

第十一章 基因组

第一节 原核基因组

第二节 真核基因组

第三节 基因重组和基因移动

原核生物只有一条染色体，包括该生物的全套基因，构成该生物的基因组。

真核生物则较复杂，体现正常细胞功能的全套染色体中的全部基因是真核细胞基因组。

那么一个DNA分子能携带多少信息？

一个多大的基因组才能携带机体所需的全部信息？

从病毒 → 哺乳动物，其基因组大大增加。

病毒DNA分子含 $10^3 \sim 10^5$ bp；

细菌基因组DNA分子含 $10^5 \sim 10^7$ bp；

人类基因组DNA的分子可至 10^9 bp。

按照平均1000bp (1kb) 编码一个蛋白质。

病毒可含4~5个基因，

大肠杆菌含 3000~4000个基因，

高等生物的染色体则可含有100万个以上基因。

2000：人超过14万 (Nature, 2000, 9. 23)；

2001：人只有3.7万个基因。

理论计算：

人染色体DNA应有200万个以上的基因，

而实际基因数3.7万个。

原因：真核DNA中含有非编码序列、基因

表达的调控序列和大量未知功能的序列。

另外，还存在大量非表达基因：

病毒：小于10%；

细菌：50-60%；

果蝇：90%；

鸡：98%；

人脑：78%（组织特异）

第一节 原核基因组

一、病毒基因组

二、细菌基因组

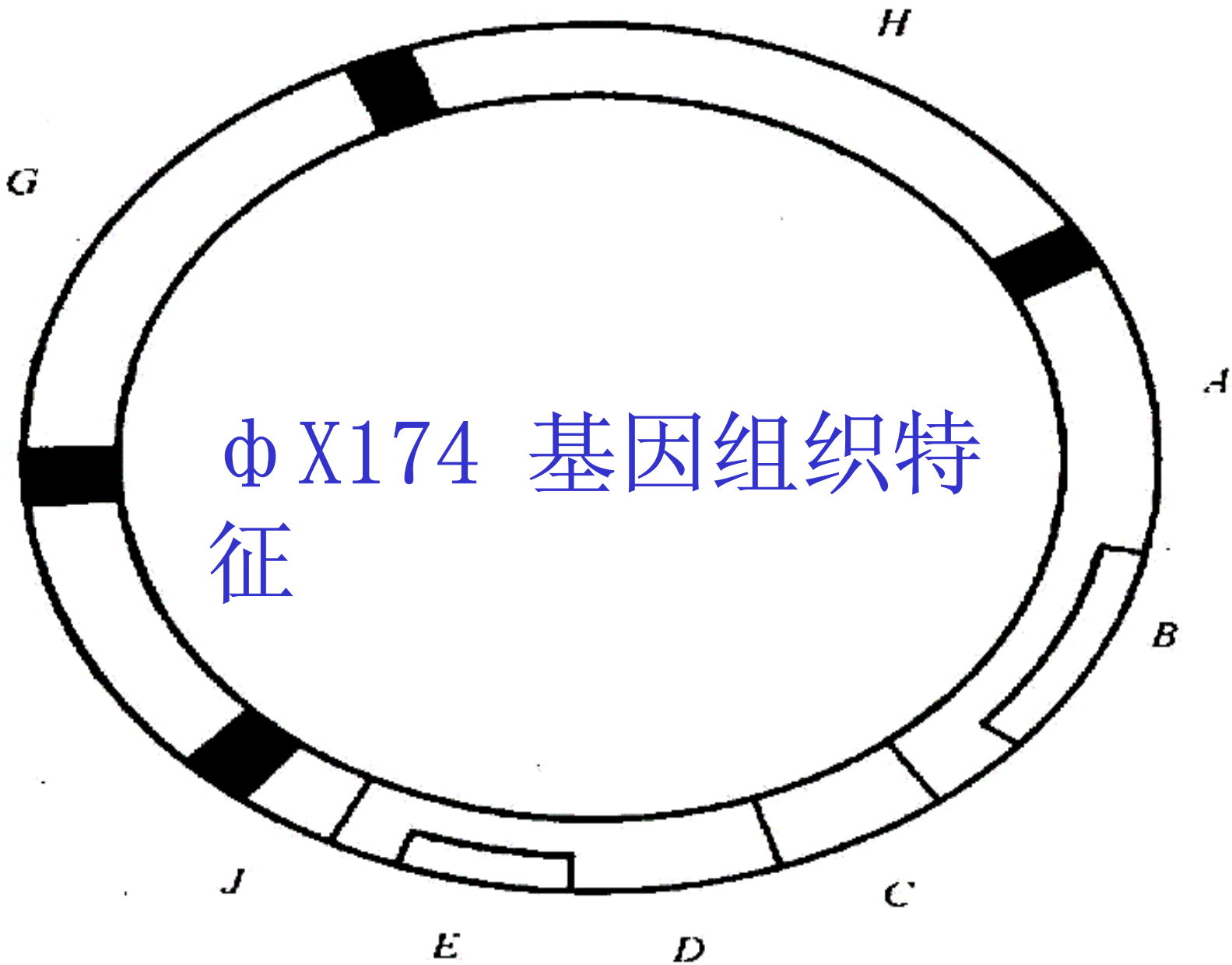
一、病毒基因组

几乎所有植物病毒、某些细菌病毒和动物病毒的基因组是由RNA组成，一般都特别小。

DNA病毒基因组范围较广，最小的病毒基因组仅有5kb左右（如 ϕ X174）；最大的200kb左右（如豆类病毒200kb）。

有些病毒的基因组不够编码自己的蛋白质，故出现基因重叠现象。

ϕ X174 基因组组织特征



二、细菌基因组

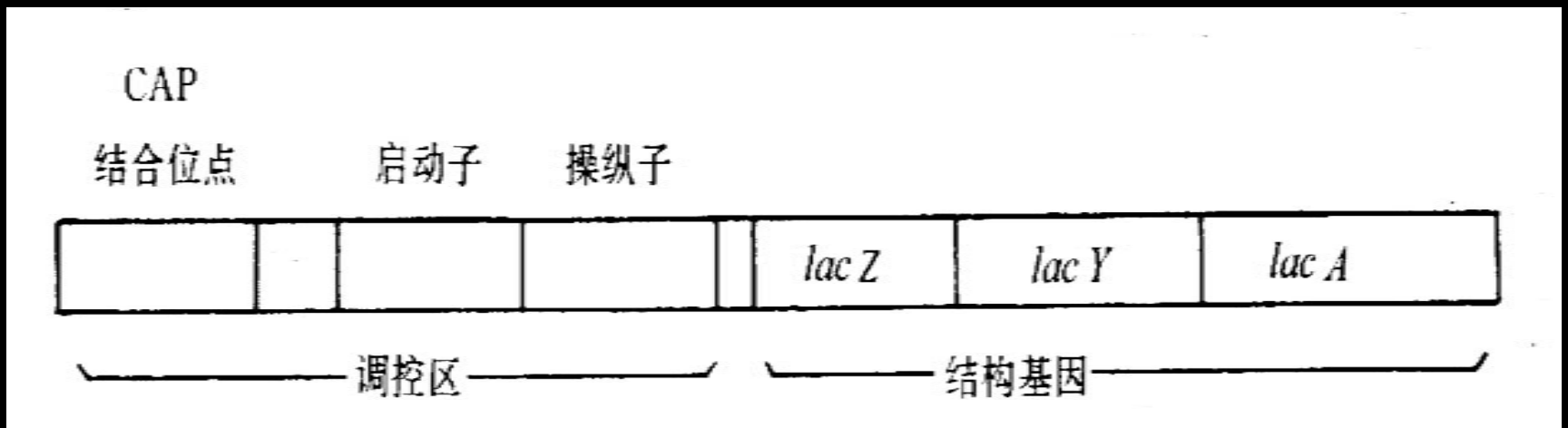
细菌含有染色体和染色体外的DNA—质粒。

大肠杆菌染色体比 λ DNA 约大200倍，是一般细菌长度的850倍，为单个双链环状DNA分子。

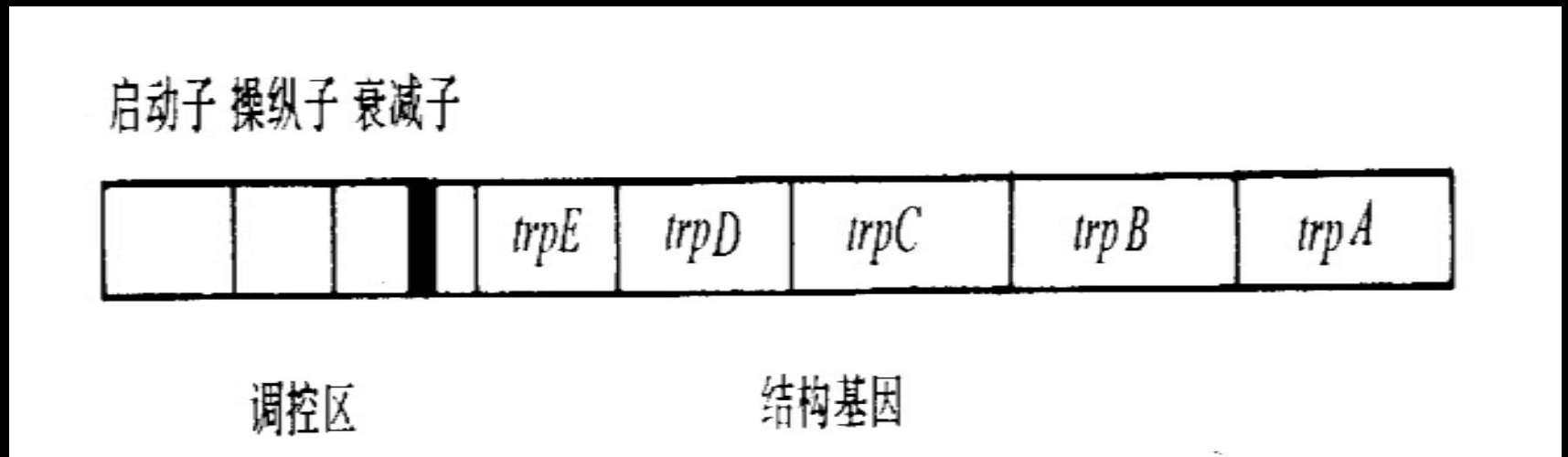
质粒 (plasmid)：存在于染色体外的细胞质环状双链DNA。它含遗传信息，能自主复制、传代细胞和表达。

细菌基因组的结构特征

- 功能上相关的基因串联在一起组成操纵子结构。
- 基因组中不存在内含子
- 基因组的绝大部分用于编码蛋白质或rRNA和tRNA，
