1.基因组注释数据库
genome：https://www.ncbi.nlm.nih.gov/genome/?term=
可以 FTP 模式下载物种基因组数据，下载格式有 asn，fasta，genbank，genbank 无序列，mfa 多重 fasta，且可以下载 gff 格式的基因组注释文件，可以提交基因组数据
gbs 格式没有序列信息，包含 contig 以及间隔的 gap 信息
mfa 格式，几乎与 fasta 格式相同，在序列描述方面包含了更多信息，如简单重复序列
首页包括包含人类基因组，微生物基因组资源，亚细胞器基因组资源，原核基因组注释，真核基因组注释，病毒基因组成对比较，基因组装配数据库资源，基因组计划数据库资源，生物样本数据库资源，人类基因组 blast 搜索快速链接，微生物基因组 blast 搜索快速链接
genomebrower：https://genome.ucsc.edu/
ensemble: http://asia.ensembl.org/有 blat/blast（输入蛋白序列与其他物种进行比对）
可看 genetree，orthologues（直系同源）可用 biomart 处理数据

小结

	启动子	转录因子
基于实验数据的数据库	EPD、ENCODE	ENCODE、TRANFAC
分析工具	TRED、NNPP、Promoter2.0、【SoftBerry】FPROM、TSSW、TSSG、UCSC Galaxy、CISTER	TFSEARCH

一、　基因组注释信息的数据存放格式
包括 gff1,gff2,gff3,gtf1/2

gff 文件除 gff1 以外均由 9 列数据組成，前 8 列在 gff 的 3 個版本中信息都是相同的，只是名稱不同：例如第一列在 gff1、gff2 和 gff3 中分別叫做”seqname″，“reference　sequence”和“seqID”，type 在 gff1、gff2 中也被稱作 feature，phase 在 gff1、gff2 中也被稱作 frame。
第 9 列 attributes 的內容存在很大的版本特異性。這 9 列信息（以 gff3 為例）分別是：

```
seqid source type start end score strand strandattributes
```

- seqid ：參考序列的 id。
- source：註釋的來源。如果未知，則用點（.）代替。一般指明產生此 gff3 文件的軟件或方法。

- type： 類型，此處的名詞是相對自由的，建議使用符合 SO 慣例的名稱（sequenceontology），如 gene，repeat_region，exon，CDS 等。
- start：開始位點，從 1 開始計數（區別於 bed 文件從 0 開始計數）。
- end：結束位點。
- score：得分，對於一些可以量化的屬性，可以在此設置一個數值以表示程度的不同。如果為空，用點（.）代替。
- strand：“＋”表示正鏈，“－”表示負鏈，“.”表示不需要指定正負鏈。
- phase ：步進。對於編碼蛋白質的 CDS 來說，本列指定下一個密碼子開始的位置。可以是 0、1 或 2，表示到達下一個密碼子需要跳過的鹼基個數。
- attributes：屬性。一個包含眾多屬性的列表，格式為“標籤＝值”（tag=value），不同屬性之間以分號相隔。

