

中国高等教育学会医学教育专业委员会规划教材

全国高等医学院校教材

供基础、临床、预防、口腔医学类专业用

医学分子生物学

主编 马文丽 德 伟

*Medicine Molecular
Biology*



北京大学医学出版社

中国高等教育学会医学教育专业委员会规划教材
全国高等医学院校教材

供基础、临床、预防、口腔医学类专业用

医学分子生物学

Medicine Molecular Biology

主 编 马文丽 德 伟

副主编 张鹏霞 李存保 万福生 刘新光

编 委 (按姓名汉语拼音排序)

德 伟 (南京医科大学)

费 嘉 (暨南大学医学院)

关一夫 (中国医科大学)

李存保 (内蒙古医科大学)

李恩民 (汕头大学医学院)

李海东 (天津医科大学)

刘新光 (广东医学院)

马文丽 (南方医科大学)

潘洪明 (齐齐哈尔医学院)

束永前 (南京医科大学)

万福生 (南昌大学医学院)

王 杰 (沈阳医学院)

王天云 (新乡医学院)

王晓华 (广州医科大学)

肖建英 (辽宁医学院)

张 杰 (中国医科大学)

张鹏霞 (佳木斯大学基础医学院)

周晓霞 (扬州大学医学院)

北京大学医学出版社

YIXUE FENZI SHENGWU XUE

图书在版编目 (CIP) 数据

医学分子生物学/马文丽, 德伟主编. —北京:
北京大学医学出版社, 2013. 12
ISBN 978-7-5659-0760-9

I. ①医… II. ①马…②德… III. ①医学—分子生
物学—高等学校—教材 IV. ①Q7

中国版本图书馆 CIP 数据核字 (2013) 第 316964 号

医学分子生物学

主 编: 马文丽 德 伟
出版发行: 北京大学医学出版社 (电话: 010-82802230)
地 址: (100191) 北京市海淀区学院路 38 号 北京大学医学部院内
网 址: <http://www.pumpress.com.cn>
E - mail: booksale@bjmu.edu.cn
印 刷: 北京地泰德印刷有限公司
经 销: 新华书店
责任编辑: 张其鹏 刘陶陶 黄 越 责任校对: 金彤文 责任印制: 张京生
开 本: 850mm×1168mm 1/16 印张: 19.75 字数: 565 千字
版 次: 2013 年 12 月第 1 版 2013 年 12 月第 1 次印刷
书 号: ISBN 978-7-5659-0760-9
定 价: 37.00 元
版权所有, 违者必究
(凡属质量问题请与本社发行部联系退换)

全国高等医学院校临床专业本科教材评审委员会

主任委员 王德炳 柯 杨

副主任委员 吕兆丰 程伯基

秘书长 陆银道 王凤廷

委 员 (按姓名汉语拼音排序)

白咸勇 曹德品 陈育民 崔慧先 董 志

郭志坤 韩 松 黄爱民 井西学 黎孟枫

刘传勇 刘志跃 宋焱峰 宋印利 宋远航

孙 莉 唐世英 王 宪 王维民 温小军

文民刚 线福华 袁聚祥 曾晓荣 张 宁

张建中 张金钟 张培功 张向阳 张晓杰

周增桓

序

北京大学医学出版社组织编写的全国高等医学院校临床医学专业本科教材（第2套）于2008年出版，共32种，获得了广大医学院校师生的欢迎，并被评为教育部“十二五”普通高等教育本科国家级规划教材。这是在教育部教育改革、提倡教材多元化的精神指导下，我国高等医学教材建设的一个重要成果。为配合《国家中长期教育改革和发展纲要（2010—2020年）》，培养符合时代要求的医学专业人才，并配合教育部“十二五”普通高等教育本科国家级规划教材建设，北京大学医学出版社于2013年正式启动全国高等医学院校临床医学专业（本科）第3套教材的修订及编写工作。本套教材近六十种，其中新启动教材二十余种。

本套教材的编写以“符合人才培养需求，体现教育改革成果，确保教材质量，形式新颖创新”为指导思想，配合教育部、国家卫生和计划生育委员会在医药卫生体制改革意见中指出的，要逐步建立“5 + 3”（五年医学院校本科教育加三年住院医师规范化培训）为主体的临床医学人才培养体系。我们广泛收集了对上版教材的反馈意见。同时，在教材编写过程中，我们将与更多的院校合作，尤其是新启动的二十余种教材，吸收了更多富有一线教学经验的老师参加编写，为本套教材注入了新鲜的活力。

新版教材在继承和发扬原教材结构优点的基础上，修改不足之处，从而更加层次分明、逻辑性强、结构严谨、文字简洁流畅。除了内容新颖、严谨以外，在版式、印刷和装帧方面，我们做了一些新的尝试，力求做到既有启发性又能引起学生的兴趣，使本套教材的内容和形式再次跃上一个新的台阶。为此，我们还建立了数字化平台，在这个平台上，为适应我国数字化教学、为教材立体化建设作出尝试。

在编写第3套教材时，一些曾担任第2套教材的主编由于年事已高，此次不再担任主编，但他们对改版工作提出了很多宝贵的意见。前两套教材的作者为本套教材的日臻完善打下了坚实的基础。对他们所作出的贡献，我们表示衷心的感谢。

尽管本套教材的编者都是多年工作在教学第一线的教师，但基于现有的水平，书中难免存在不当之处，欢迎广大师生和读者批评指正。

王德炳 柯杨

2013年11月

前言

分子生物学是从分子水平研究生命本质的一门新兴学科，它以核酸和蛋白质等生物大分子的结构及其在遗传信息和细胞信息传递中的作用为研究对象。在分子生物学的带动下，现代生物科学的众多分支学科都得到了迅速的发展，也取得了惊人的成就，使得现代生物科学成为当代成果最多和最吸引人的学科之一。医学分子生物学是分子生物学的分支，从分子水平研究人体在正常基本状态下的生命活动及其规律，从分子水平研究人类疾病的预防、诊断和治疗的一门科学。