- 少数基因有基因重叠现象。



乳糖操纵子示意图



色氨酸操纵子示意图

衰减子：一段阻遏蛋白结合的核苷酸序列

第二节 真核基因组

一、基因组大小

二、真核基因组结构组织特征

三、基因组计划及其意义

一、基因组大小

真核细胞含有更大的DNA分子。

酵母：是最小，比大肠杆菌大**4倍**。

果蝇：基因组比大肠杆菌的大**25倍**。

哺乳动物和人类：基因组比大肠杆菌大**600倍以上**。

真核细胞基因组DNA都是**线状双链DNA**，与组蛋白等多种蛋白质构成核小体。

单个人类细胞DNA约2米左右。

真核DNA的结构特征：

- 📖 编码蛋白质序列处在**不稳定的DNA序列**之中；
- 📖 大多数是**割裂基因**；
- 📖 **外显子**保持相对相似的序列，而**内含子**则存在广泛的序列变异；
- 📖 广泛存在**编码同源结构、相似功能蛋白质的基因家族**；
- 📖 **大量假基因**：保留在基因组中的无功能基因。

二、基因组的组织特征

(一) 重复序列

脊椎动物基因组90%以上是非编码序列，非编码DNA中有各种不同类型的重复序列，目前还没有发现它们有什么功能。

重复序列特点:

 重复频率呈现高度的多态性;

 重复序列的位置在相同种属的不同个体间不恒定;

 该DNA片段可移动;

 片段移动到新位置时, 可引起突变。

重序列的分类:

高度重复序列: 重复次数非常高的串联排列的短序列。

中度重复序列: 串联排列或分散存在的序列家族组成, 重复次数变化很大。

1. 高度重复序列

真核基因组DNA中G: G碱基对的分布不均一的，经等密度梯超离心分离后，出现一个主峰和一二个峰。

小峰对主峰而言，犹如主峰的卫星，所以称卫星DNA，是多种短重复序列的混合物。按照重复序列的长度，可将卫星DNA分成3类：

↓ 卫星DNA:

重复序列长度5~100bp，串联排列成总长度最长达100mb（m为 1×10^6 ）的DNA。多存在于异染色体，近中心粒和端粒。人群中多态性不强。

小卫星DNA:

重复序列长度15~70bp，串联排列
成总长度为0.5~30kb。主要存在于常
染色体。在人群中存在高度多态性。



微卫星DNA:

重复序列长度 $2\sim 6\text{bp}$ ，总长度有高度变化。存在于常染色体。也有高度多态性。

这种重复序列又称简单串联重复序列

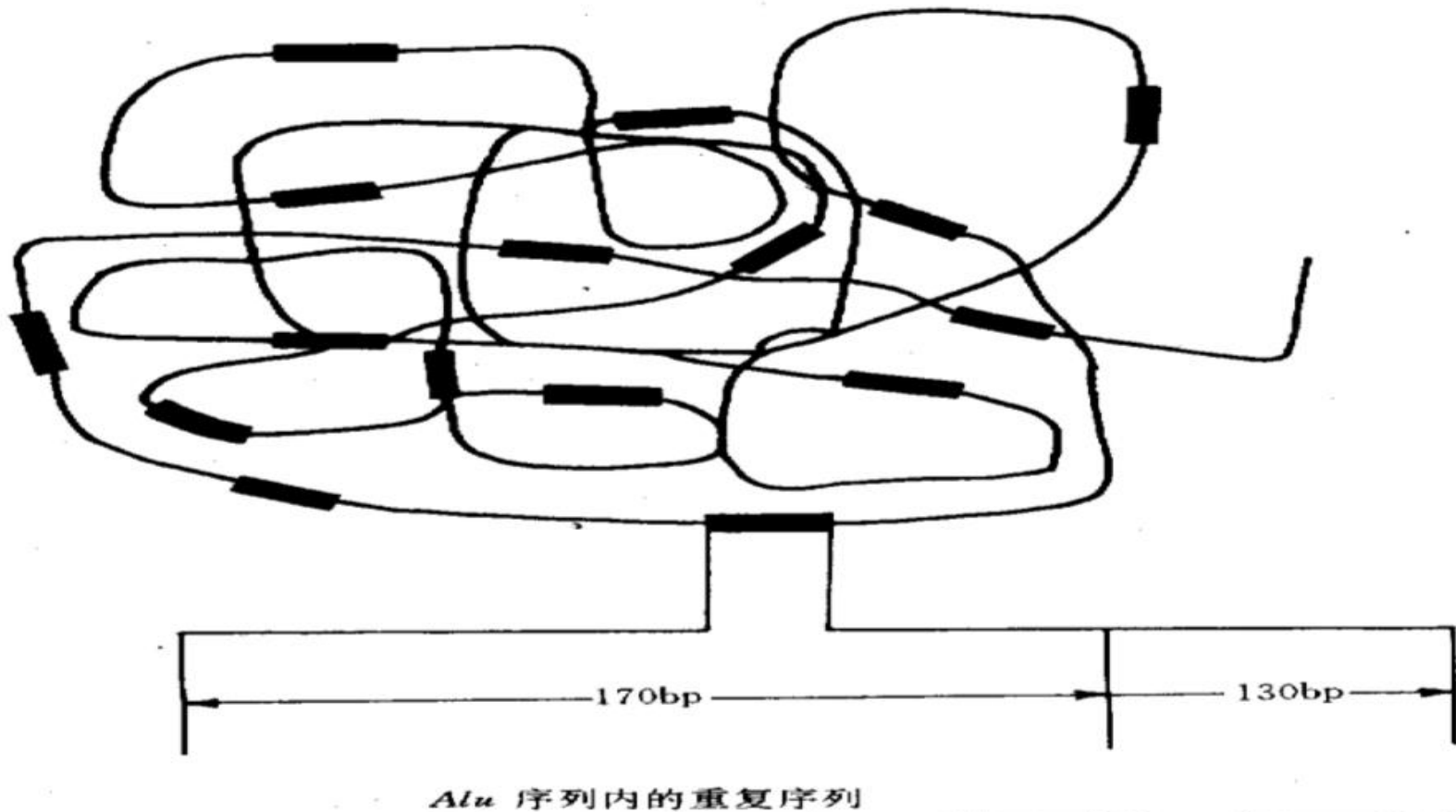
(simple tandem repeats STRs)

