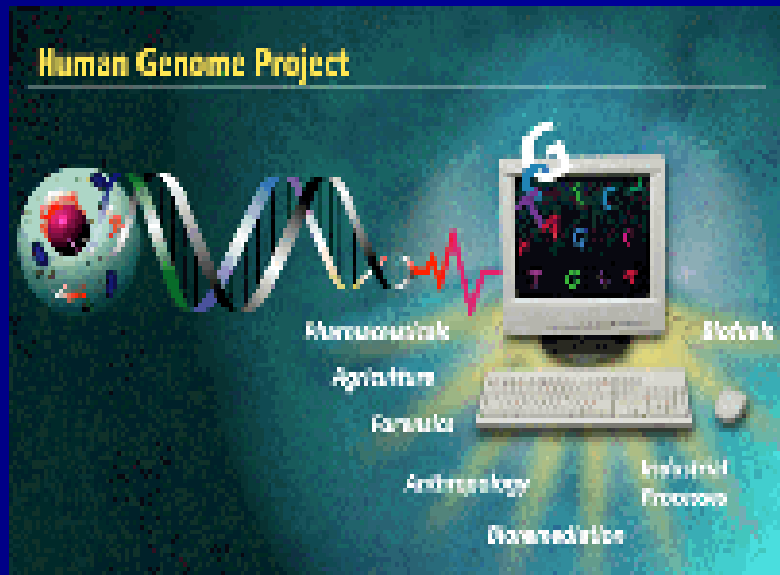
# 第一讲 基因组测序与序列组装

任科教师: 余爱丽

生命科学院 分子生物学与生物信息学系

主要内容:

什么是基因组
什么是基因
DNA测序的方法
DNA序列的组装
人类基因组方案
水稻基因组方案
后基因组学



# 1. 什么是基因组

基因组就是一个物种中所有基因的整体组成。

基因组有两层意义：遗传物质和遗传信息。

要揭开生命的奥秘，就需要从整体水平研究基因的存在、基因的结构与功能、基因之间的相互关系。

# Genome Size (Mb)

| | |
|---|---|
| Zea mays | 8,000 |
| Homo sapiens | 3,000 |
| Oryza sativa | 400 |
| Drosophila melanogaster | 165 |
| Arabidopsis thaliana | 100 |
| Saccharomyces cerevisiae | 12 |
| E.coli | 4.6 |

# 什么是C 值?

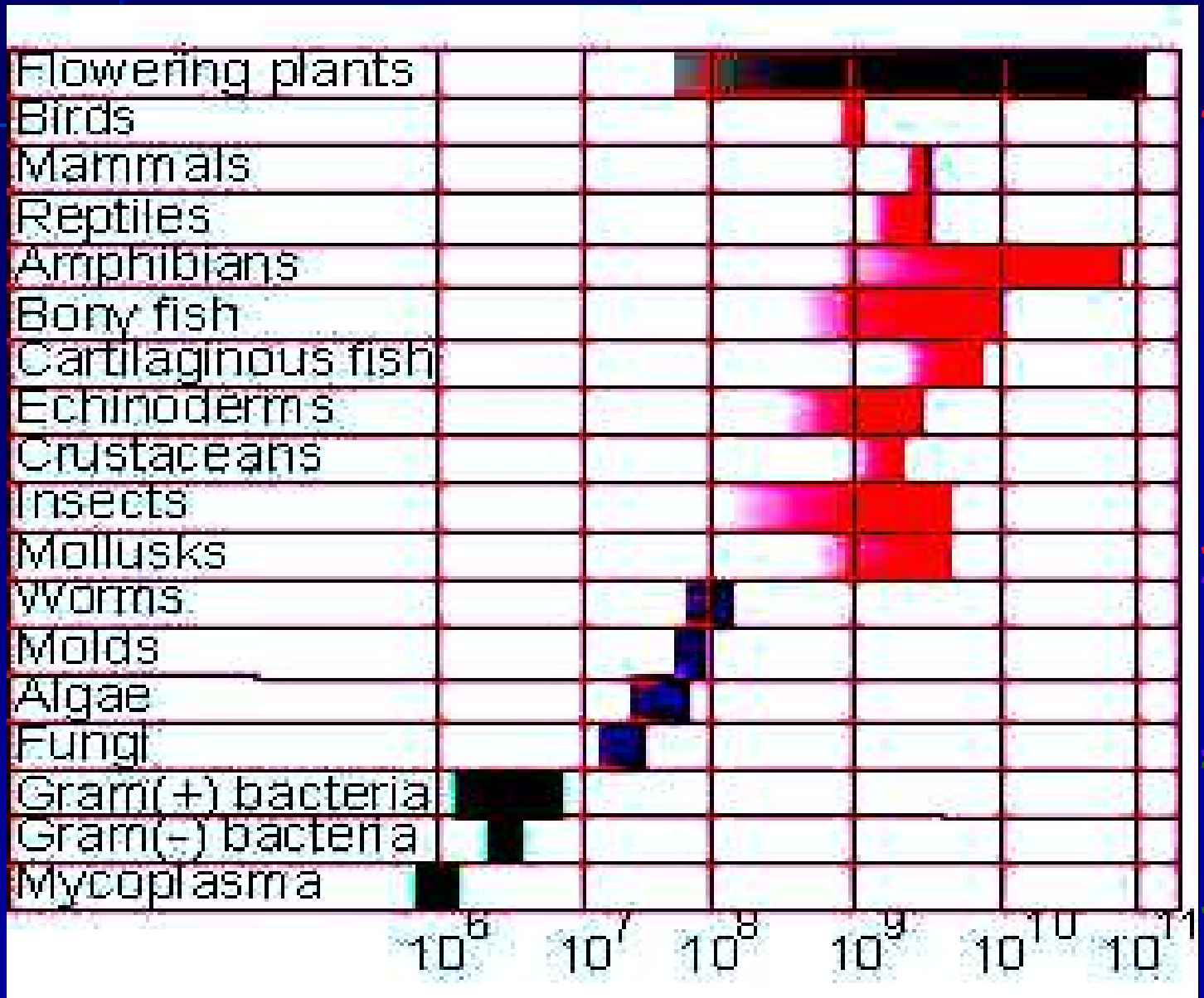
- 通常是指一种生物单倍体基因组DNA的总量.