作为一门课程，医学分子生物学涵盖了医学各专业学生必须掌握的分子生物学基础知识。医学分子生物学主要解决两方面的问题：一是理论方面，阐明疾病和亚健康状态发生和发展的分子机制；二是在解决理论问题的基础上，发展新型分子生物学技术，为疾病的预防、诊断和治疗提供新手段。系统学习医学分子生物学的知识将帮助医学生更好地面对医学发展所带来的挑战。

在本书的编写过程中，我们根据多年的教学实践和各院校教学中的经验，对教材的体例、格式和内容的编排进行了充分的讨论。本书编写内容力求丰富严谨，语言更加通俗易懂。在对学生进行基础理论、基础知识和基本技能训练的同时，更加重视分子生物学与临床知识的紧密结合。内容上注重知识的更新，引入了医学分子生物学领域的前沿进展，力求培养学生的创新意识与创新能力。

本教材适用于全国高等医学院校五年制本科生及临床医生，主要作为医学本科生的医学分子生物学必修课或选修课教材，同时可供其他院校有关专业和医学专科学校的师生参考。

本教材由全国十余所医学院校的 18 位在教学和科学研究上卓有成就的教授共同编著完成。编写过程中，得到了北京大学医学出版社的鼎力支持和帮助，得到了各参编学校的支持和鼓励，在此衷心感谢！由于编写时间仓促，可能会有不少错误，还望读者批评指正，不胜感谢！

马文丽

2013 年 12 月于广州

目 录

绪论	1	第一节 概 述	89
第一章 基因与基因组学	4	第二节 DNA 甲基化	91
第一节 基因的结构与功能	4	第三节 组蛋白修饰	95
第二节 基因组的结构与功能	6	第四节 X 染色体失活	98
第三节 基因组学	10	第五节 非编码 RNA 调控	100
第四节 基因组的复制	13	第六节 表观遗传与疾病	105
第五节 基因组学与医学	18	第七章 生物信息学	109
第二章 RNA 与 RNA 组学	21	第一节 生物信息学简介	109
第一节 RNA 的结构与功能	21	第二节 生物序列分析	114
第二节 RNA 的生物合成	24	第三节 基因表达分析	121
第三节 RNA 组与 RNA 组学	32	第四节 蛋白质结构分析	127
第四节 RNA 组学的研究方法和分析 技术	34	第五节 生物信息与医学应用	133
第五节 RNA 组学在医药研究中的 应用	35	第八章 分子生物学常用技术	137
第六节 RNA、RNA 组学与医药、 疾病的关系	38	第一节 核酸分子杂交	137
第三章 蛋白质与蛋白质组学	40	第二节 聚合酶链式反应	145
第一节 蛋白质的结构与功能	40	第三节 DNA 序列测定	151
第二节 蛋白质的生物合成与加工	45	第四节 生物芯片技术	155
第三节 蛋白质组与蛋白质组学	56	第五节 基因沉默 RNA 干扰与基因 剔除技术	160
第四节 蛋白质组学的研究方法和分析 技术	59	第六节 酵母双杂交技术	166
第五节 蛋白质、蛋白质组与医学、 疾病的关系	61	第七节 免疫印迹技术	169
第四章 代谢组与代谢组学	63	第八节 蛋白质双向凝胶电泳技术	172
第一节 代谢组与代谢组学概念	63	第九章 基因工程	175
第二节 系统生物学时代的代谢组学	64	第一节 基因工程概述	175
第三节 代谢组学的研究方法和分析 技术	65	第二节 基因工程的工具酶	176
第四节 代谢组学在医药研究中的 应用	69	第三节 基因工程的载体	179
第五章 基因表达调控	72	第四节 基因工程的基本过程	183
第一节 基因表达调控的基本原理	72	第五节 基因工程在医药领域的 应用	188
第二节 真核基因表达调控	74	第十章 基因诊断与基因治疗	192
第六章 表观遗传学	89	第一节 基因诊断	192
		第二节 基因治疗	198
		第十一章 遗传性疾病的分子生物学	207
		第一节 遗传性疾病	207
		第二节 单基因病	209
		第三节 多基因病	213

第四节	染色体病	216	生物学	270
第十二章	肿瘤的分子生物学	221	第三节	扩张型心肌病的分子生物学
第一节	癌基因与抑癌基因	221		273
第二节	肿瘤细胞增殖与凋亡的分子生物学	243	第四节	心力衰竭的分子生物学
				275
第三节	肿瘤细胞侵袭与转移的分子生物学	248	第十五章	代谢性疾病的分子生物学
				279
第四节	肿瘤血管与淋巴管生成的分子生物学	252	第一节	概 述
				279
第十三章	感染性疾病的分子生物学	256	第二节	糖尿病的分子生物学
				280
第一节	病毒致病的分子生物学	256	第三节	肥胖症的分子生物学
第二节	细菌致病的分子生物学	261		284
第三节	真菌致病的分子生物学	264	第四节	痛风的分子生物学
第四节	感染性疾病的分子生物学诊断技术	266		287
			第五节	氨基酸代谢性疾病的分子生物学
				288
第十四章	心血管疾病的分子生物学	268	第六节	骨质疏松症的分子生物学
				291
第一节	原发性高血压的分子生物学	268	第十六章	衰老与抗衰老的分子生物学
				293
第二节	动脉粥样硬化的分子生物学	268	第一节	概 述
				293
			第二节	细胞的衰老
				294
			第三节	衰老的分子生物学
				295
			第四节	抗衰老的策略
				299
			主要参考文献	302
			中英文专业词汇索引	303

1956年第一次将一种疾病的发病机制确认为分子特定改变。1940年, Pauling发现镰状细胞贫血是由于血红蛋白结构改变所致, 但未能确定变化的部位。1956年, Ingram发现珠蛋白第6位氨基酸由谷氨酸突变为缬氨酸, 这是镰状细胞贫血的致病原因。