人群中许多 STR 存在拷贝数的多态性。

2. 中度重复序列

一般是分散的，由重复序列拷贝数很大的基因家族组成。中度重复序列的重复次数不等，多的可达 5×10^5 左右，如*Alu*家族。

rRNA、tRNA等RNA基因，一些不编码蛋白质或RNA的重复序列都属于中度重复序列。部分是可移动序列。



CTAGT	CGGGA	CGGGA		CTAGT
-------	-------	-------	--	-------

A1u基因家族

人A1u长度300bp，拷贝数30万，占总DNA的5%。拷贝单位中有A1uI识别的四核苷酸序列。

编码rRNA和tRNA基因中串连重复序列拷贝数

种	拷 贝 数		
	rRNA 前体基因	5S rRNA 基因	tRNA 基因
酵 母	140	140	250
四 膜 虫	200	300	800
果 蝇	250	165	860
非洲爪蟾	450	24 000	1 150
人	300	2 000	1 300

(二) 单拷贝基因

✂ 基因割裂现象

在基因组中仅出现一次的基因称单拷贝基因，多是编码蛋白质的基因。

基因绝大多数是不连续的、内部含有非编码序列的割裂基因。

基因组中，

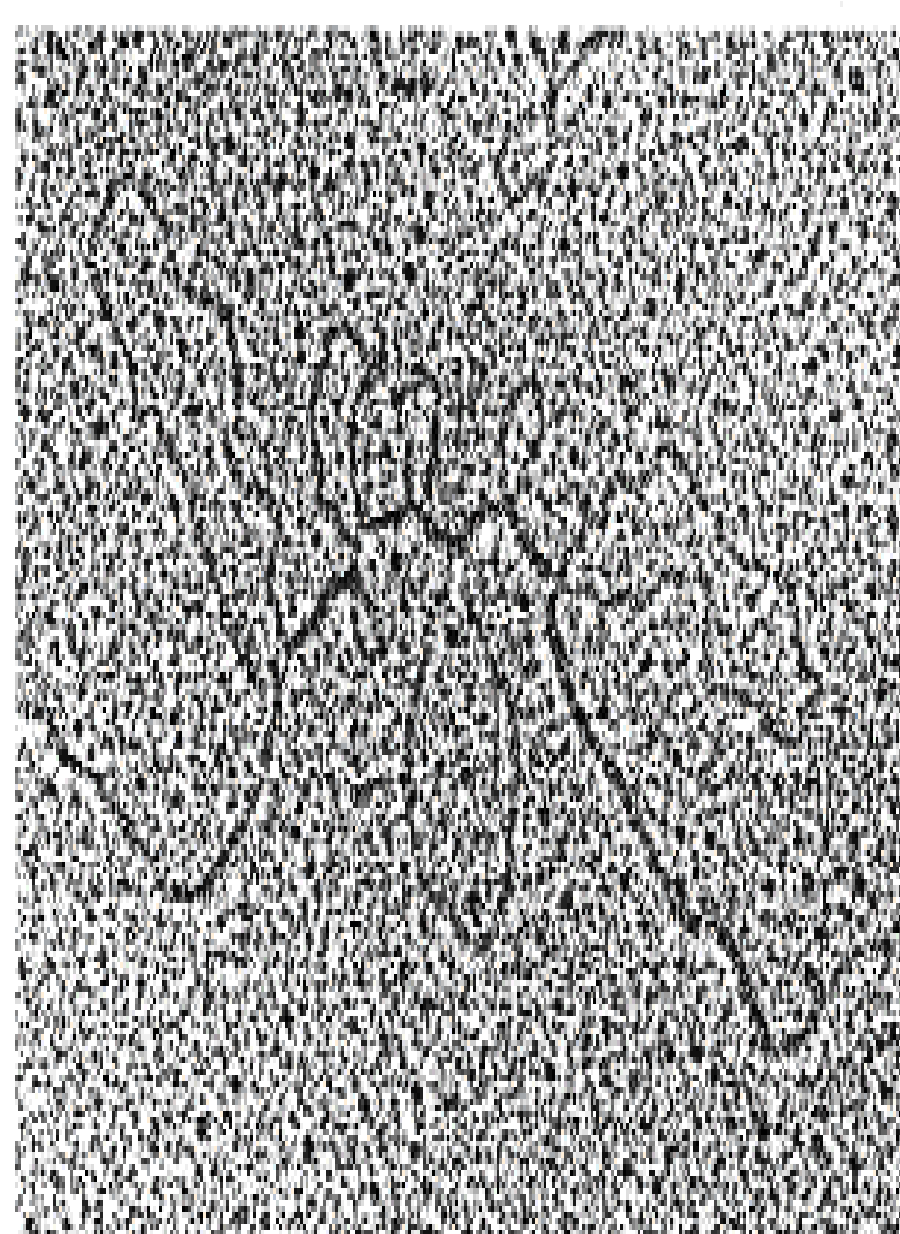
非编码部分称内含子，

编码部分称外显子。

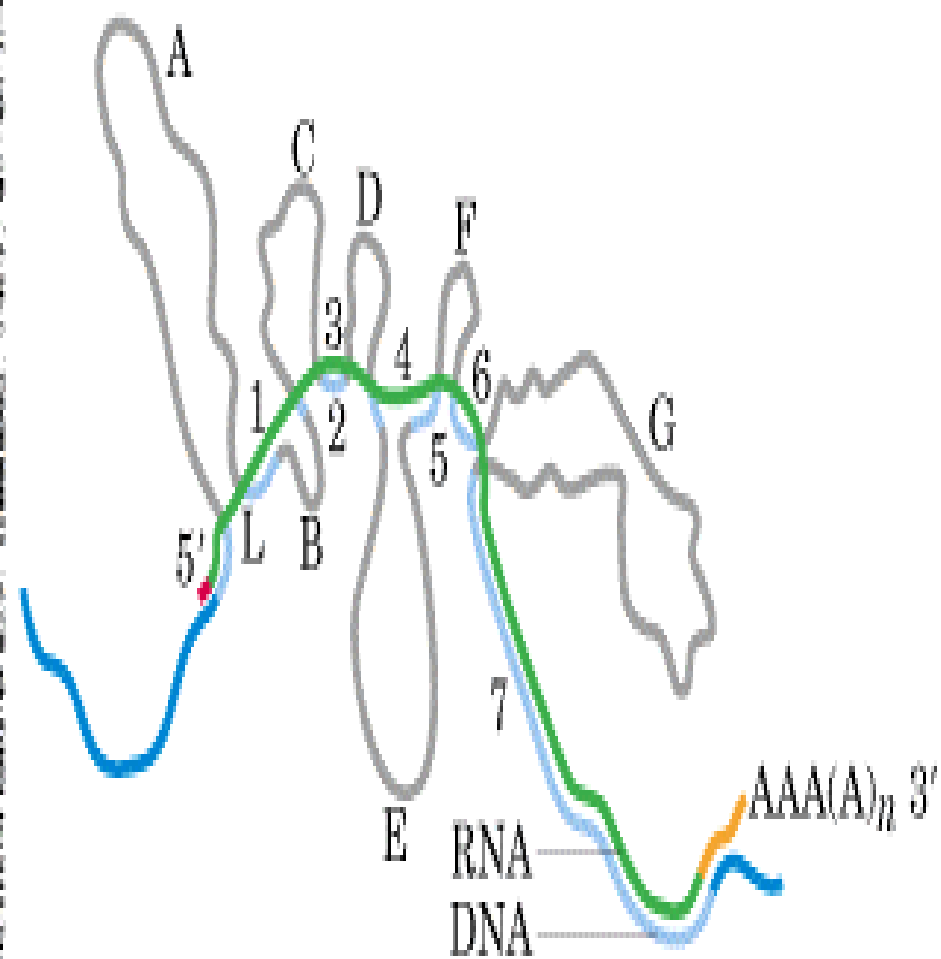
