18

基因组学与后基因组学

由于人类基因组计划(**Human Genome Project**,**HGP**)和模式生物基因组计划的提出与实施,导致产生了一个新的学科——基因组学(**genomics**)。它是研究基因组的组成、结构和功能的学科,分为结构基因组学(**structural genomics**)和功能基因组学(**functional genomics**)。结构基因组学是着重研究基因组的结构并构建高分辨的遗传图、物理图、序列图和转录图以及研究蛋白质组成与结构的学科;功能基因组学主要是利用结构基因组学研究所得到的各种信息在基因组水平上研究编码序列及非编码序列生物学功能的学科。利用各种分子标记构建的遗传图和物理图是建立基因组全序列整合图的基础。细胞遗传学图谱、辐射杂种图谱、限制酶切图谱和叠连群图谱等是一系列的物理图谱,这些图谱的整合是构建基因组框架图的基础。基因组 **DNA** 大片段文库的构建是建立基因组图的首项工作。对基因组图谱构建的根本目的是为了应用。利用基因组图谱我们可以寻找新的基因、克隆分离基因、定位基因、对基因功能进行预测、开展比较基因组学的研究等等。由于人类基因组序列图和一些模式生物的全基因组序列图的完成,生命科学研究进入了后基因组时代,即从整体水平对生物进行功能研究,从而导致了蛋白质组学的诞生。

18.1 人类基因组计划与基因组学

18.1.1 人类基因组计划

人类基因组计划是堪称与阿波罗登月计划和曼哈顿原子弹计划相比的惊世壮举，是当代生命科学中的一项伟大的科学工程。1986年美国能源部正式提出开展人类基因组的测序工作，并提出了“人类基因组计划”草案。1986年 3月诺贝尔奖获得者，美国著名肿瘤分子生物学家 Dulbecco在 Science上撰文，强调弄清人基因组序列，搞清基因组中各基因的结构和功能及其相互关系，寻找预测、预防和早期诊断人类遗传病的新方法，将有助于解决包括癌症在内的人类疾病的发病原因这一美好前景的到来。嗣后经过学术界的一番激烈争论和反复论证，美国由能源部和国立卫生研究院（NIH）合作于 1990年正式启动人类基因组计划。主要目标是计划拨款 30亿美元，用 15年时间完成人类基因组 30亿 bp全部序列的测定，在 2001年完成全部染色体的“工作草图”。

经过参与该项目的 1 000多名各国科学家的通力合作，人类基因组的工作草图已经在 2000年 6月 26日胜利绘制完成，该工作草图包含人体 90%以上碱基对的位置信息。2001年 2月 12日中、美、日、德、法、英等 6国科学家和美国 Celera公司联合公布了人类基因组图谱及初步分析结果，人类基因组由 31 647亿个碱基对组成，有 3万～3 5万个基因，远小于原先预计的 10万个基因的估计。2003年 4月 15日，上述 6国又共同宣布人类基因组序列图完成。2004年 10月，国际人类基因组测序联合体在 Nature周刊上发表了人类基因组常染色质全序列测定的论文，宣布人类基因组的常染色质部分中 99%的序列已经被测定，其精度达到每 10万个碱基中只有一个测量误差。随着人类基因组精细图的完成，研究者发现，人类基因组拥有的编码蛋白质的基因数目在 2万到 2 5万个之间，比“工作草图”的估计的基因数又低 33%。

18.1.2 人类基因组的结构特点

人类基因组是第一个被测序的脊椎动物基因组，人类基因组大小约为 3 2× 10^9 bp（3 200 Mb），其中基因和基因相关序列约为 1 200 Mb，基因间 DNA序列约为 2 000 Mb。在基因和基因相关序列中为基因编码的序列约为 48 Mb，占总基因组序列的 1 5%左右，而基因相关序列约为 1 152 Mb，占总基因组的 36%，其中包括假基因、基因片段和内含子以及非翻译区（untranslated region，UTR）；在基因间 DNA序列中散在重复序列（interspersed repeat sequences，IRS）为 1 400 Mb，占总基因组的 43 75%，包括 64 Mb的长散在核元件（LINE）、420 Mb的短散在核元件（SINE）、250 Mb的长末端重复序列（long terminal repeat，LTR）和 90 Mb的 DNA转座子。在基因间 DNA序列还含有 600 Mb的其他的基因间区域序列，包括 90 Mb的微卫星序列和 510 Mb的各种序列成分（图 18- 1）。

显然，编码基因的序列仅占人类基因组 DNA的 1%左右，98%以上的序列是非编码序列。基因中内含子的序列占基因组的 24%，因此基因组涉及与产生蛋白质有关的序列达到 25%。基因平均长 27 kb，平均具有 9个外显子，一个基因约由 1 340 bp组成编码序列，因此平均在一个基因内的编码序列仅仅只占一个基因序列碱基长度的 5%。而人类基因组 DNA中重复序列占 50%以上，主要分成 5种类型：① 转座子成分，包括有活性的和无活性的，占基因组的 45%。均以多拷贝的形式存在于基因组中。② 已加工假基因（processed pseudogene），这是一类与 RNA转录物相似的失活基因，约 3 000个，约占基因组的 0 1%。③ 简单重复序列，约占基因组的 3%。④ 大片段重复（长 10～30 kb的大片段）占基因组约 5%。只有少部分在相同的染色体上，多数分布在不同的染色体上。⑤ 串联重复，主要位于着丝粒和端粒部位。

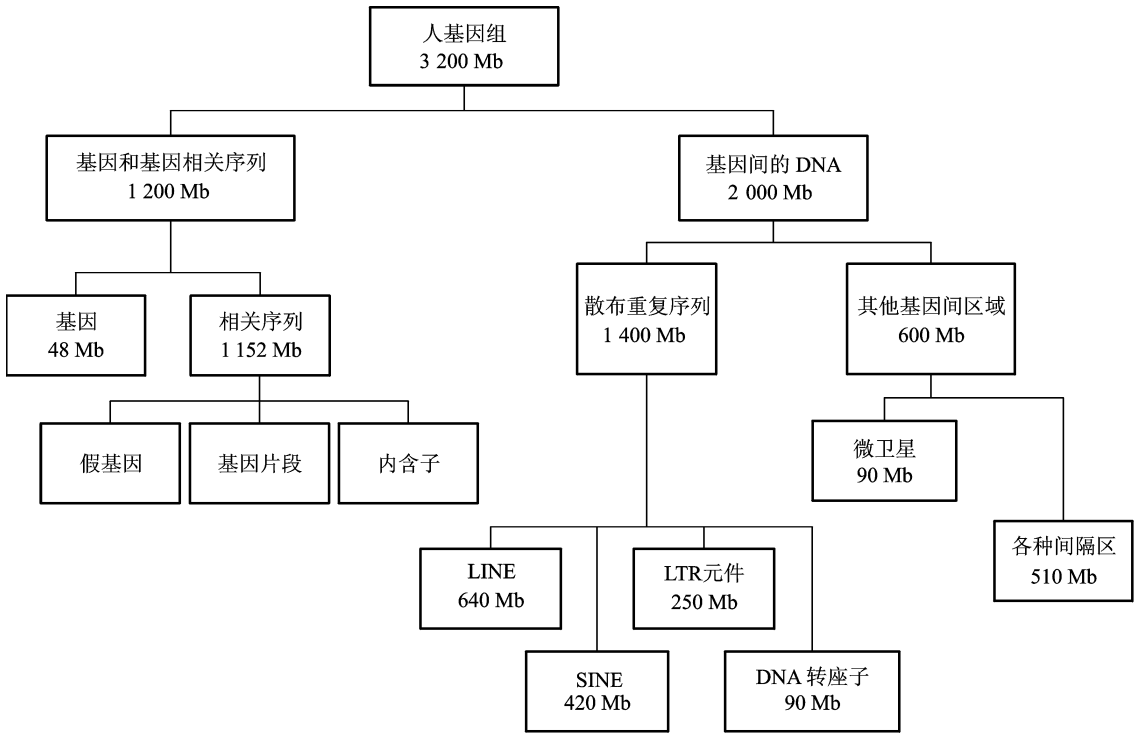


图 18-1 人类基因组的组织组成(引自 Brown, 2002)

从表 18-1可以看出(以 *Homo sapiens* Build 36.2所提供的数据为例)第 1染色体的基因数最多,达到 2 782个,该染色体也最长;除 Y染色体外,第 21染色体基因数最少,仅 352个,该染色体在物理长度上是最短的,但 22号染色体仅比 21号染色体长 6 Mb,其基因总数达 742个。同时每条染色体物理单位长度上所分布的基因数是不相等的,在染色体之间基因的分布也不均匀。第 1染色体是 10.1个基因/Mb,第 2染色体 7.7个基因/Mb,而第 4染色体 6个基因/Mb,但第 21染色体有 7.5个基因/Mb。至今我们还不能说明基因数目的分布与染色体结构的关系。

表 **18-1** 人类基因组各染色体的长度和基因的分布

染色体编号	物理长度/Mb			遗传长度/cM	已定位基因数		未定位基因数
1	263[①]	222[②]	274[③]	293[①]	2 594[②]	2 782[③]	27[③]
2	255	237	243	277	1 745	1 888	5
3	214	197	200	233	1 434	1 469	8
4	203	190	191	212	1 105	1 154	5
5	194	176	181	198	1 169	1 268	3
6	183	172	171	201	1 478	1 505	15
7	171	154	159	184	1 242	1 452	11
8	155	143	146	166	878	984	6
9	145	112	140	167	977	1 148	13
10	144	129	135	182	994	1 106	2
11	144	132	134	不完整	1 795	1 848	
12	143	134	132	169	1 360	1 370	
13	114	96	114	118	528	551	6
14	109	87	106	129	915	1 275	
15	106	79	100	110	787	945	23

续表

染色体编号	物理长度/Mb			遗传长度/cM	已定位基因数		未定位基因数
16	98 75	89		131	972	1 109	4
17	92 78	79		129	1 351	1 469	58
18	85 75	76		124	393	432	
19	67 56	64		110	1 605	1 695	10
20	72 60	62		97	728	737	
21	50 33.2	46.9		60	326	352	20
22	56 35.1	50		58	576	742	3
X	164 155	155		198	1 187	1 336	20
Y	28 11.1	58		缺乏数据	97	307	
合计	3 217.2	2 838.4	3 105		26 236	28 924	239

注:表中数据根据 NCBI网站(http://www.ncbi.nlm.nih.gov/SCIENCE96和 http://www.ncbi.nlm.nih.gov/mapview) 2007年 3月的数据综合进行整理而来。

① 为 http://www.ncbi.nlm.nih.gov/projects/genome/genemap96/chr 所提供的数据;

② 为 http://www.ncbi.nlm.nih.gov/mapview中 Celera公司所组装的结果;

③ 为 http://www.ncbi.nlm.nih.gov/mapview中 *Homo sapiens* Build 36.2(Current)所提供的数据整理而来。

18.1.3 遗传标记

遗传学中曾将可识别的等位基因称为遗传标记(genetic marker)。现代遗传学将可示踪染色体、染色体片段、基因等传递轨迹的遗传特性也称为遗传标记。除以上基因标记外,遗传标记还包括:形态标记、细胞学标记、蛋白质标记和 DNA标记。形态标记是指那些能够明确显示遗传多态的外观性状,如小麦的株高、粒色等的相对差异;细胞学标记是指能够明确显示遗传多态的细胞学特征,如染色体结构和数量的遗传多态性等;蛋白质标记主要包括非酶蛋白质和酶蛋白质。在非酶蛋白质中,使用较多的是种子贮藏蛋白。酶蛋白质主要是同功酶;DNA标记,也称 DNA多态性标记、DNA分子标记,是 DNA水平上遗传多态性的直接反映。显然,由于各自的条件,前几种遗传标记的应用受到限制。而 DNA分子标记与形态标记、细胞学标记和蛋白质标记相比,具有很多优点:不受环境条件的影响,不受发育阶段的限制,也不受个体和生物组织器官的限制;同时基因组 DNA变异非常丰富,可供选择的分子标记数量大大超过形态标记、细胞学标记和蛋白质标记的数量;此外,大多数 DNA分子标记是共显性的,符合孟德尔遗传规律,因此可供遗传标记作图。几种常用的 DNA分子标记如下:

① RFLP标记 在群体中生物个体之间,由于 DNA某一位点上的变异有可能引起该位点特异性的限制性内切酶识别位点的改变,包括原有位点的消失或出现新的酶切位点。当用限制性内切酶处理不同生物个体的 DNA时,致使酶切片段长度发生变化,个体之间出现限制性片段长度的差异,这称为限制性片段长度多态性(图 18-2)。

② VNTRs标记 真核生物基因组 DNA中含有大量的串联重复序列,在不同个体间或同一个体的同源染色体间都会产生高度的变异。一般将长 16~100个核苷酸为基本单元的串联重复序列称为小卫星,而将以 2~6个核苷酸为基本单元的简单串联重复序列,例如$(CA)_n$、$(GAG)_n$、$(GACA)_n$等,称为微卫星或简单序列重复(simple sequence repeats, SSR)。小卫星和微卫星其多态性来源于重复序列的核苷酸组成和重复的次数不同。一般又将小卫星和微卫星称作可变数目串联重复(variable number of tandem repeats, VNTRs)。VNTRs标记也呈共显性遗传,符合孟德尔遗传传递规律,因此可以用来进行遗传分析和作图(图 18-3)。而且 VNTRs在基因组中有广泛的分布,可检出的多态更丰富,出现的频率更高。

③ AFLP标记 扩增片段长度多态性(amplified fragments length polymorphism, AFLP)标记,是结合 RFLP和 PCR的优点发明的一种 DNA指纹技术。通过对基因组 DNA酶切片段的选择性扩增来

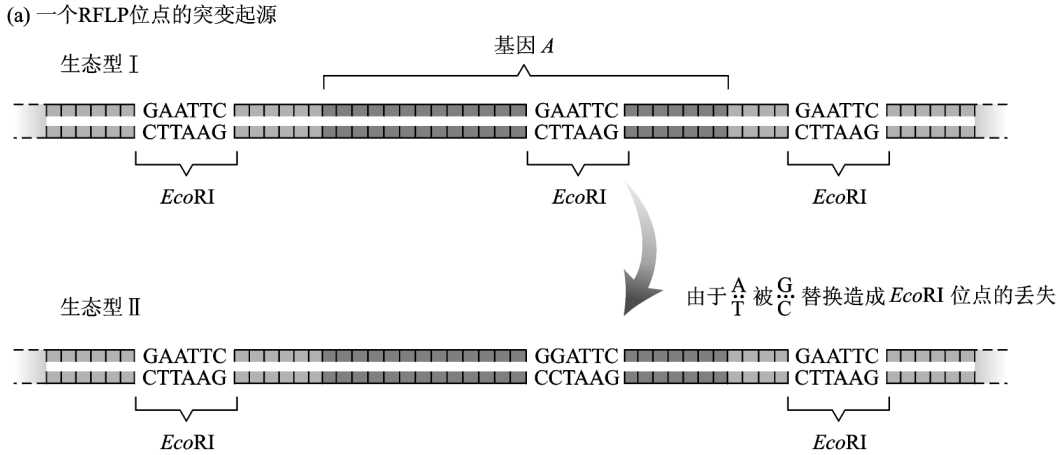


图 18-2 RFLP的产生与检测(引自 Sunstad等, 2003)

