



植物分子生物学

Plant Molecular Biology

曹仪植 主编



高等教育出版社
HIGHER EDUCATION PRESS

植物分子生物学

曹仪植 主编



高等教育出版社
HIGHER EDUCATION PRESS

图书在版编目(CIP)数据

植物分子生物学 / 曹仪植主编. —北京: 高等教育出版社, 2002
ISBN 7-04-010546-2

I. 植... II. 曹... III. 植物学: 分子生物学—高等学校
—教材 IV. Q946

中国版本图书馆 CIP 数据核字(2002)第 031357 号

策 划 林金安
编 辑 张庆波 封面设计 王凌波
版式设计 李 杰 责任印制 陈伟光

植物分子生物学
曹仪植 主编

出版发行	高等教育出版社	购书热线	010-64054588
社 址	北京市东城区沙滩后街 55 号	免费咨询	800-810-0598
邮政编码	100009	网 址	http://www.hep.edu.cn
传 真	010-64014048		http://www.hep.com.cn
经 销	新华书店北京发行所		
印 刷	北京民族印刷厂		
开 本	850×1168 1/16	版 次	2002 年 7 月第 1 版
印 张	23.5	印 次	2002 年 7 月第 1 次印刷
字 数	550 000	定 价	32.00 元

©高等教育出版社 2002

版权所有 侵权必究

前 言

分子生物学是在分子水平上研究生命现象的一门新兴学科。20 世纪下半叶以来,由于分子生物学的惊人成就,使生物学一跃成为自然科学的带头学科,它的理论和方法已渗透到生命科学的几乎每一个领域,为生命科学研究带来新的思维方式和研究手段。分子生物学取得的重大成就已使其自成体系,因此十分必要为大学本科生和研究生开设分子生物学课程。

分子生物学究其基本内容主要涉及构成生物体的生物大分子结构和功能的研究,特别是核酸、蛋白质结构和功能的研究,并在此基础上,对不同生物体以及生命现象的各个方面从分子水平上进行剖析。从中看出,分子生物学所涉及的研究内容往往与现代生物化学所研究的内容十分相似,有时很难区分,同时也说明学科交叉是当代科学研究的特征与要求。

纵观分子生物学发展的历史,据称 1938 年已第一次在正式场合使用了“分子生物学”一词,1953 年 Watson 和 Crick 提出了 DNA 结构的双螺旋结构模型,从而把分子生物学研究推向飞速发展的时期。继而,1956 年成立了第一个以分子生物学命名的实验室(即位于剑桥的医学委员会分子生物学实验室),1959 年创办了分子生物学专门杂志。世纪交替之际的人类基因组工作框架图的绘制完成更是把分子生物学研究推向自然科学研究的前沿。在人类基因组计划完成的前后不同时间内,又相继完成了啤酒酵母(*Saccharomyces . cerevisiae*)、线虫(*Caenorhabditis . elegans*)、果蝇(*Drosophila . melanogaster*)、拟南芥(*Arabidopsis thaliana*)等真核生物基因组的全序列测定。

随着人类、拟南芥等一系列基因组测序的完成,分子生物学研究将向哪个方向发展是各国科学家关心的问题。生物学家清楚地认识到,基因组测序的完成只是在研究生物体发育和功能的长征路上迈出了第一步。就基因组研究来说,以全基因组测序为目标的研究属结构基因组学(structural genomics)的范畴,而以确定基因功能为目标的研究属功能基因组学(functional genomics)的范畴,亦称后基因组学(postgenomics)。如果说结构基因组学的研究目前已取得了突破,那么下一个关键的任务是探索生命周期中合成的每一种蛋白质(complete set of proteins)其特性及其功能,即蛋白质组学(proteomics),它是后基因组时代基因功能研究最重要的目标之一,它遵循的基本思路是:基因组→转录组(transcriptomics)→蛋白质组。

随着分子生物学逐步进入后基因组时代,许多新的课题有待去探索、研究。例如,2001 年公布的人类基因组测序得到的结果显示,人体基因组中只有 3~4 万个蛋白编码基因,这与原先预测的 14 万个基因数目大相径庭。有人对此提出异议。此外,人们不禁要问,人类约 3 万个基因是否意味着 3 万种蛋白质的存在?已有科学家认为这是不可能的,因为这样少的蛋白质不可能形成人类这种复杂的生物。到底一种基因能表达出几种蛋白质是有待解决的疑问,也是蛋白质组学要确定的目标。