外显子编码蛋白质的功能结构域，

内含子共同特征：5'端以GT开始，

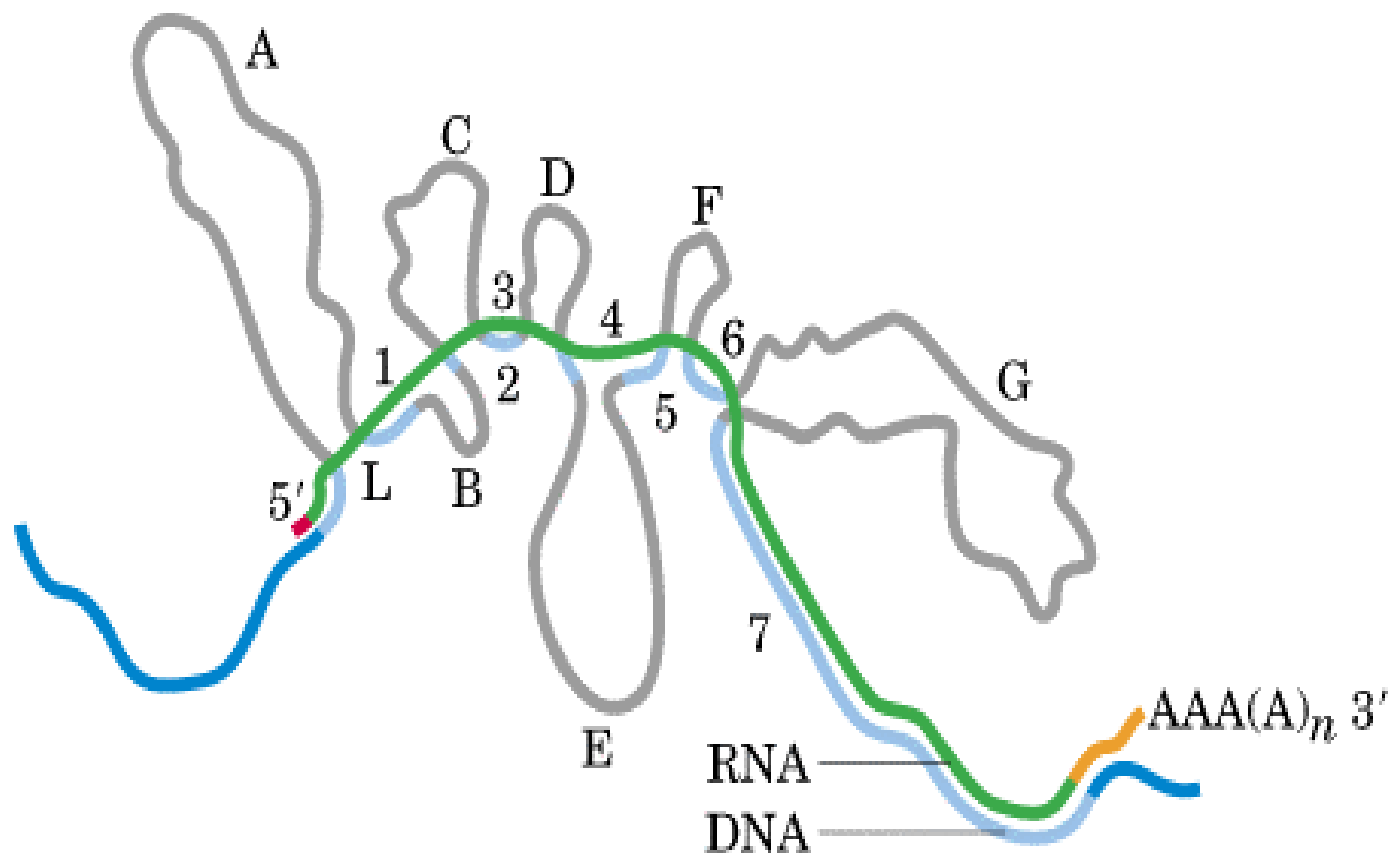
3'端以AG结束，称为GT/AG规则



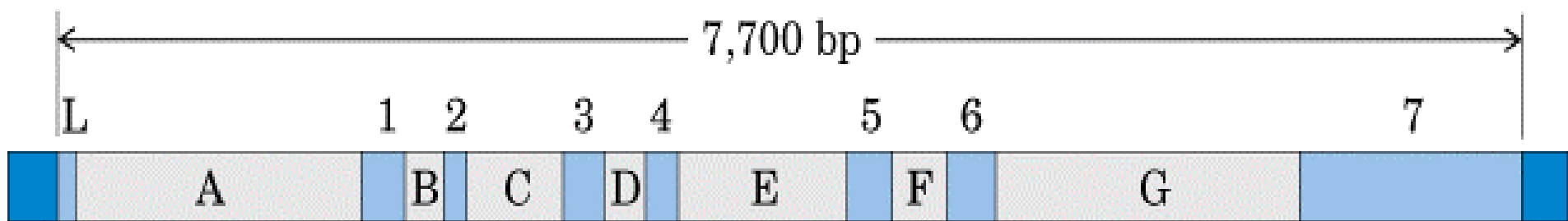
(a)



(b)



(b)



(c)

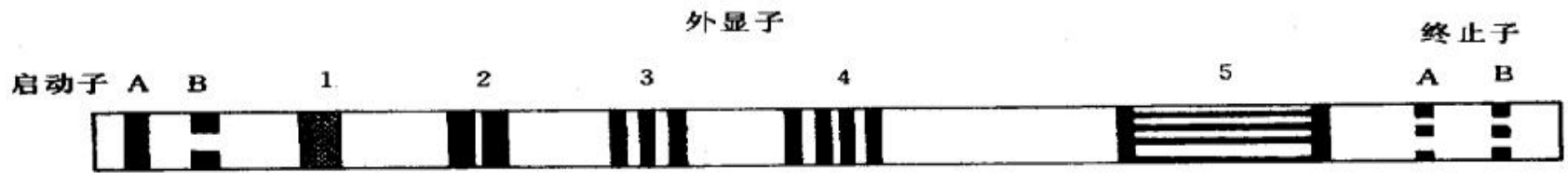
一个基因编码两种以上mRNA

📌 简单真核转录单位 (simple eukaryotic transcription units) 。

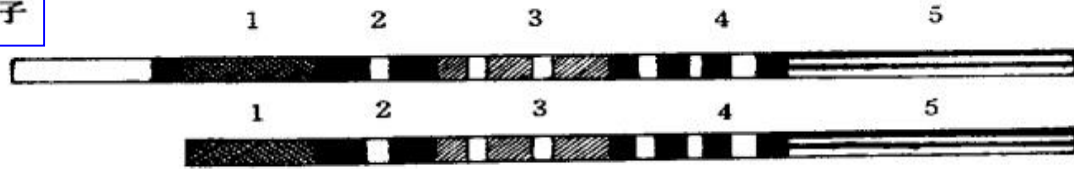
大多真核基因都单独构成一个转录单位，转录产生单顺反子mRNA，编码一种蛋白质。

📌 复杂转录单位 (complex eukaryotic transcription units) 。

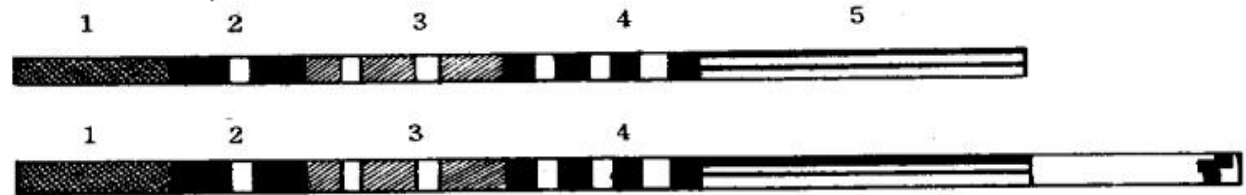
一段DNA序列可编码多种mRNA或蛋白质。



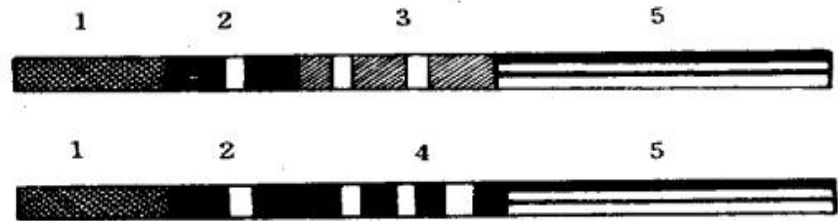
使用两种启动子



使用两种终止子

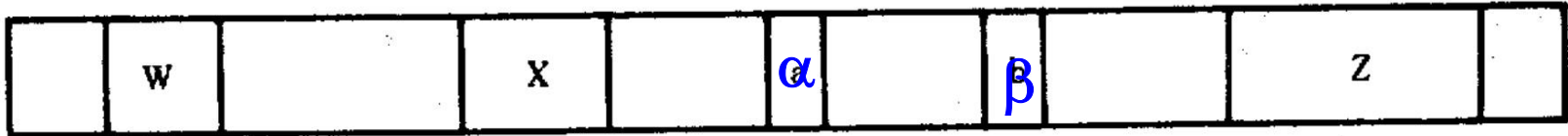
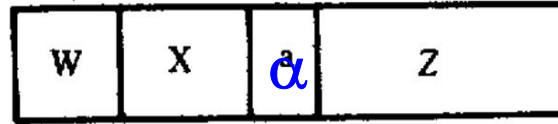