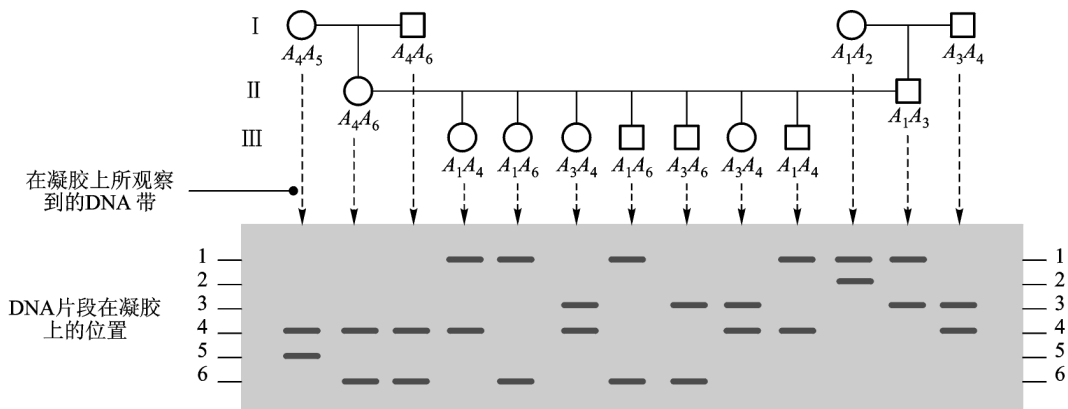


图 18-3 人 VNTRs的系谱分析(引自 Hartl等, 2001)

人类系谱分析表明 VNTR等位基因的分离。在系谱中出现 6个等位基因(A_1～A_6),但是任何一个个体仅可能有一个等位基因(纯合子)或两个等位基因(杂合子)

检测 DNA酶切片段长度的多态性(图 18-4)。AFLP揭示的 DNA多态性是酶切位点和其后的选择性碱基的变异。AFLP具有 RFLP技术的可靠性和 PCR技术的高效性。AFLP标记的主要特点是:由于在 AFLP分析中所采用的限制酶以及引物的种类、数目有较多的选择,因此在理论上能够产生的标记数目是无限的;而且扩增片段数目与引物有关而与酶切片段无关;AFLP呈典型的孟德尔遗传,

用于遗传分析；AFLP分析中所产生的大多数扩增带的片段与基因组的单一位置相对应，因此可作为遗传图谱和物理图谱的界(位)标，用来构建高密度的连锁图。

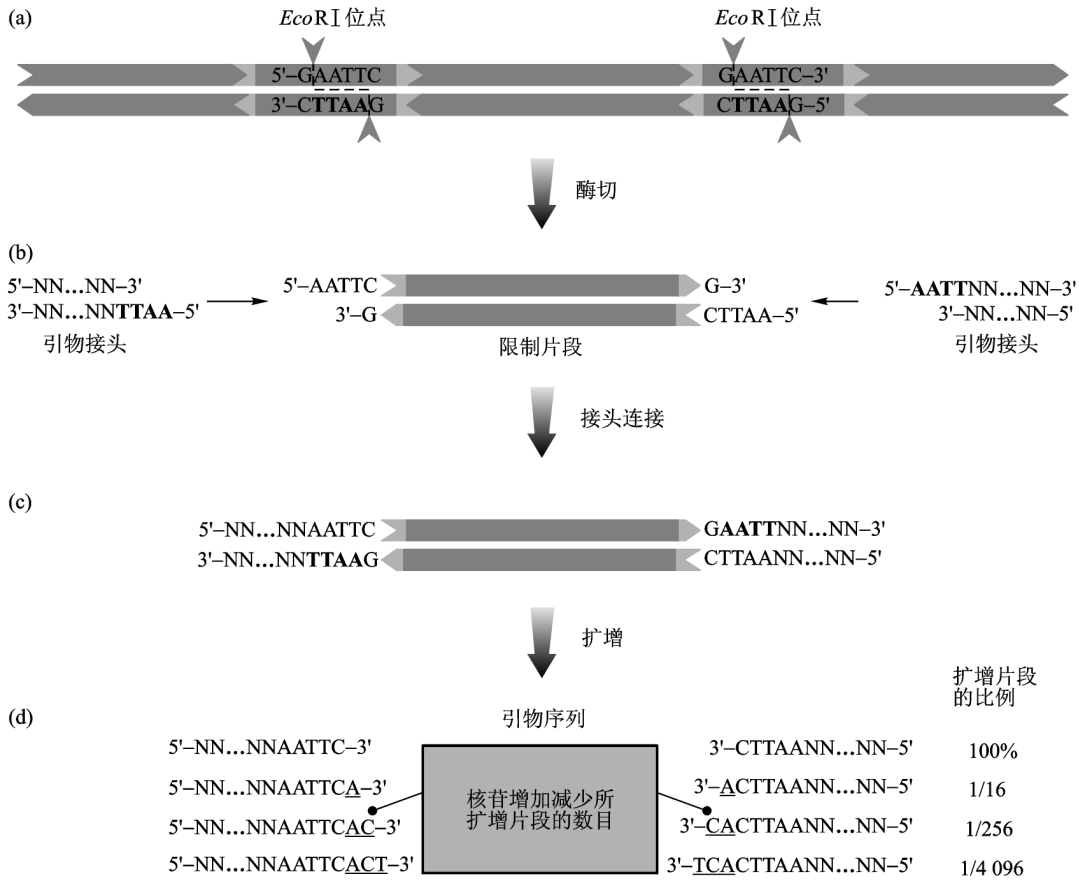
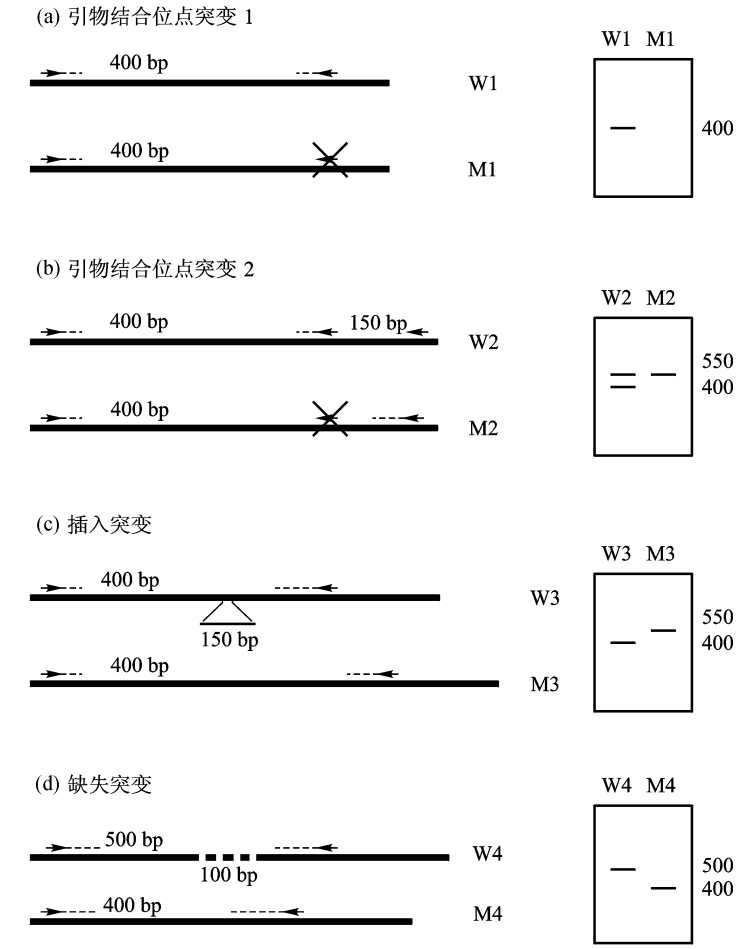


图 18-4 AFLP标记技术的原理示意图(引自 Hartl Daniel等，2001)
(a) 基因组 DNA用 1种或 2种限制酶消化(*Eco*R Ⅰ) (b) 加上引物接头
(c) 寡核苷酸接头与限制片段连接 (d) 用选择性引物进行 PCR扩增

④ RAPD标记 用寡核苷酸随机短引物(人工合成的 9～10个核苷酸组成)进行 DNA的 PCR扩增。经凝胶电泳分离，溴化乙锭染色，显示出扩增产物 DNA片段的多态性。其分子基础是模板 DNA扩增区段上引物位点的碱基序列发生了突变。因此，不同来源的基因组在该区段(座位)上将表现为扩增区段产物的有无或扩增片段大小的差异(图 18-5)。RAPD标记引物扩增产物所扩增的 DNA区段是事先未知的，具有随机性和任意性，因此随机引物 PCR标记技术可用于对任何未知基因组的研究。RAPD标记的不足之处是，一般表现为显性遗传，不能区分显性纯合和杂合的基因型，因而提供的信息量不完整。

⑤ STS标记 序列标签位点(sequence tagged site, STS)是在染色体上定位的、序列已知的单拷贝 DNA短片段。STS标记的原理，是根据单拷贝的 DNA片段两端的序列，设计一对特异引物，经 PCR扩增基因组 DNA而产生的一段长度为几百 bp的特异序列。由于不同的 STS序列在基因组中往往只出现一次，从而能够界定基因组的特异位点。用 STS进行物理作图，可通过 PCR或杂交途径来完成。STS标记可以作为比较遗传图谱和物理图谱的共同界标，因此在基因组作图上具有非常重要的作用。

⑥ SNP标记 单核苷酸多态性(single nucleotile polymorphism, SNP)标记是同一物种不同个体基因组 DNA的等位序列上单个核苷酸存在差别的现象。其比较的不是 DNA的片段长度，而是相



注: ➤ 引物 W: 野生型 M: 突变型

图 18-5 随机扩增 PCR产生多态性的分子基础(引自方宣钧等, 2001)

同序列长度里的单个碱基的差别。因此, SNP是二等位多态性,其中最少一种在群体中的频率不小于 1%;如果出现频率低于 1%,则视作点突变。 SNP在大多数基因组中存在较高的频率,估计人类基因组中有 300万个以上,平均 500~1 000 bp中就有一个。 SNP是人类可遗传变异中最常见的一种,占所有已知多态性的 90%以上。在基因组中 SNP既可存在于基因序列中,也可存在基因以外的非编码序列中。存在于编码序列中的 SNP虽然较少,但其在遗传疾病研究中却具有重要意义。相当一部分 SNP直接或间接地与个体的表型差异、人类对疾病的易感性和抵抗能力有关,因此,对 SNP的研究越来越受到人们的关注。例如国际人类基因组单体型图计划(HapMap project)的科学基础就是染色体上的 SNP的板块(block)结构。 SNPs在一段染色体上是成组遗传的,在 DNA上构成无形的区域"板块"。每个板块在进化上非常保守,在多世代的传递中没有或极少发生 DNA重组,其 SNPs的构成在单个染色体上的模式,即单体型(haplotype)。确定人类经世代遗传仍保持完整的单体型图,以及在不同族群中这些单体型的类型与分布,并将这些不同的单体型标上标签,将为人类不同群体的遗传多态性研究、疾病和遗传关联分析、治病基因和治病因子的确定、药效及副作用和疾病风险的分析、人类起源进化迁徙历史的研究等提供完整的人类基因组信息和有效的研究工具。

18.1.4 遗传图谱

人类基因组的 DNA序列分布于 22条常染色体及 X和 Y这 2条性染色体上。进行核苷酸序列

的测定时，首先应当将染色体进行分解，使之成为较易操作的小的结构区域，即对人类基因组进行标记和划分，然后将这些小的结构区域按其在染色体上的位置和先后次序进行准确的排列，这个过程简称作图。人类基因图有两类：遗传图（genetic map）和物理图（physical map）。

遗传图又称连锁图（linkage map），是指确定基因或 DNA 标记在染色体上的相对位置与遗传距离。它是通过连锁分析，计算遗传标记（或基因）间的交换频率，将某一染色体上的基因呈直线排列，确定基因之间的相对位置，一般用厘摩（cM）表示。人类基因组全长约 3 200 cM，1 cM 大约相当于 1 000 kb。遗传图的绘制需要应用多态性标记作为位标，如最早应用的 RFLP 标记，20世纪 80年代后期应用的 STR 标记。近来，多应用 SNP 标记。在遗传图中，使用遗传标记越多，越密集，所得到的连锁图谱的分辨率就越高，目前遗传图的分辨率已精确到 0.75 cM 左右。

值得指出的是在小鼠、果蝇、豌豆等动植物中，可以按照人们的意志去设计杂交方案，制作连锁图，但对于人类来讲，我们无法实施这样的方案。所以人的遗传连锁图的构建，采取系谱分析以及体细胞遗传学等特殊方法进行（详见第 4章）。利用遗传标记自祖代到后代中的传递，进行跟踪。例如，利用 RFLP 探针，通过分析人类系谱亲代和子代的 DNA，可研究 RFLP 的遗传规律，这里作为标记的 RFLP 就相当于等位基因，可分析其连锁和交换的频率，将 RFLP 标记定位在遗传连锁图谱上。

18.1.5 物理图谱

物理图谱是指各遗传标记之间或 DNA 序列两点之间，以物理距离来表示其在 DNA 分子上的位置而构成的位置图，以实际的碱基对或千碱基对或百万碱基对长度来度量其物理距离。最早的物理图谱是细胞遗传学图谱，通过原位杂交将基因定位在染色体各区带上。而现在在 HGP 中以 YAC 或 BAC 为载体构建的连续克隆系覆盖人的每条染色体的大片段 DNA。以 YAC 叠连群或 BAC 叠连群作为大尺度物理图谱，同时寻找分布于人类整个基因组的序列标签位点 STS。STS 是具有位点专一性，染色体定位明确，而且可用 PCR 扩增的单拷贝序列，是物理作图通用语言，最新制作的物理图包含了 52 000个 STS 位标。以 STS 为基础的图谱最大的优点是：适合于大规模测序，并很容易在染色体上定位。将 YAC 克隆在染色体上排序，被认为是基因组研究中最基本最关键的步骤。然后将 PCR 技术、STS 位标和 YAC 克隆以及计算机分析技术结合起来最后形成一张人类染色体的完整的物理图谱。遗传学图、细胞学图和物理图之间的相互关系如图 18-6所示。

18.1.6 模式生物基因组研究

HGP 除了对人类基因组的测序外，还包括大肠杆菌、酵母菌、线虫、果蝇、拟南芥、小鼠、水稻等模式生物体的研究计划。生物学研究早已表明，从模式生物获得的数据，对于研究和阐明人类生物学是必不可少的。在研究人类基因组的同时，研究上述模式生物的基因组的作图和测序，可以摸索、改进和提高作图和测序技术。因为这些模式生物的基因组相对于人的基因组来说都比较小、结构比较简单，易于操作；另一方面这些模式生物具有生活周期短，遗传背景清楚，易于培养繁殖，可以得到很多后代，获得很多遗传信息和各种遗传变异等优点；在进化的过程中，从低等到高等生物许多功能基因十分保守，许多核苷酸序列也十分保守，根据比较基因组学原理，相同的功能基因有着相似的生物学功能，同线群（syntenic group）的基因有类似的分布。因此，从模式生物基因组得到的数据和资料，将对分析人类基因组的组织结构及阐明一些基因和 DNA 片段的功能，具有十分重要的作用。在原核生物中已开展基因组计划并完成基因组测序的主要有流感嗜血杆菌、支原体、大肠杆菌和枯草杆菌等。在真核生物中，已经开展基因组计划并完成测序工作的有拟南芥、线虫、果蝇、酵母、水稻等。

