

现代遗传学丛书

基因及其表达

童克中 著

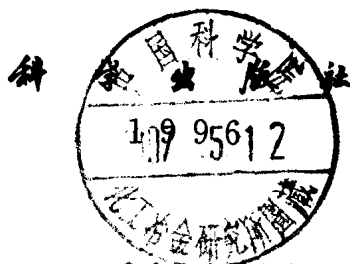
科学出版社

58.215
587

现代遗传学丛书

基因及其表达

童克中 著



(京)新登字 092 号

内 容 简 介

本书取材于近期发表的原始论文和评论,而不以任何教科书为蓝本,以简要的形式全面介绍遗传的物质基础、遗传物质的组织结构、遗传密码、基因突变、基因组复制、基因重组、转录、翻译、原核基因表达的调节、真核基因的表达及其调节等方面的最新成果;并提出许多值得进一步研究的问题。对于生物学教学、研究人员来说,阅读本书是了解分子遗传学最新进展的捷径。

本书可供生物学、医药学、农牧学等方面的科研人员、教师、研究生和高年级大学生参考。

34136/12

现代遗传学丛书

基因及其表达

童克中 著

责任编辑 刘 安

科 学 出 版 社 出 版

北京东黄城根北街 16 号

邮政编码: 100717

中国科学院印刷厂印刷

新华书店北京发行所发行 各地新华书店经售

*

1996 年 1 月第 一 版 开本: 850×1168 1/32

1996 年 1 月第一次印刷 印张: 9 3/8 插页: 2

印数: 1—4000 字数: 244 000

ISBN 7-03-004753-2/Q · 587

定价: 18.00 元

前 言

1982 年以来, 著者在中国科学院研究生院讲授分子遗传学课程, 每学年 80 学时。本人授课不以任何教科书为蓝本, 而取材于各种期刊近期发表的学术论文, 旁及各种评论、专著, 编成详细提纲约 300 打字页, 分发学生。此提纲每年进行一次增、删, 力求保持一个“新”字。由于每年修改提纲、每年授课, 对分子遗传学的最新发展, 自信是比较熟悉的。但本人并无意整理成书。分子遗传学的发展如此迅速, 分子遗传学著作的寿命必然很短。以多病之躯, 要在短时间内手写这么多字, 令人望而却步。1991 年国家自然科学基金委员会《遗传学发展战略》调研小组邀请参加撰写遗传学发展战略调研报告, 分工由本人负责遗传物质的组织结构和基因表达的调节两章。由于当时没有规定每章篇幅的长短, 一写写了约八万言。这就是本书第一章遗传的物质基础、第二章遗传物质的组织结构、第七章转录、第八章翻译、第九章原核基因表达的调节、第十章真核基因的表达及其调节等 6 章的骨架; 以后逐年增加资料。1994 年 7 月在杭州开会, 再次研究分工, 并且规定每一篇章大致字数, 结果由本人负责的两章, 总共不过二万言。我以前搭的架子显然失之过大, 干脆加一把劲, 一气写成遗传密码 (本书第三章)、基因突变 (第四章)、基因组的复制 (第五章) 和基因重组 (第六章) 等 4 章, 和前 6 章合并, 包括了分子遗传学的全部重要内容, 成为这本小书。国内从事分子遗传学教学、研究的人员很多, 但未必都有充裕的时间关心分子遗传学各方面的发展。对于这样的读者来说, 本书提供了一条捷径, 使你只要花很短时间, 就可以了解分子遗传学的最新发展。本书的优点是简要; 缺点也在于简要, 没有多用图表、没有多作解释; 但

只要稍加留意，都是可以理解的。

本书由国家自然科学基金委员会《遗传学发展战略》项目拨款资助出版，我在此表示由衷的感谢！

童克中

1995 年 5 月 29 日

目 录

前言

第一章 遗传的物质基础	1
第一节 DNA 作为遗传物质	1
第二节 RNA 作为遗传物质	1
第三节 蛋白质可以作为遗传物质吗?	3
第二章 遗传物质的组织结构	6
第一节 C 值	6
第二节 遗传信息重叠现象	8
第三节 割裂基因	10
第四节 可动基因	13
第五节 基因丰余	17
第六节 基因群	18
第七节 基因的精细结构	19
第八节 染色质	22
第九节 线粒体基因组	24
第十节 叶绿体基因组	26
第三章 遗传密码	30
第一节 摆动	30
第二节 扩展的反密码子假说	32
第三节 “副密码子”	33
第四节 起始密码子	37
第五节 终止密码子	39
第六节 密码子使用	40
第七节 密码子中的密码	43
第八节 异常密码子	46

