1. 基因组测序与组装
2. 序列比对
3. 基因预测与结构建模
4. 基因组注释数据库
5. HMM 隐马尔可夫模型
6. DNA 元件分析和预测
7. 基因组进化与比较基因组学
8. RNA 序列分析
9. 基因组数据挖掘

基因组图谱数据库：

Genbank-Genome-Map viewer

UCSC-Genome browserEnsembl-Genome browser

……. BING Search……

从头计算预测：

GENSCAN: http://genes.mit.edu/GENSCAN.html

tRNAscan-SE: http://selab.janelia.org/tRNAscan-SE/

基因结构建模：

NCBI Splign

UCSC BLAT

EMBL-EBI GeneWise

原核生物从头预测基因：

GLIMMER

GeneMark GeneMarkS GeneMarkS+

真核生物从头预测基因：

GENSCAN

Geneid

Augustus

GeneMark-ES GeneMark-ET

GlimmerHMM

mSplicer

CONTRAST

mGen

FGENESH

密码子偏好：

JAVA Codon Adaptation Tool (JCat): http://www.jcat.de/

Codon Optimization Tool: http://sg.idtdna.com/CodonOpt

检验 gff3 格式：

http://genometools.org/cgi-bin/gff3validator.cgi

https://github.com/modENCODE-DCC/validator

galaxy：https://usegalaxy.org/

Variant Effect Predictor：http://asia.ensembl.org/Tools/VEP （输入突变信息）

启动子数据库：

http://molbiol-tools.ca/Promoters.htm

https://bip.weizmann.ac.il/toolbox/seq_analysis/promoters.html

EPD:https://epd.vital-it.ch/index.php

https://cb.utdallas.edu/cgi-bin/TRED/tred.cgi?process=home

transfac:不能用

ENCODE: https://www.encodeproject.org

果蝇启动子：http://www.fruitfly.org/seq_tools/promoter.html

promoter2.0：http://www.cbs.dtu.dk/services/Promoter/

http://linux1.softberry.com/berry.phtml?topic=fprom&%20group=programs&subgroup=promoter

cister：https://zlab.bu.edu/~mfrith/cister.shtml

tfsearch:http://diyhpl.us/~bryan/irc/protocol-online/protocol-cache/TFSEARCH.html

Tssw：

http://linux1.softberry.com/berry.phtml?topic=tssw&group=programs&subgroup=promoter

Tssg：

http://linux1.softberry.com/berry.phtml?topic=tssg&group=programs&subgroup=promoter

比较基因组学数据库:

dcode：https://ecrbrowser.dcode.org/

rvista2.0 ：https://rvista.dcode.org

CoGe: **https://genomevolution.org/CoGe/**

homologene: https://www.ncbi.nlm.nih.gov/homologene/?term=

ENSEMBLE: http://asia.ensembl.org/

clustal: https://www.ebi.ac.uk/Tools/msa/clustalo/

RNA 数据库:

http://biobases.ibch.poznan.pl/ncRNA/

http://rfam.xfam.org/

https://www.science.co.il/biomedical/databases/RNA-databases.php

RNA 修饰：

http://mods.rna.albany.edu

miRNA：http://www.mirbase.org

siRNA:http://sirna.sbc.su.se

http://www.ncrna.org/

long-noncodingrna：http://lncrnadb.com/

RNA 结构预测：

https://bibiserv.cebitec.uni-bielefeld.de/rna

http://rna.urmc.rochester.edu/RNAstructureWeb/

http://mfold.rit.albany.edu/?q=mfold

http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi

tRNA: **http://lowelab.ucsc.edu/tRNAscan-SE/**

microRNA：**https://cm.jefferson.edu/rna22v2/**

sirna：**http://www.imtech.res.in/raghava/desirm/**

密码子偏好性优化：**http://www.jcat.de/CAICalculation.jsp**

绪论

- 基因组》单倍体基因组

1.核/染色体基因组
2.线粒体基因组
3.叶绿体基因组
4.病毒基因组

- 基因组学:应用重组 DNA、DNA 测序方法和生物信息学来对基因组的功能和结构进行排序、组装和分析。
- 人类基因组：核基因组：3*10^9bp（每个染色体 55-250Mb）

  线粒体基因组:16569bp(每个细胞 800 个线粒体,每个线粒体 10 个基因组拷贝)
- 可变的 DNA 结构：A-DNA, B-DNA, Z-DNA [DNA 结构取决于其环境]

①A-DNA 结构在脱水样品中占主导地位，类似于双链 RNA 和 DNA/RNA 杂交。
②水环境,包括大多数的 DNA 在细胞,BDNA 是最常见的结构。
③Z-DNA 是一种少见的结构中发现 DNA 绑定到特定的蛋白质。