在真核生物中，C值一般随着生物的进化而增加，高等生物C值一般大于低等生物。

**C值悖理：**

**生物的复杂性与基因组的大小并不完全成比例增加**

# 阴影局部为一个门内C-值的范围



# 重复顺序

- 高度重复顺序：
  - 长度：几个——几千个bp
  - 拷贝数：几百个——上百万个
  - 首尾相连，串联排列
  - 集中分布于染色体的特定区段〔如端粒，着丝粒等〕
  - 也称卫星DNA

- 中度重复顺序：
  - 一般分散于整个基因组中；　长度和拷贝数差异很大

- 单一顺序：
  - 基因主要位于单一顺序
  - 动物中单一顺序约占50％
  - 植物中单一顺序约占20％

# 顺序复杂性

- DNA 的复性 遵循二级反响动力学，可表述为：

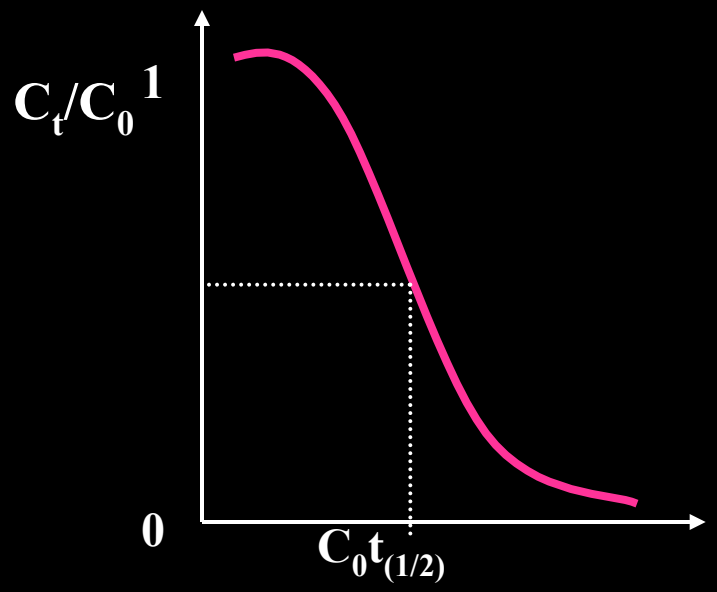
$$dC_t / d_t = -KC_0^2$$

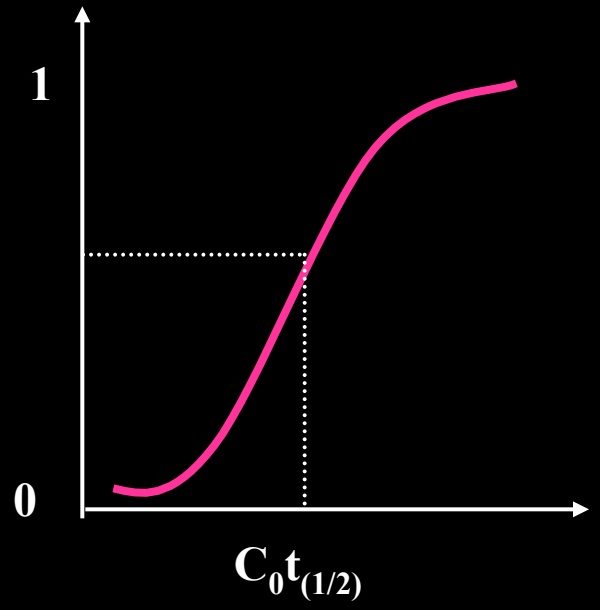
反响达 t 时，单链DNA浓度 = Ct

$C_0$ = 单链 DNA起始浓度

K＝ 复性速度常数

$C_o t_{(1/2)}$ = 1/K (mol. Sec / L)　常数




$C_0 t_{(1/2)}$值与基因组复杂性成正比。

# 2. 什么是基因？

是遗传信息的物理和功能单位，包含产生一条多肽链或功能RNA所必需的全部核苷酸序列。

