以获得大量欲研究的基因，并对许多结构基因进行了较为细致的研究。

(二)中度重复顺序

中度重复序列(moderately repetitive sequence)指在真核基因组中重复数十至数万($<10^{5}$)次的重复顺序少数在基因组中成串排列在一个区域，大多数与单拷贝基因间隔排列。依据重复顺序的长度，中度重复顺序可分为两种类型。

(1)短分散片段(short interspersed repeated segments，SINES)这类重复顺序的平均长度为300～500 bp，它们与平均长度约为1 000 bp的单拷贝顺序间隔排列。拷贝数可达10万左右。如Alu家族，Hinf家族等属于这种类型的中度重复序列。

(2)长分散片段(Long interspersed repeated segments，LINES)这类重复顺序的长度大于1 000 bp，平均长度为3 500～5 000 bp，它们与平均长度为13 000 bp(个别长几万bp)的单拷贝顺序间隔排列。也有的实验显示人基因组中所有LINES之间的平均距离为2.2 kb，拷贝数一般在1万左右，如KpnⅠ家族等。中度重复顺序在基因组中所占比例在不同种属之间差异很大，一般约占10%～40%，人类约为12%。这些顺序大多不编码蛋白质。这些非编码的中度重复顺序的功能可能类似于高度重复顺序。下面介绍几种典型的中度重复顺序。

Alu家族(Alu family)：Alu家族是哺乳动物包括人基因组中含量最丰富的一种中度重复顺序家族，约占人基因组的3%～6%，在单倍体人基因组中重复达30万～50万次。Alu序列长度约300 bp，由于每个单位170位碱基处有一个限制性内切酶Alu的切点(AG↓CT)从而将其切成长130 bp和170 bp的两段，因而定名为Alu序列(或Alu家族)。Alu序列分散在人体或其

他哺乳动物基因组中，在间隔 DNA，内含子中都发现有 Alu 序列，平均每 5 kb DNA 就有一个 Alu 序列。已建立的基因组中无例外地含有 Alu 序列。Alu 序列具有种属特异性，人 Alu 序列制备的探针只能用于检测人基因组中的 Alu 序列。有人认为 Alu 序列两侧存在着短的重复顺序，使得 Alu 序列很像转座子，因此推测 Alu 序列可能也是能够移动的。这可能是它们在整个基因组中含量如此丰富，分布如此广泛的原因之一。Alu 家族的功能是多方面的，由于在许多核内不均一 RNA(hnRNA)中含有大量的 Alu 序列，而且，Alu 序列含有与某些真核基因内含子剪接接头相似的序列，因而，Alu 序列可能参与 hnRNA 的加工与成熟。Alu 序列在人基因组中大量地存在，提示它与遗传重组及染色体不稳定性有关。另外，Alu 序列可参与附近基因的活化，包括开启或关闭基因、促进或终止转录，DNA 复制的起始等。他们的功能还有待于进一步证实。

Kpn Ⅰ家族(kpn Ⅰ family)：Kpn Ⅰ家族是中度重复顺序中仅次于 Alu 家族的第二大家族。用限制性内切酶 Kpn Ⅰ消化人类及灵长类动物的 DNA，在电泳谱上可以看到 4 个不同长度的片段，分别为 1.2，1.5，1.8 和 1.9 kb。Kpn Ⅰ家族成员顺序比 Alu 家族更长，而且更加不均一，呈散在分布，属于中度重复顺序的长分散片段型。尽管不同长度类型的 Kpn Ⅰ家族(称为亚类，subfamily)之间同源性比较小，不能互相杂交，但它们的 3′端有广泛的同源性。Kpn Ⅰ家族的拷贝数约为 3 000～4 000 个，占人体基因组的 1%，Kpn Ⅰ家族中有一部分序列是其 RNA 转录产物的 cDNA 拷贝插入到基因组 DNA 中产生的。

Hinf 家族(Hinf family)：这一家族以 319 bp 长度的串联重复存在于基因组中。用限制性内切酶 Hinf Ⅰ消化 DNA，可以分离到这一片段。Hinf 家族在单位基因组内约有 50～100 个拷贝，分散在不同的区域。319 bp 单位可以再分成两个亚单位，分别为 172 bp 和 147 bp，它们之间有 70%的同源性。