- DNA 超螺旋：

  30 亿对碱基；拉直 2m 长；6 微米宽
- 人类基因组计划：

  30 亿个碱基对，30 亿美元，1990-2005【1999.9 中国正式加入，承担人类第 3 号染色体的短臂区域 3000 万个碱基对（1%)】

  [1998 国家人类基因组南方研究中心、华大基因成立；1999 国家人类基因组北方研究中心、诺赛基因成立]

  英日法德中印

  =》国际人类基因组组织 HUGO
- 基因组进化复杂性【宏观问题】》C 值悖论
- 基因组的序列组成【微观一级结构】
  - 重复序列 k-mers：
    - 高度重复序列（卫星 DNA/小卫星 DNA/微卫星 DNA/VNTR)；
    - 中度重复序列（逆转录转座子、DNA 转座子、长末端重复(LTRs)、非长末端重复

      (non-LTRs)、长间隔重复(LINEs)、短间隔重复(SINEs)）
  - 非重复序列：单一序列（80%基因）
  - 其他 DNA 元件：CpG 岛、G-quadruplex、启动子、TFBs、转录起始位点 TSS、终止子

PS:卫星 DNA:由大段重复的非编码组成的，卫星 DNA 是功能性中心粒主要组分，形成异染色质的主要结构;

小卫星 DNA:一类可变串联重复（ variable number tandem repeat， VNTR)， 有一系列 10~60bp 重复序列组成的 DNA 片段;

微卫星 DNA:简单序列重复(SSRs)或短串联重复序列，是重复序列的 2-6 碱基对 DNA。它是一种变数串联重复(VNTR);

VNTR 是基因组中的一个位置，短核苷酸序列被组织成串联重复序列。这些可以在许多染色体上发现，并且经常显示个体之间的长度变化。

- non-codingRNA 基因：rRNA、tRNA、scRNA、snRNA、snoRNA、miRNA

- coding 蛋白基因：mRNA
- 生物基因中有哪些异常结构基因：重叠基因【基因内基因、反义基因】
- 假基因：

  指来源于功能基因但已使其活性的 DNA 序列，有沉默的假设基因，也有可转录的假基因
- 全基因组测序=》序列组装=》基因组注释

  =》信息查询、基因搜寻、启动子分析、其他 DNA 元件

  =》比较基因组学、RNA 序列分析、其他核酸分析软件使用

遗传图与物理图绘制

1. 遗传图【连锁图】：基因/DAN 标志在染色体上的相对位置与遗传距离，显示基因以及其他序列特征在基因组上位置的图。

   遗传距离通常以基因或 DNA 片段在染 色体交换过程中的分离频率厘摩 （ cM ）来表示：

   ① cM 值越大，两者之间距离越远；

   ② 一般可由遗传重组检测结果推算
2. 遗传作图方法：孟德尔遗传学、遗传重组-连锁分析
3. 人类基因组计划绘制人类基因组的四张图：遗传图、物理图、序列图、转录图
4. 匹配小片段序列在基因组（染色体）上的正确位置=》两种测序策略

   4.1.作图法测序：

   高密度分子标记遗传图和大分子 DNA 克隆重叠群 contig，

   将单个大分子 DNA 克隆逐个测序（小段），

   序列组装

   4.2.鸟枪法测序

   全基因组鸟枪法随机测序（小片段），

   搭建重叠群，并到大分子克隆内，

   以分子标记为基点将其锚定到染色体上
5. 遗传作图的标记物：

   5.1.基因标记：等位基因 allele

   5.2.DNA 标记：限制性片段长度多态性 RFLP

   简单序列长度多态性 SSLP[小卫星序列、微卫星序列]

   单核苷酸多态性 SNP
6. 遗传作图的不足：

   6.1.遗传图的分辨率有限（基因组规模子代数量）

   6.2.遗传图的覆盖较低(随机交换)

   6.3.遗传图分子标记有时会出现差错（随机取样）
7. 物理图：指标明一些界标（如：限制酶切位点 、基因等）在 DNA 上的位置，图距以物理长度为单位，例如染色体带区、核苷酸对数目等。
8. 物理作图方法：

   8.1.限制性酶切作图：

   限制性核酸内切酶；

   限制性位点长度问题（6bp 的随机概率=1/（4^6））

   8.2.基于克隆的基因组作图

   8.3.染色体细胞图

   8.4.**STS 作图**【大规模基因组物理图的主流技术】

   STS 是一已知的单一序列，根据选定的 STS 序列设计专一性引物，可对大量的

单个克隆进行 PCR 检测，能扩增出的均含有序列重叠的插入子。距离模型:最大似然法

人类基因组计划（ HGP ）的研究目标是，构建人的每条染色体的 STS 图，标记之间相距约 100kb 。获得一组组 DNA 片段的克隆，组内两两片段之间有共同的重叠序列；或是获得标记按正确次序排列、相互毗邻的片段，其连续长度超过 2000kb ，以便把染色体分段进行研究。

8.5.辐射杂交作图 X-ray breakage

9. 