第九节	遗传密码的起源和进化	46
第四章	基因突变	55
第一节	自发突变	55
第二节	DNA 损伤的修复	61
第三节	呼救 (SOS) 系统	65
第四节	化学诱变	67
第五节	诱变的特异性问题	69
第六节	适应性突变	70
第七节	反义核酸	71
第五章	基因组复制	73
第一节	单股 DNA 基因组	73
第二节	细菌基因组	77
第三节	T 奇数噬菌体	84
第四节	酿酒酵母	87
第五节	动物基因组	95
第六节	SV40	98
第七节	腺病毒	101
第八节	反转录病毒	103
第九节	乙型肝炎病毒	105
第十节	单股 RNA 病毒	106
第十一节	呼肠孤病毒	108
第十二节	线粒体基因组	109
第十三节	高等植物和叶绿体基因组	111
第十四节	其他基因组	113
第六章	基因重组	116
第一节	参与同源重组的蛋白质	116
第二节	同源重组的主要途径	132
第三节	位置特异性重组	136
第四节	转座	139
第五节	抗体基因的体细胞重组	144

第七章 转录	147
第一节 转录酶.....	147
第二节 转录因子.....	151
第三节 启动子.....	159
第四节 转录.....	165
第八章 翻译	177
第一节 核糖体.....	177
第二节 tRNA	182
第三节 蛋白质因子.....	184
第四节 翻译.....	187
第五节 翻译错误.....	196
第九章 原核基因表达的调节	203
第一节 转录调节.....	203
第二节 翻译调节.....	212
第十章 真核基因的表达及其调节	224
第一节 低甲基化.....	224
第二节 组蛋白的变化.....	225
第三节 激活蛋白.....	226
第四节 阻抑蛋白.....	232
第五节 调节蛋白的七巧板式结构.....	233
第六节 基因表达的时空控制.....	237
第七节 剪接.....	241
第八节 编辑.....	260
第九节 翻译调节.....	265

第一章 遗传的物质基础

第一节 DNA 作为遗传物质

DNA 作为遗传物质有许多优点,其中最主要的是:①信息量大¹⁾,可以缩微;②表面互补、电荷互补,双螺旋结构说明了精确复制的机理;③核糖的 2' 脱氧,在水溶液中稳定性好;④可以突变,以求进化;⑤有 T 无 U,基因组得以增大,而无 C 脱氨基成 U 带来的潜在危险。然而,如果 DNA 是最初的遗传物质,那么由于 DNA 复制需要酶,而酶是蛋白质,蛋白质又是由 DNA 的核苷酸序列编码的,这就成了一个鸡生蛋、蛋生鸡的问题。80 年代发现 RNA 拟酶,这个问题才得到解决。

第二节 RNA 作为遗传物质

RNA 拟酶集信息传递作用和酶学催化作用于一身,很可能是最初的遗传物质。在这个基础上,一个由 RNA 世界到 RNA 蛋白质世界,由 RNA 蛋白质世界到 DNA 世界的进化图景,已被科学界广泛接受。但 RNA 作为最初遗传物质的设想,仍然有许多疑难。其中最大的疑难是 RNA 本身的起源问题。根据有机化学家的意见,在模拟的前生物条件 (prebiotic condition) 下,合成核糖并非易事;即使有了核糖, D-核糖和碱基一起加热,得到的是 α -核糖核苷,而不是 RNA 中的 β -核糖核苷等等^[1],但目前没有更好的

1) 一个 1 kb 的基因,其可能的核苷酸排列顺序有 $4^{1000} = 10^{602}$ 种,而直径 10 亿光年的宇宙,其体积不过 10^{84} cm^3 或 $10^{108} \text{ \AA}^3$ 。(见 M. Eigen, Steps Towards Life, Oxford University Press, 1992)