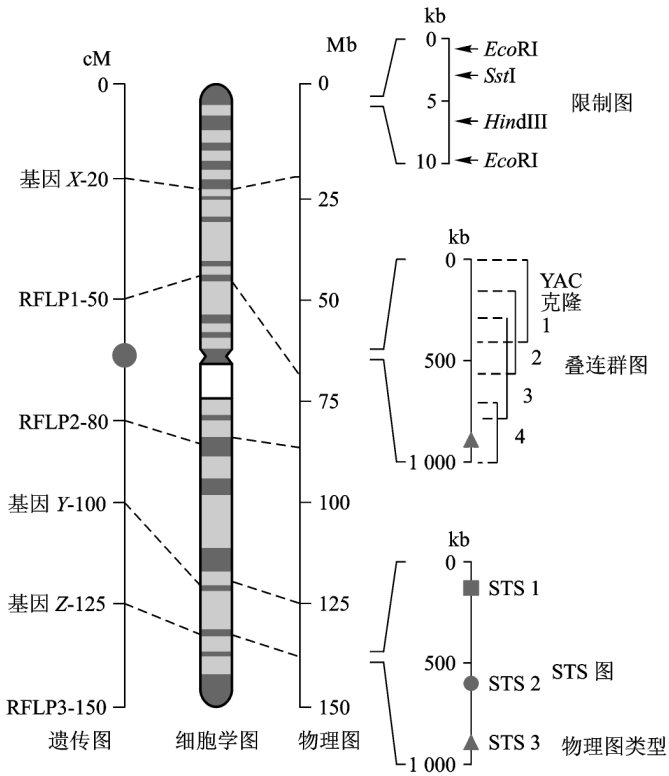


图 18-6 三种图谱之间的相互关系(引自 Sunstad等, 2003)
在现代分子标记图谱中,遗传图和物理图所表示的基因(和分子标记位点)在序列上基本相同,但在距离上是不相同的。这是由于遗传图确定的是基因或 DNA标记在染色体上的相对位置与遗传距离。物理图是指以物理距离来表示其在 DNA分子上的位置而构成的位置图

18.2 基因组测序与序列组装

18.2.1 基因组测序策略

基因组序列的测定是一种大规模的序列测定,选择适当的测序策略是很重要的。主要有两种测序策略,一种是自上而下(top-down mapping)或由长到短作图策略;另一种是全基因组鸟枪法(whole-genome shotgun method)作图策略,也称自下而上(Bottom-up approach/mapping)或由短到长作图策略(图 18-7)。

在"自上而下"测序作图策略中,首先建立连续叠连克隆系(overlapping clones)或叠连群(contig)。再对单个叠连群采用鸟枪法对其中的克隆逐个进行测序,最后在叠连群内进行拼接出全长序列。同时,需要辅以构建大 DNA克隆(100~10 000 kb),并把克隆依染色体排列构成染色体的克隆图。当对每个克隆测序完成后,就可以拼装出整个染色体的 DNA序列。

在"全基因组鸟枪法"测序作图策略中,直接将基因组 DNA随机切成 2 kb左右的小片段(BAC克隆),然后进行随机末端测序,再以基因组的分子标记为起点进行 DNA片段拼接,其余过程辅以少量大片段(约 10 kb),计算机分析串联得全序列。

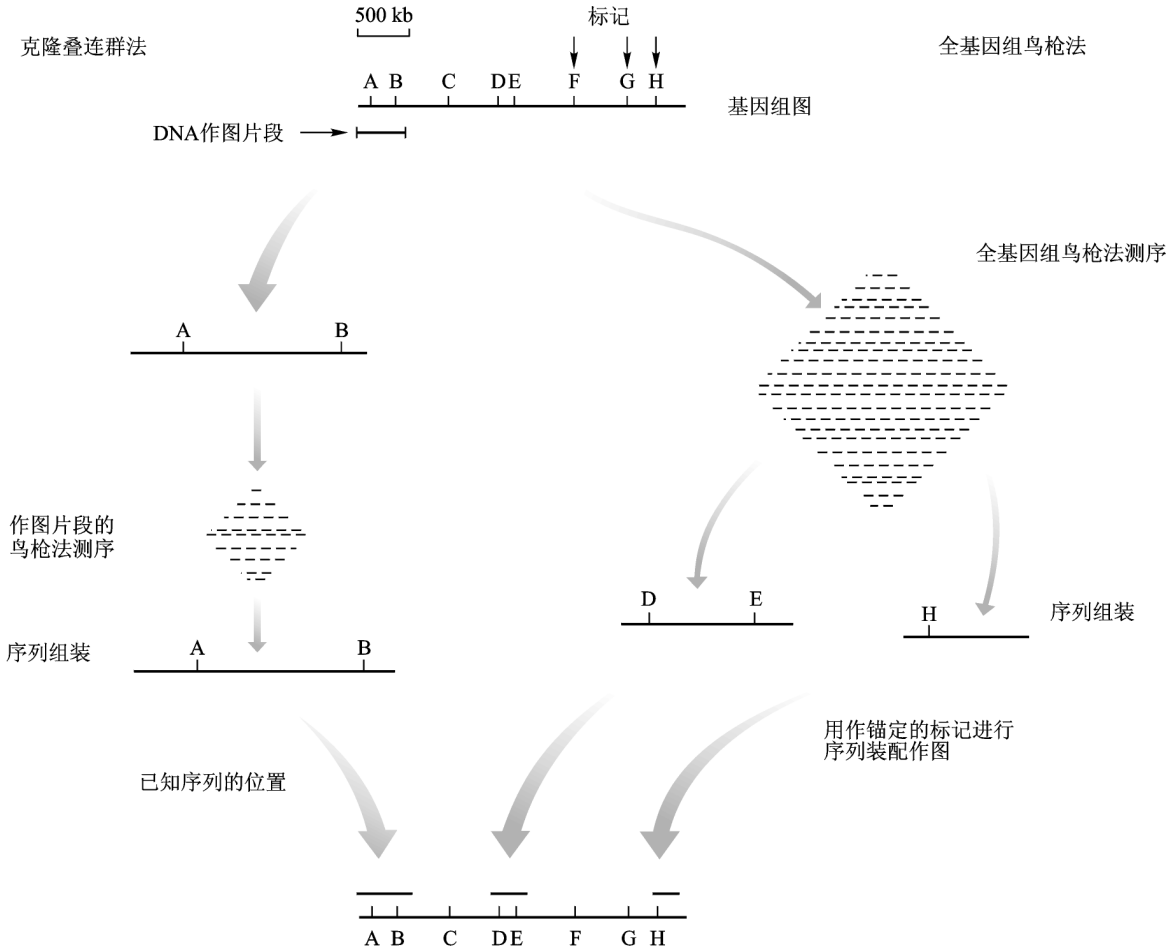


图 18-7 两种测序作图策略(引自 Brown, 2002)

一种测序作图策略为克隆叠连群法,先建连续叠连群,再对单个叠连群采用鸟枪法对其中的克隆逐个进行测序,最后在叠连群内进行拼接,拼接出全长序列;另一种为全基因组鸟枪法,直接将基因组 DNA 随机切成小片段(BAC 克隆),然后进行随机末端测序,再以基因组的分子标记为起点进行 DNA 片段拼接

18.2.2 基因组测序方法与组装

基因组序列测定技术经历了从手工到自动,从慢到快的发展过程。20世纪七八十年代 DNA 测序技术主要有 Sanger双脱氧链终止法(末端终止法)和 Maxam-Gilbert化学修饰法。1977年英国 Sanger 实验室采用末端终止法完成的 ΦX174全基因组的测定是世界上第一个被测定的基因组。由于基因组测序是大规模的序列测定,要求快速、准确,这就必须要求实现自动化测序。Sanger的末端终止法具有这些特点。因此在后来很快取代了 Maxam-Gilbert的化学修饰法,而成为 DNA 序列测定的主流技术。

荧光标记的应用、荧光自动测序仪的发明(Hunkapiller等,1991)、高质量 DNA 聚合酶的获得以及引物合成成本的降低等使现代基因组的大规模序列测定成为可能。其所采用的测序基本原理仍然是 Sanger末端终止法原理。荧光毛细管电泳技术是目前基因组序列测定的主要技术之一(图 18-8)。主要优点是:① 使用不同荧光色彩的标记化合物,分别标记 4种 ddNTP,加入到一个反应中,聚合链终止反应完成后,可以获得分别带有 A, C, T和 G的 DNA 单链。② 自动化荧光测序系统避免了人眼分辨的差错,结果准确可靠,测序质量大大提高。③ 测序速度提高,有 96个泳道,每次可同时进行 96次测序,极大地提高了测序的工作效率,1天可完成近千次反应。

DNA芯片(DNA chip)杂交测序(sequencing by hybridization)技术是近年发展起来的一种新技术。所谓DNA芯片,是指采用类似大规模集成电路的手段将寡核苷酸探针或者DNA、cDNA有规律地排列固定在指甲大小(或更小)的硅片上而形成的测序集成块。在DNA芯片杂交测序技术中,由于事先已将各种排列顺序的寡核苷酸探针固定在芯片上,每个寡核苷酸探针在排列的方阵中都有固定的位置。当待测的样品DNA分子与芯片温浴,凡是能杂交的寡核苷酸都会在确定的位置发出杂交信号,然后根据获取的信息将寡核苷酸的序列进行对比组装,拼接成完全的DNA序列。例如有一个12核苷酸序列GTAGCAGATGTA,可以被解析为5个寡核苷酸,相邻两核苷酸之间只差开一个核苷酸,而重叠其余7个核苷酸,则有以下对位方式:

```
    GTAGCAGATGTA
1       CAGATGTA
2      GCAGATGT
3     AGCAGATG
4    TAGCAGAT
5   GTAGCAGA
```

在这里每个寡核苷酸片段长8个碱基,则其所有可能的排列顺序为4^8=65 536种可能。当我们将该12 nt的核苷酸序列与一组由总数为65 536种的8 nt寡核苷酸组成的探针群体(芯片)杂交时,仅有5种探针可与该12 nt的核苷酸序列形成完全互补的双链体分子(图18-9)。

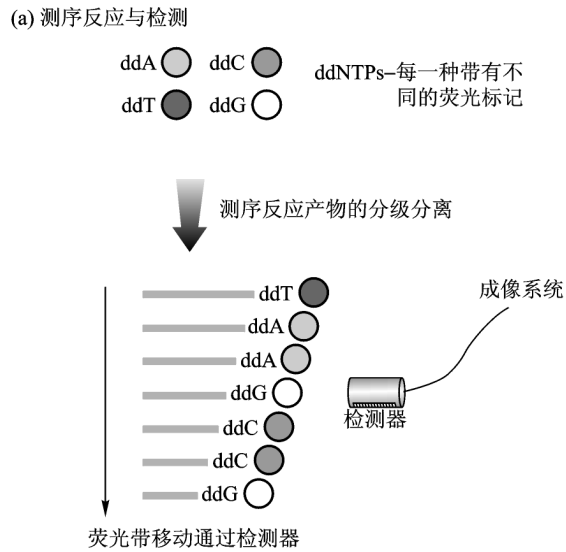


图18-8 荧光毛细管序列测定原理

(引自 Brown, 2002)

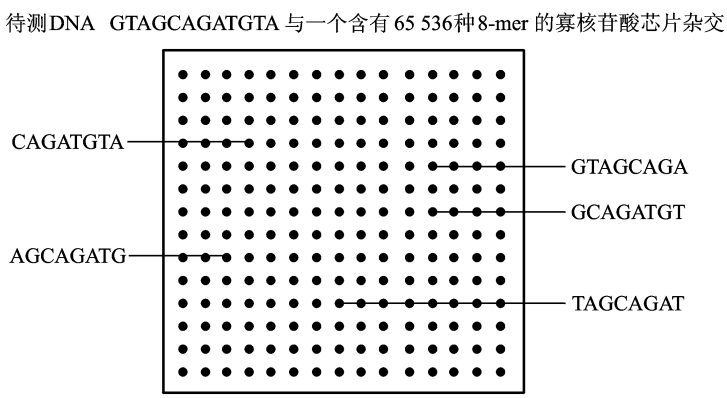


图18-9 DNA芯片杂交测序

(引自曹仪植,2002)

无论是用全基因组鸟枪法还是采用叠连群逐条克隆法测序,得到的都是成百上千万的小片段DNA序列,最后必须要将它们组装成基因组每条染色体上真实的排列顺序,这是手工操作无法完成的,需要借助计算机和数据库以及相关的软件系统。因此,DNA序列的组装是一项浩繁,技术要求高

而又精细的工作。

首先需要对 DNA序列测定的数据质量和准确性进行评估,包括:在测序仪读取信息后,进行碱基识别过程的同时,软件根据荧光曲线波形给出每个碱基的可信度;软件确定所测的每个样品的高质量部分;在序列拼接完成后,软件对拼接成的叠连群整体的可信度进行评估,给出可能的错误率。

进行序列的组装,一般采用的软件是 PhilGreen实验室建立和发展的 Phred-Phrap-Consed系统。这套软件不仅适用于大规模测序,也适用于一般的实验室。但是随着基因组计划的进行,各国各实验室都开发和建立了自己基因组测序组装的相关软件。

在基因组测序和组装方面我国也走在世界的前列,在 2002年 12月初我国率先完成了籼稻全基因组"精细图"的绘制。该图谱的特点和意义是:① 该图覆盖了 97%的基因序列,并将其中 97%的基因精确地定位在染色体上;覆盖基因组 94%的染色体定位序列准确性达到 99.99%。成为迄今为止唯一的基于"全基因组鸟枪法"构建的大型植物基因组高精度基因图。② 通过对籼稻和粳稻亚种基因组已定序列的比较分析,发现了 100多万个单核苷酸多态性,并在染色体上定位和整合在精细基因图谱上。③ 预测出约 6万个水稻基因,利用这些信息,制备出了全基因组基因芯片。④ 通过比较基因组学研究,发现水稻和拟南芥基因组在基因组结构、基因表达和基因功能方面存在广泛差异。⑤ 建立了全基因组鸟枪法测序基因组组装的计算机软件体系,使我国成为世界上少数拥有组装高精度全基因图的国家之一。

18.3 基因组图谱构建与应用

18.3.1 人类基因组遗传图谱的构建

(1) 经典遗传图谱的构建

首先要对用于作图的基因是否连锁进行判断。在人类遗传学研究中,由于通常不知道父母的基因型和父母中标记基因的连锁相是互斥还是互引,因而无法简单地通过计算重组体出现的频率来进行连锁分析,而必须通过适当的统计模型来估算重组率,并采用或然比或似然比(likelihood ratio,)检验的方法来推断连锁是否存在,即比较假设两个座位间存在连锁($r<0.5$)的概率与假设没有连锁($r=0.5$)的概率。这两种概率之比可以用或然比统计量来表示:即 $P_1(\ )/P_2(0.5)$。或然率是将连锁与独立分配区别开来的关键数据。遗传学上通常用或然率的常用对数作为标准的衡量方法,该值的对数值称为 Lod值(Lod score)或对数优势比(logarithum of the odds):根据两个非此即彼的假设,计算数据的整体或然性,以确定两个基因座或是按一定的重组率而相互连锁的可能性或是互不连锁的可能性;这两种可能性之比,是基因座实际上为连锁的可能性;这个比率的 10作底的对数就是对数优势比。为了确定两对基因之间是否存在连锁,一般要求或然比大于 1 000:1,即 Lod>3;而要否定连锁存在,则要求或然比小于 100:1,即 Lod<2。在其他生物遗传图谱的构建中,或然比的概念也用来反映重组值的可靠程度或作为连锁是否真实存在的一种判断尺度。

通常判定连锁关系是以 Lod值大小为依据。Lod值为 0,意味着连锁假设与不连锁假设的可能性是相等的;Lod值为正值,有利于连锁;Lod值为负值,表示有一定重组率的连锁。显著性的域值是+3和-2。Lod=+3时,连锁的概率为 95%。如果两个基因座都在 X染色体上,则 Lod=2.5,就足以说明两者是连锁的。