基因组测序

1. 第一代 DNA 测序【sanger 测序技术】:
   1.1. 原理：以待测 DNA 为模板，使用带有标记的碱基类似物体外合成新链，可在任意一个碱基位置终止 =》凝胶电泳时形成彼此只差一个碱基的梯形条带 =》得到序列。
   1.2. 原理：链终止法（完成了人类基因组计划）、化学降解法。
   1.3. 凝胶电泳对于信号捕捉是存在缺陷的，且不高效 =》毛细管电泳、高效毛细管电泳
2. 第二代 DNA 测序:【高通量测序平台】
   2.1. 焦磷酸测序【光点测序】
   2.2. DNA 芯片测序
   2.3. 将二代测序技术按测序原理分类：
      2.3.1. 边合成边测序 SBS：454、illumina、HiSeq/MiSeq/NextSeq、Ion torrent/proton
      2.3.2. 边连接边测序 SBL：solid【缺点：读长短】

3. 第三代测序技术:【高通量测序平台】
   3.1. 原理：单分子测序 SMS
   3.2. 特点：
      3.2.1. 测序读长：平均测序读长达到 10 ~ 18 kb ，最长可超过 60 kb ；
      3.2.2. 准确度高：测序深度达到 30× 时，准确度达到 99.999% （ Q50 ）；
      3.2.3. 敏感性强：可以检测频率在 0.1% 的 Minor Variants ；
      3.2.4. 无 PCR 扩增偏好性：样本不需要进行 PCR 扩增，避免了覆盖度不均一以及 PCR Artifacts 的产生；
      3.2.5. 最小的 GC 偏好性（ GC bias ）：在极端高 GC 和极端低 GC 区域，可以轻松测 定，从而保证序列的均匀覆盖度；
      3.2.6. 可直接检测碱基修饰：利用测序过程聚合酶反应的动力学变化，首次实现在测序的同时对碱基修饰进行直接检测
   3.3. 技术路线：
      单分子实时测序技术 SMRT：读长超过 10kb，插入缺失错误率 1%【Sparc】
      纳米孔单分子技术【Sparc】：纳米孔单分子 DNA 电流阻遏、纳米孔单分子 DNA 碱基序列的电子阅读
4. 全基因组测序注意事项：
   4.1. 基因组测序的覆盖面 P0 = e-m【P0 ：丢失概率、m：为覆盖面（单倍体基因组数）】
      m= 1，P0 = 37% ，覆盖率 63%
      m=5， P0 = 0.67% ，覆盖率 99.33%
      m=10， P0 = 0.0045% ，覆盖率 99.9955%
   4.2. 物理间隙（Physical gap） 和 序列间隙（Sequence gap）
      4.2.1. 物理间隙：构建基因组文库时被丢失的 DNA 序列
      4.2.2. 序列间隙：测序时遗漏的序列，这个序列仍保留在尚未挑选到的克隆中
   4.3. 插入片段的两端测序：
      同一个载体的两段有两个引物，每个克隆读序只有 400bp，每个克隆内部不能进行连续的测序，因为缺少引物，所以>800bp 的片段中间就存在 gap
5. 序列读取模拟软件：ART、pIRS、PBSIM、Wessim、IgSimulato
6. ART_454:
   6.1. 单末端测序 SINGLE-END SIMULATION art_454
   6.2. 双末端测序 PAIRED-END SIMULATION art_454
   6.3. 扩增子测序 AMPLICON SEQUENCING SIMULATION art_454

PS:SAM 是一种序列比对格式标准， 由 sanger 制定，是以 TAB 为分割符的文本格式。
主要应用于测序序列 mapping 到基因组上的结果表示、表示任意的多重比对结果
SAM 分为两部分：注释信息部分（header section）、比对结果部分（alignment section）

7. GenomeABC: http://crdd.osdd.net/raghava/genomeabc/

基因组序列组装

1. 序列组装的基本理论
   1.1.流程：
      多个基因组测序的副本 -》碎片化（fragments） -》长度过小的过滤 -》BAC/YAC 双末端测序序列（带有 BAC 末端的序列，reads） -》通过 overlap 寻找重叠群区域（contigs） -》锚定在染色体上的重叠群（scaffold） -》草图序列（可

覆盖测序克隆片段 3-4 倍的 DAN 序列，含间隙/没有间隙，排列方向和位置未定）-》完成序列（错误碱基数<0.01%的 DNA 序列，排列方向确定，内部不含间隙，测序覆盖率在 8-10 个单倍体基因组）