两种拼接方式

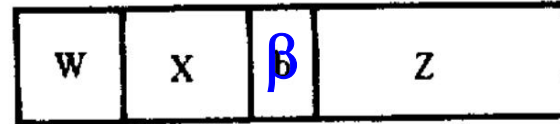


一个基因产生不同mRNA的几种方式

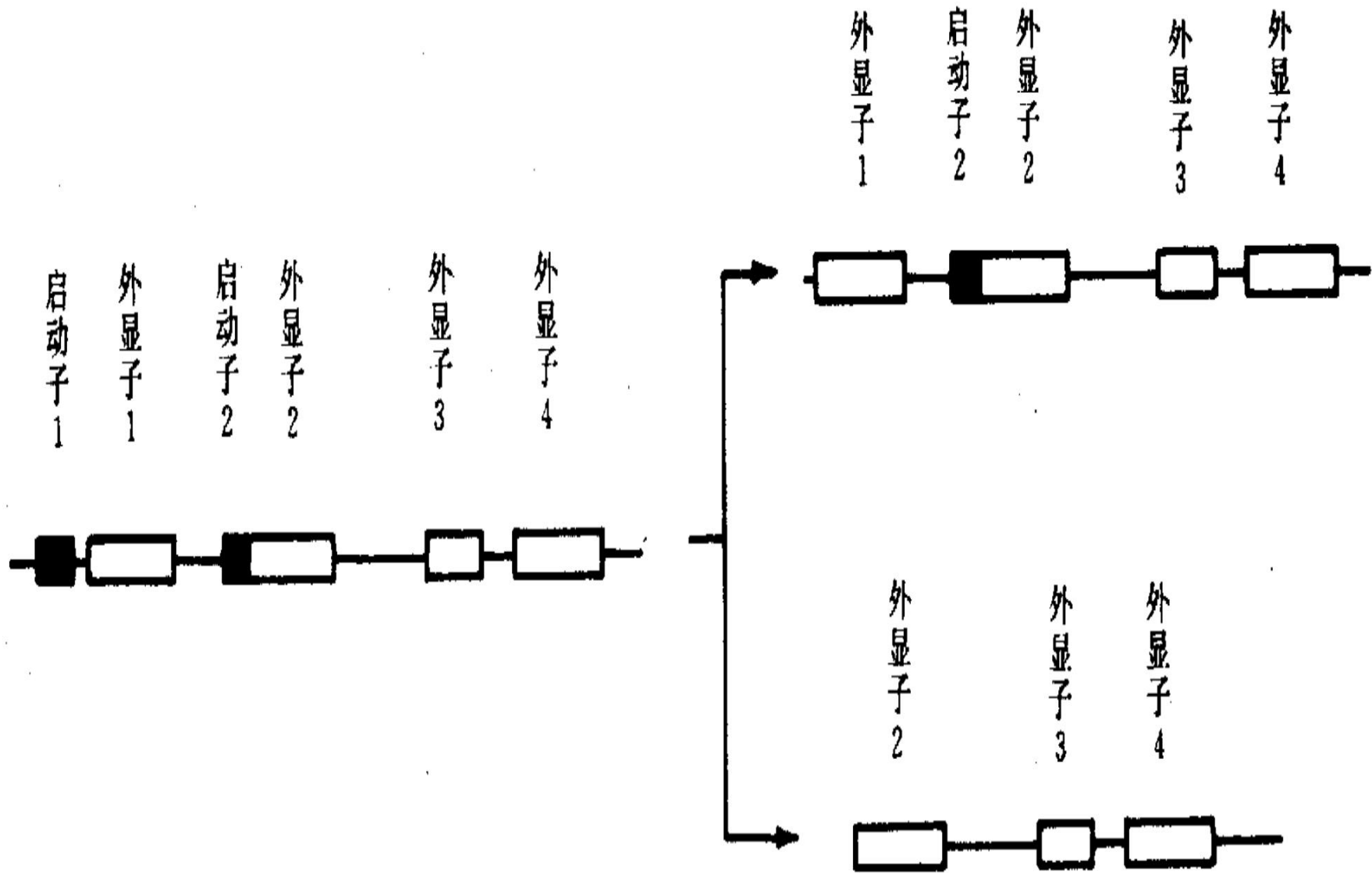
氨基酸 164 199 229 242 259



氨基酸 164 199 229 242 259



两种拼接方式产生肌钙蛋白 α 和 β



鼠淀粉酶基因使用不同启动子产生不同的mRNA

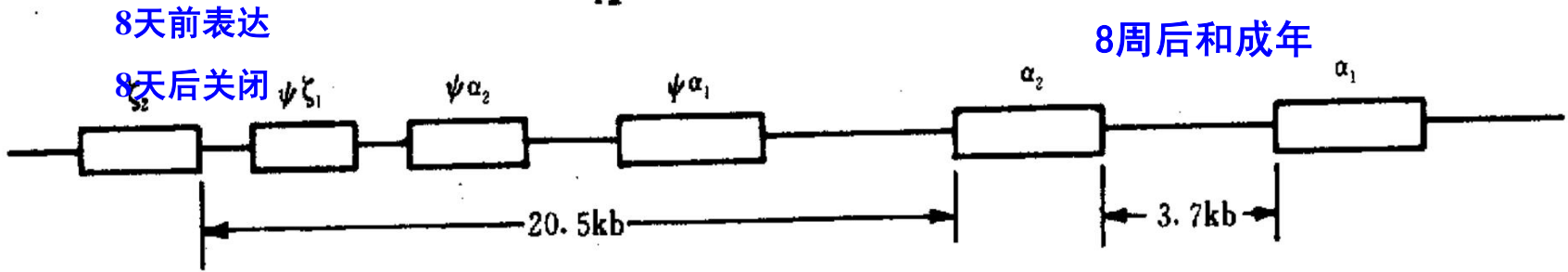
（三）基因家族

25%~50%真核编码蛋白质基因是以单个基因存在于基因组中。

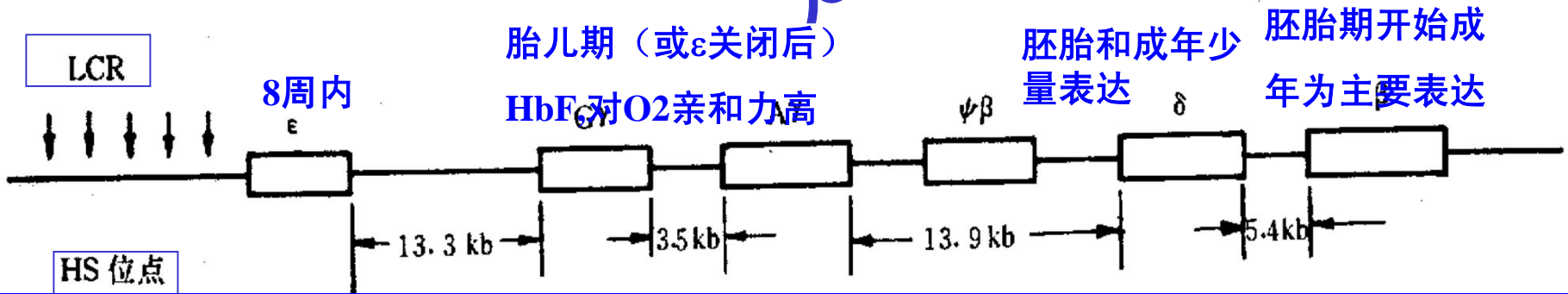
50%~75%编码蛋白基因是2个或2个以上相似基因构成的基因家族。

基因家族:来自一个祖先基因，通过“扩增”方式形成的、编码氨基酸序列相似而不相同蛋白质的一组基因。

α
A

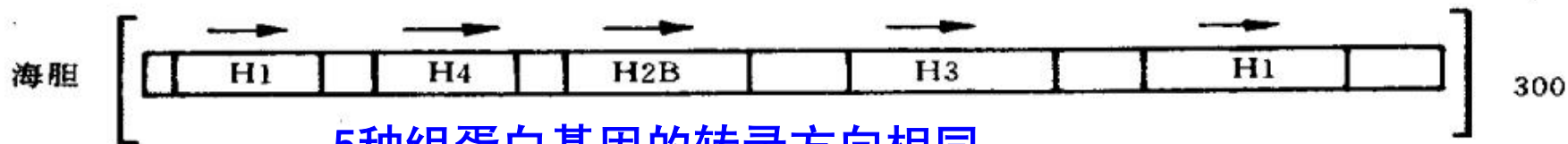
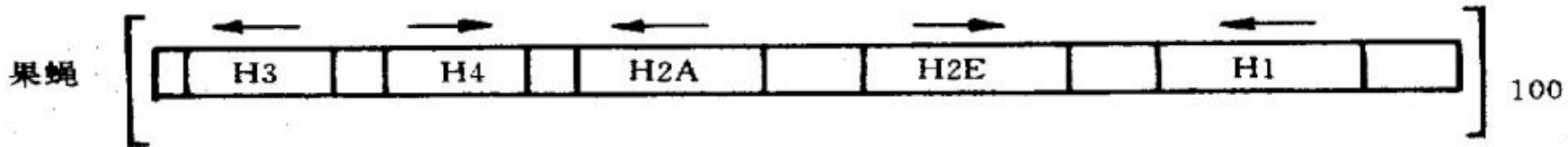