1959年鉴定出第一个染色体病。Down综合征在1866年首次被报道。1959年, Jerome Lejeune和其研究团队发现该综合征是由于21号染色体三体突变引起。

1961年第一次对新生儿代谢缺陷作出诊断。Robert Guthrie建立了检验新生儿是否患有苯丙酮尿症的新方法。

1972年第一个重组DNA分子在实验室制造成功。美国斯坦福大学Paul Berg带领的研究小组在实验室制造出第一个人工重组DNA分子。

1976年发现癌基因。Harold Varmus和Michael Bishop从1970年开始进行肿瘤病毒学研究, 他们从鸡肉瘤中分离到在体外能使鸡胚成纤维细胞转化, 在体内能使鸡患肉瘤的病毒。此后, 还确立了细胞癌基因的概念。

1978年美籍华裔科学家简悦威(Kan YM)第一次利用DNA多态性与致病基因的关联性, 成功地对镰状细胞贫血进行了产前诊断, 开创了基因诊断技术在临床医学应用的新时代。

1981年第一个转基因小鼠(生长激素)构建成功, 开创了建立疾病模型的工作基础。

1982年世界上第一个基因重组产品——人胰岛素问世。

1983年第一次在染色体定位致病基因。Huntington病的致病基因被定位在第4号染色体(4p16.3)。

1986年第一次实现了人类致病基因的定位克隆。波士顿儿童医院第一次通过定位克隆技术分离得到慢性肉芽肿病(CGD)的致病基因, 同年, 杜氏肌营养不良和视网膜母细胞瘤的致病基因也是利用定位克隆策略获得。

1987年第一张含400个RFLP标志的人遗传图谱绘制完毕。

1990年第一例真正意义上的基因治疗是用腺苷脱氨酶(ADA)基因来治疗ADA基因缺陷引起的重度联合免疫缺陷病(SCID)的患者, 正常ADA基因借助反转录病毒载体被导入患儿的T淋巴细胞, 再回输给患儿, 5年后患儿体内仍有10%的造血细胞呈ADA基因阳性。这一病例的治疗成功使基因治疗技术得以迅速开展。

1995年完成第一个病原微生物基因组测序, 即流感病毒基因组序列测定, 这也是第一个完整的生物基因组测序。

2003年4月14日, 中、美、日、德、法、英6国科学家联合宣布, 被誉为生命科学“登月计划”的人类基因组计划顺利完成, 标志着后基因组时代的开始。

2006年日本京都大学山中伸弥Shinya Yamanaka在世界著名学术杂志《细胞》上率先报道了诱导多能干细胞的研究。

2010年, 首例人造细胞宣告诞生。这项具有里程碑意义的重大科研成果是由美国科学家克雷格·文特尔领导的四百多人的专家团队用了约十五年取得的。他们用一个人工合成的全新基因组“替换”了自然支原体原有的基因组, 即“指令控制”细胞的所有活动, 使之成为“地球上第一个由人类制造的可以进行自我复制的新物种”, 向人造生命形式迈出关键的一步。

三、医学分子生物学的现状与前景

随着人类基因组计划的完成, 功能基因组研究的技术策略日新月异, 例如“RNA组学”、第三代测序技术等, 通过确定基因的功能, 通过基因工程技术克隆特定的基因并在体内表达, 可用于疾病的预防和治疗。此领域的研究进展非常迅速, 在以下方面取得了重大成就。

基因工程与疫苗: 利用基因工程技术生产有应用价值的药物是医学分子生物学发展的一个重要方向, 现在世界上已有几千家生物技术公司, 其中多数都生产医药或医药研究所需的试

剂。利用基因工程技术生产药物有两个不同的途径：一是利用基因工程技术改造传统的制药工业，例如用 DNA 重组技术改造制药所需要的菌种或创建的菌种，提高抗生素、维生素、氨基酸产量等；二是用克隆的基因表达生产有用的肽类和蛋白质药物或疫苗，虽然关于基因诊断和医药研究试剂的基因工程产品已经很多，但目前基因工程药物还只处在发展早期，至今真正被卫生部门正式批准投放市场的基因工程肽或蛋白质类药物现在还不多，但正在开发的基因工程治疗药物却有几百种，且而逐年迅速增加，可见其巨大潜力。

转基因动植物：利用转基因动物可以建立人类疾病的动物模型，为对人类疾病病因研究，以及测试新治疗方法提供了有力手段。例如用导入各种癌基因、致瘤病毒基因或其调控序列的转基因小鼠，可以观察肿瘤发生的过程和影响因素；导入相关突变基因的转基因动物可以制造出糖尿病、镰状细胞贫血、白内障等疾病模型；导入肝炎病毒基因的转基因动物可以用来研究肝炎病毒基因在肝炎中的作用，利用导入各种细胞因子基因、免疫功能基因，以及特定核酸序列的转基因动物可以从整体研究细胞因子、免疫调控、基因表达调控等问题。

转基因植物在育种上也获得成绩。1994 年比普遍西红柿保鲜时间更长的转基因西红柿投放市场，1996 年转基因玉米、转基因大豆相继投入商品生产。美国最早研制的抗虫棉花，我国科学家将自己发展的蛋白酶抑制剂基因转入棉花获得抗棉铃虫的棉桦株。与按传统孟德尔遗传规律育种比较，转基因技术显出其优越性及更大的潜力，提高光合作用、扩大固氮能力、提高营养价值、抗虫、抗病、抗旱的转基因植物都在研究中。将人的基因转入植物还可能获得医学上治疗用途的药物，例如将人抗体基因转入烟草，从烟叶中就能提取得人的抗体蛋白质。