Ps：根据确定 BAC 的排序方向以及重叠群（contigs）在支架 scaffold 中的排列方向；酵母人工染色体（*YAC*）、细菌人工染色体（*BAC*）

1.2. 问题：

测序错误

重复序列

多态性变异-》contigs

倒位 inversion

覆盖率

1.3. 组装类型：

- 从头组装（De novo） vs 基于参考的组装（ reference-based(mapping)）：

1.3.1. de novo 组装：

即使可以获得参考基因组，也应该进行从头组装，因为它可以从基因组组装中丢失的基因组片段中恢复转录的转录本。

1.3.2. mapping 组装：

对现有的主干序列进行读取，构建一个类似的序列，但不一定与主干序列相同。转录组数据主要通过对参考基因组的 mapping 进行分析

1.3.3. de novo:

即使可以获得参考基因组，也应该进行从头组装，因为它可以从基因组组装中丢失的基因组片段中恢复转录的转录本。

1.3.4. mapping 缺点：

无法解释 mRNA 转录本结构改变的原因,如可变剪接。由于基因组包含可能存在于转录本的所有内含子和外显子，因此在基因组中不连续排列的剪接变体可能被折现为实际的蛋白质亚型。

- 基因组组装 vs 转录组组装：

1.3.5. 基因组：基因组序列覆盖水平可以根据 non-codingDNA 内含子区域的重复序列随便改变；

这些重复序列也会造成基因组组装重叠群 contigs 组成的错误;

1.3.6. 转录组：转录组的覆盖水平可以表示为基因表达水平；

转录组装中的重叠群contigs区域可能是剪接的异常或者基因家族成员之间的差异

1.3.7. 基因组组装软件不能被用于转录组组装

一个基因组的基因组测序深度通常相同，但转录的深度不同；

两条链在基因组测序都是按顺序排列的，但 RNA-seq 可以是特异的；

来自相同基因的转录变异体（transcript variants）可以共享外显子，难以处理

2. 序列组装的软件：

TIGR 组装

Minimus 组装：

Minimus2 组装：

Minimo

Newbler

BioPerl -》 Module:Bio::Assembly::IO

AllPathsLG -》 DISCOVAE de novo

Velvet https://www.ebi.ac.uk/~zerbino/velvet/

SOAPdenovo http://soap.genomics.org.cn/soapdenovo.html

SOAPdenovo ：SOAPdenovo2（基因组）、SOAPdenovo-Trans（转录组）

Trinity（转录组）

CAP3

应用于转录组的序列组装：

SeqMan NGen

SOAPdenovo-Trans

Velvet（基因组很小的 reads）-》Oases（应用于转录组）

Trans-ABySS

Trinity

3. CAP3 小规模序列组装 http://doua.prabi.fr/software/cap3

使用 genebank 的 unigene 数据库中搜索的某个基因的 EST 序列进行组装

4. 获得某个基因的 EST 序列：

NCBI-UniGene -》 某基因 -》下面有 mRNA 序列和 EST 序列（fasta）

5. 序列组装算法：

5.1.从头组装（De novo）

贪婪图算法：OLC、DBG

给定一组 reads ，从中挑选一个 read 作为“种子”【规则】，用与它两端中的一段有足够数量的碱基序列相同的 read 来扩展；迭代进行，直到不可继续 扩展。再选择其他未参与拼接的 reads 序列拼 接，重复上述过程，直到所有 read 被 拼接完成。

计算多有 fragments 的成对对齐程度

-》选择最大重叠的两个 fragments

-》合并选择的 fragments

【不断迭代，中间的重叠区就是 contig】

5.1.1.OLC 方法【Sanger-data 组装】Hamilton path：

1.把每个 DNA 片段 (reads) 看成一个节点 ；

2.如果两个 DNA 片段之间存在重叠，就在相应的节点之间建立 一条边 ；

3.所有 DNA 片段通过这种重叠关联，构造出一个有向图 ；

4.通过寻找图中经过每个节点一次且仅一次的一条路径 （ Hamilton 路径） ；

5.即可获得目标 DNA 序列。

1.对参与拼接的 reads 进行比对，分析它们之间的重叠信息（ overlap ） ；

2.把存在重叠的 reads 进行组合，形成拼接结果 contigs （ layout ） ；

3.对 contigs 形成的图上的 reads 进行排列，通过在图中寻找 Hamilton 路径来确定最终序列（ consensus ）。

应用的软件 Celera Assembler, Arachne, CAP，PCAP，TIGR， PHRAP……

5.1.2.DBG 方法 Eularian path：

1. 对给定的 reads ，按照长度 k 进行连续划分（步长 =1 ），得到若干等长度段序列(k-mer)。一个长度为 l 的 read ，

将被分成 lk+1 个 k-mer

2. 对于任意两个 k-mers ： k1 和 k2 ，如果 k1 的后 k-1 个碱基序列与 k2 的前 k-1 个碱基序列相同，则建立一条从 k1 指向 k2 的有向边。通过以上两步即可构建出一个 de Bruijn 图，拼接结果序列可以通 过在图中寻找 Eulerian path 获得。