多聚 dT－dG 家族(poly dT－dG family)：这一家族的基本单位是 dT－dG 双核苷酸，多个 dT－dG 双核苷酸串联重复在一起，分散于人体基因组中。已经发现，这个家族的一个成员位于人类 δ 和 β 珠蛋白基因之间，含有 17 个 dT－dG 双核苷酸组成的串联重复顺序。在人基因组中，dT－dG 交替顺序达 10^6 拷贝，这些顺序的平均长度为 40 bp。人们推测，这样一个短的串联重复顺序可能是基因转变(gene conversion)或不等交换(unequal crossing－over)的识别信号。另外，这些嘌呤和嘧啶的交替顺序有助于 Z－DNA 的形成，在基因调节中可能起着重要的作用。中度重复顺序除了包括以上非编码区域外，许多编码区如 rRNA 基因，tRNA 基因，组蛋白基因等在基因组中也多次重复，属于中度重复顺序。

rRNA 基因：在真核生物基因组中 18S 和 28S rRNA 基因是在同一转录单位中，在高等生物中，5S rRNA 是单独转录的，而且其在基因组中的重复次数高于 18S 和 28S 基因。和一般的中度重复顺序不一样，各重复单位中的 rRNA 基因都是相同的。rRNA 基因通常集中成簇存在，而不是分散于基因组中，这样的区域称为 rDNA，如染色体的核仁组织区(nucleolus organizer region)即为 rDNA 区。18S 和 28S rRNA 基因构成一个转录单位。从转录单位上转录下来的 rRNA 前体经过酶切成为 18S 和 28S rRNA。(见图 1－1)

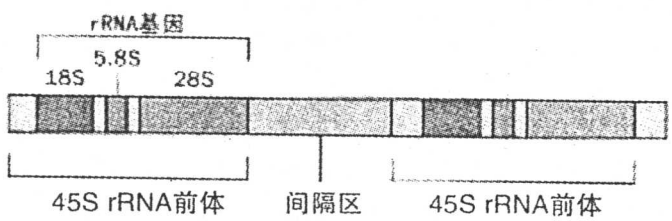


图 1－1 真核生物 rRNA 基因结构

人类的 rRNA 基因位于 13，14，15，21 和 22 号染色体的核仁组织区，每个核仁组织区平均含有 50 个 rRNA 基因的重复单位。5S rRNA 基因似乎全部位于 1 号染色体(1q42

—43)上，每单倍体基因组约有 1 000 个 5S rRNA 基因。tRNA 基因的确重复次数难以估计。在非洲爪蟾中约有 300 个拷贝由 tRNAmet，tRNAphe，tRNATrp 及其它 tRNA 基因组成的 3.18 kb 的串联重复单位。而在人体单倍基因组中约有 1 000～2 000 个 tRNA 基因，为 50～60 种 tRNA 编码，每种平均重复 20～30 次。

组蛋白基因：组蛋白基因在各种生物体内重复的次数不一样，但都在中度重复的范围内。通常每种组蛋白的基因在同一种生物中拷贝数是相同的。鸡的基因组中组蛋白基因有 10 个拷贝，在哺乳动物中为 20 拷贝，非洲爪蟾为 40 拷贝，而海胆的每种组蛋白的基因达 300～600 拷贝。不同生物中组蛋白基因在基因组中的排列不一样，组蛋白基因没有一定的排列方式，而在拷贝数高的基因组中(>100 拷贝)，大部分组蛋白基因串联重复形成基因簇。

基因组中存在大量重复序列用以编码组蛋白是有其重要意义的。DNA 复制时，组蛋白也要成倍增加，而且往往在 DNA 合成一小段后，组蛋白马上就要与其相结合，这要求在较短的时间内合成大量的组蛋白，因而需要有大量的组蛋白基因存在。人体基因组中还有几个大的基因簇，也属于中度重复顺序长的分散片段型。在一个基因簇内含有几百个功能相关的基因，这些基因簇又称为超基因(Supergene)，如人类主要组织相容性抗原复合体 HLA 和免疫球蛋白重链及轻链基因都属于超基因。超基因可能是由于基因扩增后又经过功能和结上的轻微改变而产生的，但仍保留了原始基因的结构及功能的完整性。

