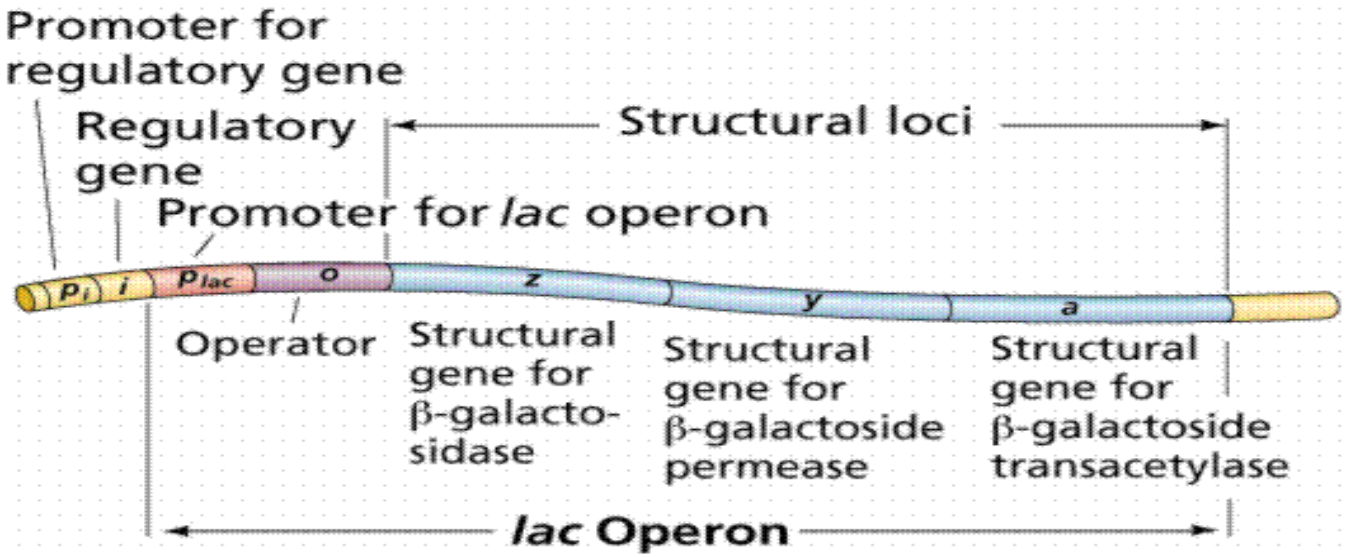
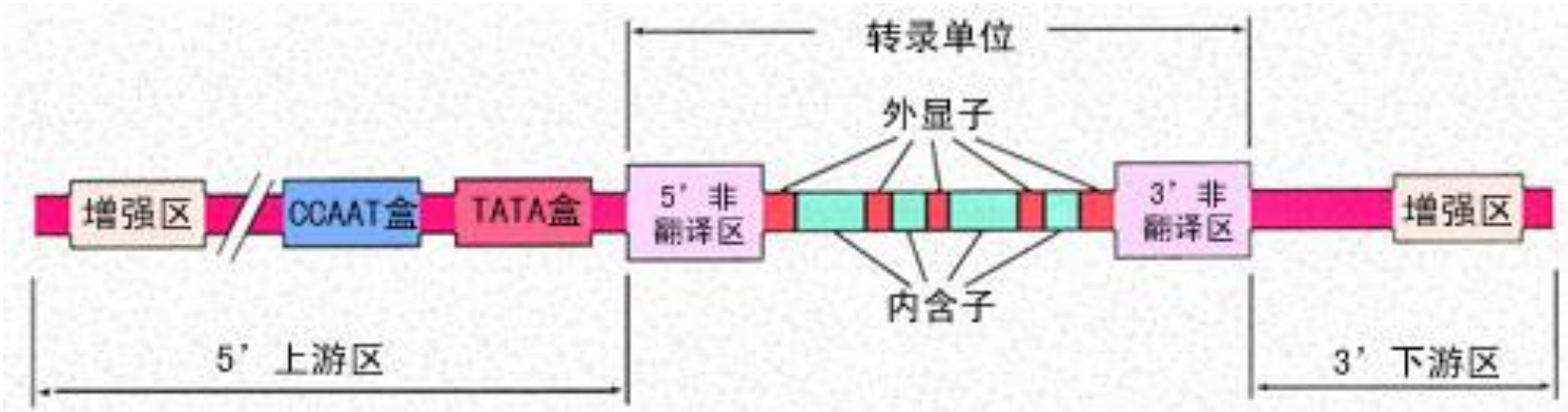
# 基因组学、转录组学
# 基因的结构分析-2