3. 第一个 k-mer 的序列全部读出，后面的每个 k-mer 只读取 最后一个碱基

5.1.3.DBG 方法的问题：

Q1.测序错误：tip 结构和 bubble 结构

错误的 reads-》错误的 k-mer 节点-》deBruijin 图大且复杂-》降低组装效率

Q2.测序 gap：

基因组覆盖不全-》k-mer 信息不全-》deBruijin 图连通性降低-》产生 dead-end 路径-》k-mer 越长问题越严重

Q3.分支问题：

数据错误/重复序列-》deBruijin 图出现分支-》无法处理

=》Solution：设定过滤标准（k-mer 出现次数）-》过滤掉可能出错的 reads；
直接删除 dead-end 路径-》可能导致 gap 问题；

5.1.4.对比 Hamilton path 和 Eularian path：

当短的 reads 数量很多时，Hamilton 图基于 reads，巨大复杂-》时间复杂度高-》空间复杂度低；Eularian 图基于 k-mer，不受影响-》时间复杂低-》空间复杂度高

5.2.基于参考的 mapping 组装

6. 大规模序列组装软件：

6.1.AllPaths-LG:短 reads

给定一个参考基因组，pipeline 能在基因组组装的不同阶段对组装过程 和结果进行评估

BASIC:基础评估，不需要参考基因组；

frag_size:小片段文库插入片段长度的均值；

frag_stddev:小片段文库的插入片段长度估算的标准偏差；

insert_size:大片段文库插入片段长度的均值；

insert_stddev: 大片段文库插入片段长度估算的标准偏差；

read_orientation:reads 的方向，小片段文库为 inward，大片段文库为 outward；

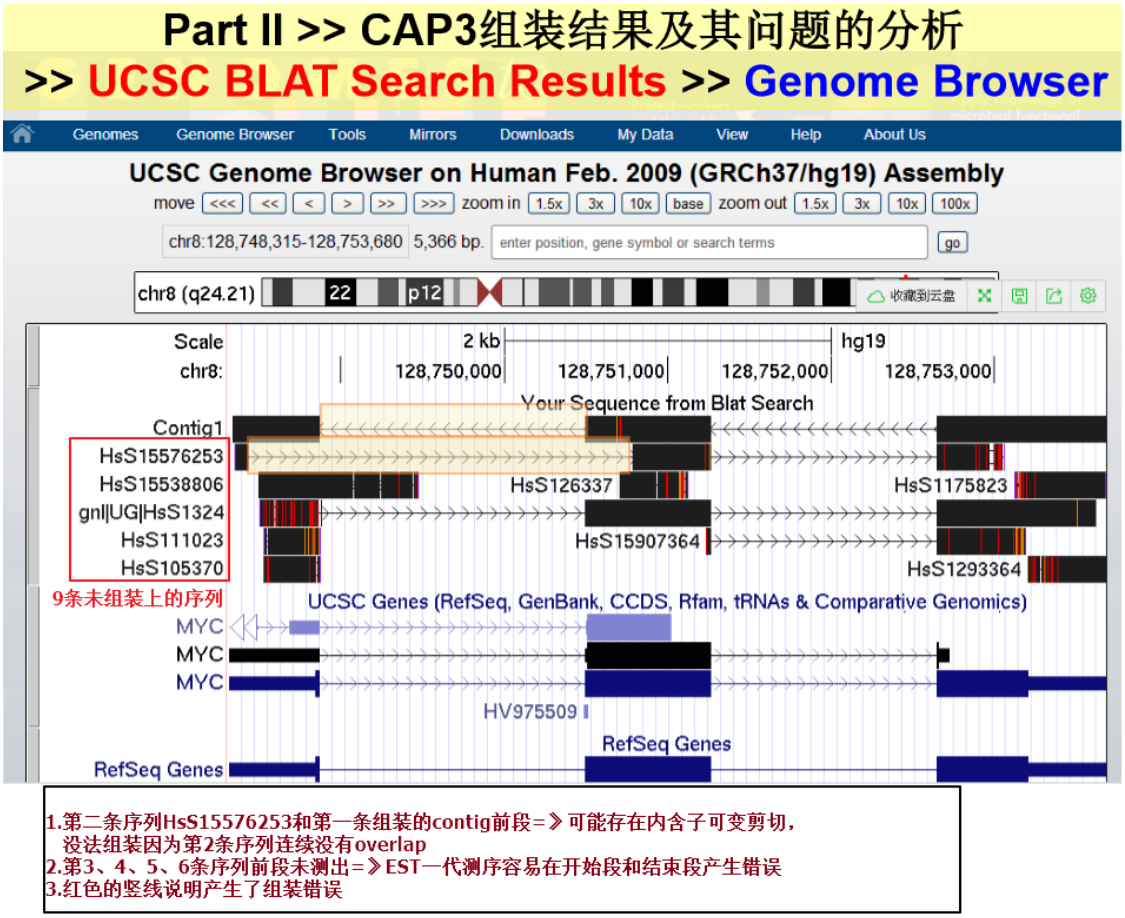
genomic_start:reads 从该位置开始，读入数据，如果不为 0，之前的碱基都被剪掉；

genomic_end:reads 从该位置开始，停止读入数据，如果不为 0，之后的碱基都被剪掉。

6.2.Velvet + SOAPdenovo

7. CAP3 序列组装结果分析：

UCSC -》 Blat -》提交 cap3 生成的 contig 结果 -》browser



1.第二条序列HsS15576253和第一条组装的contig前段=》可能存在内含子可变剪切，
设法组装因为第2条序列连续没有overlap
2.第3、4、5、6条序列前段未测出=》EST一代测序容易在开始段和结束段产生错误
3.红色的竖线说明产生了组装错误