基因分类：

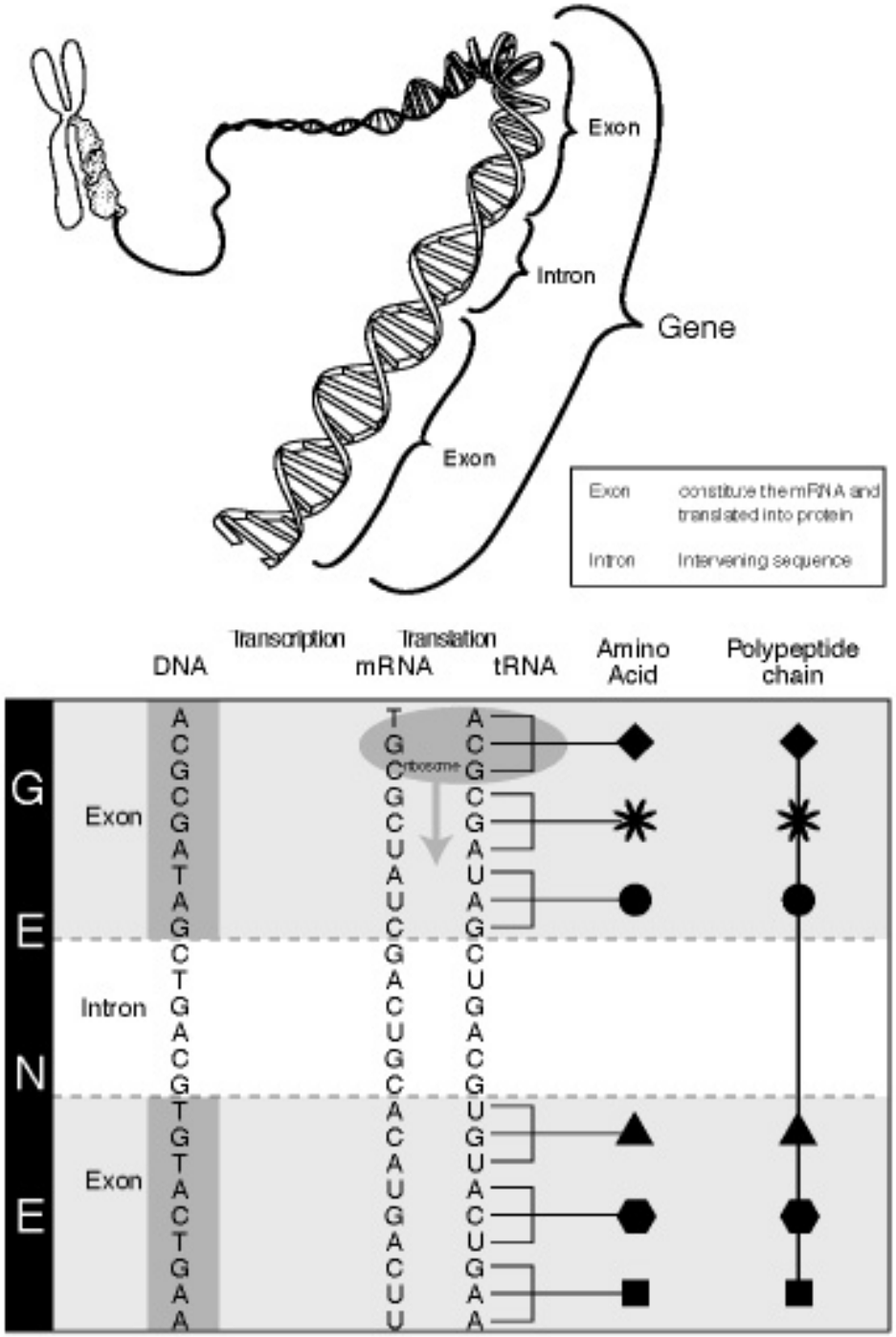
编码RNA的基因，如rRNA基因，snRNA基因等；

编码蛋白质的基因

# 基因的不连续性

**Intron 和Exon:**

**大多数真核生物蛋白质基因的编码顺序(Exon)都被或长或短的非编码顺序(Intron)隔开**



# 基因家族

一群具有一致的或相似顺序的基因,有的还担负类似的生物学功能, 可以相互补偿, 比方:E2f transcription factor

| Mouse symbol | Human Ortholog |
|---|---|
| E2f1 | E2F1 |
| E2f2 | E2F2 |
| E2f3 | E2F3 |
| E2f4 | E2F4 |
| E2f5 | E2F5 |
| E2f6 | E2F6 |

# 假基因(Pseudogene)

来源于功能基因 但已失去活性 的DNA序列

产生假基因的原因有:

1. 由重复产生的假基因;
2. 加工的假基因, 由RNA反转录为cDNA 后再整合到基因组中;
3. 残缺的基因(Truncated gene)

## 重叠基因:

同一段DNA 能携带两种不同蛋白的信息.

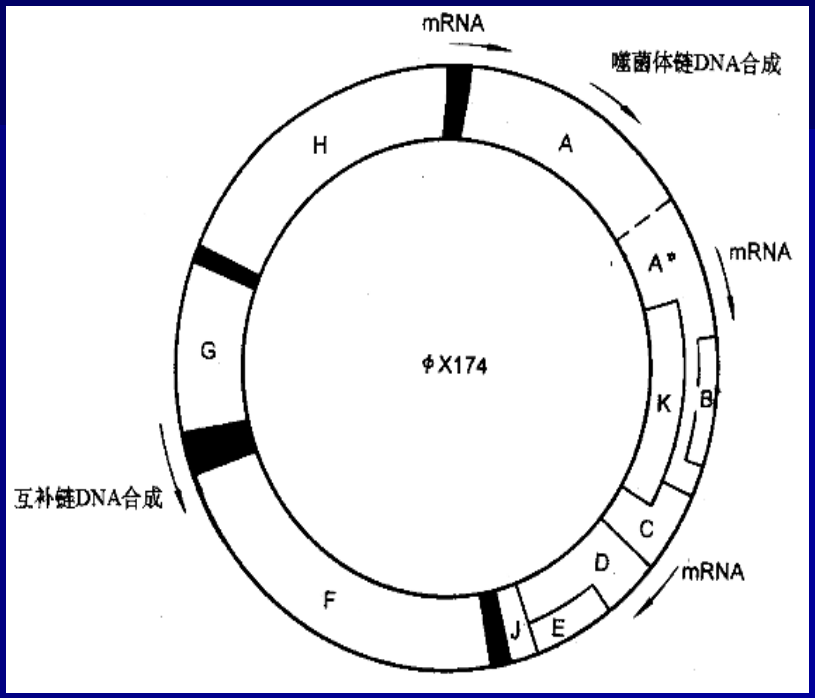
重迭基因有以下几种情况:

*一个基因完全在另一个基因内部
*局部重叠
* 两个基因共用少数碱基对

*一个基因完全在另一个基因内部

如：B和A， E和D

其读码结构互不相同




---ATG-----//------AATGCC ----//---ATAACG---//--

A* TAA----

ATGCCN----NNATAA

B

*局部重叠

如：K和C

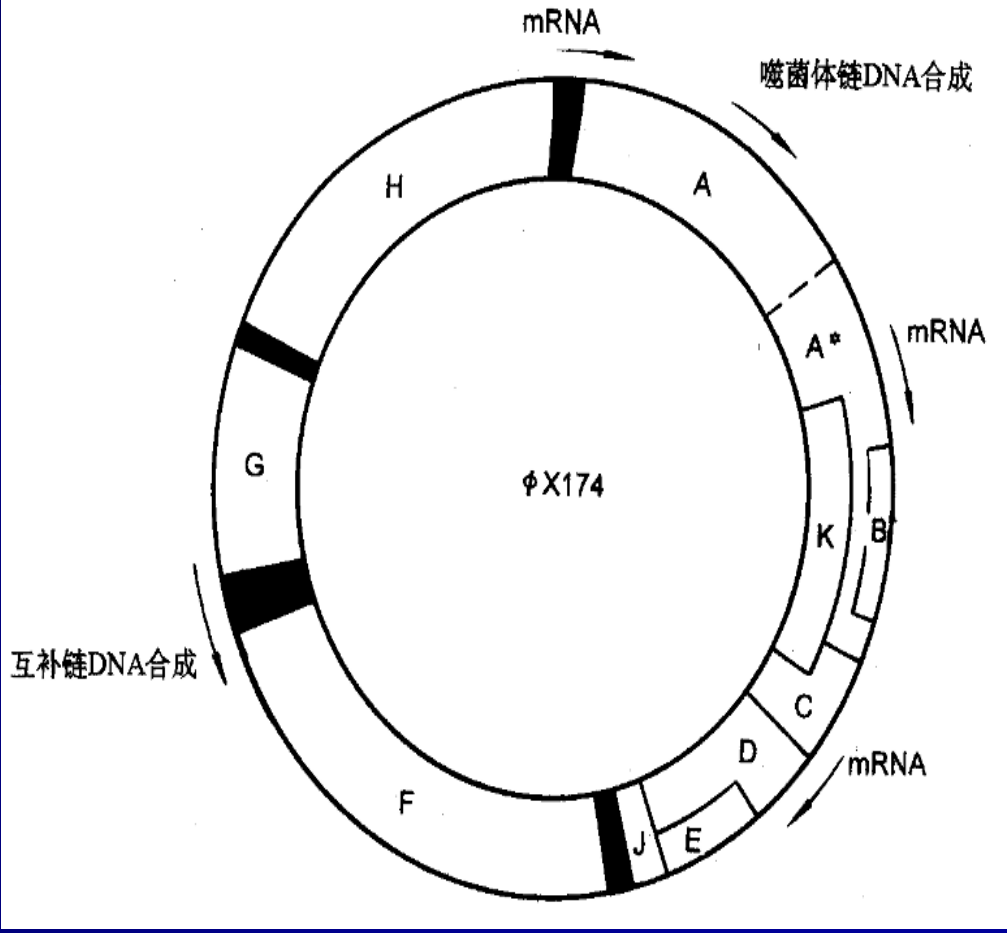
*两个基因共用少数碱基对

如：D和J

------TAATG-------

D 终止密码子

J 起始密码子



# 3. DNA测序的方法

- 链终止法测序
- 化学降解法测序
- 自动化测序
- 非常规DNA测序

# 3.1 链终止法测序(the chain termination method)

根本原理:

通过合成与单链DNA互补的多核苷酸链，由于合成的互补链可在不同位置随机终止反响，产生只差一个核苷酸的DNA分子，从而来读取待测DNA分子的顺序。

# 技术路线与要求

制备单链模板

↓

将单链模板与一小段引物退火

↓

参加DNA多聚酶

4种脱氧核苷酸

↓

分别参加少量4种双脱氧核苷酸

↓

将4种反响产物分别在4条泳道电泳

↓

根据4个碱基在4条泳道的终止位置读出基因序列

A 克隆于质粒中DNA→用碱或热变性

B M13克隆单链DNA

C 噬粒克隆DNA

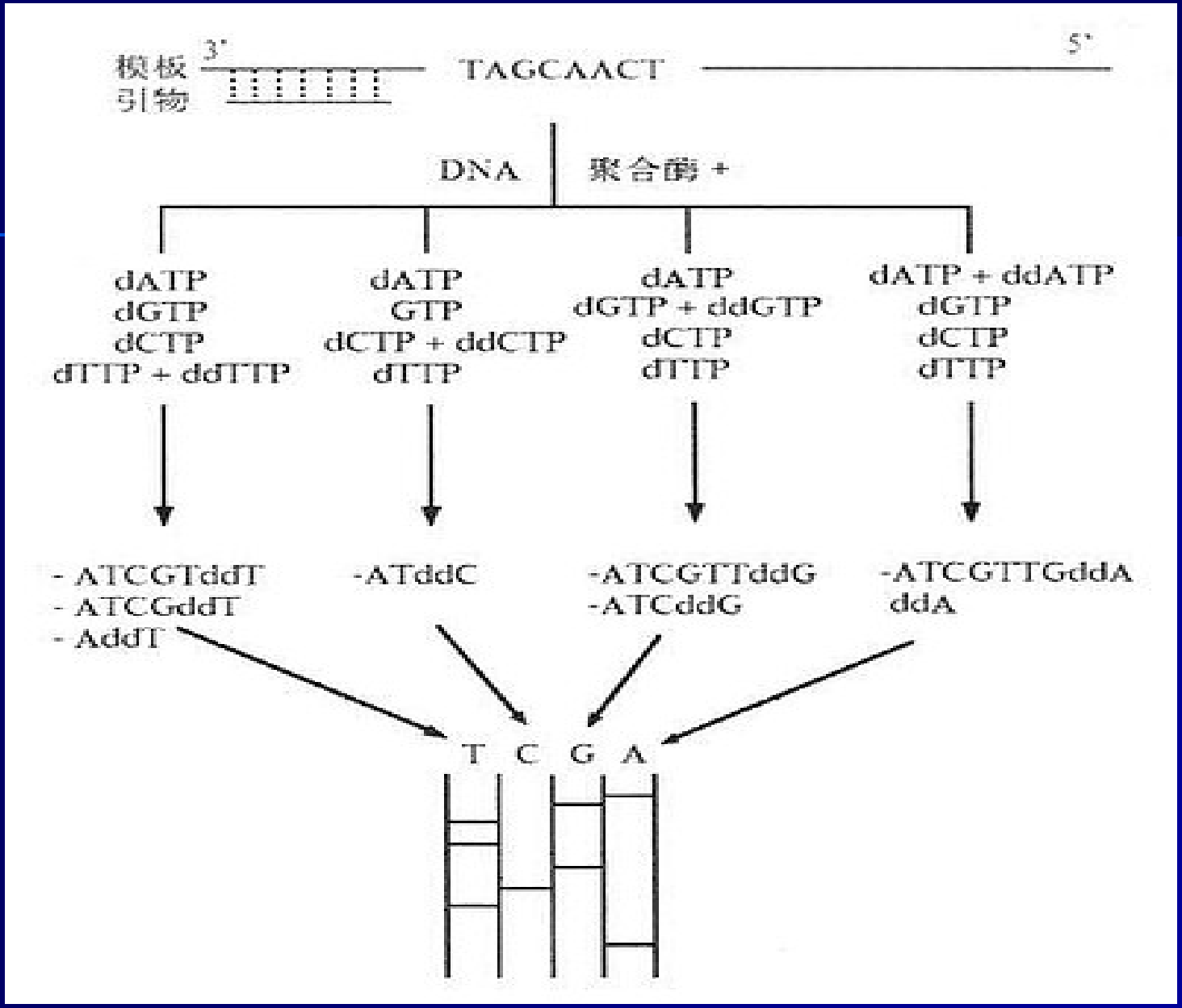
D PCR产生单链DNA

A 高酶活性

B 无5′→3′外切酶活性

C 无3′→5′外切酶活性

ddATP/ddCTP/ddGTP/ddTTP 的3′碳原子连接的是氢原子,不是羟基


模板
引物
3'
5'
TAGCAACT
DNA
聚合酶 +
dATP
dGTP
dCTP
dTTP + ddTTP
dATP
GTP
dCTP + ddCTP
dTTP
dATP
dGTP + ddGTP
dCTP
dTTP
dATP + ddATP
dGTP
dCTP
dTTP
- ATCGTddT
- ATCGddT
- AddT
-ATddC
-ATCGTTddG
-ATCddG
-ATCGTTGddA
ddA
T C G A