解释。关于前生物条件下核酸前体如何生成的研究工作，至今仍在进行。另外，目前也正在研究前生物条件下，自我复制体系是怎样产生的^[2-4]。至于有关RNA以后进化历程的研究工作，却正以稳健的步伐前进。指数富集配基系统进化法（systematic evolution of ligands by exponential enrichment, SELEX）正日趋完善。四膜虫Ⅰ类内含子的自我剪接，要以 Mg^{2+} 或 Mn^{2+} 为辅基；以 Ca^{2+} 、 Sr^{2+} 或 Ba^{2+} 为辅基，能够正确折叠，但无剪接活性；经8代突变、扩增、选择，得到能以 Ca^{2+} 为辅基而且能自我剪接的新的RNA拟酶。将T4噬菌体的*sunY* RNA拟酶的非保守部分去掉，酶活性失去，但在离体遗传选择条件下，又选得复活的*sunY*缺失突变体，其活性高于全长*sunY*^[5]。从RNA随机序列中筛选到一种新的RNA拟酶，能把核苷酸的3'-OH连接到核苷酸的5'-磷酸上，反应速度提高700万倍^[6]。将野生型基因的随机片段离体改组（shuffling），使蛋白质快速进化^[7]。酵母线粒体*cob*基因第一个Ⅰ类内含子完成剪接第一步之后，能凭磷酸酐键提供的能量，把第一步剪接反应逆转；剪下的内含子能将ssRNA切断；能把ssDNA连接在ssRNA分子上，虽然其连接活力比RNA-RNA低18倍。这头一种活性说明RNA拟酶能利用单核苷三磷酸作为底物，这后两种活性说明在反转录酶出现之前，可以由RNA世界直接进入DNA世界^[8]。根据RNA拟酶锤头状结构而设计的抗病毒药物，正在走向实用阶段^[9]。

以胸腺嘧啶T取代尿嘧啶U是DNA生物有别于RNA生物的重大步骤。胞嘧啶C经氧化脱氨基就成为U。假如DNA也和RNA一样，没有T而只有U，那末由C脱氨基而来的U和原有的U就无法区别，假李逵可以冒充真李逵，这样就不能保持遗传物质的稳定性。特别是像人这么大的基因组，其DNA总长度达30亿对碱基，C脱氨基成U的现象并不罕见，任何一个细胞都不能承受C→U带来的严重后果。以T代U根除了这种潜在的灾难，因为尿嘧啶DNA糖苷酶可以灵敏地识别DNA中的U而随时将其剔除。由于DNA中有T无U，所以DNA基因组可以大幅度扩

增,生物体结构得以进化到高级、复杂的程度,才有可能出现人类。迄今已知RNA生物,都是基因组小、结构简单的生物。

由U转变为T的代谢途径已经研究清楚。这条代谢途径充分说明U的出现在先,T在后。DNA中以T代U的进化历程尚不清楚,因为没有发现中间类型。迄今已知唯一含U的DNA生物是枯草杆菌噬菌体PBS1,其基因组相当大。自发现迄今已30余年,是实验室常用的转导噬菌体。其遗传性稳定,唯一例外的是经常出现清亮噬菌斑,不知是否是由于C脱氨基成为U之故。

第三节 蛋白质可以作为遗传物质吗?

蛋白质因其缺乏遗传表面,而且遗传嗅觉不灵敏,不能作为遗传物质,早已成为定论。prion(传染性蛋白质颗粒)问题对此带来一片疑云。把羊疯痒病(scrapie)、牛海绵状脑炎、人kuru病、Creutzfeldt-Jacob综合征、Gerstmann-Straussler-Scheinker综合征的病原传染性蛋白质颗粒(proteinaceous infectious particle)定名为prion,为的是和类病毒、拟病毒相区别。prion蛋白(prion related protein,简称PrP)由宿主的基因编码。人PrP基因位于20号染色体的短臂,鼠PrP基因位于2号染色体。健康人、畜的PrP^c分子量为33—35 kDa,它对蛋白酶敏感、无传染性。而病人、病畜PrP^{sc}的N端比PrP^c的约少67氨基酸,分子量27—30 kDa,抗蛋白酶,有传染性。迄今未见prion内有常量核酸。用种种降解核酸的办法提取传染性颗粒,传染性不会下降。据计算,即使残留有核酸,其长度亦不会超过80个核苷酸^[10]。传染时也不携带外源核酸^[11]。此病潜伏期有长、短之分,有种属屏障,有不同的株系。潜伏期长短之不同,由突变氨基酸所处位置和突变的性质而定^[12]。从十几年的研究结果来看,PrP^c转变为PrP^{sc}是由于翻译后修饰之故^[13—15]。PrP^c有 α -螺旋而无 β -片层,PrP^{sc}则多 β -片层;PrP^{sc}的N端剪短,增加 β -片层,但无损于传染性^[16]。PrP^{sc}构象的不同