一旦若干连锁基因被确定,接下来是估计基因座位的图距,并排列出这些基因的相互远近关系,从而构建出一张基因的连锁图。我们知道遗传距离的单位是厘摩,1 cM 大约相当于两个基因间平均

发生 0.01次交换而产生的距离。细胞遗传学的研究表明,在产生一个种系细胞的减数分离过程中,平均的交换次数为 33次,由此推算,人类染色体的平均遗传长度就应约为 3 300 cM。1 cM 约相当于 1 000 kb,则人类基因组约 33亿个碱基对。

由上可知,经典遗传学图谱主要用来确定生物体的基因在染色体上的排列,只能标明基因之间的相对位置,无法指明基因在染色体上的具体位置,因此无法按这种图谱直接分离和克隆基因。因而,在人类基因组计划中利用价值不大。

(2) 现代遗传学图谱的构建

现代遗传学图谱的概念是 David Botstein等于 1980年提出来的,当时由于 DNA限制性内切酶和连接酶的应用,RFLP成为一种崭新的 DNA多态性标记。他们利用 RFLP作为标记去构建多态性基因与这些标记连锁关系,进而确定多态性基因的位置。其精髓在于将单纯的表型研究深入到 DNA分子的本质上去。随后,亨廷顿舞蹈病成为第一个被确定与这些标记连锁的常染色体遗传病。

建立精细的人类遗传图谱的关键是获得足够多的高度多态性的标记。

第一代用作遗传图谱的 DNA遗传标记是 RFLP。1987年,Donis Keller等发表了第一张人类的 RFLP连锁图,由于核苷酸序列的改变遍及整个基因组,特别是在进化中选择压不是很大的非编码序列之中,RFLP的出现频率远远超过了经典蛋白质多态性。但是,由于 RFLP仅局限于 1个或少数几个核苷酸的突变,一般只能产生限制性酶切位点的"切开"与"不切开"两种情况,故所提供的"多态性"信息量较小。

第二代 DNA遗传标记利用了存在于人类基因组中的大量重复序列。短串联重复(short tandem repeat,STR)具有"多态性"与"高频率"两个突出的优点,并可以用 PCR检测,具有其他标记不可替代的优点。1996年 3月 14日 Nature发表了法、加合作获得的人类最新的遗传连锁图。该图由 5 264个微卫星(AC/TG)n 标记组成,遗传距离(分辨率)平均为 1.6 cM,已超过美国原计划到 2000年达到 2.5 cM 的指标。

第三代 DNA遗传标记是 SNP。SNP遗传标记分析完全摒弃了经典的凝胶电泳,代之以最新的 DNA芯片技术,是目前发现的最好的遗传标记,而且已成为研究基因组多样性和识别、定位疾病相关基因的一种新手段。

自人类基因组序列图提前绘制后,人类基因组研究的主战场已经转入"单体型图"。2002年 10月,中、美、英、日和加拿大五国代表在美国华盛顿正式启动了"人类基因组单体型图计划",它将整合基因组测序结果,从基因组水平检测不同族群的 SNP位点,绘制人类基因组中独立遗传的 DNA"始祖板块"及 SNP标记的完整目录,以确定不同人群之间的基因差异,从而建立人类遗传的群体信息资源库。

知识窗 **18-1**

世界首份"个人版"全基因组图谱面世

——诺贝尔奖得主"DNA之父"詹姆斯·沃森(James Watson)的"生命天书"

2007年 5月 30日,DNA之父——79岁的沃森博士获得了一张由美国贝勒医学院人类基因组测序中心(Human Genome Sequencing Center, Baylor College of Medicine)和"454生命科学公司"(454 Life Sciences Corporation)共同完成的、储存着沃森自己全基因组序列的 DVD光盘,成为世界首位"个人版"基因组图谱的拥有者(Nature, 1 June, 2007,在线发表;doi: 10.1038/news070528-10)。这是从 2003年人类基因组全序列图谱完成以来分子遗传学和基因组学研究领域的又一里程碑式的伟大成就。

首先,这标志着人类基因组测序技术的快速改进和提高,耗时与耗资均大幅下降。1990年启动的"人类基因组"计划用了 13年时间才完成全序列图谱的绘制,耗资近 10亿美元(预计为 30亿美元,

但实际支出由于技术的更新而减少很多)。绘制沃森的 DNA 图谱采用了一种新的测序技术——Genome Sequencer FLX(TM) Systems,不但极大地提高了测序效率,而且仅耗时 2个月,花费 100万美元。其次,2007年 5月 31日,沃森的基因组图谱被收入美国国家健康协会的数据库(http://www.ncbi.nlm.nih.gov/Traces/trace.cgi),并向全世界公开,研究者可以从中找到被认为与基因有关的疾病、智力、冒险精神、信仰和性格等问题的密码。沃森希望通过自己的行动带动更多的人进行基因测序。他认为,了解这些信息有助于提早预防癌症、心脏病、阿尔茨海默氏症等多种顽疾,甚至还能让人更富有同情心。这标志着个人基因组时代的到来。第三,随着首份"个人版"基因组图谱的完成,预示着不久的将来其他人的基因组图谱会陆续问世。但要真正"读懂"个人的"生命天书"上的每一条信息还有很长的路程。因为人类对生命本质的认识还处在初级阶段。随着技术的进步,制作"个人版"基因组图谱所需时间和成本会不断缩减,这无疑预示着个性化的基因组医学时代即将来临。

沃森称,他理解基于个人遗传因素方面的隐私被公开而受歧视所带来的恐惧心理,但这种担忧无疑被夸大了。从沃森"个人版"基因组图谱的研究分析中得知,沃森的基因组中包含很多变异的 DNA,其中有的基因与癌变有直接的关联。据悉,从 20多岁开始,沃森就患有皮肤癌。沃森表示,他允许将自己的基因组信息放在互联网上供科学研究,也是为了向社会证实此类信息并没有什么可怕之处。个人遗传信息的"生命天书"的隐私信息是否公之于众,是否应限制个人基因信息的公开,目前都还处于争议阶段。

18.3.2 植物基因组遗传图谱的构建

(1) 选择 DNA 分子标记

目前用于植物遗传作图的 DNA 分子标记有 RFLP、RAPD、AFLP、SCAR(sequence characteristic amplified regions)和 SSR等。具体选用何种标记,要依据标记的特点、实验条件、作图植物的生长发育特性、对该植物的研究情况及实验目的等决定。

(2) 亲本的选配

亲本的选择直接影响到构建连锁图谱的难易程度及所建图谱的适用范围。

对亲本的选择要考虑亲本间的 DNA 多态性、亲本材料的纯度和杂交后代的可育性等问题。一般,异交作物的多态性高,自交作物的多态性低。育种家常将野生种的优良性状转育到栽培种中,这种亲缘关系较远的杂交转育,DNA 多态性非常丰富;亲本材料纯度影响试验结果,纯度不高的要自交进一步纯化;亲本间的差异太大,减数分裂受到影响,后代可育性低,影响分离群体的构建,降低所建图谱的可信度和适用范围。最后,选配亲本时还应对亲本及其 F_1 杂种进行细胞学鉴定。防止亲本间有染色体易位、缺失及 F_1 染色体异常等现象对作图的不利影响出现。

(3) 分离群体类型的选择

根据其遗传稳定性分离群体分为暂时性分离群体和永久性分离群体两类。在暂时性分离群体中分离单位是个体,经自交或近交后其遗传组成会发生变化,无法永久使用,这些群体包括 F_2、F_3、F_4、BC 和三交群体等;在永久性分离群体中分离单位是株系,株系间基因组有差异,而株系内个体间的基因型是相同的(纯合的),是杂交不分离的,这类群体有重组近交系(recombinant inbred lines,RILs)和双单倍体(double haploid,DH)群体等,它们可杂交或近交繁殖后代而不改变群体的遗传组成,可以永久使用。因此,构建 DNA 连锁图谱应根据具体情况选择不同类型的分离群体。目前构建遗传连锁图谱主要应用 F_2 群体、RILs和 DH群体。

F_2 群体是常用的作图群体,迄今大多数植物的 DNA 标记连锁图谱都是用 F_2 群体构建的。其主要优点是容易建立 F_2 群体,其不足之处是对于显性纯合与显性杂合基因型无法识别,造成基因型信息简并现象的存在。

RILs群体是杂种后代经过多代自交而产生的一种作图群体,通常从 F_2 开始,采用单粒传的方法

来建立。自交的作用是使纯合的基因型增加,杂合的基因型减少,因此,RILs群体中每个株系都是纯合的,因而 RILs群体是一种可长期使用的永久性分离群体,又可以进行重复试验。它除了可用于构建分子标记连锁图外,还特别适用于数量性状基因座(QTL)的定位研究。

DH群体是由植物的单倍体经过染色体加倍形成的二倍体即加倍单倍体或双单倍体(DH)而来。DH群体产生的途径很多,最常见的方法是通过花药培养,诱导产生单倍体植株,然后对染色体加倍产生 DH植株。DH植株是纯合的,自交后产生纯系,这种纯系可以稳定繁殖,长期使用,是一种永久群体。DH群体的遗传结构直接反映了 F_1 配子中基因的分离和重组,因此 DH群体作图效率是最高的。同时,DH群体也适合于 QTL作图。

由于植物可以很方便地建立和维持较大的分离群体,所以其遗传图谱构建工作的发展速度超过了动物的同类研究。迄今为止,已构建图谱的植物多达几十种,其中包括了所有重要的农作物,如玉米、番茄、水稻、小麦、大麦、燕麦、大豆、高粱、油菜、莴苣、马铃薯等。

(4) 群体大小的确定

遗传图谱的分辨率和精度与群体的大小有密切的关系。群体越大,作图精度越高。但群体太大会增加实验工作量和费用。因此确定合适的群体大小是十分必要的。一般,构建 DNA标记连锁图谱所需的群体远比构建形态性状特别是数量性状的遗传图谱要小,大部分已经发表的分子标记连锁图所用的分离群体一般都不足 100个单株和家系。在实际工作中,构建分子标记骨架连锁图可基于大群体中的一个随机小群体(如 150个单株或家系),当需要精细地研究某个连锁区域时,再针对性地在骨架连锁图的基础上扩大群体。

(5) 连锁图谱制作的统计学原理

① 两点测验 对两个基因座位之间的连锁关系进行检测,称为两点测验。在进行连锁测验之前,必须了解各基因座的等位基因分离是否符合孟德尔分离比例,这是连锁检验的前提。在共显性条件下,F_2 群体中一个座位上的基因型分离比例为 1∶2∶1,而 BC_1 和 DH群体中分离比例均为 1∶1;在显性条件下,F_2 群体分离比例为 3∶1,而 BC_1 和 DH群体中分离比例仍为 1∶1。检验 DNA标记的分离是否偏离孟德尔比例,一般采用 χ^2 检验。而对基因座位之间的连锁关系,则采用或然比检验的方法,即 Lod值的大小来进行重组率的估计从而推断连锁是否存在。

② 多点测验 在构建分子标记连锁图谱中,每条染色体都涉及许多标记座位。要确定这些标记座位在染色体上的正确排列顺序及彼此间的遗传距离必须同时对多个基因进行联合分析,利用多个基因座间的共分离信息来确定它们的排列顺序,进行多点测验。多点测验通常也采用或然比检验法。先对各种可能的基因排列顺序进行最大或然比估计,然后通过或然比检验确定出可能性最大的顺序。在一条染色体上,经过多次多点测验,就能确定出基因的最佳排列顺序,并估计出相邻基因间的遗传图距,从而构建出相应的连锁图。

(6) DNA 标记分离数据的处理

从分离群体中收集分子标记的分离数据,获得不同个体的 DNA多态性信息,是进行连锁分析的第一步。通常各种 DNA标记基因型的表现形式是电泳带型,将电泳带型数字化是 DNA标记分离数据进行数学处理的关键。进行 DNA标记带型数字化的基本原则是,必须区别所有可能的类型和情况,并赋予相应的数字或符号。例如,用 RFLP标记构建遗传图谱,假设全部试验共 100个 F_2 单株,检验了 90个 RFLP标记,经数字化处理,可得到一个由 90(行)×100(列)的、由简单数字组成的 RFLP数据矩阵。获得有关数据矩阵后,选择适合的构建 DNA标记图谱的计算机软件对其进行分析和处理。

18.3.3 物理图谱的构建

(1) 细胞遗传学图谱

细胞遗传学图谱(cytogenetic map)是将基因或 DNA片段直观定位于染色体上的物理图谱,因此也称为染色体图谱(chromosome map)。它是把基因或其他被分离出的 DNA片段定位在它所在的染

色体区域，并且粗略地测出它们之间相距的碱基长度。其图谱的制作主要采用原位杂交技术，将目标基因或特定的 DNA片段定位到特定的染色体区带上。细胞遗传学图谱制作的关键是原位杂交探针序列与染色体目标序列的相互作用。目前用得最多的是荧光原位杂交（fluoscence in situhybridization，FISH）技术。其基本原理是将 DNA探针用特殊修饰的核苷酸分子标记（如生物素 biotin dUTP或地高辛 digoxigenin dUTP），使标记的探针通过碱基互补原位杂交到染色体切片上，再用与荧光素分子偶联的单克隆抗体与探针分子特异结合来检测该 DNA序列在染色体上的位置。近年发展起来的染色体 DNA Fiber-FISH技术，可以将几组探针带上不同的荧光标记，同时与染色体杂交，利用不同的抗体进行检测，一次可以进行多个位点的杂交，并且有较好的分辨率。一般染色体原位杂交的分辨率在 3Mb左右，即两个相距小于 3Mb的探针杂交信号会相互重叠，不能分清。改进的 FISH技术能将 DNA序列精确到 2～5Mb的范围。而 DNA Fiber-FISH结合多色荧光路标探针，与 DNA大分子限制性酶切位点进行精密排序，分辨率可达到 0.5～8 kb。

（2） 限制酶切图谱

DNA限制性内切酶酶切图谱是一种重要的 DNA物理图谱，它由一系列位置确定的多种限制性内切酶酶切位点组成，以直线或环状图式表示。限制性内切酶在 DNA链上的切口是以特异序列为基础的，核苷酸序列不同的 DNA，经酶切后就会产生不同长度的 DNA片段，由此而构成独特的酶切图谱。其主要的步骤是：制备克隆 DNA；克隆 DNA指纹分析，包括酶切、电泳分离、杂交标记或图像处理；计算机对克隆的排列和 DNA片段的排序；填补空隙，包括分离新克隆，PCR验证等。

构建 DNA限制性内切酶图谱有许多方法。通常结合使用多种限制性内切酶，通过综合分析多种酶的单酶切及不同组合的多种酶同时酶切所得到的限制性片段大小来确定各种酶的酶切位点及其相对位置。

（3） 叠连群图谱

一组相互两两头尾拼接的可装配成长片段的 DNA序列克隆群称为叠连群。分辨在基因文库克隆中叠连群插入片段的顺序关系，通过相互邻接的两个片段间存在的重叠部分，推断出各叠连群覆盖整个染色体的克隆片段在染色体上的顺序，最后构建出叠连群图谱（contig map）。其具体步骤是将染色体切割成小片段后，克隆并排序，得到由排好序的插入 DNA片段克隆组成的叠连群。克隆叠连群有多方面的重要应用，是人类基因组计划物理图谱的核心部分。使用这些叠连群图谱，可以重新确定克隆片段或其他 DNA探针在基因组中的位置，一旦构建成叠连群图谱，则整个基因组就可以以克隆形式获得，从而可在碱基序列水平上用这些克隆来分析基因组的任何区域。