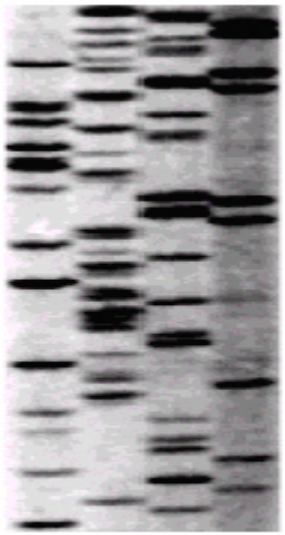


图 1.核苷酸序列原始图象

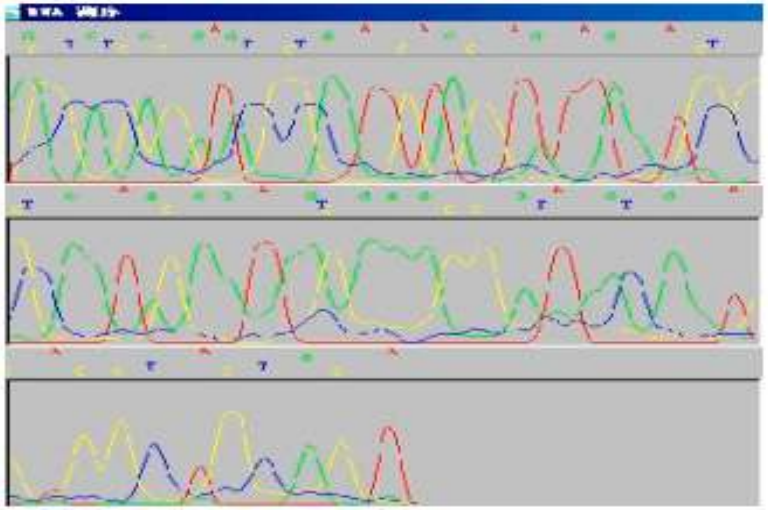


图 2. 序列扫描图

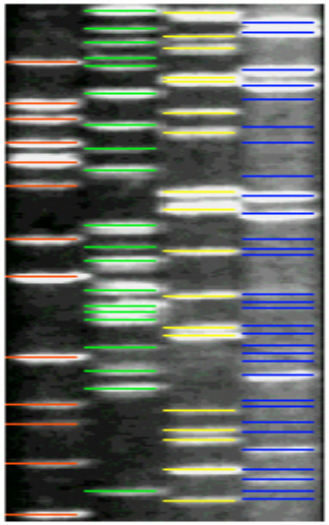


图 3. 条带密度中心


图 4. DNA 链中四种核苷酸序列

# 3.2 化学降解法测序

根本原理:

在选定的核苷酸碱基中引入化学集团,再用化合物处理,使DNA分子在被修饰的位置降解.