除了对一个基因能编码几种蛋白质产生疑问外,科学家还试图通过各种方法来解释为什么如此复杂的生命现象却只有如此少的基因在起作用,其中一些研究人员推测,人类3万个基因可能有多得多的变体,即不同个体基因组之间存在许多基因差异。也就是说,人类个体之间在遗传上具有99.9%的共性,这决定了每个人同属于一个生物物种,而恰恰剩下的0.1%的差异决定了一个人与他人的不同。最常见的人类遗传变异是基因组中散在的单个碱基的不同,即所谓的单核苷酸多态性(single nucleotide polymorphism, SNP)。SNP是一种重要的遗传工具,也是后基因组时代研究的重要目标,而要解决SNP的大规模操作,又必须采用现代生物信息学技术。

在各种生物物种全基因组序列以及表达序列标签(expressed sequence tags, EST)等大量数据取得的情况下,如何对取得的生物信息进行处理、存贮、分配、分析和解释,是解译“生命之书”的关键步骤,而这也正是现代生物信息学应运而生的内在原因。它包括必要的软件、数据库以及包括电子网络等远程通讯工具。生物信息学的崛起必将推动分子生物学的进一步高速发展。

综上所述,由于分子生物学已渗透到生命科学的各个领域,对医学、农学、林学和生物学基础理论的研究均有着深远的影响,目前不同性质的高等院校,不同的生物学专业,从不同的角度、不同的层次编写分子生物学教材,不可能有一本教材能包罗分子生物学领域的全部内容。作者根据我院课程设置的具体情况,在本书内容上尽量与其他学科避免过多重复,特别是避免与生物化学、遗传学、细胞学等学科重复,因此有关核酸、蛋白质大分子结构, DNA复制、修复、蛋白质合成等内容未编入教材。在内容上简明扼要,偏重基础,力求在阐述分子生物学基本理论的同时,有机地结合介绍分子生物学实验原理及手段。尽可能在不太多的篇幅内,使学生在较短时间内了解这一学科的基本方面,为将来从事研究工作和科技开发打下基础。对一些关键性的发现,则扼要叙述发现的经过和研究思路,以启发思考,提高分析和解决问题的能力。作者力图编写出自具特色的分子生物学教材。

本教材以讨论分子生物学基本原理和方法着手,取材本应不限于某一种生物,但鉴于许多分子生物学教材以阐述原核生物、人、动物为主,而用植物的研究资料很少,故本教材主要阐述植物分子生物学方面的基本知识。

本教材按编写结构共分21章:第1章至第4章主要讨论真核生物基因组,包括基因组的分子组成、基本结构以及真核生物特有的线粒体基因组和叶绿体基因组。对在基因表达调控中起着重要作用的真核生物转座因子进行了讨论。第5章至第10章主要介绍DNA分子操作的基本原理和方法,除了涉及一般DNA重组方法外,特别列入近年来迅速发展的PCR扩增技术、分子标记技术、DNA芯片技术以及其他功能基因组研究方法等。第11章至第14章着重讨论真核生物的基因表达。为了避免与其他相关课程重复,着重讨论转录前、转录水平和转录后水平的调控,特别是讨论DNA-蛋白质相互作用对转录的影响,而对翻译及翻译后水平的调控不专门叙述,只在某些章节有所说明。此外,专辟一章阐述外源基因转化的方法和途径。第15章至第20章以实例剖析的方式,把已学到的分子生物学理论和分子操作的知识融会到对一些植物生物学热门课题的认识中去,力图使读者能从中体会到如何利用学到的知识去解决生命科学中的实际研究课题。在教材最后一章,介绍了拟南芥和拟南芥基因组,特别是根据2000年末完成的拟南芥基因组全序列测序结果,说明拟南芥

作为模式植物在研究植物分子生物学中的重要性。

我校自 1991 年开始系统讲授分子生物学课程,本教材是在近年教学的基础上几经更改教学大纲,更新教学内容,最终编写整理而成的。它既可作为大学本科生教材,也可作为未系统学习过分子生物学课程的研究生参考书。众所周知,分子生物学研究日新月异,新方法、新思维、新概念不断更新,一本教材在它编成之时,许多概念已经又被更新,作者谨希望能以这本教材对师生学习现代植物分子生物学起到入门的作用。

在编写教材过程中,承蒙本院领导和广大师生的支持和帮助,其中特别是本院历届国内外校友的支持,他们或对教材编写提出宝贵意见;或提供精选的参考文献;或直接为本教材撰写个别章节,在此,对所有帮助完成教材编写的师生、校友表示衷心感谢。