克隆叠连群的组建通常采用染色体步移（chromosomal walking）法。首先从基因文库中随机取出一个指定的或随机克隆，然后在文库中寻找与之重叠的第二个克隆。在第二个克隆的基础上再寻找第三个克隆，依次延伸，这便是克隆叠连群图谱构建的基本原理。从实验技术上看，一个叠连群是一组克隆跑胶后的一系列泳带序列，这些泳带序列相互部分重叠，并仅属于一个叠连群，且每个叠连群至少含有一条泳带。根据重叠顺序的相对位置，通过原位杂交的方法将各个克隆首尾连接，即构成了一个连续的序列，其长度即叠连群长度，可达几百万碱基对。当今将克隆片段装配成叠连群，主要依靠计算机来处理。首先计算重叠的可能性，如果重叠的可能性在 90%以上，可认为两个克隆是相邻的，将之重叠起来，然后再将其他重叠可能性大的克隆找出，这样可逐步增长，直至形成这个叠连群。质粒、BAC、PAC和 YAC文库都能构建叠连群。叠连群图谱的绘制往往是核苷酸测序的第一步。

18.3.4 基因组图谱的应用

（1） 寻找新的基因

弄清基因组序列中所包含的全部遗传信息，即从基因组序列中进行基因搜寻、分析与基因相关的序列，这是解读整个基因组图谱的基础。

在基因组序列中查找新的基因有多种方法：一是根据已知的序列进行人工或计算机分析，来寻找

与基因相关的序列。二是通过同源查询来寻找基因。利用已存入数据库中的基因顺序与待查基因序列进行比对,从中即可查找与之匹配的碱基顺序用于界定基因。现有生物的不同种属之间具有功能或结构相似的直系基因成员,它们具有共同的起源,存在保守序列。当某一 DNA序列含有这类基因时,通过与已确定的其他基因序列比较,就可以发现其中的相似性。

(2) 基因的克隆与分离

根据饱和的基因组图谱,可以对基因进行克隆和分离。有多种克隆和分离基因的方法。通过图位克隆法(map-based cloning)或定位克隆(positional cloning)法,对基因进行克隆与分离是一种比较常见的方法。它是根据功能基因组中都有相对较稳定的基因,在利用分子标记技术对目标基因进行精细定位的基础上,用与目的基因紧密连锁的分子标记筛选 DNA文库(包括 YAC、BAC、TAC、PAC和 cosmid文库),从而构建目的基因区域的物理图谱,再利用此物理图谱通过染色体步移逐渐逼近目的基因或染色体登陆(chromosome landing)的方法最终找到包括该目的基因的克隆,并通过遗传转化试验证实目的基因的功能。

随着人类基因组计划信息资料的积累和各种数据库的建立,我国及时建立了利用生物信息学方法进行基因克隆的技术,例如 1997年夏家辉等建立了"基因家族——候选疾病基因克隆"的生物信息学克隆的技术,利用该法克隆到 9个全长 cDNA,并于 1998年 5月在国际上最早克隆到定位在 1p34的,以高频听力下降为主要特征的,神经性耳聋的疾病基因 GJB3。该基因是完全在我国本土上独立克隆的第一个疾病基因,实现了我国本土上克隆遗传疾病基因零的突破。

(3) 基因功能的预测

利用基因组图谱对基因功能进行预测,观察基因组作为一个整体如何行使其功能。基因功能的预测可以通过计算机预测和实验确定两种途径进行。计算机预测可以通过比较基因的同源性和氨基酸的一致性或相似性来进行,由已经确定了的基因和多肽的功能可以推断与之同源的片段的功能。通常氨基酸水平的比较可以更为有效地确定两个 DNA顺序是否同源,更为准确地推测基因功能。有时两个无明显同源的基因间会出现局部相似的区域,这是因为其功能域有共同的起源。

通过实验对基因功能进行预测,主要是采取反求遗传学(reverse genetics)的方法,可通过从 DNA、RNA水平上进行各种定点破坏结构基因或抑制结构基因的表达,包括定点引入碱基突变、基因敲除(knock-out)、反义 RNA技术、RNA干扰技术、转座子插入突变技术等,使基因或突变或缺失或失活来鉴别目标基因相关的表型;或在基因组内定位表达目的基因,或通过酵母双杂交系统等来研究基因的表达及其功能,或通过转基因方法来研究基因的功能,以及通过基因芯片等技术来研究和预测基因的功能。

(4) 比较基因组学研究

通过基因组比较作图(comparative mapping)可以揭示染色体或染色体片段上同线性(synteny)或共线性(colinearity)的存在,从而对不同物种的基因组结构及基因组进化历程进行精细分析。比较作图就是利用共同的分子标记(主要是 cDNA标记及基因克隆)在相关物种中进行物理或遗传作图,比较这些标记在不同物种基因组中的分布情况。

(5) 基因定位

借助基因组图谱,可使基因定位在精度、深度、广度等方面有极大的提高。基因定位有多种方法,其中全基因组扫描法(genome wide scanning)是一种新的基因定位方法。该法在对人类许多与疾病相关基因的定位中产生了很好的效果。其原理是利用人类基因组内存在的大量的短串联重复序列(STR),即微卫星标记。通过 PCR利用特定序列的引物可将某条染色体上特定位置的 STR扩增出来进行分析,探测在它的周围是否存在疾病相连锁的相关基因。该方法并不能直接寻找到具体的相关基因,而是通过研究均匀分布于整个基因组的微卫星标记来间接筛查其相关的基因。在得到阳性结果后,又可在这些阳性位置附近再加密微卫星标记,用同样的方法来确定哪一个与疾病连锁的可能性最大,这样在不断缩小分析范围后,相关基因定位的范围也越来越明确,当目标基

因的定位精确度达到约 1 cM(即 100万碱基对)时,采用或是直接在该区域进行大规模测序,找到候选基因后在该疾病及正常人群中进行突变筛选;或选择与该疾病表型呈连锁关系的微卫星位点作为标记筛选入基因组文库,以阳性克隆的 DNA为探针筛选 cDNA文库,对阳性 cDNA克隆进行测序并进行致病基因突变最终筛查出该疾病的相关基因。

18.4 基因组 DNA 大片段文库的构建

18.4.1 酵母人工染色体文库

完整的基因组文库(genomic library)的构建是进行基因组作图的先决条件。目前构建大片段基因组文库主要有 YAC、BAC和 PAC载体系统,它们的区别如表 18-2:

表 18-2 基因组克隆所采用的高容量载体

载体	容量/kb	复制子	宿主	导入细胞方式	重组子的筛选	克隆 DNA的回收
YAC	250~400	ARS	酵母	转化	ade2	脉冲场凝胶
BAC	120~300	F	大肠杆菌	电穿孔	α互补	碱提取
PAC	130~150	P1	大肠杆菌	电穿孔	sacB	碱提取

酵母人工染色体(yeast artificial chromosome,YAC)由 3种功能单位(元件)组成:着丝粒 DNA序列(centromere DNA sequence, CEN)是染色体有丝分裂和减数分裂并保证向两极运动的必需组分;端粒 DNA序列(telomere DNA sequence, TEL)能与端粒酶结合,完成染色体末端复制,防止染色体末端融合及降解;自主复制序列(autonomously replicating sequence, ARS)确保染色体在细胞周期中能够自我复制,维持染色体在世代传递中的连续性。

ARS、CEN和 TEL这 3种关键序列是真核生物染色体自主复制功能基本的,也是必需的结构基础。所有真核染色体中都含有这 3种功能元件。用重组技术将这 3种元件分别从染色体上分离出来,并按顺序重组就构成了所谓的人造微小染色体(artificial minichromsome),将这种线状的小染色体转化进入酵母细胞后可以进行正常的复制和有丝分裂。称这种"人造微小染色体"为酵母人工染色体即 YAC。人工染色体在细胞周期中可以正常复制、配对、分离。这种小染色体经过改造,组装上作为载体所必需的其他功能序列,例如多克隆位点(SmaⅠ、*Not*、*Eco*RⅠ等)和选择标记(Sup-4、URA3和 TRP等)等就可以成为人工染色体载体。

图 18-10显示了用 pYAC4载体构建植物 YAC文库的基本步骤。首先 YAC载体用 BamHⅠ和 EcoRⅠ消化和去磷酸化。同时,植物基因组 DNA用 *Eco*RⅠ部分消化,制备 Mb级大片段 DNA,利用脉冲场凝胶电泳(PFGE)进行 1或 2次选择,PFGE能分辨大至 12 000 kb的 DNA分子,对 YAC克隆片段(100~1 000 kb)的分析十分适宜。然后将 YAC载体与一定大小的基因组 DNA片段重组,将得到的重组体转化酵母宿主,依据尿嘧啶原养型和红色进行转化体筛选,红色菌落为候选克隆。将红色克隆进一步转移到缺乏色氨酸和尿嘧啶的培养基中以证实转化体中有 YAC两臂的存在。1983年 Murray等首次构建了长度为 55 kb的酵母人工染色体,1987年 Burke等构建了能克隆大片段的人体 DNA分子的载体,并构建了第一个 YAC文库。至今已对很多种真核生物构建了 YAC文库。

YAC的优点体现在:YAC载体具有灵活性,它可以作为一个质粒在大肠杆菌中增殖,因而可以用于亚克隆进行序列分析。当载体臂与基因组 DNA大片段连接导入酵母细胞后,又可以作为非人工的或天然的染色体在酵母细胞中复制。YAC的最大优点是克隆容量大,可插入 100~2 000 kb的

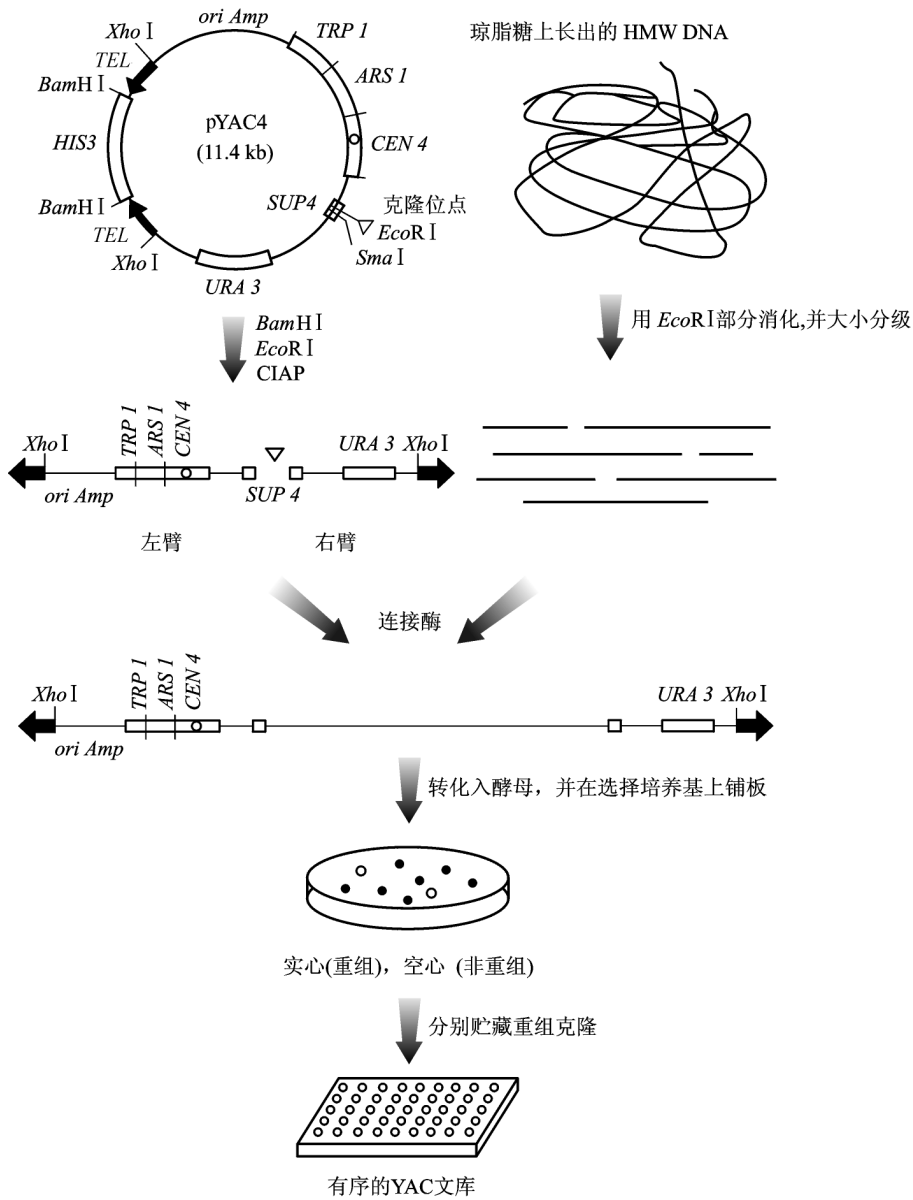


图 18-10 YAC文库构建的基本步骤

(引自曹仪植，2002)

外源 DNA片段，利用不多的克隆就可以包含特定的基因组全部序列，这样可以保持基因组特定序列的完整性，有利于通过重叠片段克隆绘制基因组物理图谱以及进行基因定位克隆等。

YAC的不足是：① 某些克隆稳定性较差，存在序列重排和插入丢失现象。② 嵌合现象严重，有5%～50%的 YAC克隆存在嵌合现象。在同一 YAC克隆中嵌合两个不连续的大片段，来自不同染色体或同一染色体的不连续的区域，这些克隆很不适合测序和作图。③ 插入 DNA片段的分离与纯化困难，因为转化细胞中 YAC的分子结构与酵母天然染色体的分子结构有一定的相似性。④ 转化效率低。

18.4.2 细菌人工染色体文库

细菌人工染色体(bacterial artificial chromosome,BAC)是以细菌 F因子为基础，人工构建的大容量细菌克隆体系。细菌人工染色体其结构为环状的 DNA分子。例如，由 Shizuya等(1992)所构建的

BAC 克隆载体 pBAC108L如图 18-11所示,其复制起点为大肠杆菌 F因子的复制起点,以氯霉素抗性基因为选择标记,在 F质粒 pMBO131基础上构建而成。为便于克隆选择,Kim等(1992)在 pFOS1质粒克隆位点引入 pGEM32的 lacZ基因片段,构建成 pBAC-lac载体,可进行蓝白斑筛选。

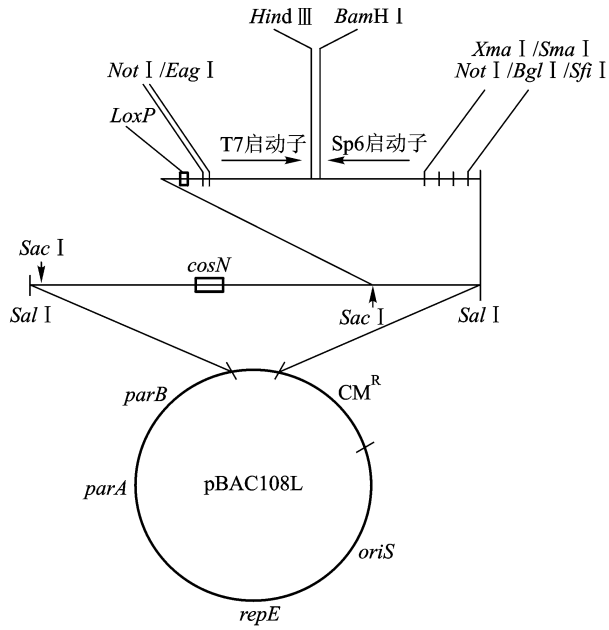


图 18-11 BAC克隆载体 pBAC108L结构(引自朱玉贤等,2002)
pBAC108L来自细菌的一个小型 F质粒,其中 oriS, repE控制了质粒的复制起始,parB和 parA控制了拷贝数