gtf 同 gff3 很相似，也是 9 列內容

seqname: 序列的名字。通常格式染色體 ID 或是 contig ID。

source：註釋的來源。通常是預測軟件名或是公共數據庫。

start：開始位點，從 1 開始計數。

end：結束位點。

feature ：基因結構。CDS，start_codon，stop_codon 是一定要含有的類型。

score ： 這一列的值表示對該類型存在性和其座標的可信度，不是必須的，可以用點“.”代替。

strand：鏈的正向與負向，分別用加號+和減號-表示。

frame：密碼子偏移，可以是 0、1 或 2。

attributes：必須要有以下兩個值：

gene_id value; 表示轉錄本在基因組上的基因座的唯一的 ID。gene_id 與 value 值用空格分開，如果值為空，則表示沒有對應的基因。

transcript_id value; 預測的轉錄本的唯一 ID。transcript_id 與 value 值用空格分開，空表示沒有轉錄本。

gtf2的內容和gff3也是很相似的，區別只在其中的2列：

	gtf2	gff3
feature/type	必须注明	可以是任意名称
attributes	名称和值以“空格”隔开	名称和值以符号“=”隔开

二、bimart 在线使用方法 http://asia.ensembl.org/biomart

1. 选择 ensemblgenes92
2. 选择 humangenes（38.p12）
3. 点击 filter，设定数据库筛选条件
4. 点击 attributes，设定筛选返回结果字段
5. 点击 result 即可

二、 隐马尔可夫模型

包含隐藏状态，可观察输出，转移概率，输出概率

概念延伸:在生物序列分析中，给你一组同源基因序列或同一家族的蛋白质序

列，构建出一个 HMM;然后再利用该模型去识别一个新序列是否属于该类同源基因或该蛋白质家族。

三、 启动子类型

1. 核心启动子：引发转录的必要部份及转录起始点，位置约为- 35;且是 RNA 聚合酶的结合位点及一般转录因子结合位点。
2. 近端启动子：基因的近端序列上游，包括一些基本的调控元件，位置约为 -250，且是特定转录因子结合位点。
3. 远处启动子：基因的远处序列上游，包括一些额外的调控元件，影响力较近端启动子弱。

四、 真核生物启动子

真核生物启动子是极端的分化及很难表现其特征。它们一般处于基因的上游及有着远离转录起始点的调控元件。转录复合物可以引起脱氧核糖核酸(DNA)向自己屈曲，以容许放置调控序列。很多真核生物启动子，但不是全部，都包含一个 TATA 盒(序列 TATAAA)会与 TATA 结合蛋白结合，以协助形成 RNA 聚合酶转录复合物。TATA 盒一般会处于非常接近转录起始点(通常于 50 个碱基对以内)

五、 聚合酶类型

1. RNA 聚合酶 I 存在：核仁 功能：合成 rRNA 前体，识别 I 类启动子，只控制 rRNA 前体基因的转录，转录产物经切割和加工后生成各种成熟 rRNA。RNA 聚合酶 I 对其转录需要 2 种因子参与，UBF1（一条 M 为 97000 的多肽链，结合在上述两部分的富含 GC 区; 1 个 TBP，即 TATA 结合蛋白）,SL1（一个四聚体蛋白，含有 3 个不同的转录辅助因子 TAFI; 在 SL1 因子介导下 RNA 聚合酶 I 结合在转录起点上并开始转录）。
2. RNA 聚合酶 II 存在：核质 功能：合成 mRNA 前体 识别 II 类启动子，催化 mRNA 和大多数核内小 RNA(snRNA)合成
3. RNA 聚合酶 III 存在：核质 功能：合成 5S rRNA 前体、tRNA 前体及其他的核和胞质小 RNA 前体 涉及一些小分子 RNA 的转录。

六、 启动子的组成

I 类启动子：核心启动子，上游控制元件

II 类启动子：

基本启动子：序列为中心在-25 至-30 左右的 7 bp 保守区，TATAAAA/T，称 为 TATA 框或 Goldberg-Hogness 框。与 RNA 聚合酶的定位有关，DNA 双链在此解开并决定转录的起点位置。失去 TATA 框，转录将在许多位点上开始。

起始子：转录起点位置处的一保守序列，共有序列为: PyPyANT(A)PyPyPy 为嘧啶碱(C 或 T)，N 为任意碱基，A 为转录的起点。 DNA 在此解开并起始转录。

上游元件：普遍存在的上游元件有 CAAT 框、GC 框和八聚体(octamer) 框等。CAAT 框的共有序列是 GCCAATCT,GC 框的共有序列 为 GGGCGG 和 CCGCCC，八聚体框含有 8bp，共有序列为 ATGCAAAT。