(三)高度重复序列

高度重复序列在基因组中重复频率高，可达百万(10^6)以上，因此复性速度很快。在基因组中所占比例随种属而异，占 10%～60%，在人基因组中约占 20%。高度重复顺序又按其结构特点分为三种。

(1)卫星 DNA(satellite DNA)

卫星 DNA(satelliteDNA)是一类高度重复序列，这类重复顺序的重复单位一般由 2～10 bp 组成，成串排列。由于这类序列的碱基组成不同于其他部分，可用等密度梯度离心法将其与主体 DNA 分开，因而称为卫星 DNA 或随体 DNA。在人细胞组中卫星 DNA 占 5%～6%。按照它们的浮力密度不同，人的卫星 DNA 可分为Ⅰ，Ⅱ，Ⅲ，Ⅳ四种。卫星 DNA 不是编码蛋白质或 RNA 的，它在染色体上的位置可用原位杂交方法进行鉴定，大部分高度重复 DNA 在真核细胞染色体的两个重要结构——着丝粒(centromere)和端粒(telomere)处。

每个染色体只有一个着丝粒，它是蛋白质的附着点，这些蛋白质把染色体和有丝分裂纺锤体的微管连接起来，对细胞分裂事使染色体有序地分配到子代细胞非常重要。真核生物中，卫星 DNA 一般都存在于着丝粒区，由一种或几种数以千计的短序列拷贝串联而成。典型的卫星 DNA 一般序列为 5～10 bp。卫星 DNA 在着丝粒中的确切功能尚不明了。人类染色体的着丝粒区有一类卫星 DNA，它的最小重复单位是 171 bp，在着丝粒区串联高度重复序列。它既有染色体特异性，又具有染色体间的同源性，在人群中还具有高度的多态性。

端粒是染色体的物理末端，人的端粒是由串联的 TTAGGG 重复序列组成。它对染色体起保护作用，防止染色体的重排、断裂等。端粒的长度随细胞分裂的不断进行而缩短，它决定着细胞的寿命，故称之为有丝分裂的生物钟。

在人和动物基因组中存在着许多串联重复排列的 DNA 序列，是由 2～70 bp 重复几次，几十次串联而成，构成数量可变的串联重复序列(variable number of tandem repeat，VNTR)。其中重复顺序为 6～70 bp 的称小卫星 DNA(minisatallite DNA)。在基因组的间隔序列和内含子等非编码区内，广泛存在着另一类 1～6 bp 的短小重复单位，称为微卫星 DNA(microsatellite

DNA)如 GACA,GATA,TCC,CT,CA 等。微卫星 DNA 又称简单重复序列,它在人基因组中出现的数目和频率不同,在基因组中分布广泛而表现出多态性,因此这些分散排列的简单重复顺序是进行限制性片段长度多态性(RFLP)分析的方便、有效的遗传标记,可在基因连锁诊断中广泛应用。

(2)倒位(反向)重复序列(inverted repeat)

反向重复序列由两个相同顺序的互补拷贝在同一 DNA 链上反向排列而成。变性后再复性时,同一条链内的互补的拷贝可以形成链内碱基配对,形成发夹式或"+"字形结构。倒位重复(即两个互补拷贝)间可有一到几个核苷酸的间隔,也可以没有间隔。没有间隔的又称回文(palimdrome),这种结构约占所有倒位重复的 1/3。若以两个互补拷贝组成的倒位重复为一个单位,则倒位重复的单位约长 300 bp 或略少。两个单位之间有一平均 1.6 kb 的片段相隔,两对倒位重复单位之间的平均距离约 12 kb,反向重复序列常出现在基因的调控区及某些蛋白质结合的区域,具体功能不清,可能与复制、转录的调控有关。

(3)高度重复顺序的功能

有关高度重复序列的功能目前尚有争论.一般认为它有几方面的功能:

a.参与复制水平的调节。反向序列常存在于 DNA 复制起点区的附近。另外,许多反向重复序列是一些蛋白质(包括酶)和 DNA 的结合位点。

b.参与基因表达的调控。DNA 的重复顺序可以转录到核内不均一 RNA 分子中,而有些反向重复顺序可以形成发夹结构,这对稳定 RNA 分子,免遭分解有重要作用。

c.参与转位作用。几乎所有转位因子的末端都包括反向重复顺序,长度由几个 bp 到 1 400 bp。由于这种顺序可以形成回文结构,因此在转位作用中即能连接非同源的基因,又可以被参与转位的特异酶所识别。