此教材错误与不足之处一定很多,望用此教材的师生批评指正。

曹仪植

兰州大学生命科学学院

2001 年 12 月

郑 重 声 明

高等教育出版社依法对本书享有专有出版权。任何未经许可的复制、销售行为均违反《中华人民共和国著作权法》。行为人将承担相应的民事责任和行政责任,构成犯罪的,将被依法追究刑事责任。社会各界人士如发现上述侵权行为,希望及时举报,本社将奖励举报有功人员。

现公布举报电话及通讯地址:

电 话:(010) 84043279 13801081108

传 真:(010) 64033424

E-mail:dd@hep.com.cn

地 址:北京市东城区沙滩后街 55 号

邮 编:100009



目 录

1	植物细胞核基因组	(1)
1.1	植物核基因组大小	(2)
1.2	重复序列	(7)
1.3	基因家族	(11)
1.4	真核基因的断裂结构	(15)
2	叶绿体基因组	(18)
2.1	叶绿体 DNA 的信息含量	(19)
2.2	叶绿体基因组的结构	(20)
2.3	叶绿体基因的组成	(22)
2.4	叶绿体基因的一些结构特征	(30)
3	线粒体基因组	(34)
3.1	线粒体基因组的大小	(34)
3.2	线粒体基因组的组织结构	(36)
3.3	线粒体基因组的组成	(39)
3.4	线粒体基因的一些特征	(42)
3.5	RNA 编辑	(44)
3.6	植物细胞质雄性不育与线粒体的关系	(45)
3.7	遗传信息在基因组之间的流动	(46)
4	植物的可移位遗传因子	(49)
4.1	可移位因子的类型	(50)
4.2	转座子	(50)
4.3	LTR 逆转录转座子	(58)
4.4	无 LTR—逆转录转座子	(63)
4.5	花椰花叶病毒中的逆转录酶	(65)
5	核酸的分子操作	(66)
5.1	限制性内切酶	(66)
5.2	操作和标记核酸的酶	(74)
5.3	分子杂交	(81)

6	分子克隆	(85)
6.1	作为受体细胞的大肠杆菌	(85)
6.2	由质粒衍生的载体	(88)
6.3	由 λ 噬菌体衍生的载体	(93)
6.4	粘粒载体	(98)
6.5	单链丝状噬菌体 M13	(100)
6.6	外源 DNA 与载体重组	(101)
6.7	重组 DNA 导入受体细胞	(103)
6.8	阳性重组体的筛选和鉴定	(105)
7	DNA 序列测定与分析	(107)
7.1	测序策略	(108)
7.2	Sanger 双脱氧终止法	(112)
7.3	Maxam-Gilbert 化学直读法	(116)
7.4	杂交测序与 DNA 芯片技术	(117)
7.5	EST	(120)
7.6	DNA 序列分析	(124)
8	PCR——聚合酶链反应	(127)
8.1	PCR 的基本原理	(128)
8.2	PCR 条件的优化	(129)
8.3	由基本 PCR 拓展的相关技术及其应用	(133)
9	DNA 分子标记技术	(142)
9.1	基于 Southern 杂交技术的分子标记	(143)
9.2	随机(或半随机)引物分子标记技术	(147)
9.3	序列标志位点技术	(152)
9.4	分子标记的应用领域	(154)
10	基因文库和目的基因的分离	(160)
10.1	基因文库的构建	(160)
10.2	克隆已知基因产物或核苷酸序列信息的基因	(168)
10.3	蛋白质功能互补克隆	(171)
10.4	定位克隆	(173)
10.5	差示减法杂交法系列	(179)
10.6	通过 DNA 全序列分析确定基因	(183)

11	转录前的基因活化	(184)
11.1	染色质结构与基因活化	(185)
11.2	核基质附着区与基因表达	(188)
11.3	DNA 甲基化在调节基因表达的作用	(193)
12	转录水平的调控	(198)
12.1	基本转录过程	(198)
12.2	顺式作用元件	(202)
12.3	反式作用因子	(206)
12.4	真核生物基因转录调控的复杂性	(211)
12.5	研究基因表达调控中 DNA-结合蛋白相互作用的实验方法	(215)
13	基因表达的转录后调控	(219)
13.1	真核生物前体 mRNA 剪接	(220)
13.2	高等植物内含子的结构特性	(225)
13.3	高等植物剪接体蛋白质的组成特点	(230)
13.4	变位剪接	(230)
13.5	RNA 编辑	(233)
14	外源基因在真核细胞中的表达	(237)
14.1	选择标记和报告基因	(237)
14.2	DNA 直接导入的基因转化方法	(241)
14.3	农杆菌作载体的植物转基因方法	(244)
14.4	转基因植物的鉴定	(250)
14.5	外源基因转化在基因表达调控研究中的应用	(251)
15	植物激素与基因表达调控	(258)
15.1	激素受体	(259)
15.2	激素作用的信号传递途径	(261)
15.3	激素与激素响应基因的表达调控	(265)
15.4	不同激素对同一基因表达的调控	(272)
16	光调节基因及其表达调控	(274)
16.1	植物光受体	(275)
16.2	光信号传递途径	(279)
16.3	光调节基因表达的转录水平调控	(281)
16.4	光敏色素基因的转录调控	(285)