应答元件：诱导调节产生的转录激活因子与靶基因上的应答元件结合。如热休克效应元件 HSE 的共有序列是 CNNGAANNTCCNNG， 可被热休克因子 HSF 识别和作用;血清效应元件 SRE 的共有序 列 CCATATTAGG，可被血清效应因子 SRF 识别和作用。

III 类启动子：

类别 III 启动子为 RNA 聚合酶 III 所识别,涉及一些小分子 RNA 的转录。

类型 1 基因内启动子：

如 5S rRNA 基因的启动子，位于转录起点下游，即在基因内部， 是 下游启动子，有两个框架序列，被 3 种辅助因子所识别。 5SrRNA 基因的启动子包括框架 A(box A)、中间元件 (intermediate element)和框架 C(box C)3 个元件组成。 TFIIIA 结合在框架 A 上，然后促使 TFIIIC 结合，后者结合导致 TFIIIB 结合到转录起点附近，并引导 RNA 聚合酶 III 结合在起点 上。TFIIIB 使 RNA 聚合酶 III 正确定位，起“定位因子” (positioning factor)作用。

类型 2 基因内启动子：

如 tRNA 基因的启动子,有两个控制元件,分别为框架 A 和框架 B。TFIIIC 结合框架 B，其结合区域包括框架 A 和框架 B，然后 导致 TFIIIB 结合到转录起点附近，并引导 RNA 聚合酶 III 结合在 起点上。

上游启动子

如 snRNA 基因的启动子,位于转录起点上游。有 3 个上游元件:OCT(八 聚体基序 octamer motif)、PSE (邻近序列元件 proximal sequence element)、TATA 元件。 在 RNA 聚合酶 III 的上游启动子中，只有靠近起点存在 TATA 元件，就能起始转录。 然而 PSE 和 OCT 元件的存在将会增加转录效率。

七、 参与 RNA 聚合酶 II 转录起始的各类因子

通用因子：作用于基本启动子上的辅助因子称为通用(转录)因子(GTF) ，或基本转录因子(basal transcription)，为任何细胞类别 II 启动子起始转录所必需，以 TFIIX 来表示

上游因子：转录辅助因子，是指识别上 游元件的转录因子

可诱导因子：生长发育不同阶段相关的基因表达调控

八、 对外界刺激信号的响应

1. 主要通过转录激活物
2. 诱导的转录激活因子与靶基因上应答元素相结合

九、研究某一基因的启动子和转录因子

1. 查文献报道（要查基因别名）
2. 找其他实验数据（USCS ENCODE 整合的 chip-seq 数据）
3. 使用启动子和转录因子分析工具进行分析和预测
4. 克隆引物设计

小结

可以利用 PubMed 数据库，查找某个基因已有研究报道的启动子信息;

可以利用 UCSC Galaxy、Genbank、TRED 等数据库，获 取某个基因的可能的启动子序列信息;

可以利用 NNPP、Promoter2.0、FPROM、TSSW、TSSG、 CISTER 等，计算分析某个基因上游可能的启动子;

可以利用 ENCODE、TRANSFAC 等数据库,查找某个基因启动子区域的转录因子信息;

可以 TRED、TFSERCH 等,计算分析某个基因启动子区域 可能的转录因子结合位点。

九、　寻找靶基因（没看懂）
十、　基因组进化研究内容
structural analysis of the genome
the study of genomic parasites
gene and ancient genome duplications
polyploidy
comparative genomics
十一、新基因产生的机制
exon shuffling
gene fission/fusion
retrotransposition
duplication-divergence
lateral gene transfer
十二、进化树绘制算法
根据距离：邻接法 UPGMA
根据进化：Maximum parsimony Maximum likelihood
十二、RNA 分类
coding RNA:mRNA
Non-coding RNA: rRNAtRNA(二级结构为三叶草，三级结构为倒 L 形) microRNA
siRNA long ncRNA