d.与进化有关。不同种属的高度重复顺序的核苷酸序列不同,具有种属特异性,但相近种属又有相似性。如人的 α 卫星 DNA 长度仅差 1 个碱基(前者为 171 bp,后者为 172 bp),而且碱基序列有 65%是相同的,这表明它们来自共同的祖先。在进化中某些特殊区段保守的,而其他区域的碱基序列则累积着变化。高度重复序列促进基因组的重新组织以及促进新物种的形成。

e.同一种属中不同个体的高度重复顺序的重复次数不一样,这可以作为每一个体的特征,即 DNA 指纹。

f.α 卫星 DNA 成簇的分布在染色体着丝粒附近,可能与染色体减数分裂时染色体配对有关,即同源染色体之间的联会可能依赖于具有染色体专一性的特定卫星 DNA 顺序。

四、多基因家族与假基因

真核基因组中许多来源相同、结构相似、功能相关的基因在染色体上成串存在,这样的一组基因成为基因家族(gene family)。多基因家族(multigene family)是真核生物基因组的一个重要特征。多基因家族是指由某一祖先基因经过重复和变异所产生的一组基因。此家族可分为两类:一类是基因家族串联排列在某一条染色体上,集中成簇的一组基因形成基因簇(gene cluster)。它们可同时发挥作用,合成某些蛋白质,如组蛋白基因家族就成簇地集中在第 7 号染色体长臂 3 区 2 带到 3 区 6 带区域内;另一类是一个基因家族的不同成员成簇地分布在不同染色体上,这些不同成员编码一组功能上紧密相关的蛋白质,如珠蛋白基因家族。

(一)假基因

在多基因家族中,某些成员并不产生有功能的基因产物,这些基因称为假基因(pseudoGene)。假基因与有功能的基因同源,原来可能也是有功能的基因,但由于缺失,倒位或点突变

等，使这一基因失去活性，成为无功能基因。与相应的正常基因相比，假基因往往缺少正常基因的内含子，两侧有顺向重复序列。人们推测，假基因可能是基因经过转录生成的 RNA 通过剪接失去内含子形成 mRNA，如果 mRNA 经反复转录产生 cDNA，再整合到染色体 DNA 中去，便有可能成为假基因，因此该假基因是没有内含子的，在这个过程中，可能同时会发生缺失，倒位或点突变等变化，从而使假基因不能表达。

（二）典型的多基因族

1. tRNA 基因

真核细胞可有几百个到一千多个 tRNA 基因。人类约有 1 300 个 tRNA 基因。不同种类的 tRNA 转运不同的氨基酸。同种 tRNA 串联在一起形成基因簇，基因间由不转录的间隔区隔开。

2. rRNA 基因

人的 rRNA 基因簇有 50～200 个。实验证明 rRNA 基因是分布在几个染色体中。次转录单位转录下来的初级转录产物是 45S RNA，经酶切后形成成熟的 18S，5. 8S，28S RNA。这三个基因顺序串联排列在同一转录单位中，在这些基因之间有短的间隔将它们分开，形成一个约 7 500 bp的转录单位。

3. 组蛋白基因

编码 H1，H2A，H2B，H3，H4 和 H5 这 6 种组蛋白基因串联排列在一起形成一个重复单位，许多这样的重复单位串联在一起，构成组蛋白的基因簇。人类组蛋白基因拷贝数有 30～40 个，分布在 7 号染色体长臂区。组蛋白基因的特点是没有内含子结构，转录后的 mRNA 也没有 polyA 尾巴。

4. 珠蛋白基因家族

珠蛋白肽链是血红蛋白的亚基，分为两类，α 和 β 类型。人类的 α 基因在第 16 号染色体短臂 p13. 3 区，由三个基因组成；β 类基因位于第 11 号染色体内，由 5 个基因组成。珠蛋白基因在红细胞中表达，在胚胎早期，血红蛋白为 $\xi_2\varepsilon_2$，$\xi_2\gamma_2$，$\alpha_2\gamma_2$ 等类型，胎儿时期为 $\alpha_2\gamma_2$ 类型。成体阶段为 $\alpha_2\delta_2$ 和 $\alpha_2\beta_2$ 类型。所以，α 基因家族的表达为 $\xi\rightarrow\alpha$。β 基因家族的表达为 $\varepsilon\rightarrow\gamma\rightarrow\delta\rightarrow\beta$。在个体发育中，珠蛋白基因逐个开启和关闭，具有发育阶段特异性。