植物细胞核基因组

我们正处于基因组计划的年代,其目标是首先搞清几个生物物种的基因组组成,例如细菌、酵母、线虫、果蝇、拟南芥、水稻、人类等模式生物物种。对这些生物物种的基因组构建遗传连锁图和物理图谱,在此基础上进行 DNA 全序列测定,从而对其所有基因进行确定和分析。

在这些基因组计划中首推人类基因组计划,该计划预计在 15 年内完成人基因组 23 条染色体基因的作图和全长 3×10^9 bp 的 DNA 测序。在此基础上进一步研究基因的结构和功能,即把一维的 DNA 序列信息如何转换成三维的结构信息,又如何把单个基因的结构和功能延伸到整个基因组及其产物的协同作用,从而调节细胞和生物体的功能活动。目前人类基因组 DNA 全序列的草图已经初步完成,最终将在 2003 年取得 DNA 全序列,下一阶段的任务将探索全套基因的表达谱(gene expression pattern)和全部基因产物谱,即蛋白质谱。人类基因组计划正进入第二个阶段,或称后基因组时代(postgenome era)。

除了人类基因组测序完成以外,1996 年完成了酵母基因组 DNA 全序列(1.25×10^7 bp)测定,1994 年日本科学家发表了水稻基因遗传图,我国科学家独立完成了水稻基因组工作框架图,线虫基因组亦已完成,1999 年报导了果蝇基因组的全序列分析。2000 年完成了拟南芥全部 5 条染色体的全序列测定。

人们为什么要化大量的人力、物力来进行基因组研究呢?这是因为从已有的某些生物物种的基因组资料了解到,不同基因组之间存在着基因序列的相似性,例如许多人类遗传病相关基因与酵母基因编码序列存在相似性,因此研究这些基因在酵母中的生理功能,将有助于了解人类遗传病的发病机理。反之亦然,人类基因组的研究结果也会促进其他生物的研究。可以相信,对拟南芥、水稻等植物中

较小基因组的研究,将会促进对其他植物基因组的研究。

对于一个生物个体来说,其遗传信息的载体是核酸,其中除极个别生物外,均为 DNA。DNA 分子所携带的遗传信息总和即为基因组(genome)。对于多数只有一个染色体的原核生物来说,如细菌、蓝藻,它的整个染色体 DNA 分子就组成了其基因组。但真核生物与原核生物不同,其基因组大部分被限制在细胞核内,受核膜的保护,而与细胞质相隔离,这部分被局限在核内的基因组称为核基因组(nuclear genome)。在核外细胞质中也存在着某些遗传信息,如植物细胞质中线粒体和叶绿体均有其自身独特的遗传信息载体 DNA,即线粒体 DNA 和叶绿体 DNA,前者组成线粒体基因组,后者组成叶绿体基因组。有关叶绿体基因组和线粒体基因组的结构与组成将在第 2 章和第 3 章中详细讨论。本章将着重分析核基因组的结构组成。

1.1 植物核基因组大小

1.1.1 核基因组大小的差异

可采用不同实验方法来估计基因组的大小,其中包括孚尔根染色显微光密度测定法、DNA 复性动力学、核体积测定和 4,6-二氨基-2-苯基吡啶核染色流式细胞计量术等。不同的测定方法得到的结果略有不同。已确定大小的基因组常作为内标来取得对其他基因组大小的估计,拟南芥、水稻等基因组序列的全部搞清,其绝对 DNA 含量将可作为其他植物更确切的内标。不同植物的核基因组大小差异可达上千倍之多(表 1-1)。已知的最小

表 1-1 一些生物物种的 DNA 质量

生物物种	DNA 质量(2C 值)/pg	染色体数(2n)	生物物种	DNA 质量(2C 值)/pg	染色体数(2n)
动 物			大麦	11.0	14
两栖鲵	168.0	24	烟草	7.8	24
蝶螈	85.3	24	金鱼草	3.6	16
赤蛙	10.9	26	油菜	3.2	38
牛	6.4	60	水稻	2.0	24
人	6.4	46	大豆	1.8	20
羊	5.7	54	亚麻	1.4	30
鼠	5.0	40	拟南芥	0.14	10
果蝇	0.2	8	真 菌	(1C 值)	(1n)
有花植物			粘菌	0.384	7
贝母	196.7	24	酵母	0.026	15
百合	72.2	24	细 菌 每个细胞中 DNA 质量		
小麦	34.6	14	大肠杆菌	0.0040	
洋葱	33.5	16	肺炎双球菌	0.0020	
蚕豆	28.0	12	病 毒 每个细胞中 DNA 质量		
黑麦	18.9	14	T2	0.000220	
玉米	11.0	20	λ	0.000055	