第二条=》可能是转录突变体
第三条前段=》可能有污染

8. 序列比对基本原则：

相似性、同源性

8.1. 相似性：是指序列比对过程中， 用来描述检测序列和目标序列之间，相同 DNA 碱基或氨基酸 残基顺序所占比例的高低。 《= 统计

8.2. 同源性：是指从某一共同祖 先经趋异进化而形成的不同序列。 《=进化

8.3. 当相似程度高于 50%时，比较容易推测检测序列和目标 序列可能是同源序列；

当相似性程度低于 20%时，就难以确定或者根本无法 确定其是否具有同源性；

无论相似程度有多高或多低，都不能 100%确保两个序列 之间一定同源，它只能作为推测的依据之一。

8.4.

8.5. 在蛋白质家族或超家族中经常存在，不同的蛋白质序列之间只有局部区域存在高度相似性（保守性 Motif），而这些局部区域在整个序列中所占的比例有时很低！

序列比对过程中需要在检测序列或目标序列中引入空位， 以表示插入(Insertions)或删除(Deletions) [Indel]

序列比对的数学模型大体可以分为两类： 一类从全长序列出发，考虑序列的整体相似性，即整体比对/全局比对； 第二类考虑序列部分区域的相似性，即局部比对。

局部相似性比对的生物学基础，是蛋白质功能位点 往往是由较短的序 列片段组成的，这些部位的序列具有相当大的保守性，尽管在序列的 其它部位可能有插入、删除或突变。此时，局部相似性比对往往比整 体比对具有更高的灵敏度，其结果更具 生物学意义

8.6.序列比对时的打分模型：

突变数据矩阵 MD、模块替换矩阵 BLOSUM

8.6.1.MD:

建立在已知的同源蛋白质/蛋白质家庭 的多序列比对的基础之上的;

统计某位点出现各种氨基酸的比例(%) [Pj, j=1 To 20，对应 20 种 Aa，伪随机概率;

该位点一个氨基酸发生改变，改变成另一个 Aa 的概率 P1,2 = Pj1/Pj2;

可接受点突变(Point Accepted Mutation,PAM) 1 个 PAM 的进化距离表示 100 个残基中发生一个残基突变的概率;

PAM 是在蛋白质高度相似的基础上选择的。蛋白质对齐要求显示至少是 85%

8.6.2.BLOSUM:

以序列片段为基础;基于蛋白质模块数据库 BLOCKS;从蛋白质模块数据库 BLOCKS 中找出一组替换矩阵，用于解决序列的远距离相关。

8.6.3.PAM 和 BLOSUM 换算：

PAM 后的数字——越大 越适用于序列相似性低的序列之间的比对

PAM 后的数字——越小 越适用于序列相似性高的序列之间的比对

Blosum 正好相反

PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM60
PAM200	BLOSUM52
PAM250	BLOSUM45

PAM	BLOSUM
To compare closely related sequences, PAM matrices with lower numbers are created.	To compare closely related sequences, BLOSUM matrices with higher numbers are created.
To compare distantly related proteins, PAM matrices with high numbers are created.	To compare distantly related proteins, BLOSUM matrices with low numbers are created.

PAM	BLOSUM
Based on global alignments of closely related proteins.	Based on local alignments.
PAM1 is the matrix calculated from comparisons of sequences with no more than 15% divergence but corresponds to 99% sequence identity.	BLOSUM 62 is a matrix calculated from comparisons of sequences with a pairwise identity of no more than 62%.
Other PAM matrices are extrapolated from PAM1.	Based on observed alignments; they are not extrapolated from comparisons of closely related proteins.
Higher numbers in matrices naming scheme denote larger evolutionary distance.	Larger numbers in matrices naming scheme denote higher sequence similarity and therefore smaller evolutionary distance.