5. 生长激素基因家族

人类生长激素基因家族包括 3 种激素基因，人生长激素（hCg）基因、人胎盘促乳素（hCS）基因和催乳素（OT）基因。它们之间的同源性很高，hCg 和 hCS 之间的氨基酸有 85%同源，说明他们来源于同一祖先基因。但三个基因并不排列在一起，hCg 和 hCS 基因位于第 17 号染色体长臂，而 OT 位于第 6 号染色体。

6. 超基因

超基因（supergene）是指一组由多基因和单基因组成的更大的基因家族。在高等真核细胞中，一个基因簇内含有数百个功能相关基因，他们可能是由基因扩增后结构上轻微变化而产生的。这些基因在结构上有程度不等的同源性，功能上仍保持原始基因的基本功能，或者进化成具有相关而不同的新功能，这样的一簇基因称为超基因家族。目前发现的超基因家族很多，最经典的是免疫球蛋白超基因家族、核受体超基因家族和细胞因子超基因家族等。

（三）自私 DNA（selfish DNA）

在哺乳动物包括人体基因组中，存在着大量的非编码顺序，如前述的高度重复顺序，内含子，间隔 DNA 等。这些顺序中，只有少部分具有重要调节功能，绝大部分没有特殊功用。这些 DNA 序列积累了大量缺失，重复或其他突变，但对生物并没有影响，它们的功能似乎只是自身复

制，人们称这类 DNA 为自私 DNA 或寄生 DNA(parasite DNA)。自私 DNA 可能有重要的功能，但目前我们还不了解。

第二节 真核结构基因

基因是遗传的基本单位，研究发现，真核生物在基因结构上，表达方式和过程等方面都远比原核生物复杂，DNA 的含量也比原核生物的大得多。噬菌体基因组很小，但又要编码一些必不可少的蛋白，碱基显然不够用，这样几乎所有的碱基都参加编码，而且在进化中出现了“重叠基因”，以有限的基因编码更多的遗传信息。真核基因组正好相反，DNA 十分富余，不仅无需“重叠基因”，而且很多序列不编码，如重复序列、间隔序列(spacer) 和间插序列(intervening sequence) 即内含子(intron)等。但不编码并不等于没有功能。有的我们可能还不了解，如重复序列。间隔区和间插序列这两个概念是不同的，间隔区是指基因之间不编码的部分，但有的转录称转录间隔区(TS)，有的不转录称为非转录间隔区(NTS)。间插序列是指基因内部不编码的区域，也称内含子，在初始转录本中存在此序列，但在加工后将被切除掉，所以常不作为翻译的信息。间隔区常常含有转录的启动子和其他上游调节序列。有的内含子也可以编码，如成熟酶和内切酶等。

一、真核生物的断裂基因

在 20 世纪 70 年代前，人们一直认为遗传物质是双链 DNA，在上面排列的基因是连续的。Robert and Sharp 彻底改变了这一观念。他们以腺病毒作为实验对象，因为它的排列序列同其他高等动物很接近，包括人。结果发现腺病毒基因在 DNA 上的排列由一些不相关的片段隔开，是不连续的。他们的发现改变了科学家以往对进化的认识，对于现代生物学以及生物进化论的研究具有重要的奠基作用，对于肿瘤以及其他遗传性疾病的医学导向研究，亦具有特别重要的意义。

通常将能编码蛋白质的基因称为结构基因。真核生物的结构基因是断裂的基因。断裂基因又称不连续基因，是指蛋白质的基因中间存在一些不能表达为蛋白质肽链的核苷酸序列，一个结构基因分成或断裂成若干部分，这样的基因叫做断裂基因。一个断裂基因能够含有若干段编码序列，这些可以编码的序列称为外显子。在两个外显子之间被一段不编码的间隔序列隔开，这些间隔序列称为内含子。不连续基因具有外显子和内含子交替排列的结构，在初始转录产物 hnRNA 加工产生成熟的 mRNA 时，被切除的非编码序列称为内含子(intron)，在成熟的 mRNA 或蛋白质中存在的序列称为外显子(exon)。基因的不连续性是真核基因所特有的。但不是所有真核基因都一定具有这种不连续性。每个断裂基因在第一个和最后一个外显子的外侧各有一段非编码区，有人称其为侧翼序列。在侧翼序列上有一系列调控序列。