植物基因组是拟南芥基因组, 只有 125 Mb, 最大的基因组之一是百合科贝母属的 *Fritillaria assyriaca*, 其大小为 100 000 Mb(图 1-1)。

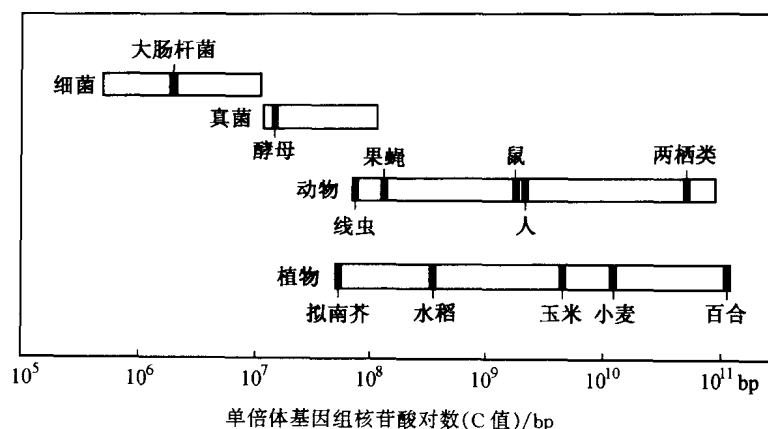


图 1-1 不同有机体的 C 值大小

表中 C 值(C value)代表每一个生物种属的单倍体基因组的 DNA 含量, 一个生物物种 C 值是恒定的。C 值可以用 pg 或碱基对数来表示。由于细胞核 DNA 的实际重量是十分小的, 故基因组的 C 值常以 pg 计, 即 10^{-12} g。一个 DNA 双链分子为 1 pg 的话, 意味着它长约 31 cm, 含有 10^9 个碱基对, 相对分子质量为 6.1×10^{11} 。以玉米二倍体核为例, 它的双链 DNA 为 11 pg 重, 这意味着这个 DNA 分子全长 341 cm。因为每个染色体含有一个 DNA 分子, 而二倍体玉米细胞有 20 个染色体(10 对), 每个染色体含有约 17 cm 长的 DNA 分子, 这比玉米分生组织细胞的平均直径(0.002 cm)长约 8 000 多倍。

C 值也可用碱基对数表示, 以 1C 值进行单倍体基因组大小衡量时, 拟南芥为 125 Mb, 水稻为 500 Mb, 番茄为 950 Mb, 玉米为 6 600 Mb, 大麦为 4 900 Mb, 小麦为 16 000 Mb, 百合为 123 000 Mb。

由 C 值大小可看出, 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, 病毒基因组在 $10^4 \sim 10^5$ bp 大小范围内, 细菌平均为 10^6 bp, 线虫基因组为 8×10^7 bp, 而要进化到昆虫, 基因组必须大于 8×10^8 bp, 进化成哺乳类更要具有大于 2×10^9 bp 的基因组。DNA 含量与有机体复杂度之间这样的关系是不难理解的, 随着有机体变得复杂, 它们需要更多的核 DNA。

然而, 我们也常常看到以生物已知的进化地位很难来解释某些生物种类之间的基因组大小差异, 即所谓的 C 值矛盾(C-value paradox)。例如, 许多高等植物有比人类多得多的核 DNA。有着相似复杂度的有机体, 如同为被子植物, DNA 含量可相差几个数量级。拿藻类来说, C 值可变动于 $10^8 \sim 10^{11}$ bp, 很难理解某些藻类虽然在形态、发育和生物化学上比有花植物简单得多, 但却需要比许多被子植物大 100 倍的 DNA。又一些十分相近的物种之间有着十分不同的 C 值, 如同为豆科的两个种, 蚕豆(*Vicia faba*)和百脉根(*Lotus tenuis*), 有着同样的结构特性和 6 个染色体, 但蚕豆 C 值是百脉根的 100 倍(图 1-2)。