β
B

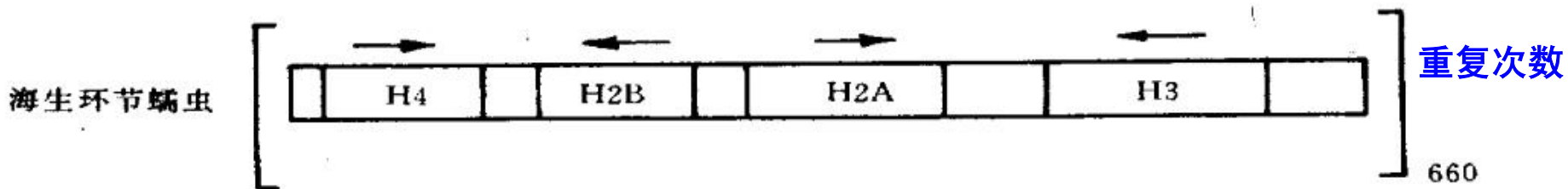
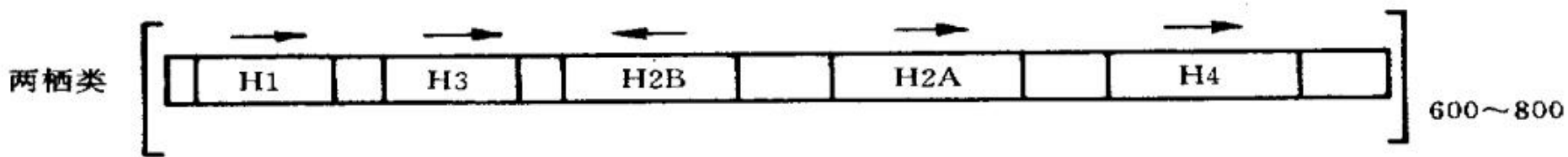


α 和 β 珠蛋白的基因家族

LCR: (locus control region)局部调控区域。一种顺式元件, 可调控一组基因的转录, 对DNA酶I高度敏感, 如 β 珠蛋白基因家族 ϵ 基因上游20kb处有5个LCR调控区。**HS:**核酸酶切高敏位点



5种组蛋白基因的转录方向相同



组蛋白基因家族

（四）假基因

进化过程中，基因‘扩增可发生片段的丢失，失去某些调控信号，无转录功能；无拼接加工信号，转录产物不能正确拼接，或在编码区产生终止信号，产生不完整的肽链，但仍然保留在基因家族中。

三、基因组计划及其意义

(一) 研究内容:

- ① 建立基因组高分辨率的遗传图谱。
- ② 完成全部染色体物理图谱及某些模型生物的DNA物理图谱。
- ③ 测定人类DNA和模型生物DNA全部序列。
- ④ 建立收集、储存、分类和分析所有有关资料和数据的工作系统。
- ⑤ 建立完成以上目标所需的新技术、新方法。

(二) 遗传图谱

遗传图谱 (genetic map) 又称**连锁图** (linkage map) : 指出遗传标记在染色体上的相对位置。

遗传标记可以是: 遗传特征、分子特征, 如限制性片段长度多态性(**RFLP**), 简单串联重复序列多态生 (**STRP**) 。

（三）物理图谱

物理图谱：是以已定位的DNA序列作界标，以DNA的实际核苷酸个数（bp）为图距的基因组图谱。在物理作图中最重要的界标是限制性内切酶位点。

（四）酵母人工染色体

酵母人工染色体（yeast artificial chromosome, YAC）的出现，使得绘制整个基因组物理图谱成为可能。YAC载体含有自主复制序列（*ARS*），着丝粒（*CEN*），端粒（*TEL*），选择标记（*Trp, URA*）和多克隆位点等。

YAC的多克隆位点可插入酶切产生500~1000kb大小的片段，制成人类基因组YAC库。然后以“步移”及其他技术反复筛选YAC库，建立人类基因组重叠群（contig）。

（五）建立新技术和新方法

脉冲电场电泳：分离大到 4×10^6 bp的DNA片段。

大分子克隆技术：YAC载体可克隆500-1000kb的DNA片段。重组的YAC载体可转化酵母细胞，在酵母细胞内复制。

新DNA测序技术：荧光标记核苷酸-DNA外切酶法
寡核苷酸探针分子杂交法，以及扫描隧道显微镜法等

第三节 基因重组和基因移动

一、基因重组

二、基因移动

一、基因重组

过去认为基因组只是在进化规模上有变化，其实基因重组和基因移动是生物界普遍现象。是生物进化的动力，所以基因重组具有重要的生物学意义。

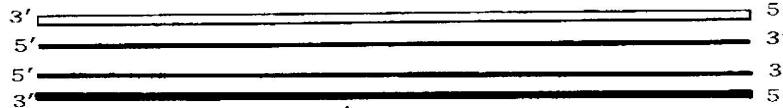
(一) 同源重组

同源重组 (homologous genetic recombination) : 具有同源序列的两个基因DNA序列之间的交换。

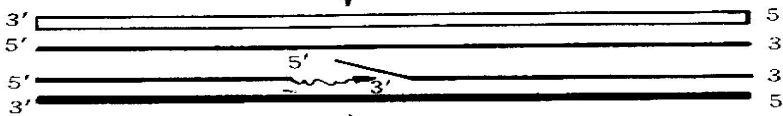
同源区序列愈长，重组发生率愈高。同源区序列太短，很难发生重组。

同源染色体 I

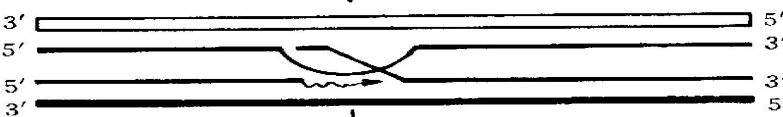
同源染色体 II



染色体 II 的一条 DNA 链断裂
在断裂链的 3' 端合成一新的片
段, 取代原来的片段

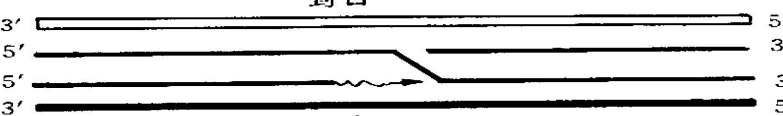


染色体 I 的一条 DNA 链与染
色体 II 中被取代的片段发生碱
基配对, 使得 I 的另一条链的
片段形成不配对的袢状结构



封口

切除袢状结构, 并连接染色体 I
与 II 中一条链的切口(封口)



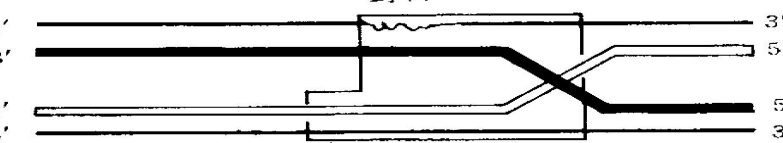
旋转

两条染色体旋转, 产生两条相
互交叉的链



封口

交叉点沿着双链移动, 致使两
条染色体上均形成一异源双
螺旋区



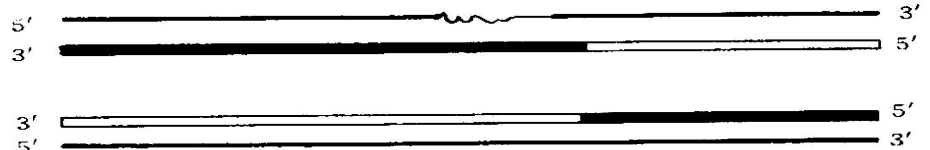
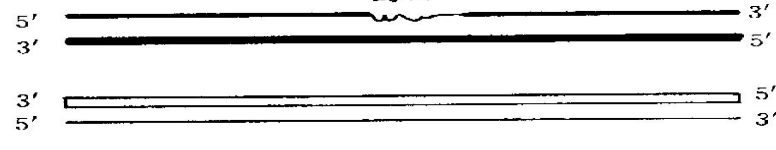
非交换