基因诊断：目前在临床医学诊断、遗传病预防，以及法医学个体认定等领域有了广泛的应用，人类遗传病尤其是单基因病的确诊是目前基因诊断技术应用领域的成功典范。许多新发现的小分子非编码 RNA 不仅是细胞增殖、分化和凋亡的重要调控因子，而且与细胞的异常表型和人类疾病密切相关。在许多肿瘤中可检测到微 RNA (miRNA) 基因的异常表达或信使 RNA (mRNA) 异常可变剪接体。一些动物病毒也编码用于逃逸宿主细胞免疫攻击的 miRNA。比较分析正常生理和疾病发生过程中的非编码 RNA 的表达及其作用，将从 RNA 调控的角度揭示疾病发生机制并为疾病诊断和治疗提供新的基因靶点和分子标记。基因诊断未来在多基因常见病的发病风险预测、个体化用药指导和疗效评价等方面将显示出巨大的应用潜力。

基因治疗：医学分子生物学的研究手段，如疾病基因组学、药物基因组学、代谢基因组学、疾病蛋白质组学、疾病动物模型、RNA 组学等将为药物的发现和开发提供更多的新靶点，也将提供更多的药物筛选技术平台。同样，个体基因组学信息将评价和给出个体最佳的用药方案。miRNA 和 RNA 干扰 (RNAi) 技术作为人类重大疾病新的治疗技术正在迅速发展，miRNA 药物及靶点的运用前景十分广阔。

DNA 重组技术和基因工程使人类进入了能动改造的生物界新纪元，使医学发展到分子生物学的新阶段。但由于人类对生物基因组的结构、基因表达调控等认识还很有限，因而分子生物学的成果在医学上的应用还处在初级阶段。新的基因工程药物虽然不断涌现，但已应用的还是少数，而且由于对基因产物的整体效应等研究还不够充分，即使已批准投入市场的基因工程药物，有些疗效还不很理想。基因诊断应用的范围尚有待扩大，基因治疗成功的例子还不多。由于基因导入后在基因组上的定位整合等知识和技术尚不成熟，因而现在转基因的工作还很盲目、成功率还很低。这些都有待于进行许多扎实的基础研究，了解更多分子遗传学方面规律，并改进和创建新技术，才能得到提高。然而探索生命本质的分子生物学已经指出了光明的前程，随着科学的进步，肯定将逐步按人们的意志去获得理想的结果，可以说“前途光明灿烂，道路曲折而遥远”。

第一章 基因与基因组学

从简单的病毒、细菌到高等的动、植物细胞，决定 RNA 和蛋白质结构的信息是以基因为基本单位贮存在 DNA（有些病毒为 RNA）分子中的。基因及基因组的结构和功能研究是现代分子生物学的核心内容。人体的生长、发育、衰老、死亡以及多种疾病的发生都与基因的结构和功能密切相关。本章简要介绍基因、基因组、基因组学、人类基因组计划和基因组复制等内容，为认识生命及疾病的本质奠定基础。

第一节 基因的结构与功能

一、基因的概念

1865 年，遗传学奠基人 Mendel 通过豌豆杂交实验发现生物体的遗传性状是由“遗传因子”决定的。1909 年，Johannsen 将孟德尔提出的遗传因子改称为“基因 (gene)”。1910 年，Morgan 通过果蝇实验建立“染色体-基因”学说：染色体是孟德尔式遗传性状传递机制的物质基础，基因既是遗传的功能单位，能产生特定的表型效应，又是一个独立的结构单位，进一步创立了基因学说。1944 年，Avery 通过肺炎链球菌转化实验证实基因是由 DNA 组成的，这正式确定了基因的化学本质。1952 年，Alfred Hershey 和 Martha Chase 通过放射性核素标记实验进一步证实了 DNA 是遗传物质的携带者。1953 年，Watson 和 Crick 提出了著名的 DNA 双螺旋结构模型，基因就是 DNA 分子的一个区段，每个基因由成百上千个脱氧核苷酸组成，一个脱氧核糖核酸 (DNA) 分子可以包含几个乃至几千个基因。

1941 年，Beadle 和 Tatum 在 PNAS 上发表了题为“链霉素中生化反应的遗传控制”的研究报告，提出了“一个基因一种酶”的假说，认为基因对性状的控制是通过基因控制酶的合成来实现的，并因此获得 1958 年诺贝尔生理学 and 医学奖。但该学说对异源多聚体酶的编码基因难以解释，因此人们对“一个基因一种酶”的假说进行了修正，提出“一个基因一条多肽链”的概念。

随着对遗传学、生物化学和分子生物学领域的研究不断深入和完善，人们对基因的定义又进行了进一步的探讨：发现有的基因只编码 RNA 却不生成蛋白质，也发现基因的序列不一定是编码序列，也包含非编码序列。另外，在某些特定的生物体内，例如病毒 (virus) RNA 也可以作为遗传信息的携带者。因此逐渐形成了基因的现代概念。不同领域对相关概念的解释不完全相同，但本质是一样的。

从遗传学角度讲，基因就是遗传上相对独立的基本单位或单元，决定遗传性状的表达，基因定位于染色体和线粒体 DNA 分子上，呈线性排列并世代相传。从分子生物学角度看，基因是负载特定遗传信息单位的核酸分子片段，是贮存核糖核酸 (RNA) 序列信息、有功能的蛋白质多肽链序列信息，以及表达这些信息所必需的全部核苷酸序列。在绝大部分生物中，构成基因的核酸分子是 DNA，极少数生物如 RNA 病毒为 RNA。

二、基因命名的基本原则

1979年,人类基因命名委员会被授权批准人类基因名称和符号,首次发表了《人类基因的命名指南》,并于1987年、1995年、1997年和2001年四次对此指南进行修订。与此同时其他工作者也对酵母、小鼠、大鼠、斑马和鱼等不同生物的基因命名及其基因符号的使用提出了相应的规定。在约定俗成的基础上,基因名称以及基因符号的使用将向着越来越规范的方向发展。

(一) 基因命名的基本原则

人类基因命名委员会有关基因命名的基本原则为简明、独特、能够表达基因的特征或功能。基本符号的命名也应遵循独特、简短,仅含有拉丁字母和阿拉伯数字,不应含有标点符号的基本原则。