在动物体中也有类似的 C 值矛盾, 例如两栖类动物, C 值小的可以低至 10^9 bp 以下, 大

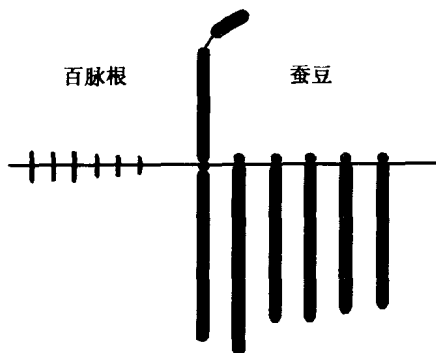


图 1-2 同为 6 个染色体的
蚕豆和百脉根的染色体比较
(修改自 Stebbins, 1971)

的可高达 10^{11} bp 左右,而作为哺乳动物(包括人类)的 C 值在 10^9 bp 的数量级,很难相信两栖动物的结构和功能会比哺乳动物,甚至人类更复杂。

由上所述,对于较简单的有机体,一般符合有机体越复杂,需要基因组越大的趋势。而对于较复杂的基因组,特别是真核生物基因组为什么出现 C 值矛盾呢?即为什么不能以生物的进化程度来体现基因组的 DNA 含量?为此需要对各种大小不同的基因组进行复性动力学分析。

1.1.2 复性动力学和基因组的不同序列组成

对染色体水平上基因组组成的研究早期是采用细胞遗传学技术,主要研究一些具有大染色体的植物种,如贝母、豌豆、黑麦等,其中最理想的植物是玉米。通过玉米细胞遗传学研究得到几个重要的概念:染色体连锁群的联会、遗传交换点的交叉、转座子等。利用光学显微镜观察了染色体数目、倍性水平、着丝粒位置和核仁组织区分布,进而用电镜又把研究深入一步。

在植物基因组复杂度方面取得的重要进步则是通过 20 世纪 70 年代对 DNA 复性动力学的研究。当 DNA 双螺旋稳定的非共价键力在加热和高浓度盐处理时,氢键会断开,双螺旋分离,产生变性。由于组成 DNA 的核苷酸杂环在紫外光区有很强的光吸收,当发生变性时,光吸收会发生相应变化。当 DNA 以双链形式存在时,质量浓度为 $50 \mu\text{g/ml}$ 的光吸收值 $A_{260} = 1.00$,变性形成单链后,相同浓度的光吸收值增至 $A_{260} = 1.37$ 。

与变性相反,在适宜的条件下,如逐渐降温,两条分开的互补链可重新形成稳定双螺旋,即 DNA 复性。复性过程是两条互补链之间碱基逐渐配对的过程。当任何两个互补的核酸序列之间形成双螺旋结构,这个过程称退火(anneal)。

在大规模测序能够进行之前,复性动力学方法通过测定变性后的基因组序列,进行复性的化学反应过程来认识基因组的 DNA 复杂性。一般是通过对基因组 DNA 进行剪切或超声波处理,使其变成几百至几千碱基对长的大小一致的 DNA 片段,然后进行变性、复性过程。对复性过程进行检测一般有两种方法:测定紫外光区 260 nm 处的光吸收的减少值,减色性与复性呈线性相关;在一定温育期内检测能与羟基磷灰石(hydroxyapatite)柱结合的数量,这是基于此柱能选择性地结合双链 DNA,而使单链 DNA 从柱中洗脱掉。

DNA 复性取决于两条单链的随机碰撞,因此遵循二级反应动力学,其反应方程式为:

$$\frac{dc}{dt} = -kc^2$$

式中, c 为时间 t 时单链 DNA 浓度, k 为复性反应常数。当 $t=0$ 时, 单链 DNA 浓度为 c_0 , t 时间后为 c , 积分上式得:

$$\frac{c}{c_0} = \frac{1}{1 + kc_0 t}$$

当复性完成一半时, t 为 $t_{1/2}$, $c/c_0 = 1/2$, 则代入上式

$$\frac{c}{c_0} = \frac{1}{2} = \frac{1}{1 + kc_0 t_{1/2}}, \text{即 } c_0 t_{1/2} = \frac{1}{k}$$

从中看出, 控制复性的是单链 DNA 浓度(c_0)和反应培育时间(t)的乘积, 称为 $c_0 t$ 值。而复性一半时的 $c_0 t$ 值为 $c_0 t_{1/2}$, $c_0 t_{1/2}$ 的值越大, 表明达到复性一半所需的时间越长, 复性反应越慢。图 1-3 绘制出不同基因组的复性动力学曲线。