是否和株系不同有关, 仍有待研究。PrP^c的原有功能不明^[17]。将小鼠编码 PrP^c的基因敲掉(knock out), 然后用传染性蛋白质颗粒侵染, 在众多感染小鼠中, 起码能活上 2 年, 而且健康, 无疯痒病征; PrP^c基因健在的小鼠, 经传染性颗粒侵染后, 半年之内, 统统死亡^[18]。此实验充分证明 PrP^{sc}确由 PrP^c译后修饰而来, 也排除了微量外源核酸的作用问题。最近报道酵母线粒体伴娘蛋白 Hsp60 (由核基因编码) 缺失突变体中, 输入的野生型 Hsp60 不能折叠、组装成为有功能的 14 聚体。共生蛋白(symbionin)和 GroEL (细菌 Hsp60) 有 85% 同源, 能使叶绿体中未折叠的核酮糖二磷酸羧化酶亚基在离体条件下正确折叠起来^[19]。根据这些现象, 可以推测病人、病畜的 PrP^{sc}指导健康人、畜 PrP^c折叠、组装, 使人、畜罹病。如果确实如此, 那么第一个 Hsp60 14 聚体是由什么指导折叠组装的? 第一例羊疯痒病是怎样发生的呢?

主要参考文献

- [1] Orgel, L. E., in *Evolutionary Tinkering in Gene Expression*, 1989, 215—224, ed. M. Granberg-Manago, Plenum.
- [2] Orgel, L. E., *Nature*, 1992, 358: 203—209.
- [3] Li, T. and Nicolaou, K. C., *Nature*, 1994, 369: 218—220.
- [4] Sievers, D. & von Kiedrowski, G., *Nature*, 1994, 369: 221—224.
- [5] Green, R. & Szostak, J. W., *Science*, 1992, 258: 1910—1915.
- [6] Bartel, D. P. & Szostak, J. W., *Science*, 1993, 261: 1411—1418.
- [7] Stemmer, W. P. C., *Nature*, 1994, 370: 389—391.
- [8] Morl, M. et al., *Cell*, 1992, 70: 803—810.
- [9] Barinaga, M., *Science*, 1993, 262: 1512—1514.
- [10] Kellings, K. et al., *Phil. Trans. R. Soc. Lond. B*, 1994, 343: 425—430.
- [11] Prusiner, S. B., *Phil. Trans. R. Soc. Lond. B*, 1994, 343: 447—463.
- [12] Carlson, G. A. et al., *Phil. Trans. R. Soc. Lond. B*, 1994, 343: 363—369.
- [13] Prusiner, S. B., *Science*, 1991, 252: 1515—1522.
- [14] Carlson, G. A. et al., *Trends Genet.*, 1991, 7: 61—65.
- [15] Aiken, J. M. & Marsh, R. F. *Microbiol. Rev.*, 1990, 54: 242—246.
- [16] Baldwin, M. A. et al., *Phil. Trans. R. Soc. Lond. B*, 1994, 343: 335—341.

- [17] Collinge, J. et al. , Nature, 1994, 370: 295--297.
- [18] Sailer, A. et al. , Cell, 1994, 77: 967--968.
- [19] Hightower, L. E. , Cell, 1991, 66: 191--197.

第二章 遗传物质的组织结构

第一节 C 值

前章论及DNA中有T无U,是DNA生物基因组扩大、向高级、复杂进化的重要前提。从大尺度看,这是对的,某些噬菌体如MS2、Q β 、R17的基因组很小,人的基因组却很大。从另一个角度看,却又不尽然。人为万物之灵,其单倍基因组的数值(即C值)只有 10^9 bp,而肺鱼的C值居然比人更高。人和肺鱼亲缘关系相去甚远,很难相互比较。在亲缘关系相近的物种之间,C值仍然相差很大,两栖类的不同物种之间,C值可相差100倍,被子植物不同物种之间,C值相差达600倍。这就确凿证明C值的大小并不说明遗传复杂性的高低,而只说明基因组中自在DNA的多少。