(二) 人类基因命名法

基因名称应使用美式英语的拼写方式,应简明、独特,并且能够表达基因的特征或功能,要避免使用组织特异性、分子量,以及家庭成员关系的描述。另外还应注意:

1. 除使用人的名字来描述一种疾病或一种表型外,如 *AHDS* “Allan-Herndon-Dudley syndrome”,基因的名字以小写字母开始。
2. 在对基因名称进行限时,限定词应在名称的后面,并以逗号与主名称分开,如: *ACO 1*, “aconitase 1, soluble”。
3. 当使用替代名称作为基因名称的一部分时,替代名称应在括号内,如 *IDS* “iduronate 2-sulfatase (Hunter syndrome)”。
4. 物种的名称应在括号内,并应放在基因名称的后面,如: *LFNG* “lunatic fringe homolog (*Drosophila*)”。

(三) 基因符号的命名

多数情况下,基因符号应使用大写拉丁字母或小写英文字母与阿拉伯数字的组合来表示。基因符号的第一个字母应与基因名称一致,以便于按照字母来排序。符号应简短,最好不超过6个字母。应避免使用基因符号表示基因的全部信息,避免拼写成为单字或已知的缩写。在使用新的基因符号时,不能与已经存在或已获得认可的符号重复。此外,以下几点规则也应遵循:

1. 符号应以字母开始,伴随字母或阿拉伯数字。
2. 避免在符号中使用字母的上标和下标。
3. 避免在符号中使用罗马数字,曾经使用罗马数字的符号应改为相应的阿拉伯数字。
4. 避免在符号中使用希腊字母,所有的希腊字母应改变为相应的拉丁字母。
5. 基因名称的希腊字母前缀应改变为相应的拉丁字母,并放在符号的后面,如: *GLA* “galactosidase, alpha”; *GLB* “galactosidase, beta”。
6. 除人类白细胞抗原 (HLA)、免疫球蛋白 (immunoglobulin) 和 T 细胞受体 (T cell receptor) 外,不要在基因符号中使用标点符号。
7. 应避免在基因符号中标明其组织特异性和分子量。
8. 基因符号的前缀和后缀应与其特定的含义保持一致。

(四) 基因名称在文章中的表达方式

基因以及基因符号在文章中的表达方式应遵守上述规则。此外,在文章手稿中(包括打印稿),基因以及基因符号因加下划线,在印刷成文后应为斜体。

随着分子生物学技术的不断发展,人们对基因的认识也在不断加深。基因及其符号的命名也必将顺应这一发展并不断完善。

三、基因的结构与功能

构成基因的 DNA 片段按功能大致可区分为结构基因和调节基因两部分。

(一) 基因的结构

1. 结构基因 (又称为可转录序列) 在构成基因的 DNA 片段中, 一段 DNA 序列贮存着某一特定 RNA 分子全部序列的信息, 在转录时该片段 DNA 的一级结构全部序列编码该 RNA 分子的一级结构序列。有的可转录序列只编码一些具有特定功能的 RNA 分子, 而不编码蛋白质多肽, 如核糖体 RNA (rRNA)、转移 RNA (tRNA)、核酶 (ribozyme)、微 RNA (miRNA) 等。有的可转录序列通过第一步转录成信使 RNA (mRNA), 由 mRNA 分子中的密码子指导蛋白质多肽的合成。在多肽编码的可转录序列中, 不仅包含决定多肽链一级结构氨基酸残基顺序的 DNA 序列, 还包括部分非编码序列一级结构的 DNA 序列, 这些 DNA 序列调控多肽链的翻译和加工。真核基因可转录序列中还包括外显子和内含子序列, 在这些序列前还可能有序列 mRNA 的 5' 端非翻译区对应。由此可见, 可转录序列并非全部为多肽链编码。

2. 调节基因 (又称为调节序列) 在构成基因的 DNA 片段中, 可转录序列以外的具有调节基因转录功能的 DNA 序列, 包括启动子、增强子等特异 DNA 序列。这些序列多位于编码序列的上游, 个别调节元件例外, 近年来发现有些调节序列可以控制染色体调节相关基因的表达。

(二) 真核基因结构特点

1. 不连续性 真核基因之间基因大多是不连续的, 在第一个和最后一个外显子的外侧都有一段不被转录的序列区, 它们也是分割不同基因的转录间隔序列, 称为旁侧序列 (flanking sequence)。其中可能存在一些调控序列, 包括启动子、增强子和终止子等, 这些序列对基因表达都起着调控作用。

2. 单顺反子 与原核基因显著不同, 真核生物基因都是由一个可转录序列和与之相关的转录调控序列组成, 转录物为单顺反子 (monocistron), 即一个编码基因转录生成一个 mRNA 分子, 经翻译生成一条多肽链。很多真核蛋白质由几条不同的多肽链组成, 因此需要多个基因协调表达。而在原核生物中, 功能相关的多个基因常常串联在一起, 并由同一组调节序列调节其转录, 构成一个转录单位, 其转录物为多顺反子 (polycistron), 指导多个蛋白质的合成。

3. 断裂结构 真核基因内部由编码序列和非编码序列相间排列组成。DNA 初级转录物 (primary transcript) 亦称 mRNA 前体 [核内不均一 RNA (hnRNA)]。初级转录物借剪接 (splicing) 机制去除非编码区的转录间隔序列的内含子 (intron) 后拼接为连续的编码序列, 最终形成成熟 mRNA (mature mRNA)。不同剪接方式可形成不同的 mRNA, 翻译出不同的多肽链, 因此转录后的剪接过程是真核基因表达调控的另一重要环节。间断序列可能参与分割编码信息, 并有利于以重组的方式实现 DNA 的重新排列及促进生物的进化。真核基因由内含子、外显子组成, 即基因断裂结构是真核基因的重要特点之一。原核基因的编码序列是连续的, 没有内含子, 转录后不需要剪接。