张百芳

# 基因结构分析的方法

# 基因结构分析包括基因编码区和调控区结构分析




真核生物基因的结构

# 基因结构的研究技术

## DNA或基因的结构：

**DNA序列/一级结构；拷贝数；在染色体上的位置；**

**编码区(外显子/内含子)；调控区(启动子、转录起始点)**

DNA 研究技术

- **DNA Sequencing** 揭示基因的一级结构
- **Southern blot** 分析基因的拷贝数变化
- **PCR** 也可分析基因的拷贝数
- **FISH** fluorescent in situ hybridization 分析基因在染色体上的位置
- **cDNA 克隆**
- **RACE/Deep-RACE、5’SAGE/CAGE**
- **DNA footprinting**

# 一、基因结构的生物信息学分析

对基因序列进行生物信息学检索和比对分析，可了解基因的相关信息：基因序列的编码区（外显子/内含子）、调控区（启动子）及其在染色体上的定位，并可对基因表达产物的结构和功能进行预测。

生物信息学的发展

利用电脑对基因序列进行检索、比对、拼接及结构分析和预测等

# 常用数据库


- NCBI数据库

1988，NCBI（美国国立生物技术信息学中心）首先创建GenBank数据库

- 于1991年开发了Entrez数据库检索系统，该系统整合了GenBank、EMBL、PIR和SWISS-PROT等数据库核酸及蛋白质序列信息、蛋白三维结构数据、基因组图谱以及MEDLINE有关序列的文献信息，有机链接成检索系统。

- NCBI提供多种数据库及分析工具，包括在线人类孟德尔遗传(OMIM)、三维蛋白结构的分子模型数据库(MMDB)、人类基因序列集成(UniGene)、人类基因组基因图谱(GMHG)、生物门类(Toxonomy) 等数据库及BLAST程序。

# 常用数据库

## 1. Nucleotide 数据库

该数据库由NCBI 的GenBank、日本DNA数据库(DDBJ)和英国Hinxton Hall的欧洲分子生物学实验室数据库(EMBL)三部分数据组成，三个组织数据交换与共享。

## 2. Genome 数据库

即基因组数据库，提供了多种基因组、完全染色体、重叠序列图谱以及一体化基因物理图谱

## 3. BLAST 数据库

Basic Local Aligment Search Tool是一套核苷酸数据库和蛋白质数据库相似性比较的搜索程序，还可作为鉴别基因和遗传特点的手段。

| 程序 | 数据库 | 查询 | 内容 |
|---|---|---|---|
| BLASTn | 核苷酸 | 核苷酸 | 寻找较高分值的匹配，对较远关系不太适用 |
| BLASTp | 蛋白质 | 蛋白质 | 利用蛋白质序列搜索相似的蛋白质序列， |
| BLASTx | 翻译核苷酸 | 蛋白质 | 对于新DNA序列和EST的分析极为有用 |
| tBLASTn | 蛋白质 | 翻译核苷酸 | 用于寻找数据库中没有标注的基因编码区 |
| tBLASTx | 翻译核苷酸 | 翻译核苷酸 | 特适用于EST分析 |

## 4. Structures 数据库

即结构数据库或称分子模型数据库(MMDB)，包含来自X线晶体学和三维结构的实验数据

NCBI已经将结构数据交叉链接到书目信息、序列数据库和NCBI的Taxonomy中运用NCBI的3D结构浏览器和Cn3D，可以很容易地从Entrez获得分子的分子结构间相互作用的图像

## 5. Taxonomy 数据库

即生物学门类数据库，可以按生物学门类进行检索或浏览其核苷酸序列、蛋白质序列、结构等

# 常用数据库

## 6. PopSet 数据库

包含研究一个人群、一个种系发生或描述人群变化的一组组联合序列

PopSet既包含了核酸序列数据又包含了蛋白质序列数据

## 7. 文献数据库

PubMed：生物医药科学的检索系统，该数据库包括原文信息、参考信息，并可链接*MEDLINE*数据库中相关文献和序列信息

OMIM：孟德尔遗传学数据库是人类基因和基因疾病的目录数据库

其他：书目，杂志，文章引用匹配等

# 通过数据库查找和定位基因序列

## (一)检索/比对已知基因序列

在已有的数据库中对基因或DNA序列进行比对分析，以预测其结构、功能及在进化上的联系.

- 比对方法：
  1. 两两比对
  2. 多序列比对

序列比对目的：

- 判断两个或多个序列间是否具有足够的相似性

从而判断二者之间是否具有同源性

直接的数量关系

进化上曾具有共同祖先

## 序列比对的结果：

- 取代
- 插入
- 缺失

保守序列：

- 可能是共同进化的标志
- 可能并不代表功能的重要性

缺失?