空位开放罚分

空位延伸罚分

v(g) = - d - (g-1)e

【g 代表连续空位数 d 代表空位开放罚分 e 代表空位延伸罚分】

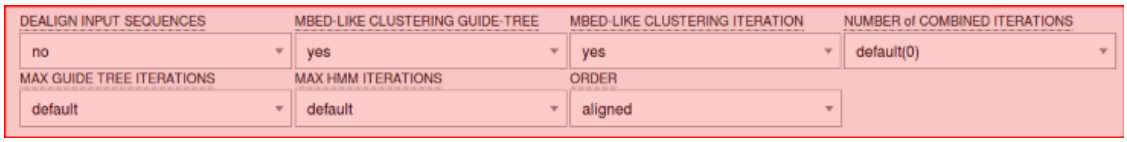
8.7.序列比对原则：动态规划算法(最长共同子序列)、启发式搜索算法（最优方案、完整性、准确度精密度、执行时间）

启发式算法应用：神经网络、自学习。

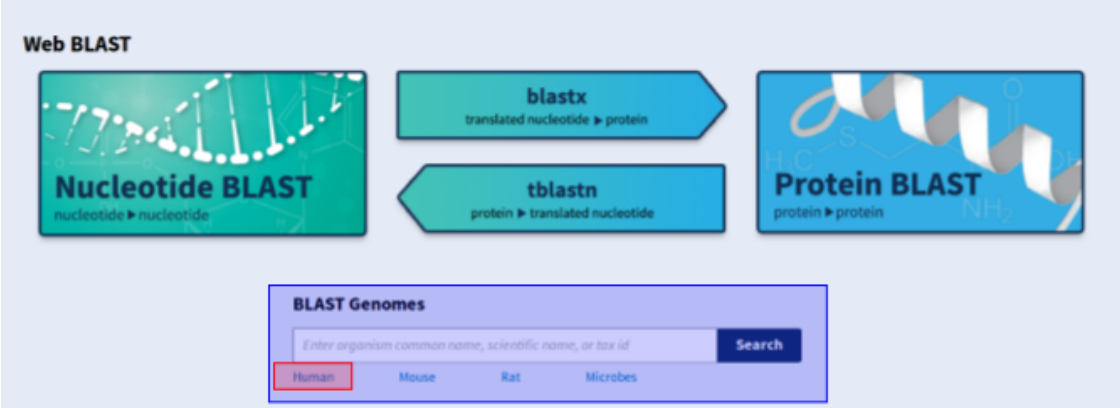
启发式算法应用在序列比对 ：等长的高分片段对-》通过延伸或连接优化结果-》运用动态规划算法引入空位