第二节 基因组的结构与功能

一、基因组的概念

基因组 (genome) 是指一个生物体中一套完整单倍体遗传物质的总和, 即包含在该生物体内 DNA (部分病毒是 RNA) 中的全部遗传信息。例如, 生物个体体细胞中的二倍体由两套染色体组成, 其中一套 DNA 序列就是一个基因组。基因组一词可以特指整套核 DNA (例如核基因组), 也可以用于包含自己 DNA 序列的细胞器基因组, 如线粒体基因组或叶绿体基因组。

人类基因组包含了 22 条常染色体 DNA、2 条性染色体 (X 染色体和 Y 染色体) DNA 以及线粒体 DNA 中的所有遗传信息。

不同的生物体的基因组的大小、结构以及所贮存的遗传信息量有显著差别。进化程度越高的生物, 其基因组的结构与组织形式越复杂。部分已测基因组的大小列表如下 (表 1-1)。

表 1-1 原核生物和真核生物基因组大小比较

原核生物		真核生物	
生物体种类	基因组 (bp)	生物体种类	基因组 (bp)
病毒 SV40	5.2×10^3	阿米巴变形虫	6.7×10^{11}
噬菌体 Φ -X174	5.4×10^3	贝母	1.3×10^{11}
噬菌体 λ	5×10^4	酵母	1×10^8
大肠埃希菌	4×10^6	秀丽隐杆线虫	8×10^7
		黑腹果蝇	2×10^8
		人	3×10^9

二、病毒基因组的结构特点

病毒基因组的大小和结构有比较大的差异。如乙型肝炎病毒 DNA 只有 3 千碱基 (kb), 编码 4 种蛋白质, 而昆虫痘病毒的基因组 DNA 达 300kb, 可以编码几百种蛋白质。但总体比原核生物和真核生物要小得多。另外病毒基因组有的由 DNA 构成, 有的由 RNA 构成。

1. 病毒基因组可以由 DNA 或 RNA 组成 病毒的核酸通常是 DNA 或 RNA, 很少有二者共存。病毒的 DNA 分子或 RNA 分子可以与蛋白质共同组成病毒颗粒。病毒颗粒的组成有简有繁, 有的病毒自身带有基因组复制所需要的部分酶, 有的则完全需要依赖于宿主细胞。

2. RNA 病毒基因组有单、双链和正、负链之分 大多数 RNA 病毒的基因组是单链 RNA 分子, 但也有一部分 RNA 病毒的基因组为双链 RNA 的结构。在单链 RNA 病毒中, 根据基因组 RNA 是否可以作为 mRNA 模板指导蛋白质的生物合成, 分为正链 RNA 病毒和负链 RNA 病毒两大类。如脊髓灰质炎病毒 (poliovirus)、SARS (重症急性呼吸综合征) 冠状病毒和鼻病毒 (rhinovirus) 感染宿主后, 可直接作为 mRNA, 是蛋白质合成的模板, 翻译出病毒所编码的蛋白质, 包括衣壳蛋白和病毒 RNA 聚合酶。然后在病毒 RNA 聚合酶的作用下以基因组 RNA 为模板合成出负链, 再以负链为模板复制病毒 RNA, 以复制的病毒 RNA 和衣壳蛋白进行自我装配成为成熟的病毒颗粒。这些病毒成为单股正链 RNA 病毒。如滤泡性口腔炎病毒、流行性感病毒、副流行性感病毒等 RNA 病毒, 它们的 RNA 序列与 mRNA 序列互补, 进入宿主后不能直接作为模板指导蛋白质合成, 而是先以基因组 RNA 为模板转录生成互补 RNA, 再以这个互补 RNA 作为 mRNA 模板翻译出遗传密码所决定的蛋白质, 再进行包装。种类病毒称为负链 RNA 病毒。负链 RNA 病毒没有感染性, 需要转录生成 mRNA 后才具有感染性。

肠易孤病毒 (reovirus)、轮状病毒 (rotavirus) 的基因组是双链 RNA 病毒, 都含有正、负两条 RNA 链。这两条链都有经过复制产生互补链的功能, 但只有正链是 mRNA, 具有编码能力。

部分 RNA 病毒基因组可以反转录为 DNA, 该类病毒通常引起人和动物的肿瘤。如人类免疫缺陷病毒 (human immunodeficiency virus, HIV)、白血病毒 (leukemia virus)、肉瘤病毒 (sarcoma virus) 等。

3. DNA 病毒基因组有环状 DNA 分子和线性 DNA 分子之分 DNA 病毒基因组也有单、双链和正、负链之分。腺病毒 (adenovirus) 是线性双链 DNA 病毒, 其基因组 DNA 长度为 35.8~36.2kb, 编码两大类蛋白质。乳头瘤病毒 (papillomavirus) 的基因组 DNA 为双链环状 DNA, 其长度在 7.2~8.0kb, 主要编码与病毒复制、转录、调控和细胞转化有关的蛋白质。

噬菌体 Φ -X174 病毒基因组为单链环状 DNA 分子, 有 5.387kb。

三、原核基因组的结构特点

原核基因组比真核基因组小得多, 结构简单, 通常由一条环状的双链 DNA 分子组成, 大小在 6×10^5 bp (如支原体) 到 8×10^6 bp (如固氮菌), 与蛋白质以复合物的形式存在。可转录序列与调控序列以操纵子的形式将多个功能上相关的基因串联在一起, 共用一个启动子, 其转录物为多顺反子。编码蛋白质的基因多为单拷贝基因 (编码 rRNA 的基因是多拷贝基因)。基因组中重复序列少, 非编码序列少, 编码序列在基因组中所含的概率远大于真核基因组。

1. 原核基因组通常由环状双链 DNA 分子组成。
2. 操纵子结构是原核基因组的结构特点之一。

原核生物的绝大多数结构基因按功能相关性成簇地串联排列在染色体上, 连同其上游的调控区 (包括启动子和操纵元件) 以及下游的转录终止信号共同构成一个基因表达单位, 即操纵子 (operon) 结构。如乳糖 (*lac*) 操纵子、阿拉伯糖 (*ara*) 操纵子等。(详细参考基因表达调控一章)。其调控主要以阻遏蛋白的负调控为主。