BAC 文库构建的基本思路与 YAC 文库相似,只是需要采用 BAC 载体和大肠杆菌宿主系统,例如,可用 pBeloBAC11作为载体,以大肠杆菌 DH10B作为宿主来构建 BAC 文库。首先用限制酶 Hind Ⅲ分别对载体和基因组 DNA进行适当消化,然后将一定大小的基因组 DNA片段与载体重组,经电穿孔法转化大肠杆菌。通过蓝/白颜色进行筛选,白色菌落为阳性克隆。

虽然 BAC载体的插入片段长度一般为 100~350 kb,小于 YAC载体的克隆片段,但是其优势在于:BAC在大肠杆菌宿主中具有相当的稳定性;BAC 文库中没有嵌合现象;采用电激法转化大肠杆菌,转化效率比 YAC的转化效率高 10~100倍;由于 DNA以环状超螺旋形式存在于大肠杆菌,BAC克隆易于操作和 DNA提取;采用菌落原位杂交方法,使 BAC文库筛选方便;BAC载体上的插入片段可以直接进行测序以获得末端序列。

18.4.3 P1噬菌体衍生人工染色体文库

P1噬菌体衍生人工染色体(P1-derived artificial chromosome)是将 BAC 和 P1噬菌体克隆体系(P1-clone)的优点结合起来所产生的克隆体系。P1噬菌体作为载体,可克隆 95 kb长的 DNA片段。该体系将克隆的 DNA及载体包装λ噬菌体颗粒后注入大肠杆菌中,其 DNA通过 P1 loxP重组位点和在受体菌中表达的 P1 Cre重组酶的作用而环化,形成环形质粒。P1-clone载体含有卡那霉素抗性基因,便于筛选。该环形质粒在宿主细胞中以单拷贝存在,可以避免因多拷贝所造成的克隆的不稳定性。

在 P1克隆及 BAC基础上,Loannou等(1994)结合 F因子及 P1噬菌体的特点构建了 PAC载体及 PAC文库,PAC载体的插入片段平均为 130~150 kb。PAC克隆通过电激法转化大肠杆菌,以单拷贝形成稳定遗传,没有嵌合体。由于 PAC无嵌合现象,可以作为 YAC连续克隆文库的重要补充。例如在日本水稻基因组计划(Rice Genome Program, RGP)中,为了克服 YAC克隆的局限性,又以

PAC为载体构建了水稻 Nipponbare(日本晴)基因组文库,此文库由 72 000个 Sau3AⅠ酶切克隆组成,平均插入片段长 120 kb,覆盖水稻基因组的 16倍。

高质量的大片段基因组文库的建成极大地方便了基因、基因组的研究工作。在大片段基因组文库基础上开展了大规模的物理作图、测序、图位克隆基因及基因转化、分子标记的发掘、着丝粒的研究、基因的定位及比较基因组的研究工作等。

18.5 比较基因组学和功能基因组学研究

18.5.1 比较基因组学

比较基因组学(comparative genomics)是一门通过运用数理理论和相应计算机程序,对不同物种的基因组进行比较分析来研究基因组大小和基因数量、基因排列顺序、编码序列与非编码序列的长度、数量及特征以及物种进化关系等生物学问题的学科。最重要也是最能体现比较基因组学的学科特点的是不同生物间全基因组的核苷酸序列的整体比较。

随着人类基因组计划的完成和 676种生物全基因组序列的测定以及 3 109种基因组测序的即将完成(GOLD, Genome Online Database, 2007-10-30),爆炸式增加的基因组数据需要进行比较,只有进行基因组的比较分析,我们才能认识蕴藏其中的遗传信息或了解这种序列和表型的关系,获取更多有效信息。这是由于生物在进化上是相互关联的,对一种生物的研究可以为其他生物提供有价值的信息。比较基因组学的重要作用之一是它能根据对一种生物相关基因的认识来理解、诠释甚至克隆分离另一种生物的基因。

① 通过比较基因组学的研究加深了人们对基因组结构和基因功能的认识　例如对基因组大小和基因数目的比较(表 18-3),使我们更进一步加深了对 C值悖理的理解。基因组大小与遗传复杂性并非线性相关,果蝇和线虫的基因组大 8~9倍于酵母,但其基因数目却仅为酵母两倍多。黑青斑河豚基因数目比人类多,但它的基因组大小却仅为人类的 1/10。这是因为编码区只代表总 DNA的很小一部分,非编码序列的存在使我们不能从基因组总大小来推测其基因数目。一般大的基因组中有非编码 DNA的大量增加,而编码序列的大小及数目的差别不如非编码序列显著。同一基因组中基因的分布也是不均的,甚至某些区段的基因密度远远高于全基因组中的平均密度,形成基因岛(gene island)。已经发现一些基因岛中的基因群通常具有功能上的相关性,可能是造成这些基因紧密连锁的重要原因。

表 18-3　部分物种的基因组大小与基因数目

物　种	碱基对/bp	预测基因数	备　注
ϕX174	5 386	10	E. coli的病毒
米米病毒 Mimivirus	1 181 404	1 262	迄今基因组最大的病毒
甲烷球菌 Methanococcus jannaschii	1 664 970	1 783	第一个被测序的古细菌
超耐温寄生虫 Nanoarchaeum equitans	490 885	552	迄今发现的最小基因组
生殖道支原体 Mycoplasma genitalium	580 073	485	最小的生物体
大肠杆菌 E. coli	4 639 221	4 377	4 290编码蛋白,其余编码 RNA

续表

物 种	碱基对/bp	预测基因数	备 注
蓝色链霉菌 Streptomyces coelicolor	6 667 507	7 842	用来生产全世界 2/3的抗生素药以及其他药物的放线菌
裂殖酵母 Schizosaccharomyces pombe	12 462 637	4 929	
酿酒酵母 Saccharomyces cerevisiae	12 495 682	5 770	
拟南芥 Arabidopsis thaliana	115 409 949	28 000	开花被子植物
水稻 Oryza sativa	3.9×10^8	37 544	
秀丽隐杆线虫 Caenorhabditis elegans	100 258 171	19 427	第一个被测序的多细胞真核生物
黑腹果蝇 Drosophila melanogaster	122 653 977	13 379	
疟蚊 Anopheles gambiae	278 244 063	13 683	携带疟疾寄生虫蚊虫
黑青斑河豚 Tetraodon nigroviridis	3.42×10^8	27 918	
人 Humo spaiens	3.3×10^9	约 25 000	
大鼠 Mouse	2.7×10^9	25 000左右	
玉米 Zea mays	4.5×10^9	?	
蝾螈 Amphiumaspp.	7.65×10^{10}	?	
两栖动物 Amphibians	$10^9\sim10^{11}$	?	

注:表中数据来源于 http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/G/GenomeSizes.html并经过修改。表中数据截止日期为 2007年 7月 13日。

② 比较基因组学的研究表明在亲缘关系较近的物种中存在保守的连锁群或染色体“板块”结构 人类与小鼠可能有 181个不同的保守连锁群,平均大小为 9 cM 左右。如小鼠的 11号染色体很可能由多个祖先染色体片段“拼凑”而成,而这些片段在人类基因组中分布于 22q、7p、2p、5q、17p和17q等染色体区域(图 18-12)。有人将人类的基因与小鼠的基因进行比较,发现有 1 886个基因同源。而大鼠的 775个基因中,有 542个与小鼠同源,530个与人类同源。2002年对小鼠物理图谱的绘制证实小鼠与人的基因组存在着很高的相似性。研究发现,小鼠基因组共有约 27亿个碱基对,比人类少 15%,但其包含的基因数目约为 3万个,与对人类基因数的最新估计非常接近。人类和小鼠约 40%的基因组完全相同,80%的人类基因组可以在小鼠身上找到对应的基因组部分。大鼠基因组测序结果(2004)也揭示,大鼠染色体上有 27.5亿个碱基对,与人类染色体上碱基对相当接近,大鼠约包含了 25 000个基因,其中 90%的基因与小鼠以及人类的基因相匹配。

近缘物种如人类和黑猩猩相似性高达 98.7%(Enard等,2002);远缘物种如酿酒酵母 S. cerevisiae 也有多至 30%的基因与人类基因相近。借助这种相似性,我们可以通过对不同物种的基因组的比较追踪它们在进化长河中的共同起源。除了基因组成的相似性,在不同基因组中基因在染色体片段上排列顺序也具有一致性,即共线性,更能体现基因组的共同起源,例如禾谷类作物之间具有广泛的共线性,加之得到一些相应的遗传图和物理图,可以将某作物的共线性区域的标记作为相关作物进行精细定位和鉴定候选基因。作为模式作物水稻的小基因组为其他禾谷类基因组研究提供了基础,包括鉴定高效直系基因、调控区域、基因功能和便利其他禾谷类基因组的测序。Goff等报道几乎每个禾谷类蛋白质与水稻都有一个相关基因,80%~90%禾谷类基因与水稻有同源性。禾谷类作物中大部分基因是保守的,它们的表型差异是由于少数不同基因或相似基因的功能差异引起的。通过基因组共

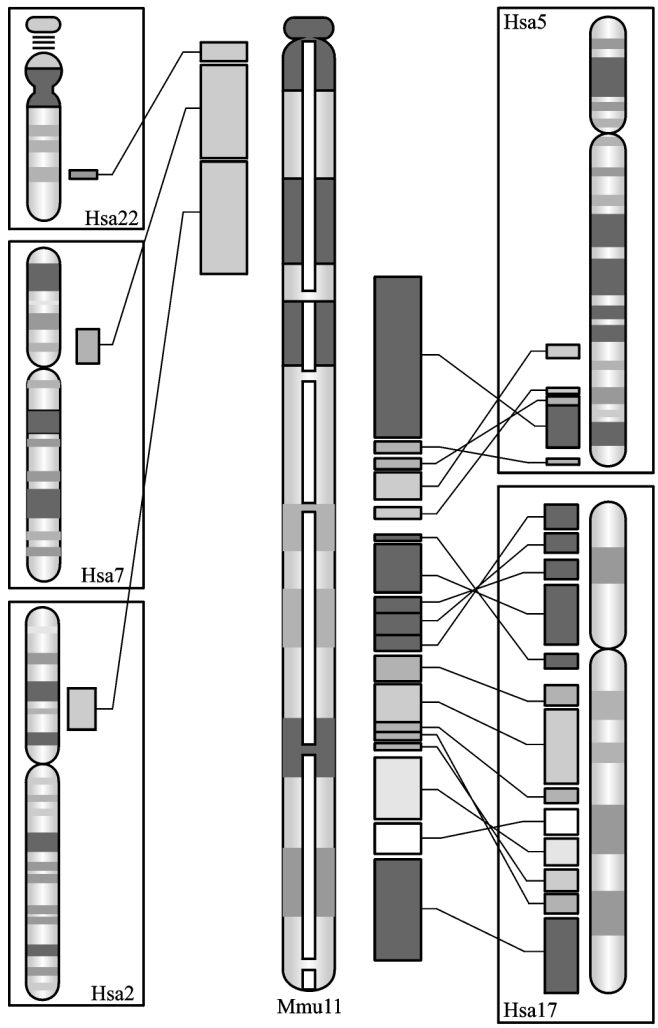


图 18-12 小鼠 11号染色体和人染色体保守区段的比较(引自 Simon等, 2002)
小鼠的 11号染色体,很可能由多个祖先染色体片段"拼凑"而成,而这些片段在人类基因组中分布于 22q、7p、2p、5q、17p和 17q等染色体区域

线性比较,有利于鉴定其他禾谷类定位的性状与水稻相关的基因。

从上述事例可见,基因共线性既可出现在不同基因组的对应区段,也可以出现在同一基因组内部的不同染色体区段。宏观水平的共线性系指遗传连锁图上锚定标记排列次序的一致性,而微观共线性则指物理图上基因顺序的一致性。在进化过程中,基因共线性被转座、染色体片段的插入、缺失、重排或加倍等各种因素所破坏,使得物种间进化距离越远,基因共线性越差,如水稻与小麦基因组的共线性远大于水稻与拟南芥间的。对于保守性高的区段,共线性的破坏往往会产生致命的结果。近缘物种之间较为精确的微观共线性使得在基因组较小的模式生物中分离被精确定位在大基因组中的基因成为可能,如致病基因、重要功能基因等。但微观共线性往往被各种各样的因素破坏,只在相当有限的距离内保持,因而有的基因如植物中与抗病基因相关的区段由于共线性较低而很难实现跨种分离。

18.5.2 功能基因组学

随着多种生物全基因组序列的获得,基因组研究正在从结构基因组学(structural genomics)转向功能基因组学(functional genomics)的整体研究。在功能基因组学的研究中通常运用高通量技术

(high-throuput techniques),如 DNA微阵列(DNA microarrays),反求遗传学技术如基因打靶(gene targeting),转基因(transgene)以及反义 mRNA(antisense mRNA)和 RNA干扰(RNA interference,RNAi)等技术来系统地分析基因功能及基因间相互作用、基因组的时空表达以及发现和寻找新基因等。

DNA微阵列技术又称 DNA芯片(DNA chips)或基因芯片(gene chips)技术,它通过将对应于不同基因或 cDNA的 DNA片段或寡聚核苷酸点样于微芯片上形成高密度的矩阵,与荧光标记的总 mRNA进行杂交,然后通过激光共聚焦扫描检测并运用计算机软件对杂交信号进行自动化定性定量分析,具有高通量、实时、灵敏、准确等特点。

基因打靶是指通过转染的 DNA序列与细胞内同源的基因组序列(靶序列)之间进行同源重组,以改变靶序列来研究其结构和功能或进行基因治疗的技术。基因打靶技术包括基因敲除(knock-out)和基因敲入(knock-in),前者是用无功能的 DNA序列与靶序列重组,破坏原基因组的遗传功能,后者是用有功能的 DNA序列与受到破坏的靶序列重组使其恢复遗传功能。基因打靶技术是一种从基因到表型的新研究方法,属于反求遗传学范畴。

反义 mRNA技术是通过向细胞导入一段与特定编码 mRNA互补的非编码 RNA链,使其与该段 mRNA特异性结合而定向阻抑靶基因表达的技术。这一技术的成熟,为功能基因组学的研究和基因治疗提供了新的思路。在反义 mRNA技术的研究过程中,科学家们意外发现导入正义 mRNA(sense mRNA)与导入反义 mRNA具有等效的阻抑效应。而更令人吃惊的是如果导入相应双链 RNA(dsRNA),其阻抑效应比导入任一单链 RNA强十倍以上,dsRNA若经纯化则阻抑效应更强。这种双链 RNA特异性地作用于与其序列配对的基因而抑制其表达的现象称为 RNA干扰。RNA干扰技术作为一种新的定点基因敲落(gene knockdown)技术,赋予了功能基因组学、基因治疗等全新的思路,堪称生命科学近年来的革命性的突破,因而发现 RNA干扰机制的两位美国科学家安德鲁·法尔和克雷格·梅洛荣获 2006年诺贝尔生理学或医学奖。可见该项成果的重大科学意义(详见第 15章)。

18.5.3 蛋白质组学

2003年 4月 15日美、英、德、日、法、中 6国共同宣布人类基因组序列图完成。至此,生命科学研究进入了后基因组时代(post genome era)。后基因组时代的基本任务是从整体水平上对生物进行功能研究,从而导致了蛋白质组学(proteomics)的诞生。