# 技术路线

将双链DNA样品变为单链

↓

每个单链的同一方向末端都用放射性同位素标记,以便显示DNA条带

↓

分别用不同方法处理,获得只差一个核苷酸的降解DNA群体

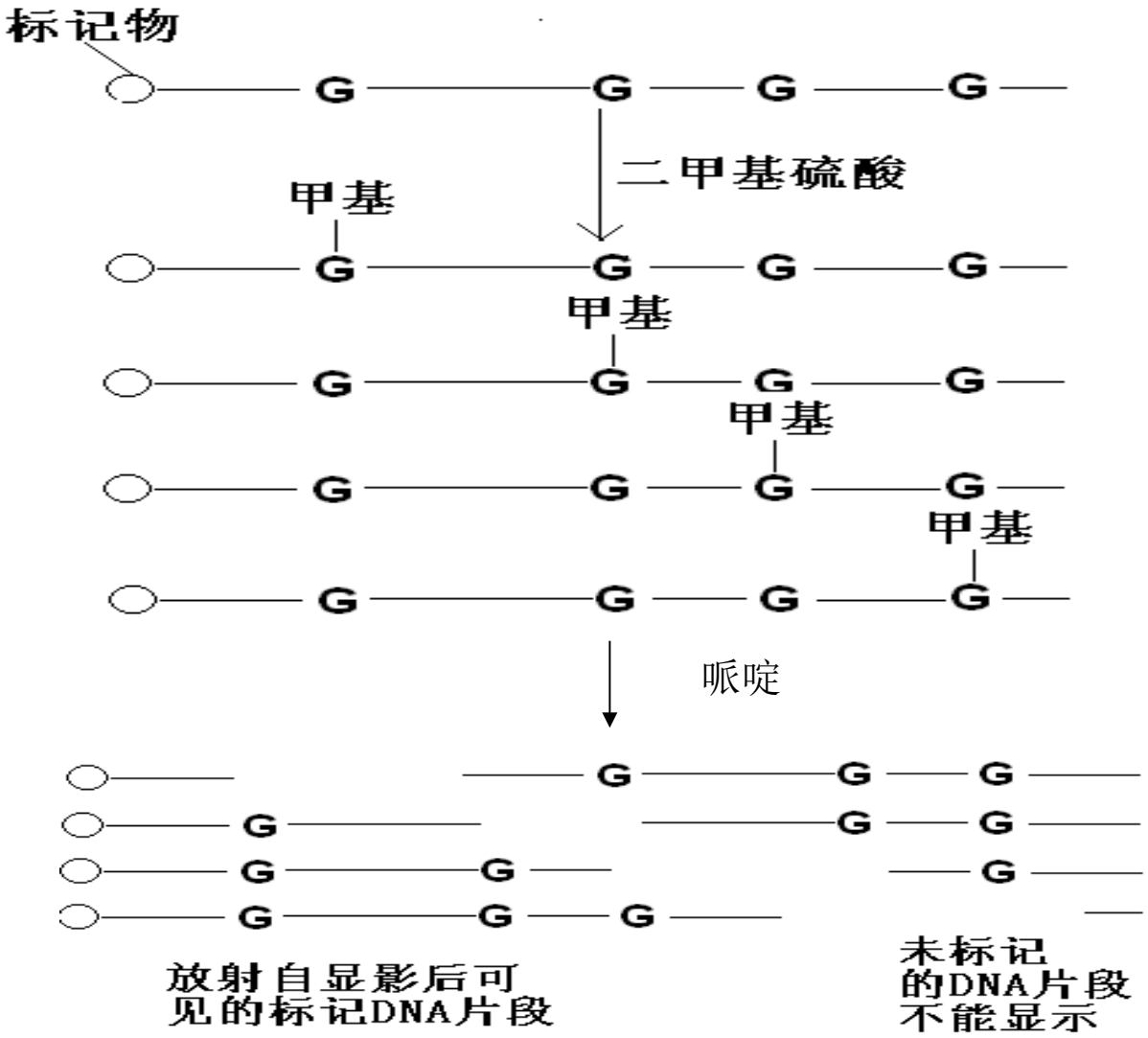
↓

电泳,读取DNA的核苷酸顺序

# Maxam-Gilbert 法所用的化学技术

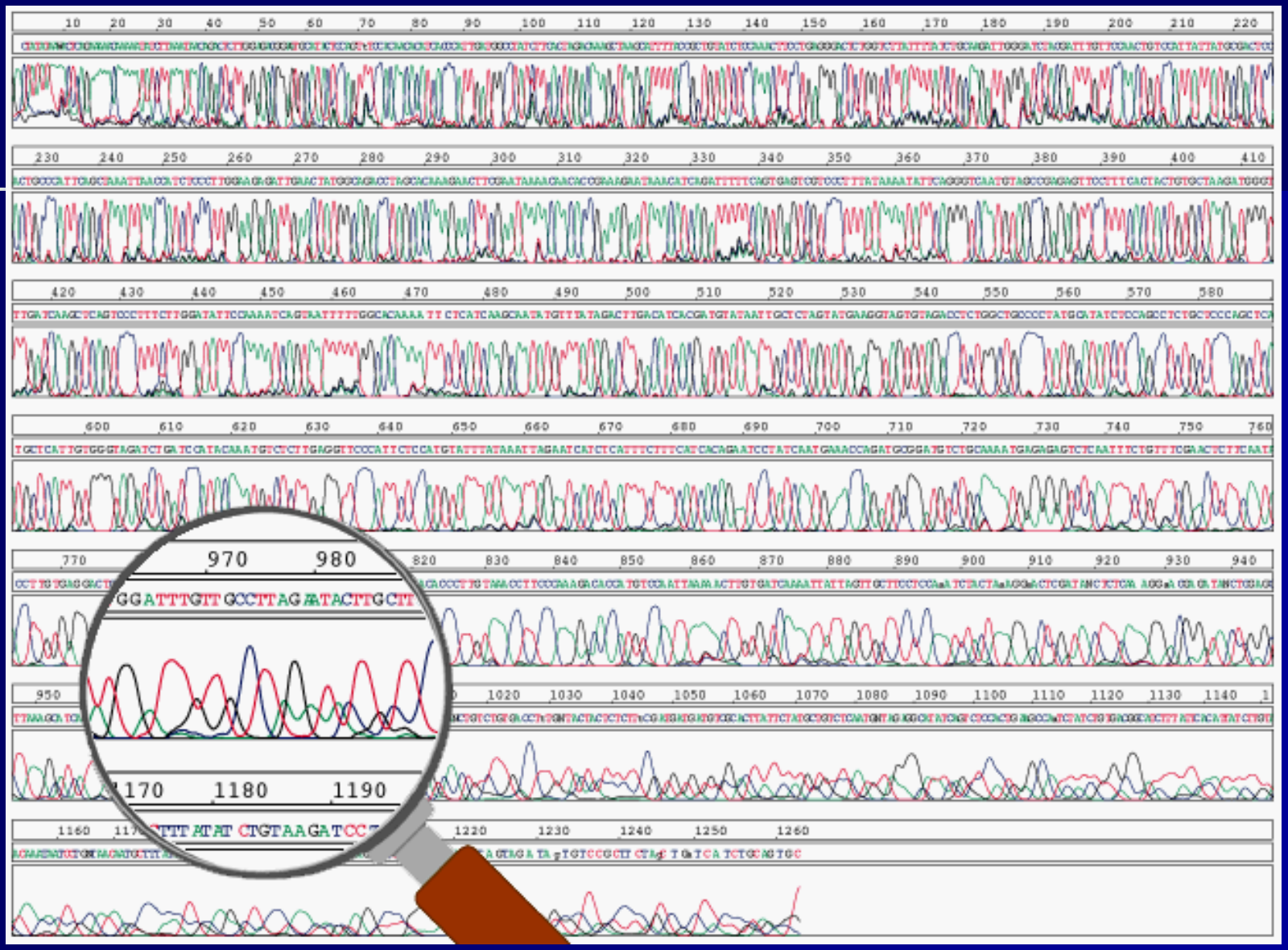
| 碱基 | 特异修饰方法 |
| --- | --- |
| G | Ph8.0,用硫酸二甲酯对 N7进行甲基化,使C8-C9键对碱基裂解有特殊敏感性 |
| A+G | pH2.0 哌啶甲酸可使嘌呤环的N原子化,从而导致脱嘌呤,并因此消弱腺嘌呤和鸟嘌呤的糖苷键 |
| C+T | 肼可打开嘧啶环,后者重新环化成五元环后易除去 |
| C | 1.5mol/L NaCl存在时,可用肼除去胞嘧啶 |

# 化学法测序实例



# 3.3 自动化测序

- 根本原理
- 　　与链终止法测序原理相同,只是用不同的荧光色彩标记ddNTP,如ddATP标记红色荧光,ddCTP标记蓝色荧光, ddGTP标记黄色荧光, ddTTP标记绿色荧光.由于每种ddNTP带有各自特定的荧光颜色,而简化为由1个泳道同时判读4种碱基.