3. 原核基因组中的基因密度非常高, 编码区所占的比例较大。多为单拷贝, 没有内含子。
4. 基因组中的重复序列很少, 几乎没有类似于真核基因组内广泛存在的重复序列。
5. 不同原核基因组中的 GC 含量 (GC content) 变化很大, 其范围从 25% 到 75%。因此, 测量基因组的 GC 含量可以用来识别细菌种类。
6. 原核基因组的非编码区内主要是一些调控序列。这些调控序列包括复制起始区 (ori), 复制终止区 (ter), 转录启动区和终止区等。大多数具有一定的共有序列。

四、真核基因组的结构特点

真核生物基因结构庞大, 基因信息的组织形式非常复杂, 含有巨大遗传信息的 DNA 分子与组蛋白结合, 以高度折叠、紧密卷曲的方式存在于细胞核的染色体内。

1. 真核生物核基因的结构 真核生物染色质 (chromatin) 由双链 DNA、组蛋白、少量非组蛋白与 RNA 构成。非组蛋白部分为与 DNA 复制有关的酶类, 如拓扑异构酶、RNA 聚合酶复合体。组蛋白为一组碱性蛋白质 (H1、H2A、H2B、H3 和 H4)。(H3/H4)₂—(H2A—H2B)₂ 形成的八聚体构成核小体的核心, DNA 分子环绕其上形成核小体 (nucleosome)。H2A—H2B 起稳定并保护环绕在 (H3/H4) 表面的 DNA 的作用。核小体直径为 10nm, 高 5nm, 由 30bp 长的 DNA 连接, 并以每圈 6~7 个核小体进一步盘绕成直径为 30nm 的超螺旋。H1 组蛋白起稳定超螺旋结构的作用。超螺旋被进一步压缩成更紧密复杂的结构。

对组蛋白共价修饰影响基因的功能, 修饰作用包括乙酰化、甲酰化、磷酸化、ADP (腺苷二磷酸) 核糖基化与泛素化。在全部 5 种组蛋白中, 只有 H2A 接受泛素化修饰作用。各种共价修饰作用影响染色质的结构与功能。例如, H3 与 H4 乙酰化修饰可激活或灭活基因转录, H1 组蛋白的磷酸化修饰与 DNA 复制过程中染色体的浓缩有关; 而组蛋白经 ADP 核糖基化修饰可能与 DNA 的修复有关。

在染色质 DNA 结构中存在活化区与非活化区。染色质活化区基因与非活性区有很多不同。活性染色体的核小体对脱氧核糖核酸酶 (DNase) I 敏感, 提示与转录相关。DNA 共价修饰是影响 DNA 转录的重要因素。在活性染色质区域内, 一些短核苷酸序列 (100~300bp) 构象的改变导致了对 DNase I 的高度敏感。高度敏感区域通常位于活化区基因的上游, 通常是在与非组蛋白转录调节因子的结合区域。

染色单体的组成具有特征性。在细胞分裂中期, 哺乳动物染色体具有相同的染色单体, 并由长度为 $10^2 \sim 10^6$ bp 的着丝粒连接, 此区域富含腺嘌呤——胸腺嘧啶 (A-T)。染色体的末端

为富含 TG 重复序列的端粒。人类单倍体基因组 (3×10^9 bp) 由大约 1.7×10^7 个核小体组成, 平均包含 1.3×10^8 个核苷酸。

2. 真核生物核基因的重复序列 核基因组结构庞大, 编码序列所占的概率远小于非编码序列。

人类单倍体基因组有 23 条染色体, 全部单倍体基因组可编码约一百五十万等长基因。利用核酸杂交测定哺乳类细胞含 5000~10 000 种 mRNA, 由此计算哺乳类基因组有四万个以上的基因。按每个编码基因含 1500 个核苷酸计算, 这些基因约占全部基因组的 6%。在非编码序列中, 一些 DNA 序列承担基因表达的调控作用; 更多的非编码序列由功能未知的重复序列构成, 可能没有直接的遗传学功能, 这是真核基因组与原核基因组的不同之处。

人类基因组 DNA 至少由 30% 的重复序列组成。重复序列长短不一, 重复频率也不尽相同。短重复序列在十个核苷酸以下, 长重复序列达数百乃至上千个核苷酸。根据重复频率可将重复序列分为高度重复序列 (highly repetitive sequence)、中度重复序列 (moderately repetitive sequence) 及单拷贝序列 (unique sequence)。高、中度重复序列统称多拷贝序列; 单拷贝序列在整个基因组中只出现一次或很少的几次。重复序列及基因重组均与生物进化有关。重复序列有种属特异性, 基因组愈大, 重复序列含量愈丰富。某些重复序列发生在调控区, 如转录终止区、衰减调控区及某些酶或蛋白因子的结合位点, 可能对 DNA 复制、转录调控具有重要意义。

(1) 高度重复序列 通常由很短的碱基序列组成, 长度为 2~300bp, 重复频率可达 10^6 次。这类高度重复序列在基因组 DNA 中所占的概率随着种属不同而有差异。一般占 DNA 碱基对的 10%~30%。高度重复序列可分为两种类型: ①卫星 DNA 是最主要的高度重复序列, 含 G-C 少, A-T 较多。它的重复单位一般由 2~10bp 组成; ②回文顺序是一类特殊形式的序列, 又称反向重复序列 (inverted repetitive sequence), 即在 DNA 分子的某些片段, 两条链碱基排列顺序相同而方向相反的重复。通常, 这些序列在单倍体基因组内成簇地串联排列在丝粒与端粒内, 拷贝数在 $1 \sim 10^6$ 。此类序列通常不具转录活性, 组成染色体的结构成分。