蛋白质组学是研究细胞内全部蛋白质的组成、结构与功能的学科。蛋白质组(proteome)是指由一个基因组所表达的全部相应蛋白质。因此,蛋白质组与基因组相对应,也是一个整体的概念,是基因组表达的全部蛋白质。两者的根本区别在于:一个有机体只有一个确定的基因组,组成该有机体的所有不同细胞基因组都相同;但基因组内各个基因表达的条件和表达的程度则随时间、地点和环境条件的不同而不同,因而它们表达的模式,即表达产物的种类和数量随时间、地点和环境条件也是不同的。所以,蛋白质组是一个动态的概念。由于蛋白质的种类和数量总是处在一个新陈代谢的动态过程中,同一细胞的不同时期,其所表达的蛋白质是不同的。同一细胞在不同的生长发育阶段和生长条件下(正常、疾病或外界环境刺激),所表达的蛋白质也是不同的。正是这种复杂的表达模式表现了各种复杂的生命活动。DNA序列并不能提供这些信息,所以仅用核酸语言不足以描述整个生命活动。再加上由于基因剪接、蛋白质翻译后修饰和蛋白质剪接等,使从基因到蛋白质的遗传信息的表现规律变得更加复杂,不再是经典的一个基因一个蛋白对应关系的理念,一个基因可以表达的蛋白质的数目可能远大于一。如现在预测人类基因总数为 20 000~25 000,但是已发现的蛋白质总数就已高达 200 000种。

蛋白质组学是蛋白质组概念的延伸,是在整体上研究细胞内蛋白质组的结构与功能及其活动规律的科学,包括分析全部蛋白质组所有组成成分及它们的数量,确定各种组分所在的空间位置、修饰方法、互作机制、生物活性和特定功能等。与传统对单一蛋白研究相比,蛋白质组学研究所采用的是高通量和大规模的研究手段。

双向电泳(two-dimensional electrophoresis, 2-DE)技术、计算机图像分析与大规模数据处理技术以及质谱(mass spectrometry, MS)技术是蛋白质组学研究的三大基本支撑技术。双向凝胶电泳是目前分离蛋白质最有效的方法,是蛋白质组技术的核心。在双向电泳中,第一相是以蛋白质的电荷差异为基础进行分离的等电聚焦(isoelectric focusing, IEF),第二相是以蛋白质分子量差异为基础的 SDS-PAGE (sodium dodecyl sulfate-polyacrylamide gel electrophoresis)。近年由于第一相采用 IPG(immobilized pH gradients)胶条,分辨率可达 0.001 pH单位,大大提高了分辨率和重复性。在目前情况下双向电泳的一块胶板(16 cm×20 cm)可分出 3 000~4 000个,甚至 10 000个可检测的蛋白斑点。通常在银染条件下,灵敏度可以达到 10^{-18}~10^{-15} mol水平,基本满足了对蛋白质组分析的要求。双向电泳分离得到的图谱经扫描输入计算机,数字化处理,确定每个蛋白质点的等电点和相对分子质量,并进行图谱间的比较。提供蛋白质鉴定的初步信息,一旦蛋白质点经过分析鉴定,就可以建立起蛋白质组数据库。

蛋白质组分析技术有多种选择,质谱分析以其快速、准确、灵敏而成为蛋白质组的主要鉴定分析技术。目前在蛋白质组鉴定分析中以电喷雾离子化(electrospray ionization, ESI)质谱仪和介质辅助的激光解吸/离子化飞行时间质谱(matrix-assisted laser desorption/ionization time-of-flight mass spectrometry, MALDI-TOF-MS)技术应用最为广泛,这是因为这两种质谱仪在离子化和质量分析方式上适应于蛋白质大分子的性质。它们都是"软电离"方法,即样品分子电离时,保留整个分子的完整性,不会形成碎片离子。现在质谱技术可在几分钟内完成一个蛋白质整个肽谱的鉴定,得到完整的蛋白质全序列,经计算机数据库查询,可以很快地鉴定蛋白质。所谓"肽质量指纹图谱"(peptide mass fingerprinting, PMF)就是首先用蛋白酶部分消化 2-DE凝胶上的蛋白质,获得多肽,用质谱分析后得到的一套多肽分子量质谱。由于每个蛋白酶有相对固定的酶切位点,因此不同的蛋白质消化后获得的肽链长度及数目、肽链的氨基酸组成是不同的,所以各蛋白质的多肽分子量谱是特异的,称之为"肽质量指纹图谱"。

目前,在蛋白质组的研究中有很多新的发展,例如 ESI质谱可以很方便地与高压液相色谱(HPLC)、毛细管电泳(CE)等分离仪器在线联用。运用 X射线衍射晶体分析(X-ray crystallography)和核磁共振(nuclear magnetic resonance, NMR)分析蛋白质或多肽的三维结构;运用亲和色谱(affinity chromatography)、酵母双杂交(yeast two hybridization)、荧光共振能量传递(fluorescence resonance energy transfer, FRET)和表面胞质团共振分析技术(surface plasmon resonance, SPR)分析蛋白质-蛋白质、蛋白质-DNA相互作用等。

蛋白质组研究中最重要的一个方面是数据库的建立,包括蛋白质序列数据库、质谱数据库、双向电泳数据库等等。著名的蛋白质组数据库有 Swiss-Prot and TrEMBL(http://us.expasy.org/sprot/), PIR (International Protein Sequence Database, http://www.pir.org/), Pfam(Protein families database of alignments and HMMs, http://www.sanger.ac.uk/Software/Pfam/)和 PDB(Protein Data Bank, http://www.rcsb.org/pdb/)等。

最使人振奋的是人类蛋白质组计划(Human Proteome Project)是继人类基因组计划之后最大规模的国际性科技工程,也是 21世纪第一个重大国际合作计划。其首批行动计划包括"人类血浆蛋白质组计划"和"人类肝脏蛋白质组计划"(Human Liver Proteome Project, HLPP)。其中 HLPP是第一个人类组织/器官蛋白质组计划,也将是我国第一次领导的重大国际协作计划。拟在 2010年前后完成。HLPP 的科学目标是:构建蛋白质表达全谱和蛋白质修饰谱,绘制蛋白质相互作用连锁图和细胞定位图,建立符合国际标准的肝脏标本库,发展规模化抗体制备技术并建立肝脏蛋白质抗体库,建立完整的肝脏蛋白质组数据库,寻找药物作用靶点和探索肝脏疾病防治诊治的新思路和新方案。

18.5.4 生物信息学与后基因组学

生物信息学(bioinformatics)是运用计算机技术和信息技术开发新的算法和统计方法,对生物实验

数据进行分析，确定数据所含有的生物学意义，并开发新的数据分析工具以实现对各种信息的获取和管理的学科。生物信息学产生于 20世纪 80年代，它是生物学与计算机科学以及应用数学等学科相互交叉而形成的一门新兴学科。生物信息学以计算机为主要工具，开发各种软件，对日益增长的DNA和蛋白质的序列和结构等相关信息和生物学实验数据进行收集、储存、发行、提取、加工、分析和研究，同时建立理论模型，指导实验研究。它由数据库、计算机网络和应用软件三大部分构成，在基因组计划中发挥不可替代的作用。

自动化测序技术和基因组计划催生了大规模的 DNA、蛋白质序列实验数据，使得 GenBank, EMBL (European Molecular Biology Laboratory nucleotide sequence database)，DDBJ (DNA Data Bank of Japan)，PIR (Protein Information Resource) 和 Swiss-Prot等数据库也如雨后春笋般呈现眼前；在基因组计划实施过程中和实施后产生的结构基因组学、功能基因组学、蛋白组学、转录组学(transcriptomics)、代谢组学(metabolomics)等使得分子生物学数据越来越多样化。没有强大的数据存取、分析技术，它们将是一堆垃圾。随着计算机数据库技术的长足的发展，特别是互联网的出现，信息技术、信息基础设施的各个层面如数据库技术、数据的访存接口技术、出版技术及软件的开发与使用都发生了革命性变化，这为生物信息学的产生和发展提供了强有力的动力。

目前生物信息学主要被应用在以下一些方面：大规模基因组测序中的信息处理与分析；基因组相关信息的收集、存取与管理；新基因和新 SNP的发现与鉴定，在“国际人类基因组‘单体型图’计划”实施中，进一步确立世界上主要人群基因组的遗传变异图谱；基因识别及编码序列、非编码序列分析；序列比对，遗传密码的起源和生物分子进化；完整基因组的比较研究；大规模基因功能表达谱的分析；生物大分子的结构模拟与药物设计和药物开发；生物信息学分析方法的研究；建立国家生物医学数据库与服务系统等。

由于 DNA与蛋白质的不完全对应性，因此提出了转录组的概念，指特定环境下某生物一个或一类细胞中由一套基因组转录产生的全套 RNA分子。对于某种特定的生物，其基因组是固定不变的，而转录组却是变化的。转录组学是研究基因组转录产生的全部转录物的种类、结构和功能的学科。尽管 RNA的表达水平与蛋白质实际表达水平存在差异，但它仍是走近生命活动真实过程的重要环节。我们将特定条件下、特定时间点的特定细胞、组织、体液、器官或生物体代谢物——基因表达终产物的总和称为代谢组(metabolome)，而代谢组学即研究代谢组的学科。细胞内许多生命活动如细胞信号转导、能量传递、细胞通信等都发生在代谢物层面且受代谢物调控，因而通过代谢物的种类和浓度可以分析细胞瞬时的生理生化功能状态，因而代谢谱的研究是极其直观而重要的。其主要技术手段是核磁共振、质谱、色谱(high pressure liquid chromatography, HPLC)，其中以核磁共振为主。在完成基因组图谱构建以及全部序列测定的基础上继基因组学后产生的功能基因组学、蛋白质组学、转录组学、代谢组学，以及其他旨在全基因组水平研究基因功能、相互关系及调控机制为主要内容的学科，统称为后基因组学。

知识窗 **18-2**

解码人类基因组的蓝图——ENCODE计划

The Encyclopedia of DNA Elements Project即“DNA元件百科全书计划”，简称 ENCODE计划，是在完成人类基因组全序列测定后的 2003年 9月由美国国立人类基因组研究所(National Human Genome Research Institute, NHGRI)组织的又一个重大的国际合作计划，其目的是解码基因组的蓝图，鉴定人类基因组中包括基因、启动子、增强子、抑制子(repressor)/沉默子(silencer)、内含子、复制原点、复制终止位点(sites of replication termination)、转录因子结合位点(transcription factor binding sites)、甲基化位点(methylation sites)、DNaseⅠ高敏感位点(DNaseⅠ hypersensitive sites)、染色质修饰(chromatin modification)和还未知功能的多个物种的保守序列等在内的所有功能元件。ENCODE计划中提出的

每一类元件都是已经被发现过的，所不同的是现在要在全基因组的范围内进行系统的研究。

ENCODE计划的实施分为 3个阶段：试点阶段（a pilot phase）、技术发展阶段（a technology development phase）和生产阶段（a production phase）。在试点阶段中，首先集中对按一定标准选择的人类基因组中约 1%的序列，共长 30 Mb，分布在不同染色体上的 44个靶区（ENCODE targets）序列，例如α-和β-珠蛋白基因簇、囊性纤维化跨膜传导调节蛋白基因 CFTR等进行解码注释，并评估现有各种鉴别基因组元件的策略方法正确与否，确定一套有效程序，促进发展高通量更准确的基因组功能元件的鉴别技术方法。而技术发展阶段是与试点阶段同时进行的，其主要目标是设计发明新的实验方法和计算方法来改进鉴别新的功能序列和发现新的基因组功能元件的能力。生产阶段是利用以上两个阶段所建立的成熟的技术和方法，对人类基因组其余 99%的基因组序列进行高效益的全面分析（Science，2004，306：636—640）。

ENCODE计划联合体（Consortium）由 11个国家 80家科研机构 35个小组的研究人员组成。在 2007年 6月分别在 Nature和 Genome Research上报道了他们 4年来研究的主要成果：研究表明人类基因组的大多数 DNA都会被转录成 RNA，即基因组中的碱基大多会出现在原始转录物中，包括非蛋白编码转录物和重叠转录物等，因此，人类基因组实际上是一个非常复杂的网络；对转录调控的研究，确定了许多以前不为人知的 DNA转录起始位点，在基因之外的调控区域新发现了 4 491个转录起始位点，对转录起始位点有了新的认识，包括它与特异性调控序列、组蛋白修饰和染色质可接近性等之间的联系。推翻了传统观点的认识，调控区域也有可能位于 DNA转录起始位点的下游；进一步认识了染色质结构，以及它与 DNA复制、转录调控之间相互关系的复杂性；通过哺乳动物种间和种内的序列比较，对人类基因组在功能与进化上又有了新的认识，研究表明大约一半人类基因组中的功能元件在进化过程中不会受到很大限制。

以上这些新的成果挑战了关于人类基因组的传统理论，说明人类基因蓝图不是由孤立的基因和大量“垃圾 DNA片段”组成的，而是一个复杂的网络系统，单个基因、调控元件以及与编码蛋白无关的其他类型的 DNA序列一道，以交互重叠的方式相互作用，共同控制着人类的生理活动。

与人类基因组计划相比，“ENCODE计划”的显著特点是：采用综合性研究策略，重视新技术的研究与开发，将计划向学术界和公司开放。ENCODE计划的实施为进一步认识人类基因组的功能蓝图开辟了道路。

ENCODE数据的主要门户网址：www.genome.ucsc.edu/ENCODE

ENCODE原始数据的门户网址：http://www.ncbi.nlm.nih.gov/projects/geo/info/ENCODE.html和 http://www.ebi.ac.uk/arrayexpress/

ENCODE联合体的信息网址：http://www.genome.gov/ENCODE

18.6 基因组的进化

18.6.1 基因组进化的分子基础

生命多样性的表现特点之一是遗传的多样性。遗传变异在自然界中是十分普遍的现象。进化的基础是遗传的变异，遗传变异的存在是进化的必需条件。遗传物质的改变主要包括基因突变、遗传重组和染色体畸变，是进化的基础。遗传物质改变所发生的分子事件是基因组进化的分子基础（详见第 12和 13章）。

基因组随时间而进化，由突变引起小规模序列改变，而重组则使其产生 DNA序列大规模重排。

重组和突变是遗传变异的两种截然不同的机制,重组是已经存在的信息重排,而突变是在基因组中导入新的信息。在真核生物中重组涉及真核细胞减数分裂时同源染色体之间的交换。随后又在分子水平观测到细菌的接合、转导和转化等都与外源 DNA 的重组有关。重组可直接改变基因组的遗传组成。重组主要包括同源重组、位点专一重组、转座和异常重组这四种方式,其分子机制是互不相同的(详见第 6、7、8和 11章)。

18.6.2 基因组的起源

对基因组的起源与进化的研究,是探索生命起源和进化的重要组成部分。DNA 分子既是遗传信息的载体,又能进行自我复制,被认为是生命大分子进化的起点。而现存地球上所有生物的基因组可分为核基因组(包括真核和原核生物)、线粒体基因组和叶绿体基因组。从进化上追踪寻源,这 3类基因组都应起源于一种原始基因组。那么原始基因、原始生命从何而来?