**Mouse:**

**GGKDSCQGDSGGPVVCNG----QLQGVVSWGDGCAQKNKPGVYTKVYNYVKWIKNTIAAN**

**Crayfish:**

**GGKDSCQGDSGGPLAASDTGSTYLAGIVSWGYGCARPGYPGVYTEVSYHVDWIKANAV--**

保守序列

插入?

- 当两个序列非常相似时，是否一定说明它们具有相似的功能？

## 序列同源性比对：

比对表达序列标签（**expressed sequence tag, EST)**

- 可利用**cDNA**及基因组序列的特征等对可能的基因组外显子、内含子进行注释
- 可拼接发现新基因序列

## 比较基因组学：

- 可提供不同生物种属基因**/DNA**结构与功能相似性信息
- 可确定真核基因组的进化机制

例如：

人EPO基因序列检索

在检索框中输入检索词，检索词间默认逻辑关系为AND，检索规则基本同PubMed

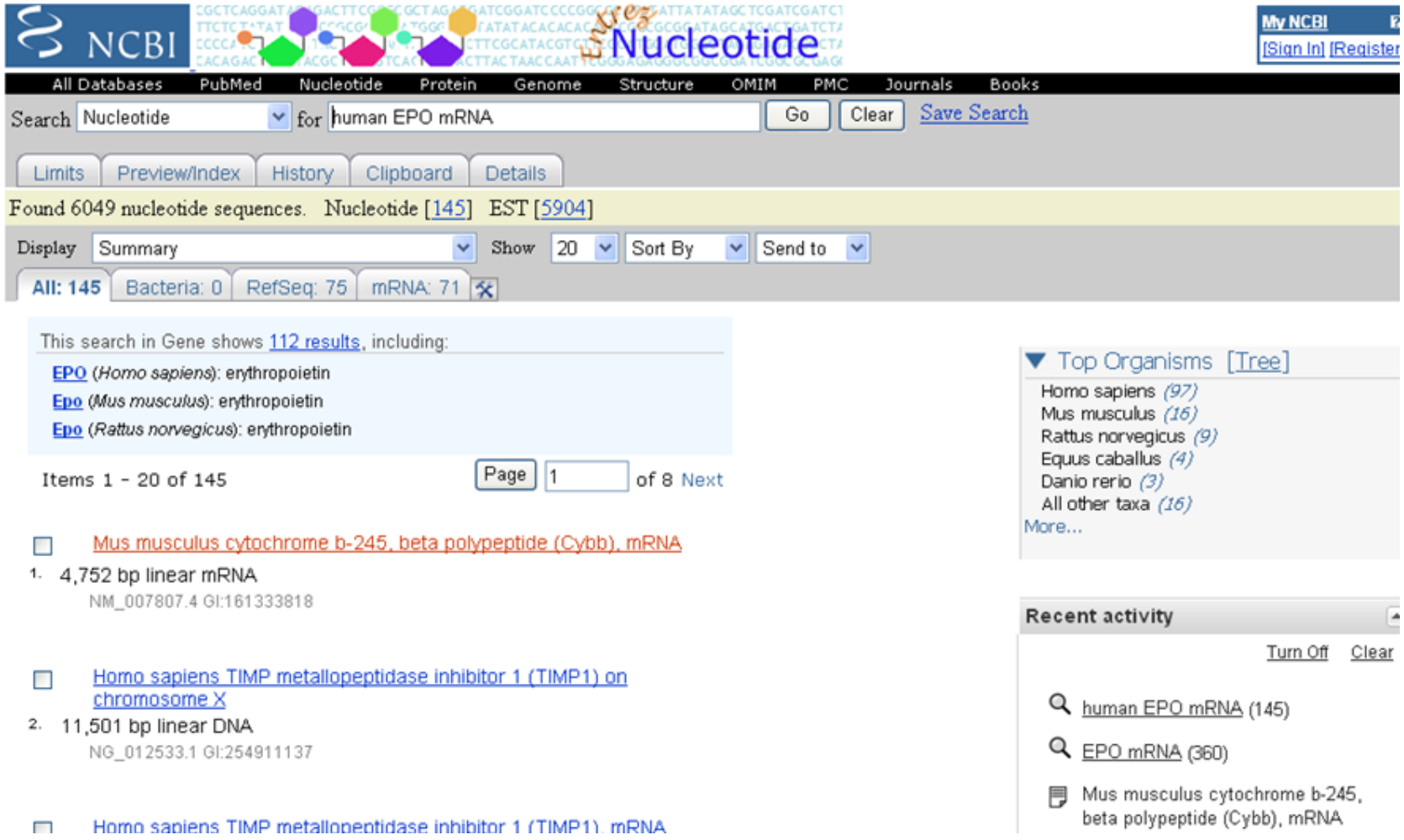
通过下拉菜单选择记录的显示格式，*GenBank Report*格式显示较完整的基因记录；*FASTA Report*格式仅包括检出序列的简要特征描述。