异源双链区

交换

封口

封口

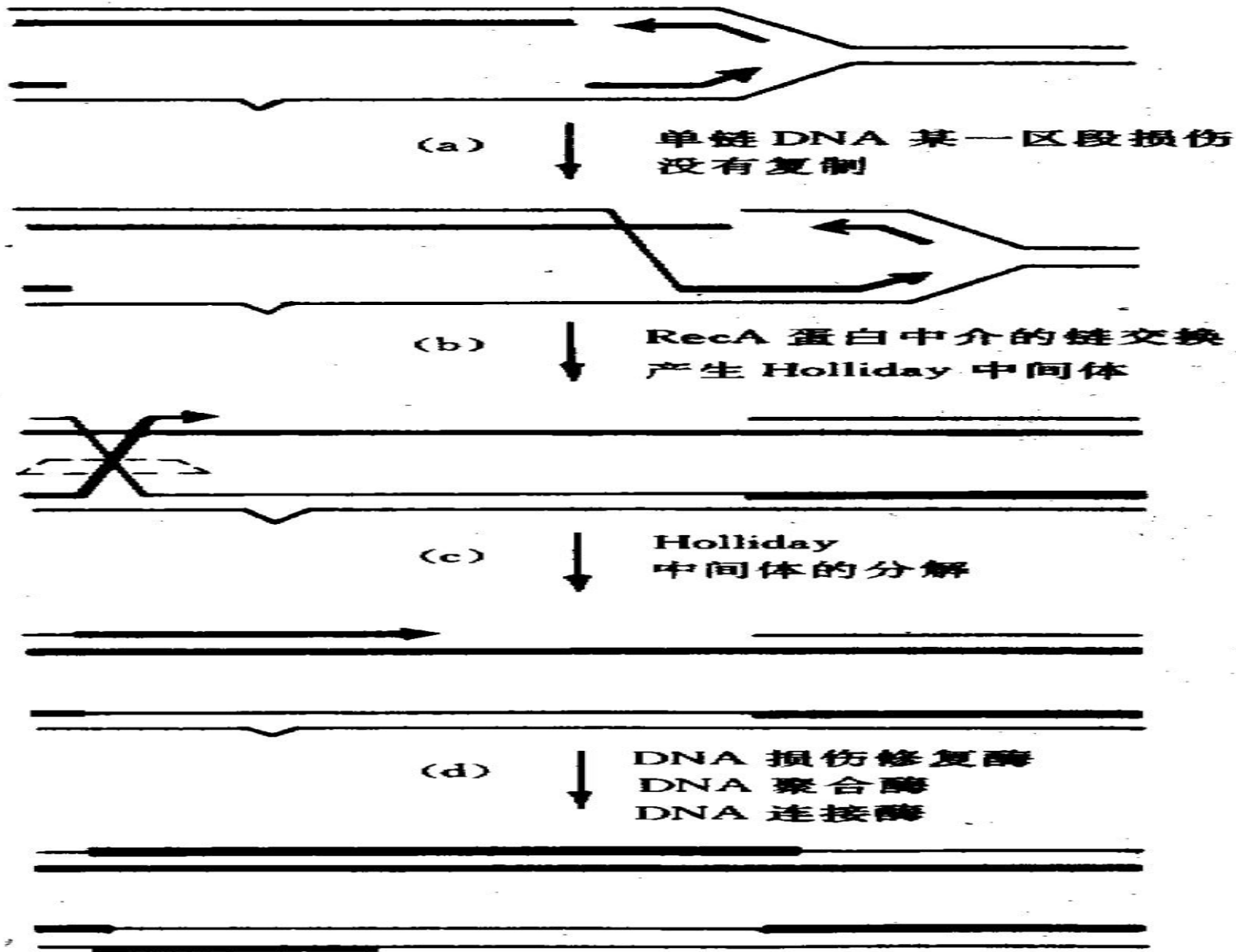


基因同源重组模型

同源重组也是DNA损伤修复的重要机制。

当一条链损伤，而另一条互补链又不能作为修复模板时，就会利用其他同源染色体，进行同源重组修复损伤链。

例如，二条链都发生断裂，或发生链交叉的损伤等。



重组修复模型

(二) 位点特异性重组

位点特性重组与同源重组不同，它发生在位点特异的短序列区，重组时发生精确的切割和连接反应，DNA既不丢失，也不合成，重组是相互的。

位点特异重组系统由**两部分**
组成：

一个是**重组酶**（**recombinase**，
或整合酶），

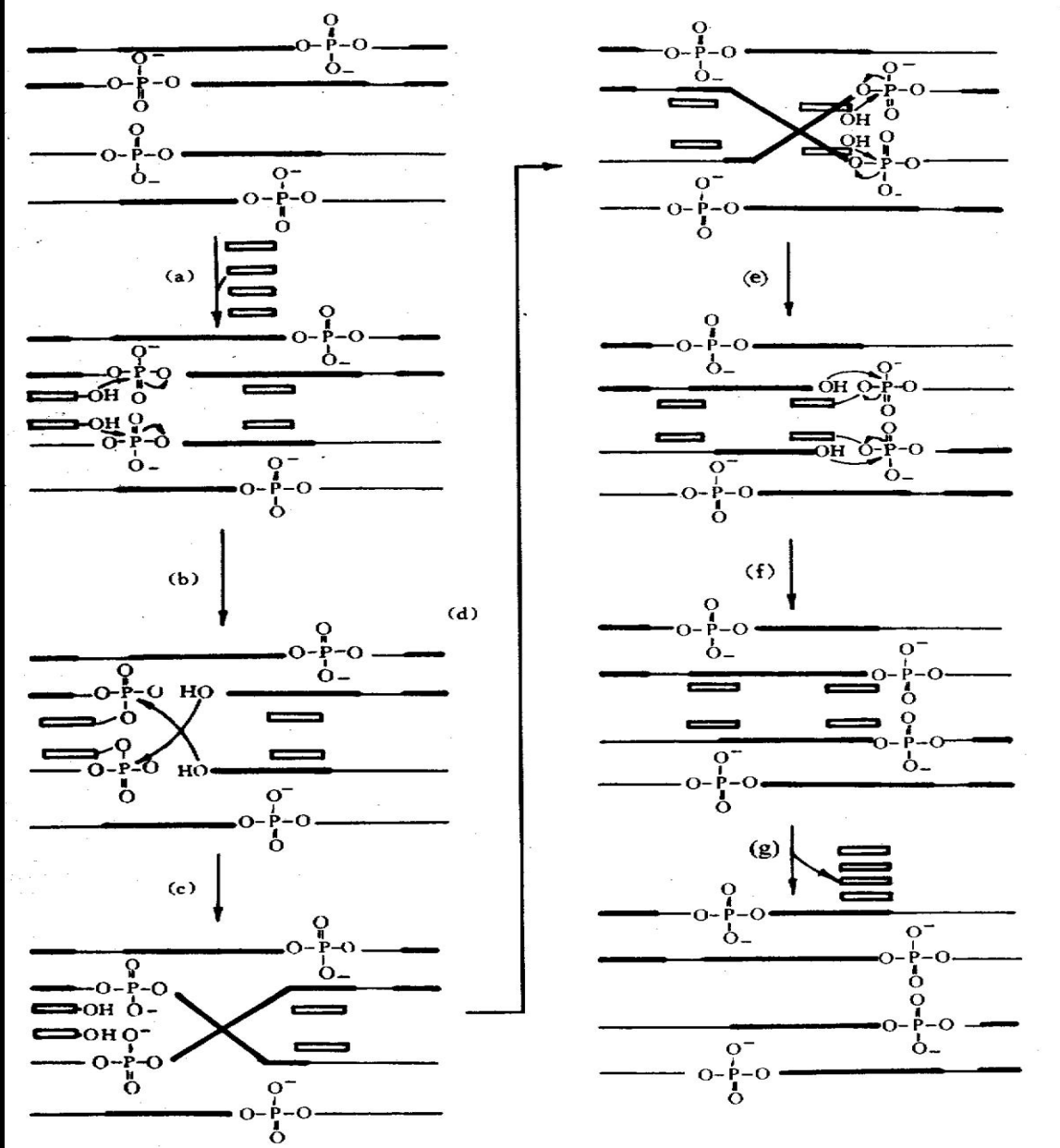
另一个是**特异的DNA序列**（**重组**
发生位点，约20~30bp）。

有些位点特异性重组系统还需要一些**辅助蛋白**。

重组酶结合
到重组位点

重组酶切开
DNA链，酶
活性中心的酪
氨酸羟基亲核
进攻磷原子，
形成磷酸酪
氨酸共价连接

重组序列的两
部分连接到新
链上形成新的
双链配对，第
一次切开/交
换形成Holliday
中间体



异构化反应
将交叉移到
第二次切开/
交换位点

后面重复
最初三步的
逆过程，重
组位点的原
有序列从新
产生

位点特异性重组反应模式

λ 噬菌体感染细菌后，有**两种复制方式**：

①是在细菌内**利用宿主的条件进行复制**，

产生新的噬菌体，破坏宿主细胞；

②是与**宿主染色体重组**，**整合到宿主基**

因组中，再一起复制，繁殖多代。

λ 噬菌体基因组整合入宿主基因组是通过位点特异重组。在 λ 噬菌体重组酶 (Int)、宿主编码的整合宿主因子 (IHF) 的参与下完成。

λ 噬菌体的重组位点 (*attP*) 和大肠杆菌重组位点 (*attB*) 的同源区有15bp的相同序列:

5' -GCTTTTTTATACTAA-3'

3' -CGAAAAAATATGATT-5'

二、基因移动

一个基因以各种方式从一处移动到另一处，称为**基因移动**，其基因称为**可移动基因**。可移动基因一般通过**转座子**形式进行移动，**几乎所有生物细胞中都存在可移动基因**。

(一) 细菌中的转座子

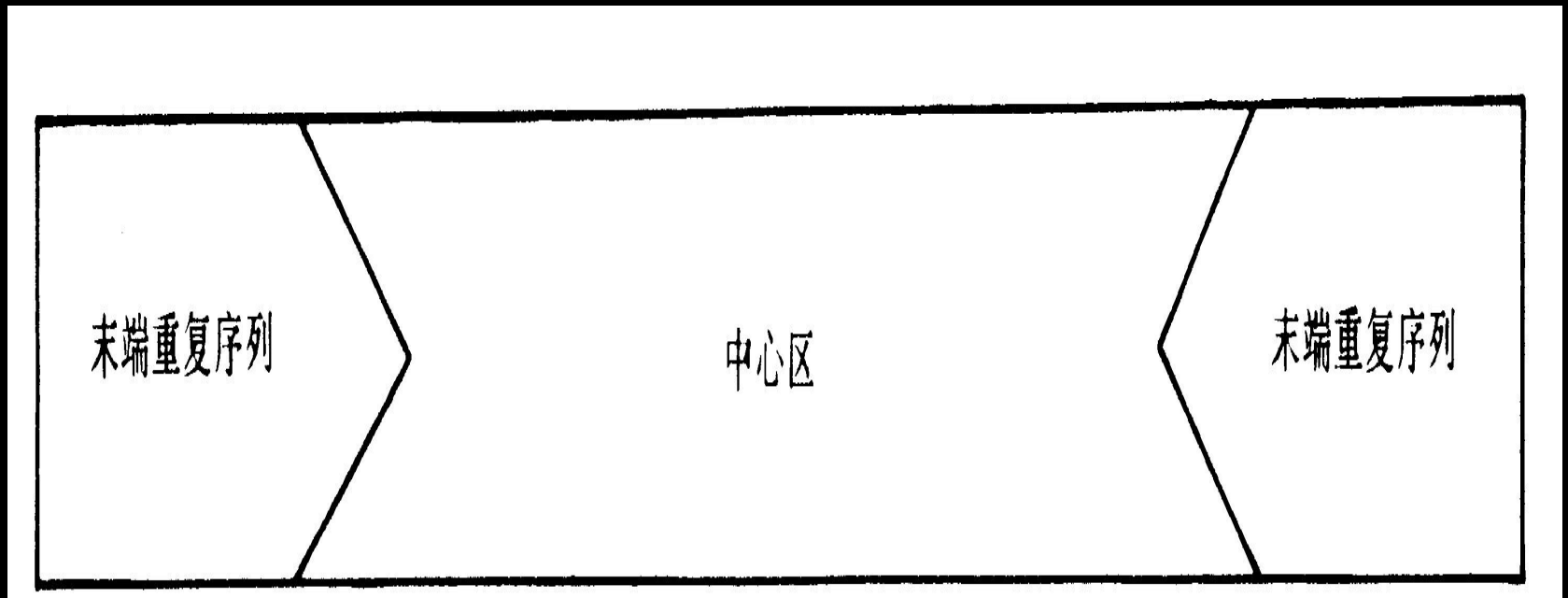