二、真核基因的外显子与内含子

1. 外显子与内含子的关系

真核生物中的断裂基因由一系列外显子和内含子交替组成。但在成熟的 mRNA 分子中，只有外显子序列，内含子序列在 mRNA 成熟或翻译之前已被切除。切除内含子的过程称为 RNA 剪接。一般来说，剪接包括从初始转录产物中去掉内含子，然后切点两侧的 RNA 末端重新连接，形成完整的外显子。

生物的不同基因拥有内含子的数量和大小相差非常悬殊，且具有一定的物种特异性。例如，胶原蛋白基因，长约 40 kb，至少具有 52 个内含子，其中短的只有 50 bp，长的可达到 2 000 bp。卵清蛋白基因含有 7 个内含子。β—珠蛋白基因的 α 和 β 基因家族之间具有共同的结构特征，即

都具有两个内含子，而且在家鼠、兔、人、鸡、绵羊、山羊及非洲爪蟾等生物中，β－珠蛋白基因的内部结构具有一定的相似性；但第二个内含子的长度在上述物种中有较大的变异。少数基因，如组蛋白和α型、β型干扰素基因，根本不含内含子。

2. 外显子与内含子的连接区

断裂基因的一个重要特点是内含子与外显子连接区(intron－exon junction)的高度特异性和保守性。通过比较分析人、羊、鼠、兔、鸡等生物的部分基因的基因组与mRNA的序列，发现连接区有两个重要特征：①各种内含子的两端序列之间没有广泛的同源性和互补性；这说明在RNA剪接之前，内含子上游序列和下游序列不可能通过碱基配对形成二级结构。②内含子与外显子连接区是高度保守的序列，尽管很短，但却似乎是剪接反应所必需的。每个内含子5′端和3′端的碱基都有很强的规律性，通常有1～4 bp的碱基序列在内含子的每个末端是重复的，而且几乎每个内含子5′端起始的两个碱基都为gT，3′端最后两个碱基为Ag。这种交感序列的规律性，即5′GT……AG3′，称为GT－AG法则。这种内含子的共同碱基序列在真核生物的许多断裂基因里都可以发现，说明这类基因剪接成mRNA遵照一种共同的机制。

在断裂基因中，内含子与外显子的关系并不是完全固定不变的，有时会出现这样的情况，即某基因的一条mRNA链的内含子可以是该基因另一条mRNA链的外显子，而表现出内含子与外显子的重叠。其结果是同一段DNA序列可以加工生成两条或两条以上的mRNA链。这是断裂基因的另一个重要特点。

小鼠淀粉酶蛋白可以同时在肝脏和唾液腺中合成。它仍由同一基因编码，但基因的启动子在肝脏和腮腺中不同，使肝脏和腮腺的mRNA从不同的外显子开始转录，具有不同的5′端，小鼠也就产生了两种淀粉酶的mRNA。肝脏中mRNA的5′端最初161个核苷酸由位于外显子2上游约4 500 bp处的外显子L编码；腮腺中mRNA的5′端最初50个核苷酸则由位于外显子2上游约7 300 bp处的外显子S编码。外显子S和L分别提供了mRNA的开始序列。外显子L是腮腺里表达的基因中很长内含子的一小部分，这部分内含子序列在mRNA前体剪接时被剪去。

一个基因通过不同方式剪接，扩大了内含子存在的意义。通过一个外显子替代另外一个外显子，改变了氨基酸序列，产生了有不同功能的相关蛋白质，使基因成为一个复杂的转录单位，转录和加工出丰富多彩的蛋白质产物，以适合细胞、组织和发育特异性的需要。

3. 内含子的功能

在高等真核生物中内含子普遍存在。在断裂基因中内含子的数目可以多得惊人，其长度在50～20 000 bp以上范围内，而外显子的长度均在100～250 bp范围内，在一个典型的脊椎动物的结构基因中外显子的序列仅占20%左右。目前对内含子的功能还没明确的认识。当然不能说内含子在生物学功能上就不重要，如果失去内含子，真核基因将失去活性。但目前研究结果表明大多数基因的内含子确实又没有什么明显的功能。Gilbert提出的关于内含子功能的假说，认为结构基因是通过内含子序列之间的重组，将外显子聚集在一起而产生的，即内含子是原初基因重新组合过程的残留物。这一假说得到了许多实验证据的支持。