## *NCBI*数据库检索

- 输入关键词，选择合适的程序
- 向下拉寻找符合目标的条目
- 点击此条打开连接
- 向下拉寻找关注的内容
- 可以直接拷贝保存相关内容

# 例如：人EPO基因序列检索

- 输入关键词，选择合适的程序



# 例如：人EPO基因序列检索

- 向下拉寻找符合目标的条目

Mus musculus cytochrome b-245, beta polypeptide (Cybb), mRNA

1. 4,752 bp linear mRNA
   NM_007807.4 GI:161333818

Homo sapiens TIMP metallopeptidase inhibitor 1 (TIMP1) on chromosome X

2. 11,501 bp linear DNA
   NG_012533.1 GI:254911137

Homo sapiens TIMP metallopeptidase inhibitor 1 (TIMP1), mRNA

3. 931 bp linear mRNA
   NM_003254.2 GI:73858576

Homo sapiens erythropoietin (EPO), mRNA

4. 1,340 bp linear mRNA
   NM_000799.2 GI:62240996

Homo sapiens suppressor of cytokine signaling 1 (SOCS1), mRNA

5. 1,216 bp linear mRNA
   NM_003745.1 GI:4507232

Homo sapiens eosinophil peroxidase (EPX) on chromosome 17

6. 19,447 bp linear DNA
   NG_013020.1 GI:260593660

More...

**Recent activity**

Turn Off Clear

- human EPO mRNA (145)
- EPO mRNA (360)
- Mus musculus cytochrome b-245, beta polypeptide (Cybb), mRNA
- EPO (1239) Nucleotide

» See more...

# 例如：人EPO基因序列检索

- 点击此条打开连接

NCBI Reference Sequence: NM_000799.2

## Homo sapiens erythropoietin (EPO), mRNA

Comment Features Sequence

```
LOCUS       NM_000799               1340 bp    mRNA    linear   PRI 01-NOV-2009
DEFINITION  Homo sapiens erythropoietin (EPO), mRNA.
ACCESSION   NM_000799
VERSION     NM_000799.2  GI:62240996
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1340)
  AUTHORS   Rundqvist,H., Rullman,E., Sundberg,C.J., Fischer,H., Eisleitner,K.,
            Stahlberg,M., Sundblad,P., Jansson,E. and Gustafsson,T.
  TITLE     Activation of the erythropoietin receptor in human skeletal muscle
  JOURNAL   Eur. J. Endocrinol. 161 (3), 427-434 (2009)
   PUBMED   19515792
  REMARK    GeneRIF: Signaling through EPOR is involved in exercise-induced
            skeletal muscle adaptation, thus extending the biological role of
            EPO into the skeletal muscle.
REFERENCE   2  (bases 1 to 1340)
  AUTHORS   Labonte,L., Li,Y., Yang,L., Gillingham,A., Halpenny,M., Giulivi,A.,
            Sills,T., Evans,K., Zanke,B. and Allan,D.S.
  TITLE     Increased plasma EPO and MIP-1 alpha are associated with
            recruitment of vascular progenitors but not CD34(+) cells in
            autologous peripheral blood stem cell grafts
  JOURNAL   Exp. Hematol. 37 (6), 673-678 (2009)
   PUBMED   19463769
  REMARK    GeneRIF: At baseline, increased levels of MIP-1 alpha were
            associated with increased graft vascular progenitors while higher
            EPO concentrations on the day of PBSC collection predicted higher
            graft vascular progenitors levels
REFERENCE   3  (bases 1 to 1340)
```

Change Region Shown

Customize View

**Sequence Analysis Tools**

- BLAST Sequence
- Pick Primers

**Articles about the EPO gene**

- Activation of the erythropoietin receptor in human skeletal mus [Eur J Endocrinol. 2009]
- Increased plasma EPO and MIP-1 alpha are associated with recruitm [Exp Hematol. 2009]
- Low oxygen concentration as a general physiologic regulator of [Exp Hematol. 2009]

» See all...

**RefSeq Protein Product**

See the reference protein sequence for erythropoietin precursor (NP_000790.2).

**More about the EPO gene**

This gene is a member of the EPO/TPO family and encodes a secreted, glycosylated cytokine composed of four alpha helical bundles. The protei...

Also Known As: EP, MGC138142, MVCD2

**Homologs of the EPO gene**

The EPO gene is conserved in chimpanzee, dog, cow, mouse, rat, and zebrafish.