(1) RNA 起源假说

1982年 Cech发现四膜虫的 rRNA前体可自我剪接生成成熟的 rRNA,表明 RNA 具有催化活性,改变了传统的只有酶蛋白具有催化活性的观念。随后,经过一系列的实验发现具酶活性的 RNA(RNA 催化剂)还具有核苷酸转移酶、磷酸二酯酶、RNA 限制性内切酶、磷酸转移酶和磷酸酯酶等多种催化活性。虽然 DNA 能编码氨基酸和自我复制,但所有的反应必须依赖蛋白质的参与才能进行;酶蛋白虽能催化生化反应,但它不能自我复制,其合成需由多核苷酸编码。而 RNA 分子既能催化生化反应,又能自我复制,这些正是生命起源物质所要求的特点。据此,Gilbert于 1986年提出了"生命起源的 RNA 世界"说。根据该学说,人们推论生命进化似乎按以下进化路线进行:存在于原始地球上的小分子物质经过一系列化学进化首先形成一种具有自体催化能力的 RNA 催化剂系统。RNA 催化剂不仅催化自身的复制,还会催化环境中其他分子的复制。一些 RNA 可能结合氨基酸生成原始 tRNA,一些 RNA 则可能促进邻近两个 tRNA 的结合,而在 RNA 模板上催化肽链的形成,并通过某种机制合成了最初的蛋白质。这些蛋白质反过来又催化产生更多的 RNA 催化剂分子,这些分子则更多地聚合在一起。自然选择和 RNA 分子的突变使核糖在第二位脱氧生成脱氧核糖,因而形成了 DNA;或者 RNA 与蛋白质结合以 RNA 为模板合成 DNA。由于 DNA 比 RNA 更具稳定性,因此经过一代又一代的调整,最初 RNA 的编码功能由 DNA 取代,催化功能转移到蛋白质,RNA 自身则只起着遗传信息表达的中介作用。遗传信息就由 DNA 承载,并通过指导蛋白质合成来体现生物的性状。

(2) 原基因组的形成

根据以上假说及其思路,原始基因应来自 RNA,形成所谓的原基因组(protogenome)。原基因组发展成为 DNA 基因组时,最重要的一个变化是:出现了酶蛋白,以取代酶性核酸(RNA)的催化活性。最初的 DNA 基因组由许多单个 DNA 分子组成,一个 DNA 分子只确定一种蛋白质,相当于一个基因,以后在进化过程中连接起来生成最初的染色体。

最初的生命世界很有可能只是由种类单一的可自我复制的原始 RNA 大分子组成。当这些原始 RNA 大分子复制到具有一定的数量后,通过形成重复序列,不但可以使生物大分子成十倍百倍地增加其大小并有可能形成更稳定的结构,而且更重要的是获得了比单独存在的原始 RNA 大分子大得多的进化潜力和进化可能性。在这种意义上,可以认为最原始的"基因组"就是由可自我复制的原始 RNA 大分子组成的重复序列。重复序列构成的最原始的基因组的最有可能的进化产物是断裂基因。根据这种观点,原始的基因组是由重复序列和原始的断裂基因构成的。那么,这种原始的基因组又如何进化成为现代的各种基因组呢?核(类核)基因组可以比较直接地从原始基因组进化而来,而从原始基因组进化成线粒体基因组和叶绿体基因组还要经历"内共生"的过程以及一系列包括将一部分基因转移到核基因组的基因组变迁。由于内共生是发生在生物进化和基因组进化的比较原始的阶段,所以进入内共生状态初期的线粒体和叶绿体各自祖先的基因组是仍然含有重复序列和内含子的。对此,特别是对含有内含子这一点,已有不少共识。从而可以认为,核(类核)基因组、线粒体基因组和叶

绿体基因组的进化起点在结构特征上是相似的。而通过对现代核(类核)基因组、线粒体基因组和叶绿体基因组的结构分析发现,这三种基因组都存在着“小基因组”型和“大基因组”型以及与之对应的结构特点。

通过对各种生物进行基因组测序和比较,使我们对基因组的起源与进化有了更深的了解。根据基因组测序结果发现,人与果蝇之间的种间同源基因有 2 758个,人与线虫之间的种间同源基因有 2 031个,人、线虫和果蝇 3者之间共有的种间同源基因有 1 523个,这是跨越了无脊椎动物与脊椎动物之间的种间同源基因。同时,科学家通过几个完整基因组的比较,统计出维持生命活动需要的最少基因为 250个左右。同样,当我们比较鼠和人的基因组就会发现,尽管两者基因组大小和基因数目相似,但基因组的组织却差别很大。例如存在于鼠 1号染色体上的基因已分布到人的 1、2、5、6、8、13、18号 7条染色体上了。由此推测鼠与人的表型差异主要来自基因组的组织。完整基因组分析的最新结果还发现,同源基因的百分比与它们的亲缘关系紧密相关。人们试图通过比较基因的排列顺序来研究物种间的系统发生(phylogeny)关系。亲缘关系越近,基因排列顺序越相似。通过对禾本科几种重要作物作图比较,发现它们在染色体或染色体片段上遗传图之间存在广泛的同线性或共线性。水稻和玉米的基因组大小相差 9倍,染色体数目也不同,它们从共同的祖先分化已有 5千万年。水稻与小麦分化更早,有 6千万～7千万年,小麦基因组为水稻的 35倍。但它们在较大染色体片段上存在遗传标记顺序的一致。如今,对水稻、小麦、大麦、黑麦、燕麦、珍珠谷、玉米、高粱、甘蔗等进行比较作图发现,尽管它们在一些基因组中有大量重复(玉米和小麦中更明显),并通过易位和其他突变方式而进行基因重排,但仍然可识别出具有一套相同的祖先基因。

18.6.3 基因组的进化

基因组的进化表现为编码区组分 DNA与非编码区组分 DNA的进化。编码区组分 DNA的进化主要表现在新基因的获得。

(1) 基因组大小的进化

基因组大小的进化主要反映在基因组 DNA含量、基因组的结构以及基因组所含基因的数量多少等几个方面。

DNA含量:不同种的生物,甚至近缘种中单倍体的平均 DNA含量变化很大。一般而言,真核生物基因组比原核生物基因组大,DNA含量高。基因组的大小大体上反映了一种进化趋势,但却存在C值悖理,例如玉米和蝾螈的 DNA含量远高于人的。

基因组结构:原核生物基因组一般是环状,再形成拟核高级结构,DNA含量仅数百万碱基对。其基因组都是由单拷贝或低拷贝的 DNA序列组成,基因的排列比较紧密,基因数目较少,为几百个到几千个,一般以多顺反子为转录单位,较少非编码序列,更缺少非编码的重复序列和内含子,因此,在大小和结构上属于所谓的“小基因组”。真核生物的核基因组,一般是线状,再形成染色体或间期核的高级结构,比原核生物拟核基因组大许多倍,基因数目多,为几千个到几万个,且变化范围大。在结构组成上,普遍存在非编码区、重复序列以及内含子结构,重复序列和内含子的含量直接影响了基因组的大小。因此相对原核生物来说,真核生物的核基因组被称为“大基因组”。由 RNA而来的原始基因首先是由重复序列组成的,在此基础上发展成具有原始的外显子、内含子及重复序列的原始基因组。再向着小基因组与大基因组两个方向发展。如果向小基因组进化,则由于繁殖效率、空间结构等的约束,使其在进化过程中丢失重复序列和内含子结构;如果向大基因组进化,那么重复序列、内含子作为残遗结构都能继续存在,并具结构上和进化上的功能。

基因数目:从原核生物到真核生物,基因数目增加的一个先决条件是要获得新的基因。现在认为可以通过两条途径获得新的基因,一条途径是基因组中现有基因的全部或一部分实现倍增,另一条途径则是从其他物种那里获取。在基因组进化中现有基因的加倍毫无疑问是最重要的方式之一。在进化进程中,可以发生整个基因组倍增、一条染色体或一条染色体的一部分倍增以及一个基因一组基因

倍增。整个基因组倍增使基因数目突然增加,这是增加基因数目最迅速的途径,但并未改变基因组的复杂性,仅增加了基因的拷贝数。从进化角度看,更多考虑的是单个基因或一些基因的倍增,而不是整个基因组的倍增。基因的重复倍增,例如珠蛋白基因家族,在进化过程中是经常发生的,由于突变和选择,产生新的基因,因而促进基因组的进化。核苷酸顺序的变化往往会造成基因的失活,从而成为假基因。但偶尔也有一些突变会使加倍的基因具有新的功能,这些基因对生物的进化会有所贡献。单个基因以及基因群加倍在进化过程中经常出现。从其他物种的基因组中获取基因的方式,主要有转化、转导、转座等基因转移的途径。另外染色体重排和染色体畸变等,也是其中重要的途径。

(2) 基因组的分子进化

DNA序列的进化:通过研究不同物种的同源基因序列表明:有的 DNA序列比别的 DNA序列进化速度快。像前胰岛素原被加工成胰岛素时,抛出的无功能或近乎无功能的序列,其进化速度比编码有功能的蛋白质序列快。又如小鼠和人类的生长激素基因,二者的序列相差 20个核苷酸,这 20个核苷酸是通过了 6 500万年的进化过程发生歧化而形成的,其进化速率为每年每个位点取代 4×10^{-9}个核苷酸。对众多基因中核苷酸序列的研究揭示了基因不同部分以不同的速率在进化。基因的不同区域,其进化速率不同。内含子中的碱基对趋异进化速率大于外显子;在有功能的基因中,编码序列涉及同义突变的进化速率快于非同义突变的速率,因为这种改变不会影响到蛋白质的功能。编码序列中的非同义突变的进化速率最低,这些核苷酸发生突变会改变蛋白质中氨基酸的序列,因此这种突变大部分会被自然选择所淘汰。没有功能的假基因进化速率最高,如人类的珠蛋白基因中的假基因,其核苷酸的进化速率是有功能珠蛋白基因编码序列进化速率的 10倍。这是由于这些假基因不再编码蛋白,而且这些基因的改变并不影响人的适应性,所以也没有被自然选择淘汰。如果我们发现了某段序列有很高的进化速率,那么意味着此序列可能没有什么功能。

多基因家族的进化:在真核生物进化过程中,由一个祖先基因经重复而产生一系列相关的基因,这样的一组基因称为多基因家族(multigene family)。基因的进化机制主要是通过基因扩增和不均等交换形成多拷贝基因,再通过突变积累、基因重排和自然选择等因素形成多成员的基因家族或形成新的基因。例如,珠蛋白的基因家族就是一个多基因家族,在人类的第 16号染色体上有 α-珠蛋白基因簇,在第 11号染色体上有 β-珠蛋白基因簇。在动物中也有珠蛋白基因。在多种动物中几乎所有有功能的珠蛋白基因结构都相同,由 3个外显子组成,中间间隔两个内含子。但珠蛋白基因的数量和次序在各种动物中是不同的,由于所有的珠蛋白基因的结构和顺序都是相似的,因此它们存在着一个原始的珠蛋白祖先基因。肌红蛋白基因和珠蛋白基因相似。植物的豆血红蛋白基因也和珠蛋白基因相似,它们都由 3个外显子结构组成,因此可将 3个外显子结构看成它们共同的祖先。通过对哺乳

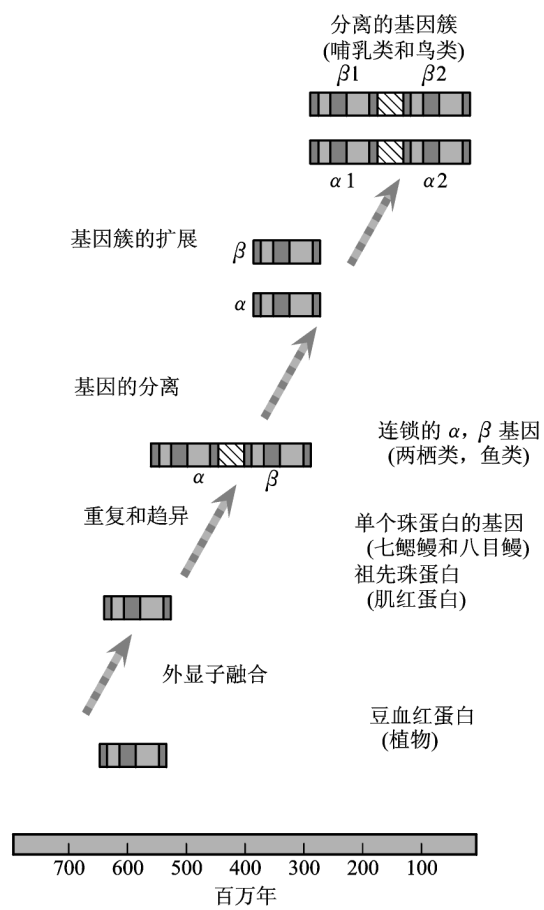


图 18-13 珠蛋白家族的进化(引自 Lewin, 2004)
所有的珠蛋白经过一系列重复、转化和突变从一个祖先基因进化而来。最先从植物豆血红蛋白基因开始,经过外显子融合变为单个珠蛋白基因(肌红蛋白基因),再经过重复和趋异,变为连锁的 α和 β基因(两栖类和鱼类),然后经过基因的分离变为 α和 β两个基因,又经过基因族的扩展,成为分离的 α和 β两个基因家族(哺乳类和鸟类)

动物肌红蛋白单个基因的研究，推测它大约在 8亿年以前和珠蛋白基因在同一进化枝上分歧。根据对各个物种的珠蛋白基因组的分析，推测珠蛋白祖先基因的可能进化途径是哺乳类和鸟类的 α基因簇和 β基因簇各自独立，而两栖类和鱼类的 α基因和 β基因仍然是连锁的。哺乳类和鸟类从爬行类祖先歧化出来的时间大约在 2亿 7千万年前，而爬行类从两栖类祖先歧化出来的时间大约在 3亿 5千万年前。这样，α基因和 β基因失去连锁的时间在 2亿 7千万至 3亿 5千万年前。α基因和 β基因分离开可能是由于转座作用造成的。某些原始圆口类和原始鱼类只有一种珠蛋白基因，它们从进化路线上歧化出来大约在 5亿年前，而所有这些珠蛋白基因，大约起源于 5亿年前，由一个祖先基因通过外显子的融合和内含子的插入反复进行最后重复和歧化而形成(图 18-13)。

此外，外显子混编、基因水平转移(gene lateral transfer)等现象都可以产生新的基因，促进基因组的进化。在外显子混编中，来自不同基因的 2个或多个外显子相互接合，或基因内部的外显子产生重复而形成新的基因结构(详见第 11章)。基因水平转移是指遗传物质从一个物种通过各种方式转移到另一个物种的基因组中。在原核生物中，转化、转导、接合和转染等机制的基因转移是频繁发生的。因此，基因水平转移对原核生物的基因组进化的贡献是相当大的。这些通过水平转移产生的外源基因在选择的作用下，经过突变积累，功能分化，可能形成新的基因。

思考题

1. 你认为人类基因组计划的意义是什么？近年基因组研究有哪些重要进展？

2. 人类基因组的结构特点有哪些？与其他模式生物基因组有哪些不同？

3. 有哪些 **DNA**分子标记？这些 **DNA**分子标记的特点和用途有哪些？

4. 何为遗传图？何为物理图？两者的关系如何？阅读相关书籍或专著了解除本章所述的人类基因组图谱之外还有一种什么图谱，其图距单位有何不同？

5. 基因组测序策略和方法有几种？各自的特点是什么？

6. 为什么说人类经典遗传图谱在人类基因组计划中利用价值不大？经典遗传学图谱与现代遗传学图谱的主要差别是什么？

7. 植物基因组遗传图谱的构建有哪些步骤？各要注意哪些事项？

8. 物理图谱有哪些类型？它们的特点是什么？基因组物理图谱的利用价值有哪些？

9. **YAC**文库、**BAC**文库和 **PAC**文库各自的特点是什么？如何利用这些文库进行基因组研究？

10. 试述比较基因组学和功能基因组学各自的研究特点在技术方法上有什么不同？有什么新进展？

11. 何为蛋白质组学？有哪些研究蛋白质组学的新技术、新方法？

12. 生物信息学在后基因组学研究中的重要作用是什么？

13. 基因组进化的分子基础是什么？谈谈你的看法。