对于任何一个基因组来说,其核酸都可以分为编码区和非编码区两类序列。编码区为信使RNA、核糖体RNA、转移RNA以及其他各种RNA编码。非编码区又可分为信号序列和非信号序列两类。复制起点、增强子、启动子、终止子及一切由调节蛋白识别和结合的位点,都属于信号序列。间隔区则属于非信号序列。但这种分法带有很大的随意性,例如细菌启动子-35区和-10区之间 17 ± 1 pb的核苷酸顺序不求一致,但过长、过短都是不合适的。把它看作间隔区固无可,但它含有信息。这信息不表现为核苷酸的顺序,而表现为序列的长度。又如MS2外壳蛋白结构基因上游的非编码序列不得少于30核苷酸;否则,翻译效率要降低10倍。真核基因的内含子并非编码序列,但其中含有重要信息,如5'剪接位点、3'剪接位点、3'剪接位点上游20—40核苷酸及套索中间体分枝点附近的序列,都有严格要求;其余部分含有的信息较

少,不易划出明确的界线。真核基因组含有大量的重复序列,其中高度重复序列可重复达百万次。细胞学上称为异染色质的部分,如端粒、中心粒都属于高度重复序列。高度重复序列的多少,对C值的影响可能最大。而中等重复序列的影响,也不可忽视。一般认为重复序列是无用的。但也不尽然,例如构成端粒的重复序列,既是稳定染色体并避免和其他断裂染色体或天然染色体发生融合之所需,又是保证线状染色体末端完整复制所不可忽缺的。重复序列还是进化的源泉之一。例如人的 *Alu* 家族重复序列,由于剪接或其他原因插到 mRNA 中,使蛋白质发生变化。从人 DNA 序列数据库中寻找编码区的 *Alu* 序列,结果在 15 种不同的编码序列中,找到 17 个插入的 *Alu* 片段^[1]。在脊椎动物中, *RMSA-1* (regulator of mitotic spindle assembly) 基因较保守。*RMSA-1* 基因的编码区有 2 份 *Alu* 元件,占编码区的 40%。第一份位于 168—278 位密码子,与全长 *Alu* 有 80% 同源;第二份两侧有 9 bp 的同向重复^[2]。鸡有 *RMSA-1* 而无 *Alu*。照此推理, *RMSA-1* 应该比 *Alu* 更为古老。重复序列的全面生物学意义尚有待阐明。

亚麻在氮、磷、钾、钙不同配比的条件下经一季栽培,其株高、分支状况、果壳茸毛等表型性状发生明显的变化^[3]。分析其 DNA 含量,发现多种重复序列,其中包括 rDNA 和 5S rDNA 都发生了变化^[4]。葱属 (*Allium*) 若干物种之间染色体数目相同,但 C 值相差达 4 倍,前述裸子植物不同物种之间 C 值相差可达 600 倍。不知道这种差别由何而来,这种差别在进化上有什么意义。

短散布重复序列长 100—300 bp,每个基因组中可达数十万份。例如人 *Alu* 基因家族共 14 成员,总拷贝数达 30—50 万份。灵长类其他物种亦然。啮齿类的 *B1* 基因家族、偶蹄类的 *C* 基因家族和人的 *Alu* 家族同源。以上这些重复序列,也和 7SL RNA 以及某些 tRNA 同源。果蝇和爪蟾的基因组没有 *Alu* 家族,而有 7SL。7SL 最初在反转录病毒中发现,所以推测 *Alu* 基因家族等是由于其一再反转录扩增而来。插入前的 *Alu* 序列相同,称为 *CONSBT* (conserved before insertion)。目前已测序的 *Alu* 超过 1 000 种。据

推算, 插入在 300—500 万年前发生, 300 万年以来很少变化^[5]。

灵长类和啮齿类还有长散布元件 (long interspersed nucleotide elements, 简称 *LINE*), 全长 5—7 kb, 称为 *L1*。以后在果蝇、锥虫、粗糙脉孢菌中也相继发现相关的序列, 是进化上保守的一个超基因家族, 其可读框 1 和反转录病毒的 *gag* 基因 (group specific antigen gene) 相似, 可读框 2 和反转录酶同源, 所以可把它看作反转录转座子。果蝇的 *LINE* 研究得最多, 其活跃转座会引起杂种劣势。哺乳动物基因组有 1—10 万份 *LINE*, 但全长并有转座活性的 *LINE* 很少见。*LINE* 在基因组的组织结构和进化中的作用如何, 为什么分布很广, 都值得研究^[6]。