970
980
1170
1180
1190

## 3.4 非常规测序

### 毛细管电泳

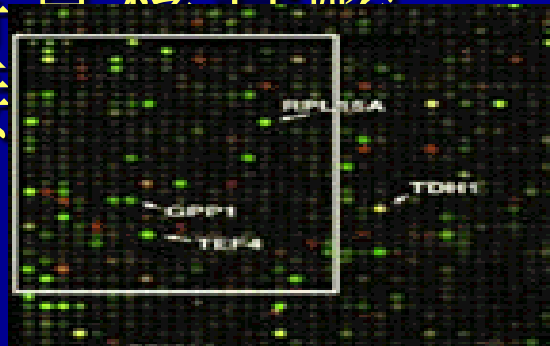
用毛细管电泳取代聚丙烯凝胶平板电泳,节省时间,加快测序进程,其他程序同链终止法或化学测序法.

### 光点测序

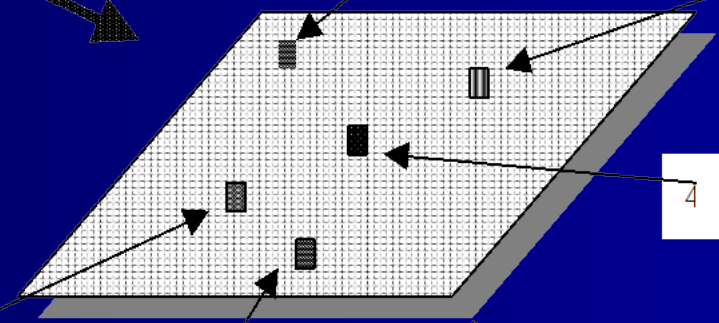
脱氧三磷酸核苷酸连接到DNA 3′-末端时会释放1个焦磷酸(PPi) ,焦磷酸在磷酸化酶的作用下转化为化学能,并发出光亮.由此,往反响液中每次只参加1种核苷酸,当参加的核苷酸结合时,反响液发出亮点,并记录核苷酸种类;当核苷酸未结合时,反响液中的核苷酸酶迅速分解此核苷酸,由此来测定DNA序列.

- DNA芯片测序

- 根本原理
- 将各种排列顺序的寡核苷酸点播在芯片上，每个点播的寡核苷酸在排列的方阵中都有指定的位置.待检测的DNA分子与芯片温浴,但凡能杂交的寡核苷酸都会在确定位置发出信号,然后根据获取的信息将寡核苷酸的顺序进行比照组装,拼接DNA顺序.



DNA 样品 TATGCAATCTAG 与基因芯片上 65,000 种可能的八聚体进行杂交从而形成特定的结合图形




计算机分析杂交图象并由探针的重叠情况推导样品的核酸序列

1 ATACGTTA
3 TACGTTAG
4 ACGTTAGA
2 CGTTAGAT
5 GTTAGATC

互补序列为：ATACGTTAGATC
样品序列为：TATGCAATCTAG

# 利用基因芯片进行杂交测序的原理

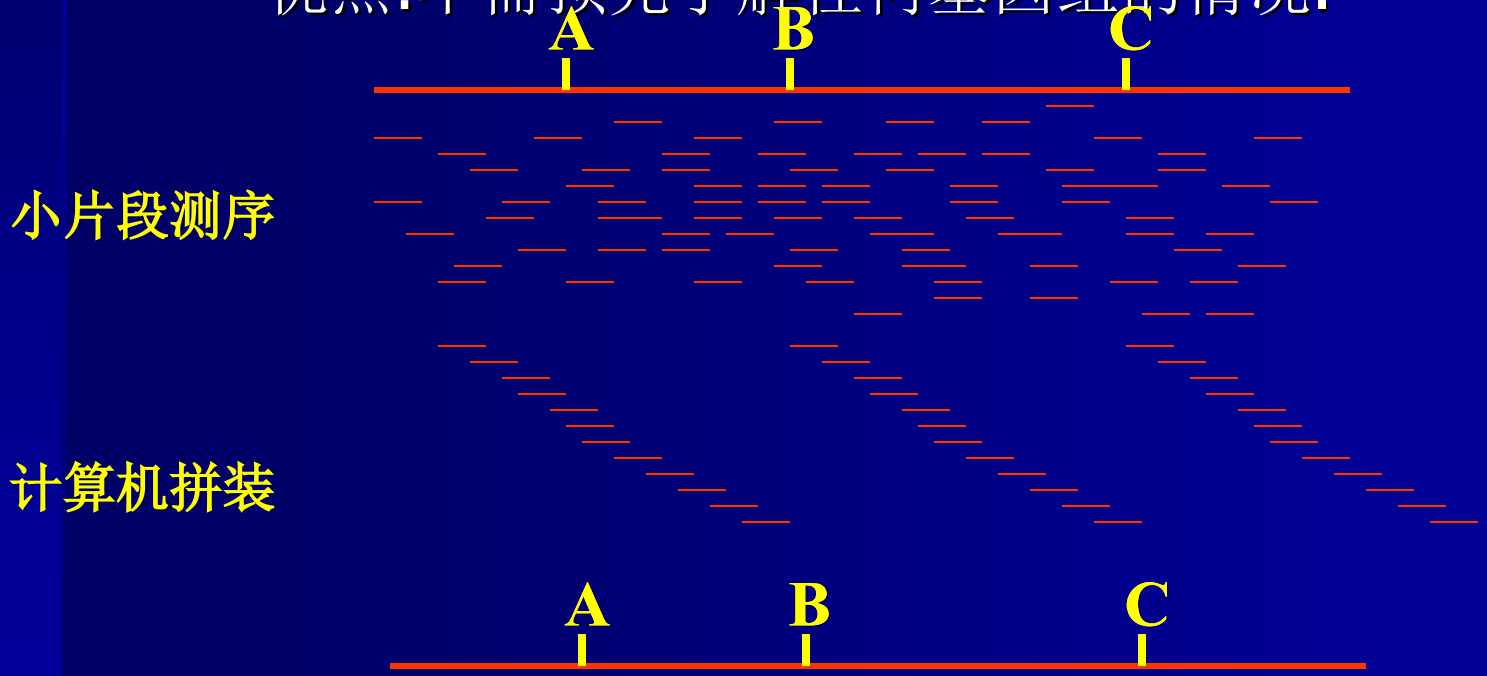
# 4 序列的组装

## 4.1 随机测序与序列组装

随机测序也称"鸟枪法".

序列组装原理:直接从已测序的小片段中寻找彼此重叠的测序克隆,然后依次向两侧邻接的序列延伸.