# 例如：人EPO基因序列检索

- 向下拉寻找关注的内容

```
gene            1..1340
                /gene="EPO"
                /gene_synonym="EP; MGC138142; MVCD2"
                /note="erythropoietin"
                /db_xref="GeneID:2056"
                /db_xref="HGNC:3415"
                /db_xref="HPRD:00586"
                /db_xref="MIM:133170"
exon            1..194
                /gene="EPO"
                /gene_synonym="EP; MGC138142; MVCD2"
                /inference="alignment:Splign"
                /number=1
STS             104..852
                /gene="EPO"
                /gene_synonym="EP; MGC138142; MVCD2"
                /db_xref="UniSTS:483133"
CDS             182..763
                /gene="EPO"
                /gene_synonym="EP; MGC138142; MVCD2"
                /note="epoetin"
                /codon_start=1
                /product="erythropoietin precursor"
                /protein_id="NP_000790.2"
                /db_xref="GI:62240997"
                /db_xref="CCDS:CCDS5705.1"
                /db_xref="GeneID:2056"
                /db_xref="HGNC:3415"
                /db_xref="HPRD:00586"
                /db_xref="MIM:133170"
                /translation="MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLL
                EAKEAENITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEAVL
                RGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPDAASAAPLRTI
                TADTFRKLFRVYSNFLRGKLKLYTGEACRTGDR"
sig_peptide     182..262
                /gene="EPO"
                /gene_synonym="EP; MGC138142; MVCD2"
mat_peptide     263..760
                /gene="EPO"
```

# 例如：人EPO基因序列检索

- 可以直接拷贝保存相关内容

```
                     /inference="alignment:Splign"
                     /number=4
     exon            608..1330
                     /gene="EPO"
                     /gene_synonym="EP; MGC138142; MVCD2"
                     /inference="alignment:Splign"
                     /number=5
     STS             795..1124
                     /gene="EPO"
                     /gene_synonym="EP; MGC138142; MVCD2"
                     /standard_name="SHGC-12325"
                     /db_xref="UniSTS:37062"
     polyA_site      1330
                     /gene="EPO"
                     /gene_synonym="EP; MGC138142; MVCD2"
ORIGIN
        1 cccggagccg gaccggggcc accgcgcccg ctctgctccg acaccgcgcc ccctggacag
       61 ccgccctctc ctccaggccc gtggggctgg ccctgcaccg ccgagcttcc cgggatgagg
      121 gcccccggtg tggtcacccg gcgcgcccca ggtcgctgag ggaccccggc caggcgcgga
      181 gatgggggtg cacgaatgtc ctgcctggct gtggcttctc ctgtccctgc tgtcgctccc
      241 tctgggcctc ccagtcctgg gcgccccacc acgcctcatc tgtgacagcc gagtcctgga
      301 gaggtacctc ttggaggcca aggaggccga gaatatcacg acgggctgtg ctgaacactg
      361 cagcttgaat gagaatatca ctgtcccaga caccaaagtt aatttctatg cctggaagag
      421 gatggaggtc gggcagcagg ccgtagaagt ctggcagggc ctggccctgc tgtcggaagc
      481 tgtcctgcgg ggccaggccc tgttggtcaa ctcttcccag ccgtgggagc ccctgcagct
      541 gcatgtggat aaagccgtca gtggccttcg cagcctcacc actctgcttc gggctctggg
      601 agcccagaag gaagccatct cccctccaga tgcggcctca gctgctccac tccgaacaat
      661 cactgctgac actttccgca aactcttccg agtctactcc aatttcctcc ggggaaagct
      721 gaagctgtac acaggggagg cctgcaggac aggggacaga tgaccaggtg tgtccacctg
      781 ggcatatcca ccacctccct caccaacatt gcttgtgcca caccctcccc cgccactcct
      841 gaaccccgtc gaggggctct cagctcagcg ccagcctgtc ccatggacac tccagtgcca
      901 gcaatgacat ctcaggggcc agaggaactg tccagagagc aactctgaga tctaaggatg
      961 tcacagggcc aacttgaggg cccagagcag gaagcattca gagagcagct ttaaactcag
     1021 ggacagagcc atgctgggaa gacgcctgag ctcactcggc accctgcaaa atttgatgcc
     1081 aggacacgct ttggaggcga tttacctgtt ttcgcaccta ccatcaggga caggatgacc
     1141 tggagaactt aggtggcaag ctgtgacttc tccaggtctc acgggcatgg gcactccctt
     1201 ggtggcaaga gcccccttga caccggggtg gtgggaacca tgaagacagg atgggggctg
     1261 gcctctggct ctcatggggt ccaagttttg tgtattcttc aacctcattg acaagaactg
     1321 aaaccaccaa aaaaaaaaaa
//
```

- 凡是链接的地方都可以点击查看

# 通过数据库查找和定位基因序列

## (二)查找/检索未知基因序列

以一段DNA序列作为“检索探针”