除了内含子、重复序列对 *C* 值有很大影响外, 其他如丰余基因份数的变化、转座子、基因的缺失或重复, 也会对 *C* 值有影响。研究亲缘关系相近的物种之间 *C* 值不同的原因及其在进化上的作用, 是一项工作量很大的任务。许多种细菌也有重复序列。已知有两类, 其一称为 *REP* (repetitive extragenic palindrome), 处于两个操纵子之间或操纵子之尾, 一般 <200 bp, 大肠杆菌和鼠伤寒沙门氏菌都只有 38 bp, 500—1000 拷贝。另一类称为 *IRU* (interspersed repeat unit), 零散分布。细菌重复序列的功能不明, 分布各处的道理不明, 如何保持序列同源的机制不明。*REP* 结合促旋酶和 DNA 聚合酶 I, 而且 Hu 蛋白促进 *REP* 与促旋酶结合, 所以有人推测 *REP* 可能有助于把 DNA 折叠为超螺旋环。

第二节 遗传信息重叠现象

自从发现 $\phi \times 174$ 噬菌体基因重叠^[7]以来, 已经积累了许多关于遗传信息重叠的资料, 其中包括重叠操纵子、异相位重叠基因、同相位重叠基因、反向重叠基因等等。

大肠杆菌 *frd-ampC* (延胡索酸还原酶-氨苄青霉素) 中, *frdA B C D* 中的 *frdD* 基因 *C* 端 10 氨基酸编码区伸入 *ampC* 基因的启动子内, 但这两个操纵子在功能上毫不相干。

异相位重叠基因中最常见的一种是上游基因的终止密码子和下游基因的起始密码子重叠，例如 *trpE* 和 *trpD* 两基因交界处是…… UAAUG……，其中第二个 A 使用两次，翻译时 70S 核糖体不必重新组装。*galT* 的终止密码子在 *galK* 的核糖体结合序列之中，造成两个基因的翻译耦联，此时表达水平最高；加长终止密码子和核糖体结合序列之间的距离，表达水平下降；*galT* 终止密码子伸入到 *galK* 编码区以内，表达水平下降；伸入到 +98 以下，则 *galK* 不能翻译。

MS2、f2、R17 和 Q β 等 RNA 噬菌体的 *L* 基因跨 *C* 基因的 3' 端和 *R* 基因的 5' 端、以及其间的间隔区 36 核苷酸。*L* 基因要由翻译 *C* 基因的核糖体作 +1 移格而译出。*C* 基因有插入，*L* 基因不表达；*C* 基因发生琥珀色突变，*L* 不表达；宿主大肠杆菌发生抗链霉素突变 (*strA*)，*L* 不表达。+1 移格实际上是翻译错误。这种歪打正着的进化机制，很难理解。反转录病毒 *gag-pol* 的 -1 移格，与此相仿。

根据遗传密码起源的催化超循环 (catalytic hypercycle) 学说，原始密码子的碱基顺序应该是 RNY (R=A 或 G, Y=U 或 C, N=A 或 G 或 U 或 C)，+1 移格所得的 NYR 是以后进化的。NYR 多编码疏水氨基酸，所以这种重叠对改进蛋白质的性能，会起很大的作用。但迄今为止，可供统计、比较的数据仍嫌太少。

同相位重叠基因是重叠基因中最简单的一种，其所编码的蛋白质只是长短不同，其氨基酸顺序完全一样。RNA 噬菌体 MS2、Q β 的 *C* 基因因终止密码子读通得 A1 蛋白质；*StrA* 宿主不能读通，无 A1。*R* 基因因终止密码子被读通，得 66 kDa 的蛋白质。用动物细胞的无细胞体系翻译 MS2，则无 66 kDa 蛋白质。A1 和 66 kDa 的机能不明。Tn5 转座子的 P2 蛋白质比 P1 蛋白质 N 端少 40 个氨基酸，P1 是转座酶，P2 是转座酶的竞争性抑制物。这类重叠基因很多，往往接受共同的控制。

反向重叠基因两股 DNA 都有编码机能。IS5 由右向左编码一个较大的蛋白质，由左向右编码一个较小的蛋白质；两者读框相