内含子除了使新基因的进化易于发生外，就目前人们的认识来看，还具有以下功能：①影响基因的表达调控。有证据表明鼠B细胞κ链基因内含子的增强子序列通过诱导去甲基化和促进组织专一转录而调节B细胞的分化；免疫球蛋白K基因的内含子有调控体细胞超突变的作用。在许多动物的血红蛋白基因中，内含子通过启动子、起始位点的精确碱基配对，来阻止或增强RNA聚合酶的作用，从而调控该基因的转录表达。有些基因的表达要求内含子的一些序列存在，例如人的apoB基因在转基因小鼠中的高水平、专一性表达，除了需要增强子之外，还需要一

些内含子的一定序列。②调控 RNA 的剪接。内含子具有各种剪接信号，不同细胞能选择不同的剪接点，将 mRNA 前体进行不同的加工，对外显子进行有选择的拼接，最终产生不同的成熟 mRNA 分子。③含有可阅读框架。内含子的可阅读框架可能编码酶或者蛋白质，已有报道某些内含子的可读框架(open reading frame，ORF)可以表达为成熟酶、逆转录酶和核酸内切酶等。蛋白质的有些可阅读框架可能独自存在于内含子内部，也可能与上游外显子合框。④保护基因家族。在基因家族中，内含子可以保护基因家族中相邻近的外显子不被不等交换(unequal crossing over)所消除。

4. 断裂基因的变异

用 DNA 分子杂交检测重复基因的内含子和外显子时，发现在不同的基因中，内含子之间的关系比外显子之间的关系差异更大。内含子序列几乎都不是重复的，不具有同源性；而外显子序列在两个重复基因之间却保持很大的同源性。这可能是外显子受相应编码蛋白质的氨基酸序列或者有各种功能的 RNA 的制约而表现进化上的保守性；而内含子不编码蛋白质不受此限制，可随机积累更多的变异，在进化上表现得更迅速。

内含子的变异包括由缺失、插入、易位等原因造成的碱基序列的变异。这些变异表明，内含子不具有特异的功能，不产生特异的影响，因此，可以随机自由地积累各种变异。据研究表明，内含子和外显子发生碱基序列变异的频率相等，只不过在外显子中这些变异可通过逆向选择而消除，得不到保留；相反内含子可以完全保留各种变异的结果。许多实验结果证明，内含子与外显子连接区的碱基序列发生改变，将会影响正常的剪接方式而使该基因所编码的蛋白质结构发生显著的变异。

三、跳跃基因

跳跃基因(jumping gene)又称为移动基因(movable gene)或转座子(transposon)。它可以从染色体基因组上的一个位置转移到同一条染色体或另一条染色体的另一个位置。这种基因是生物基因组中的一些特殊 DNA 序列，其长度变异很大，两端带有一段重复序列。

跳跃基因最早是由美国女科学家 McClintock 于 20 世纪 40 年代在玉米中发现的。当时称为控制基因(controlling gene)。她在研究玉米性状的遗传时注意到玉米粒的颜色会经常改变，于是提出了一种假说，认为是由一种控制基因在玉米的基因组中移动的结果。根据她的研究，这种控制基因可以插入到玉米染色体上编码色素的基因中，并改变色素基因的表达活性，而使玉米粒颜色发生变化。但这种基因插入的位置会不断地改变，似乎可以沿着染色体分子移动，所以造成玉米粒的颜色成斑驳状。控制基因在插入玉米染色体之后的适当时间内又可以被重新删除，此时被影响基因的功能也会恢复。

转座子在生物界中广泛存在，所有转座子的分子结构均具有一些共同的特征：

①在转座子的两端有 20～40 bp 的末端反向重复序列(ter minal inverted repetitive sequence，TIR)，即一个末端序列如果为 AGCT，那么另一个末端序列为 TCGA。如果缺失任何一端，都会妨碍转座功能。