🌀 插入序列 (insertion sequence, IS)

也称简单转座子 (transposon)。

仅含有转座所需的序列和促进转座过程的蛋白质 (如转座酶) 基因。有多种类型的插入序列, 都用 IS 表示, IS 后以阿拉伯数字表示其型或分离先后

共同特征：

- 它们都是**短序列**（768~1531bp）；
- 两端是**短的反向重复序列**。



🌀转座子： 也称复杂转座子（Tn）

除了含有转座所需的序列和基因外，中心区还含有一个或更多个抗性基因。两边含有方向相同或相反的重复序列臂。

抗性基因在细菌群体中传播，就是通过转座子。转座子序列比插入序列长，通常有几千到上万个碱基对。

转座机制

复制转座

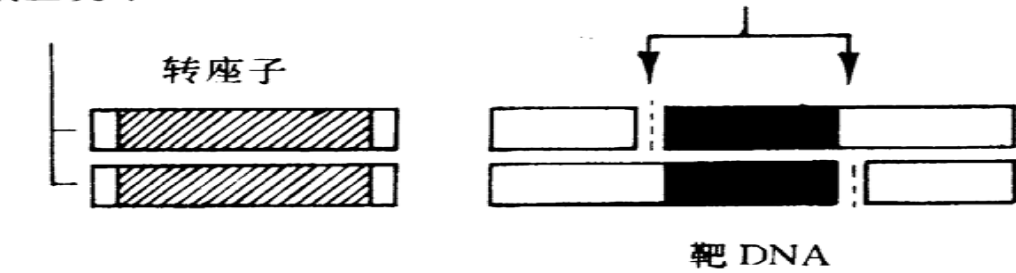
在复制型转座中，转座元件在反应中被复制，一份留在原有位置，另一份插入新位点。因此，转座是通过增加拷贝数来完成的。复制转座需两种酶：

转座酶（transposase）作用于原有转座子的末端；

分解酶（resolvase）作用于复制的拷贝。

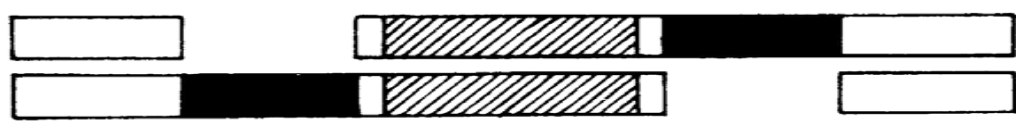
末端重复序列

转座酶切开靶位点



在靶DNA插入位点产生交错切口

转座子插入切口位点



转座子同靶位点单链末端连接

通过复制填满空隙转座子两侧序列被复制



补平在靶位上的缺口（一般仅5~10bp），并连接缺口

复制转座机制

(2) 非复制转座机制

过程与复制转座机制相似，不同的是：

转座元件不复制，直接转移至另一处。

靶位上增加转座元件，而原有位置上就丢失一元件，留下一段缺口。

这一缺口可能被修复机制修复，大肠杆菌能耐受这种缺失。

非复制转座仅需要一种酶，即转座酶。

（二）真核细胞转座子

部分真核转座子与细菌类似，但已发现一些与细菌不同的转座机制。

真核转座机制涉及一个RNA的中间体，这种转座子在结构上类似逆转录病毒，故称为逆转录转座子。

转座子：含一个编码类似逆转录酶的编码区，编码区两侧是长末端重（LTR）。

通过**RNA中间体**转座子从基因组一处转座到另一处，再通过逆转录酶合成一个RNA的**DNA拷贝**，随后在**新的靶位点整合入基因组**。

大多数真核转座子都使用这种机制进行转座。与细菌转座的**差别**：**通过RNA中间体，而不是DNA**。

逆转录转座子**缺乏外壳蛋白基因**，**不能形成病毒颗粒**，所以**不会在细胞之间转移**。