- 通过EST进行电子克隆
- 利用不同生物基因数据库进行同源性比对
- 通过DNA序列的推导产物预测查找基因序列

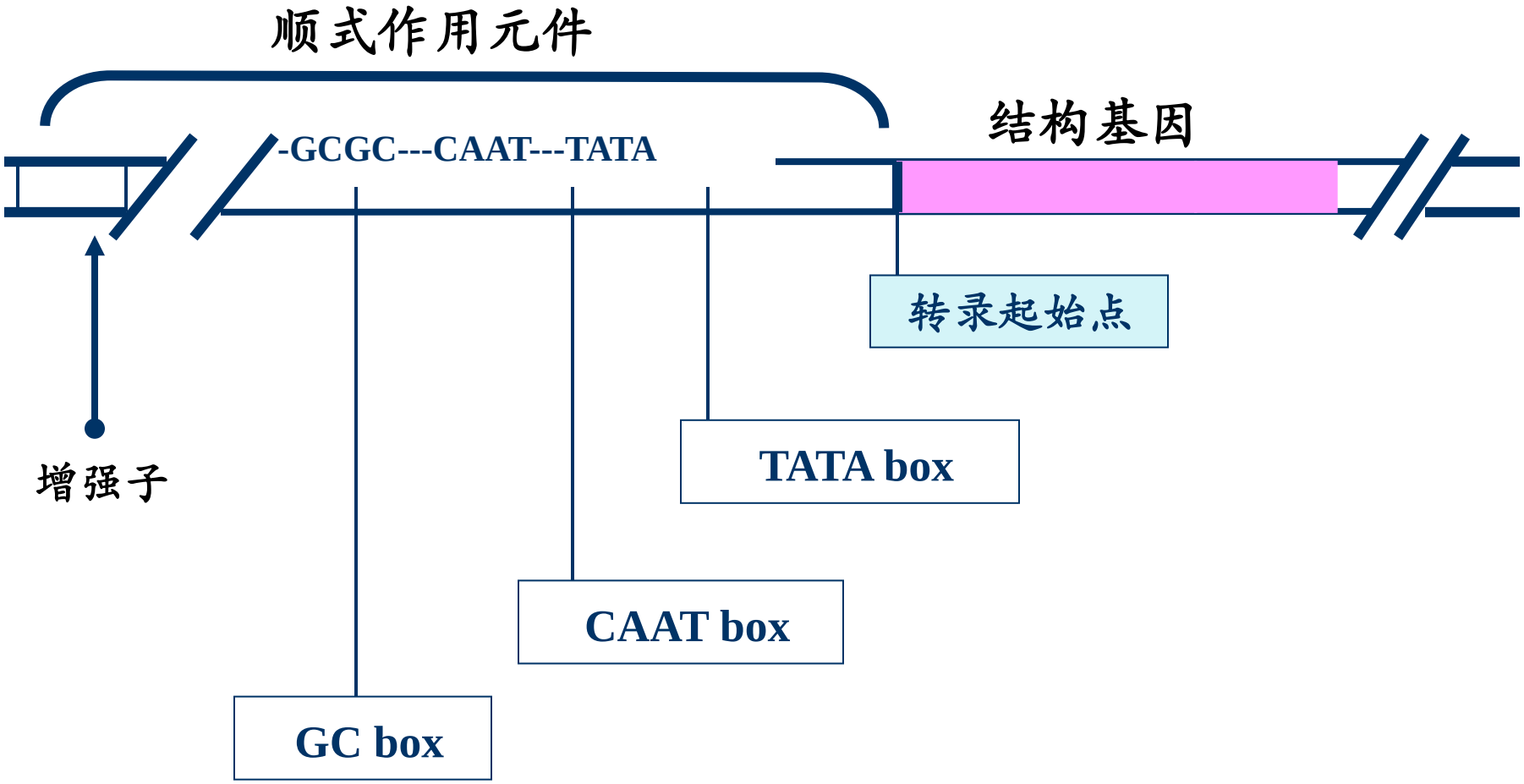
## (三)基因序列的染色体定位

- 利用不同生物基因组数据库，以基因编码区有高度同源性的特征进行基因作图及基因定位，从而确定在染色体中的位置。
- 基因定位到染色体的相应位置，可根据基因图谱对上下游的基因进行观察以精确定位。

# 二、基因启动子及调控序列的结构分析方法

# 基因启动子与转录起始点的序列特征

## 真核基因及其调控元件



# 一）生物信息学预测启动子及转录起始点

## 生物信息学预测启动子(promoter)

- 真核基因组的测序正在以不断增长的速度进行着，目前已经可以获得大约50个完整真核生物基因组的序列信息；
- 对基因组注释工作中最难的就是精确鉴定和描绘启动子，因此，启动子的预测就显得非常重要