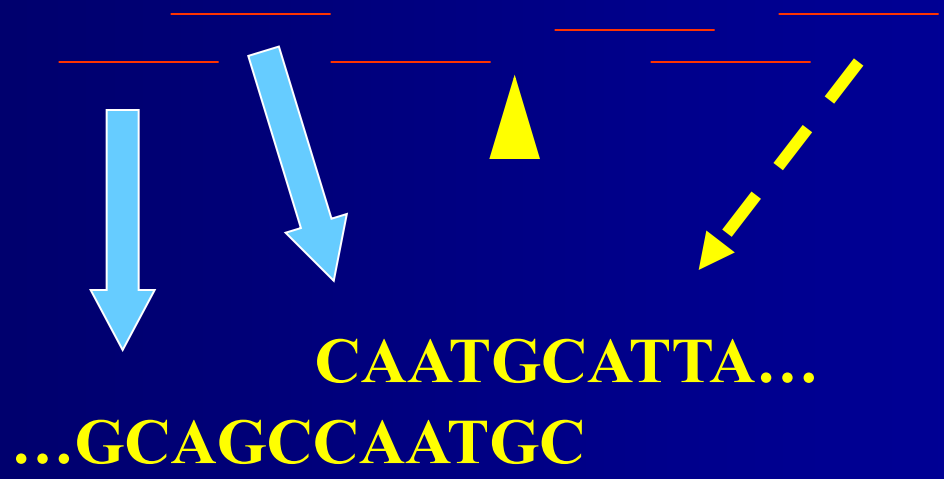
优点:不需预先了解任何基因组的情况.



# 鸟枪法(Shotgun)测序的问题


小片段测序

计算机拼装




# 实例:流感嗜血杆菌基因组的测序及顺序组装

超声波打断纯化的基因组DNA

↓

琼脂糖电泳收集1.6～2.0Kb的区段、纯化

↓

构建到质粒载体中

↓

随机挑选19687个克隆,进行28643次测序,得到可读顺序为11 631 485 bp

↓

组装成140个覆盖全基因组范围的独立的顺序重叠群,

↓

各重叠群间仍有间隙

顺序间隙

↓

测序时遗漏的测序

↓

解决方法:通过相邻顺序作为探针筛选已有的基因组文库

物理间隙

↓

载体或宿主菌选用不当而被丧失的顺序

↓

解决方法:利用其它宿主菌与载体重新构建文库

# 4.2 限制测序

- 限制测序：是指将一段染色体区段的DNA 顺序进行组装.
- 一些已绘制了遗传图与物理图的微生物基因组测序中也采用这一方法.
- 如高等植物拟南芥基因组的测序完全依据克隆重叠群,先进行各个BAC克隆的随机测序,再进行序列组装;
- 水稻基因组测序方案采取得策略与此相同.

# 4.3 指导测序与序列组装

建立在基因组图谱根底上的"鸟枪法",即所谓"指导鸟枪法"或"指导测序"。

在人类基因组进入测序组装阶段就采用此方法,其根本步骤如下:

A 构建平均为2Kb的人类基因组质粒文库,进行双向测序;

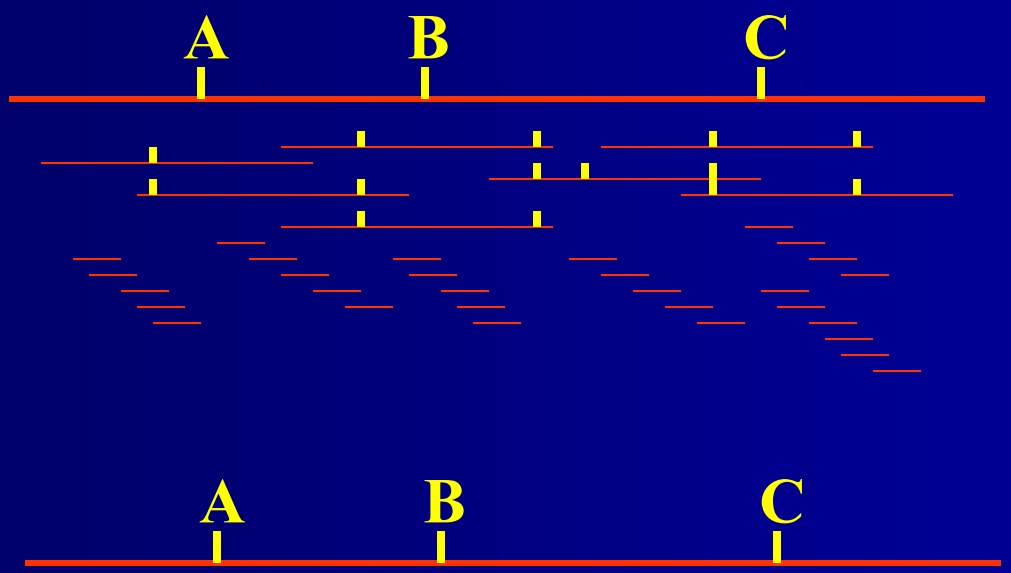
B 构建平均10Kb的人类基因组质粒文库,进行双向测序,读取2个端部顺序;

C 参考人类基因组图,特别是大量的STS位标作为基点,进行序列组装，排成重叠克隆群.

先将染色体打成比较大的片段(几十-几百Kb), 利用分子标记将这些大片段排成重叠的克隆群(Contig), 分别测序后拼装. 这种策略叫基于克隆群(contig-based)的策略.

大片段contig

小片段测序拼装



# 两种策略的比较

| 鸟枪法策略 | 指导测序策略 |
| --- | --- |
| 不需背景信息 | 构建克隆群<br>(遗传、物理图谱) |
| 时间短 | 需要几年的时间 |
| 需要大型计算机 | |
| 得到的是草图(Draft) | 得到精细图谱 |

# 4.5 其他测序路线

- 重要区域优先测序

  人们对感兴趣的基因或与疾病相关的基因优先测序.

如:人类主要组织相容性复合区位于第6号染色体,与人类免疫系统有关，因而优先测序.

- EST (Expressed sequence tag) 测序

- EST是一种重要的基因组图分子标记,以EST为探针很容易从 cDNA文库中筛选全基因,又可从BAC克隆中找到其基因组的基因序列.

- 优点:
- A mRNA 可直接反转录成cDNA,而且cDNA文库也比较容易构建;
- B 对cDNA文库大量测序,即可获得大量EST的序列;
- C EST为基因的编码区,不包括内含子和基