图 1-3(a)中, 箭头指向 $c_0 t_{1/2}$ 值(即指向每条曲线的拐点), 同时也标示出某基因组的核苷酸对数。如大肠杆菌 DNA 的 $c_0 t_{1/2}$ 值为 9, 它的基因组大小为 4.2×10^6 bp。对于像 MS2、T4 和大肠杆菌 DNA 来说, 它们均为单拷贝序列, 随着 $c_0 t_{1/2}$ 值增大, 意味着复性速率越慢, 基因组越大。从中看出在不存在重复序列的情况下, $c_0 t_{1/2}$ 值与基因组大小成正比, $c_0 t_{1/2}$ 值可用于衡量一个基因组大小的尺度。但对真核生物来说, 其基因组中存在重复序列, $c_0 t_{1/2}$ 反映的基因组状况复杂得多, 图 1-3(b)显示假设的真核生物基因组的复性动力学曲线。

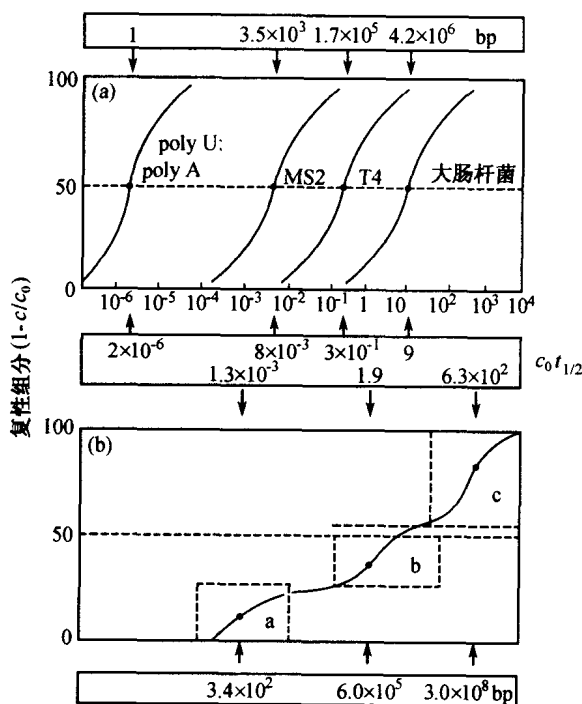


图 1-3 不同基因组的复性动力学

(a) 简单基因组; (b) 假设的真核生物基因组

从图 1-3 可以看到,真核生物的 DNA 复性曲线比原核生物要复杂得多,复线曲线带多个拐点, c_0t 值跨越的范围也大。这是因为在真核生物基因组复性动力学中,复性的快慢不仅决定于基因序列的大小,还决定于重复序列的拷贝数多少。当基因组中某一序列多于一个拷贝时,即重复序列的数目越多,互补序列相碰撞的概率越高,复性的速率越快,反之亦然。从图 1-3(b)中看,第一个 a 虚线框中的曲线拐点 $c_0t_{1/2}$ 为 1.3×10^{-3} , b 框曲线拐点为 1.9, c 框为 6.3×10^2 , 与这 3 个拐点相对应的核苷酸对分别为 3.4×10^2 , 6.0×10^5 和 3.0×10^8 bp。通过进一步对化学方法测定得到的长度比较和运算(此处不作推导),可得到 3 个虚线框部分其序列的重复频率分别为 5.0×10^5 , 3.5×10^2 和 1。从中看出,假设的真核生物基因组其序列组成可有 3 个组分, a 框处为快复性组分, b 框处为中等复性组分, c 框处为慢复性组分。

对真核生物基因组来说,慢复性组分其序列在基因组中只出现一次,即只有一个拷贝,故称为非重复序列,即单拷贝序列。在原核生物中均为非重复序列。中等复性组分在基因组中每序列有较多的拷贝,故称为中等重复序列。而序列拷贝数最多的快复性组分,称为高度重复序列。图 1-4 为不同模式生物各序列组分所占的比重。