预测启动子的切入点

- 启动子的结构特征
- 启动子在染色体上的位置

# 1. 启动子的结构特征

- Ⅱ型启动子通常位于结构基因的上游
- **共有序列(consensus sequence)**是其特征性序列

典型启动子包括：

**核心启动子**：一般在TSS上游-35区域以内，如TATA盒

**启动子上游元件**：一般涉及TSS上游几百个碱基，如GC盒

**启动子远端调控序列**：一般涉及TSS上游几千个碱基

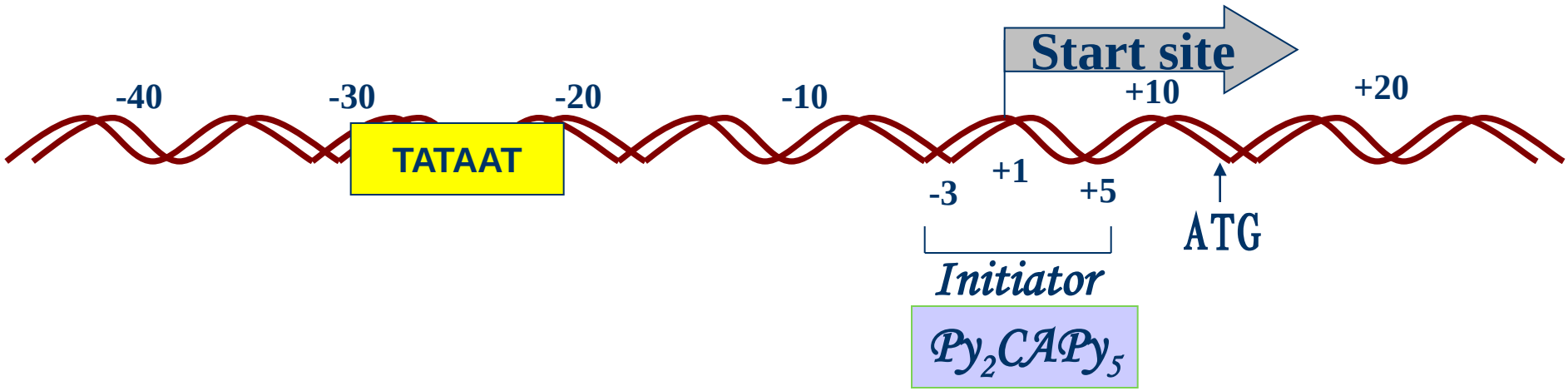
含有增强子或沉默子

- 真核基因启动子在-50区域附近（大约5%~30%基因启动子在-25~-30区域）有TATA box（TATAAA序列）

# 1. 启动子的结构特征

- 一些特征性的结构：TSS附近的CG岛经常出现在启动子中；转录因子结合位点（TFBS）密度及碱基组成也有特点

**共有序列和启动子所处的位置是研究启动子的重要线索**



## 2. 启动子的预测分析

- 启动子数据库
- **EPD (Eukaryotic promoter databases)**
- **TRRD (Transcription regulatory regions databases)**
- 基因转录起始点数据库 **(DBTSS)**

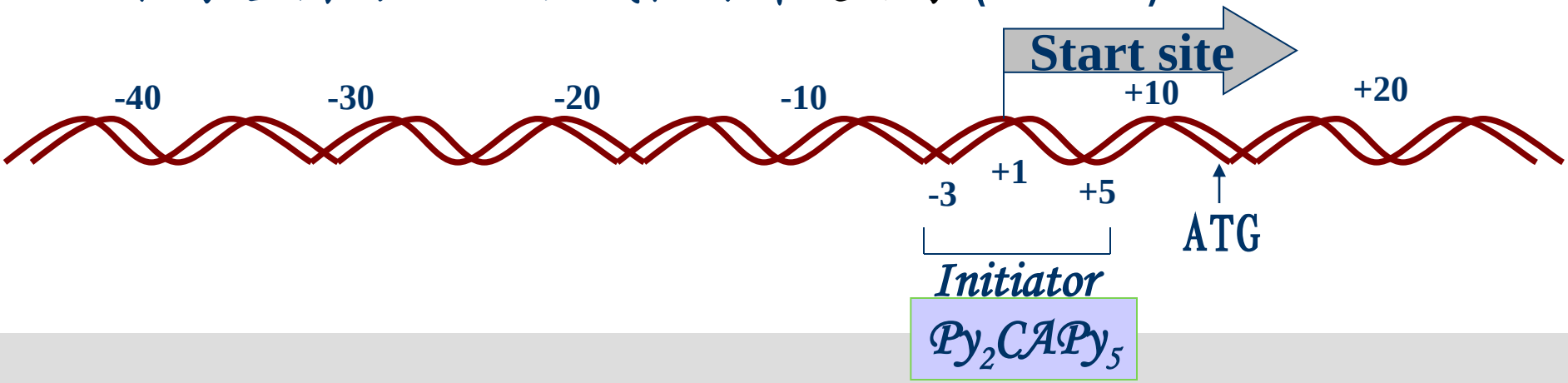
这些数据库主要通过计算机识别、判断及分析，在数据库中寻找启动子的特征结构。

# 利用数据库搜索基因转录起始点

**II 型启动子的转录起始点(transcription start site, TSS):**

**与mRNA第一个碱基对应的DNA序列**

- **无明确的保守序列**
- **mRNA 的第一个碱基倾向是A，其侧翼碱基一般是嘧啶**
- **与mRNA第一个碱基对应的位置标记为+1区**
- **位于基因的-3～+5区域被称作起始子 (initiator)**