微卫星重复序列 (microsatellite repetitive sequence) 是一类散在的串联排列重复序列, 长 2~6bp, 重复可达 50 次。微卫星重复序列常在一条 DNA 链中由 AC 构成, 另一条 DNA 链对应为 TG 重复序列。除 AC 重复序列外, 其他类型为 CG、AT 与 CA 重复序列。在任何基因座上, 同一种微卫星重复序列在两条染色体的重复数目完全不同, 因此决定了特定微卫星拷贝数的杂合性 (heterozygosity)。杂合性是一种遗传特征, 通过聚合酶链式反应 (PCR) 扩增 AC 重复序列已经成为建立基因连锁图谱的基本方法。很多基因与一种或几种微卫星标记相关, 通过测定标记的位置可确定染色体上致病基因的相对位置。

(2) 中度重复序列 中度重复序列指单倍体基因组内少于 10^6 次拷贝的重复序列, 不成簇分布, 仅仅散在分布于 DNA 序列中。不同的中度重复序列的长度和拷贝数差别较大, 由几百至几千个碱基对组成, 平均长度约为三百 bp。在基因组 DNA 中可重复 $10^2 \sim 10^5$ 次。中度重复序列中有一部分是 rRNA、组蛋白和 β -珠蛋白的编码序列。很多情况下, RNA 聚合酶 II 催化合成这些长重复序列, 并具有与 mRNA 同样的“帽子”结构。

依照中度重复序列的长度, 将其分为长散在重复序列 (long interspersed repetitive sequence, LINE) 与短散在重复序列 (short interspersed repetitive sequence, SINE)。长散在重复序列与短散在重复序列都属于反转录子 (retroponson)。反转录酶催化反转录子的生成, 以 RNA 为模版生成 DNA 后转座 (transposition) 到基因结构中。哺乳动物基因组含 2 万~5 万个长度为 5~7kbp 的长散在重复序列。长度为 70~300bp 的短散在重复序列在基因组内拷贝数超过 1 万。人类基因组一种名为 Alu 家族的短散在重复序列, 每个单倍体基因组的拷贝数为 50 万, 占人类基因组的 5%~6%。中等重复序列具有的转座功能可导致严重的结果, 如 Alu 序列插入

基因后导致基因的突变是多发性神经纤维瘤的直接原因。

(3) 单拷贝序列 单拷贝序列又称非重复序列, 基因组 DNA 只有单一的拷贝或少数几个拷贝。一般由 800~10 000bp 组成。它们是编码细胞中的各种蛋白质和酶的基因。

3. 多基因家族 多基因家族 (multigene family) 是由某一祖先基因经过重复和变异所产生的一组基因, 它们可成簇地分布在同一染色体上, 亦可成簇地分散在不同的染色体上, 不同的家族成员编码一组功能密切相关的蛋白质。

4. 假基因 1977 年, G. Jacp 在对基因家族的研究后提出了假基因 (pseudogene) 的概念。中等重复序列的 Alu 家族在结构式上与反转录病毒的末端相似, 具有进出哺乳动物基因组的能力。小 DNA 单元进出人类基因组的现象得到证实。现已发现, 一些与免疫球蛋白有关的加工基因 (processed gene) 具有转座的能力。此类加工基因由类似 mRNA 的序列组成, 具有 5' 端非转录区与 3'-polyA 结构。这些具有转录功能的基因转录生成 DNA 初级转录物, 经剪接加工去除内含子后反转录成互补 DNA (cDNA), 最终整合到染色体 DNA 中。在长期的进化过程中, 这类基因已经被随机修饰改变了结构。由于这类基因含有无意密码子而不能够编码功能性蛋白质, 因此它们被命名为假基因。大多数真核基因组中都存在有假基因, 假基因一般不含内含子, 具有使自身结构进出宿主基因组的能力, 由此影响相邻 DNA 序列的功能。假基因可影响物种的进化过程。

第三节 基因组学

21 世纪被誉为生命科学的世纪, 作为生命科学研究热点, 基因组学是一门系统的研究某物种全部基因及其产物所构成的生理结构和生命活动的学科。1986 年美国著名 Thomas Rodrick 教授创立了基因组学 (genomics) 的概念, 并被广泛接受, 而今基因组学已成为生命科学领域最活跃、最有影响的前沿学科。

一、基因组学的概念

基因组学 (genomics) 是研究生物基因组的组成, 组内各基因的精确结构、相互关系及表达调控的科学, 是用于概括涉及基因作图、测序和整个基因组功能分析的遗传学分支。该学科提供基因组信息以及相关数据系统利用, 试图解决生物、医学、和工业领域的重大问题。基因组学的研究内容主要包括: 结构基因组学 (structural genomics)、功能基因组学 (functional genomics) 和比较基因组学 (comparative genomics)。近年来, 随着上述基因组学研究的不断深入, 一系列的基因组学也应运而生, 如环境基因组学 (environmental genomics)、行为基因组学 (behavioral genomic)、营养基因组学 (nutritional genomics) 和药物基因组学 (pharmacogenomics) 等, 反映了不同学科交叉融合而产生的新的研究方向。

二、结构基因组学

结构基因组学是以全基因组测序为目标, 确定基因组的组织结构、基因组成及基因定位的一个基因组学分支, 研究蛋白质组成和结构的学科, 为基因功能的研究奠定基础。结构基因组学是继人类基因组之后又一个国际性科学热点, 主要目的是试图在生物体的整体水平上 (如全基因组、全细胞或完整的生物体) 测定出 (以实验为主, 包括理论预测) 全部蛋白质分子、蛋白质-蛋白质、蛋白质-核酸、蛋白质-多糖、蛋白质-蛋白质-核酸-多糖、蛋白质与其他生物分子复合体的精细三维结构, 以获得完整的、能够在细胞中定位, 以及在各种生物学代谢途径、生理途径、信号传导途径中全部蛋白质在原子水平的三维结构全息图。在此基础上, 使人们有