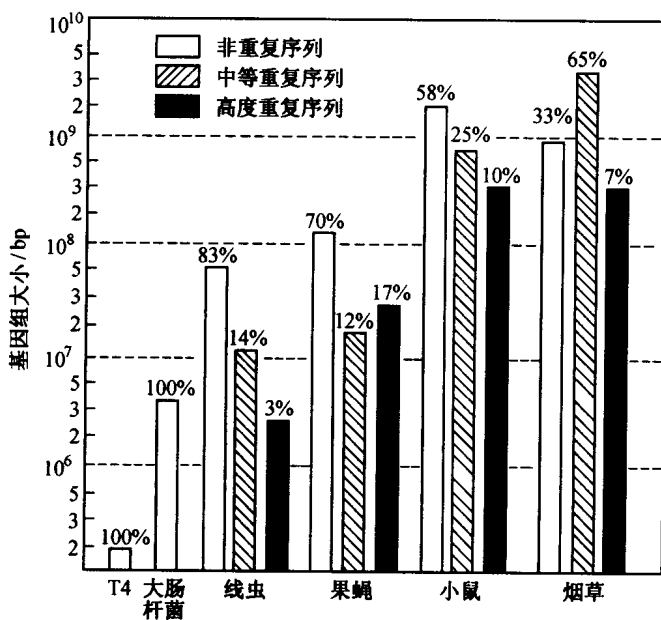


图 1-4 不同生物基因组中不同序列组分的比例

从图 1-4 看出,从原核生物至真核生物,随着基因组增大,非重复序列的核苷酸序列绝对含量增加,在达到约 2×10^9 bp 时不再增加。但在原核生物中,100% 的序列均为非重复序列。低等真核生物,重复序列所占比例不到 20%,而高等真核生物重复序列所占比例高,一般可超过 50%,如烟草中可达 72%。在更大的高等植物基因组中重复序列甚至可占基因组的 95%。从中看出,为什么不同植物基因组大小差异有时会十分大,而从进化角度很难来解释的一个重要原因是重复序列的存在。许多这样的重复序列没有编码功能,但由于它

们的大量存在而使不同植物在基因组大小上存在很大差异。

除了重复序列的存在会影响基因组的大小差异外,植物 DNA 含量的差异还由于多倍性,在植物发育过程中会产生多倍性。据估计,50% 的被子植物是多倍体,它们或者是由于一个种内染色体加倍而成,或者是两个种的杂交但没有染色体减少而致。分子图谱证实,某些植物种,如玉米,含有两个几乎完全的基本基因组的拷贝(见 9.4.2)。因此,单倍体染色体不含有“初级基因组(primary genome)”,而是含有一个隐蔽性多倍体的基因组。多倍体只是导致了基因组大小的增加,但没有引起有机体复杂度的增加。

1.1.3 植物基因组中序列构成层次

在叙述了根据复性动力学把基因组分为 3 个组分后,我们换个角度来理解基因组的构成,即从基因组全序列的庞大核苷酸序列逐渐解析到最基础的千碱基对单位的不同层次。

构成基因组序列的最大单位是染色体 DNA,其次是亚染色体水平,它是由带保守基因组合的核苷酸片段组成。比亚染色体单位小一些的序列是多兆碱基单位,如卫星序列或特殊的重复序列。亚兆碱基大小的序列是由非甲基化位点确立的扩增序列组。而一些重复因子(如转座子、逆转录转座子)、基质附着区(MAR)和基因是处在千碱基水平上的序列。最后许多其他的调节序列是由十分少的碱基对组成的。

1.2 重复序列

根据复性动力学等的研究分析表明,植物核基因组中存在有重复序列,即某一核苷酸序列在基因组中不只一个拷贝,这个序列可以重复许多次。这些重复序列是否由完全相同的核苷酸序列组成的呢?重复序列 DNA 是由许多相似而不完全相同的碱基序列家族组成,正如一个大家庭一样,家族的每一个成员都具有某些相似的特征,但每个成员之间具有相似而不完全相同的碱基序列。一个植物基因组中有多少重复序列家族存在,这取决于植物种,不同植物之间差异很大。根据复性动力学估计,棉花基因组有 1 000 个不同的重复序列家族,烟草为 4 000 个,大豆为 40 000 个。这些家族各包括 20~300 成员,其重复序列的大小平均约 200~400 bp,但较大的重复序列可达 1 000 bp 或 1 000 bp 以上。

从重复序列的编码状况来分,有些重复序列是有编码功能的,例如重复频率较高的编码 rRNA 基因、tRNA 基因和组蛋白基因的重复序列 DNA,重复频率相对较低的一些基因家族,如一些贮藏蛋白基因家族、光敏色素基因家族等。另有一些重复序列是没有编码功能的,如哺乳动物中重复频率高的 Alu 序列家族、真核生物中的卫星 DNA、转座子、逆转录转座子等。这类重复序列虽然没有编码功能或很少编码序列(如编码转座酶的转座子),但它们在基因组的结构与整合可能仍起重要作用。

从重复序列在基因组中的排列状况来分,可有两种基本的排列分布类型:串联重复序列(tandemly repeated sequences)和散布重复序列(interspersed repeated sequences)。串联重复序列的排列方式故名思义是指重复序列在基因组中以串联形式,在某些 DNA 区域头尾相连排列在一起,成簇存在。在成百成千个串联一起的重复序列之间被一些间隔序列分开。散布重复序列分散在基因组中,各重复序列拷贝之间可相隔几千个碱基对。