②绝大多数转座子含有开放阅读框架(ORF)，它可能编码转座酶，这种酶催化转座子插入新的位置。

③由于转座子的插入，原有 DNA 上的靶序列在转座子的两侧形成正向重复序列。序列的长短对转座子是特异的。

四、假基因

假基因(pseudo gene)具有与功能基因相似的序列，但由于有许多突变以致失去了原有的功

能，所以假基因是没有功能的基因，常用 ψ 表示。1977 年在爪蟾的 5S 基因系统中发现了假基因，以后在珠蛋白基因簇、免疫球蛋白基因簇以及组织相容性抗原基因簇中都发现有假基因，而且通常是散布于有活性的功能基因之间。关于假基因的来源一般认为是由 mRNA 反转录成 cDNA，然后整合在基因组中。假基因同 cDNA 一样没有内含子序列，也没有启动基因转录的启动子序列，而在 5′端都有 mRNA 分子特有的多聚腺苷[poly(A)]序列。由于假基因没有生物学功能，所以不再受到进化的选择压力，因此在假基因中可以积累许多突变，并常常同时存在三种终止密码子序列。假基因是由功能基因演变而来，可以看做是进化的一种遗迹。假基因是一种非自主的反转录转座元件。由于它来源于 RNA 聚合酶Ⅱ的转录产物，因此假基因必定是没有转录活性的。这是因为反转录产生假基因的 RNA 不包含位于转录起点上游的启动子序列。假基因通常具有 mRNA 的特征，所以称为已加工的假基因(processed pseudo gene)。假基因不含内含子序列，末端有很短一段 A－T 碱基对，两端又各有一个短的正向重复序列，在染色体上的位置与其原先所在位置无关等，这些结构特征都说明了假基因的反转录转座起源假说。当然，目前对此也还存在不同的看法，主要是由于假基因本身并不编码反转录酶，也不产生转座酶等。因此，有人提出，假基因的出现可能是以反转录病毒为中介的转座过程，被转座序列的末端只是偶然地同转座子的末端相似而已。

在真核生物中，常见的假基因主要有两类：

①重复的假基因　这类假基因是已有的重复基因在结构上发生较大的变化而失去了功能后形成的。其特点是在假基因两侧有同向重复序列。例如，位于小鼠第 11 号染色体上的 α 珠蛋白基因簇中的 α－ad1 和 α－ad2 基因，只是所编码蛋白质的第 68 位上一个氨基酸的差异。

②加工的假基因　这类假基因都没有启动子和内含子，但在基因的 3′端都有一段延伸的短 AT 碱基对序列，恰似 mRNA 分子 3′－末端的 poly (A)尾巴。根据这些结构上的特点，一般认为此类假基因可能是来源于加工的 RNA 之 DNA 拷贝，故称为加工的假基因。例如，加工假基因 Mt－IIB 是人金属硫蛋白功能基因 Mt－IIA 的 mRNA 的准确拷贝。它是从帽位点开始的，但已失去了相应的内含子序列，并终止在 poly(A)位点。同时在假基因的两侧还有 21 bp 的同向重复序列。但由于无启动子，因此，加工的假基因是无转录活性的。

第三节　人类基因组计划

一、人类基因组计划的诞生及意义

人类基因组计划(Human Genome Project，HGP)是人类生命科学史上最伟大的工程之一，是人类第一次系统、全面地解读和研究人类遗传物质 DNA 的全球性合作计划。人类基因组是指合成有功能的人体各类细胞中蛋白质及(或)多肽链和 RNA 所必须的全部 DNA 顺序和结构，包括人类的 23 对染色体上全部的 DNA 所携带的遗传信息的总和，即 30 亿个碱基对的序列，最初估计含约 10 万个基因。

1985 年 5 月，美国能源部提出“人类基因组计划”草案；经过一番讨论后于 1986 年 3 月宣布实施这个草案；1986 年 3 月 7 日，Dulbecco R 在 Science 上发表了一篇有关开展人类基因组计划的短文，引起了全世界的强烈反响，不仅推动了美国，也推动了全世界的人类基因组计划的发展；美国国会正式批准的“人类基因组计划”到 1990 年 10 月 1 日才正式启动，其规模在世界上是最大的，计划在 15 年内投入 30 亿美元以上的资金进行人类基因组的分析。其最初的目标是，通过国际合作，用 15 年时间(1990～2005)，构建详细的人类基因组遗传图和物理图，确定人类 DNA 的全部核苷酸序