- 利用寡核苷酸帽法构建cDNA文库，进行测序，建立转录起始点数据库（DBTSS）
- 寡核苷酸帽法加平行测序法，开发了TSS测序法，建立了TSS测序文库的TSS标签（$3*10^8$）。

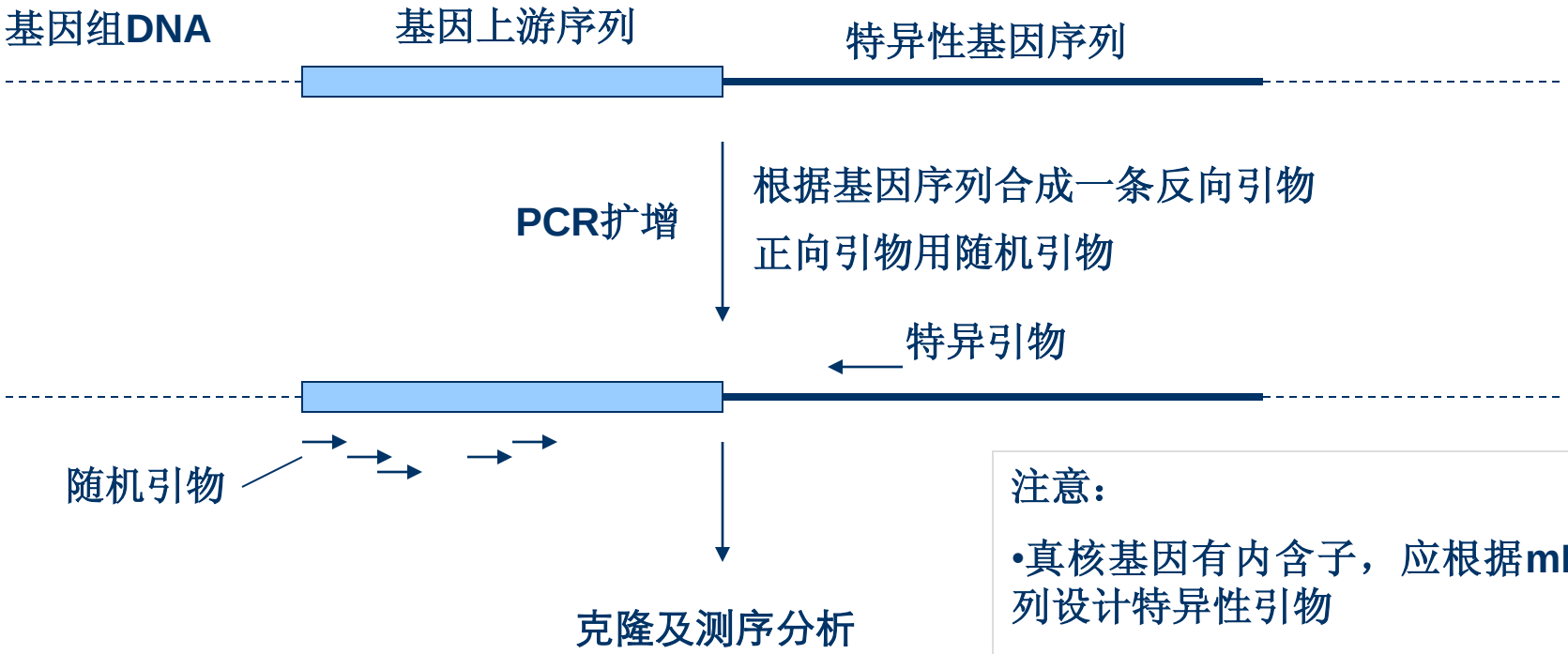
# 二）研究启动子结构的实验方法

主要方法

- 利用PCR技术克隆启动子
- 利用核酸-蛋白质相互作用方法研究启动子
- 利用报告基因研究启动子活性

# 利用PCR技术克隆启动子

## 1. 根据已知基因序列直接进行PCR扩增




注意：

- 真核基因有内含子，应根据mRNA序列设计特异性引物
- 特异性引物尽可能靠近基因的5′端

# 2. 利用环状PCR钓取启动子