8.8.回溯：从最高的 分数开始 ，进行 traceback 直到碰到 0 为止。

8.9.序列比对软件：EBI ClustalW2【动态规划算法 >> 全局联配工具】

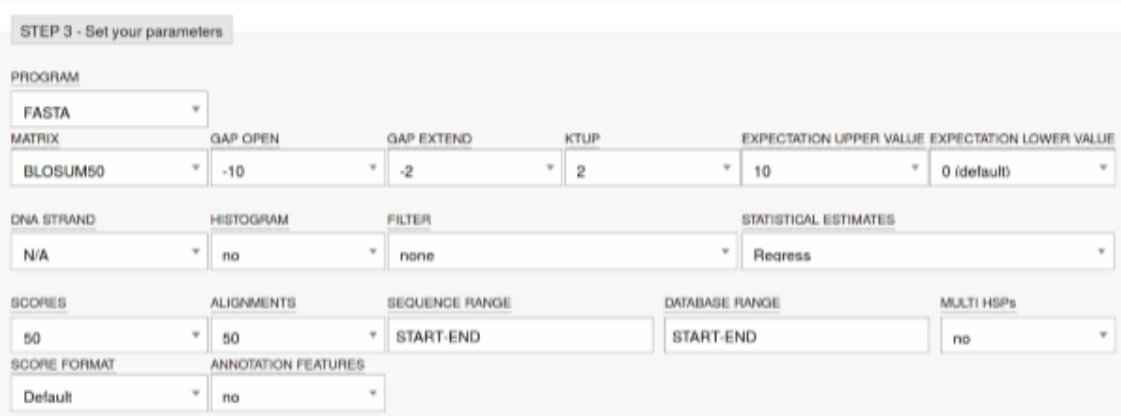


8.10.BLAST 和指定的物种进行序列比对：【启发式搜索算法（局部打分策略）:】



8.11. EBI >> FASTA【启发式搜索算法（局部打分策略）:】

https://www.ebi.ac.uk/Tools/sss/fasta/

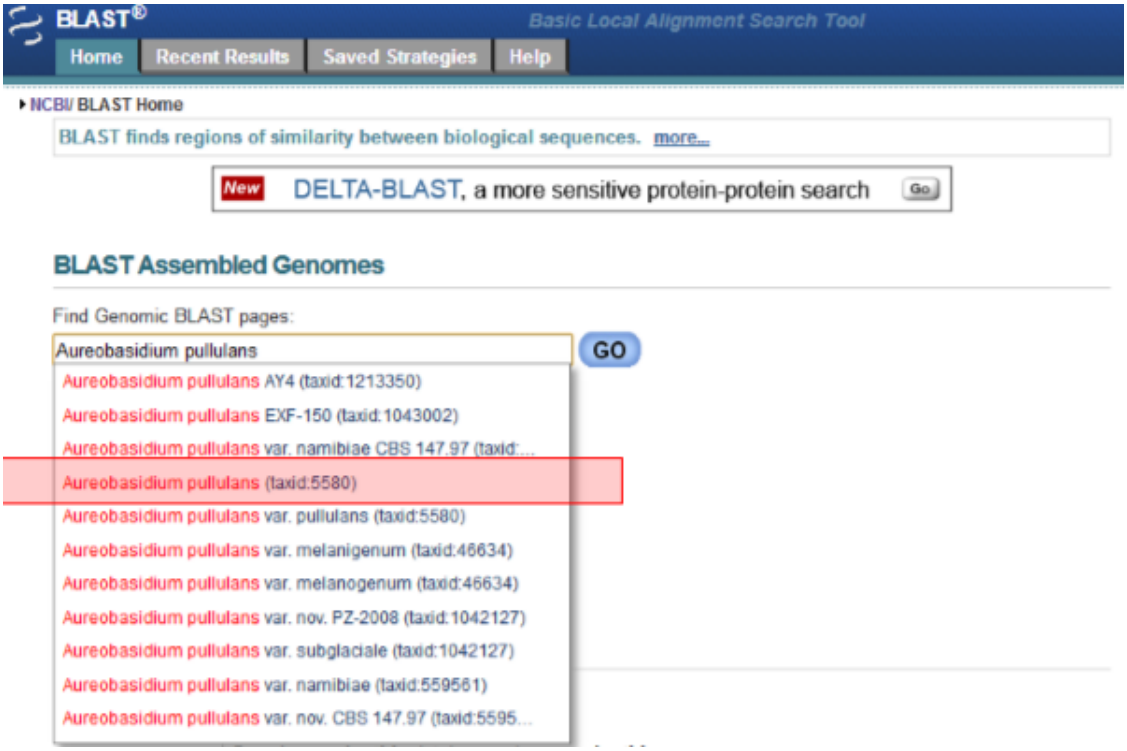


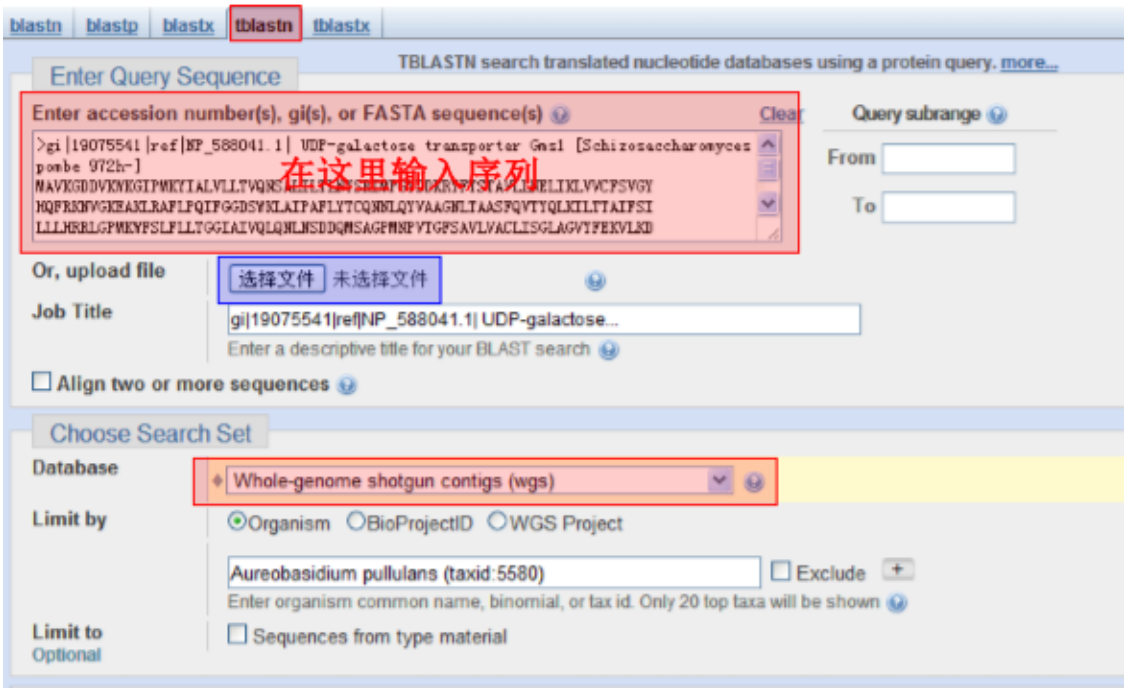
8.12. 二代测序 NGS 比对（4）

BWA、Bowtie、Bowtie2(=》illumina,solid)、samtools、

新基因发现与基因结构建模

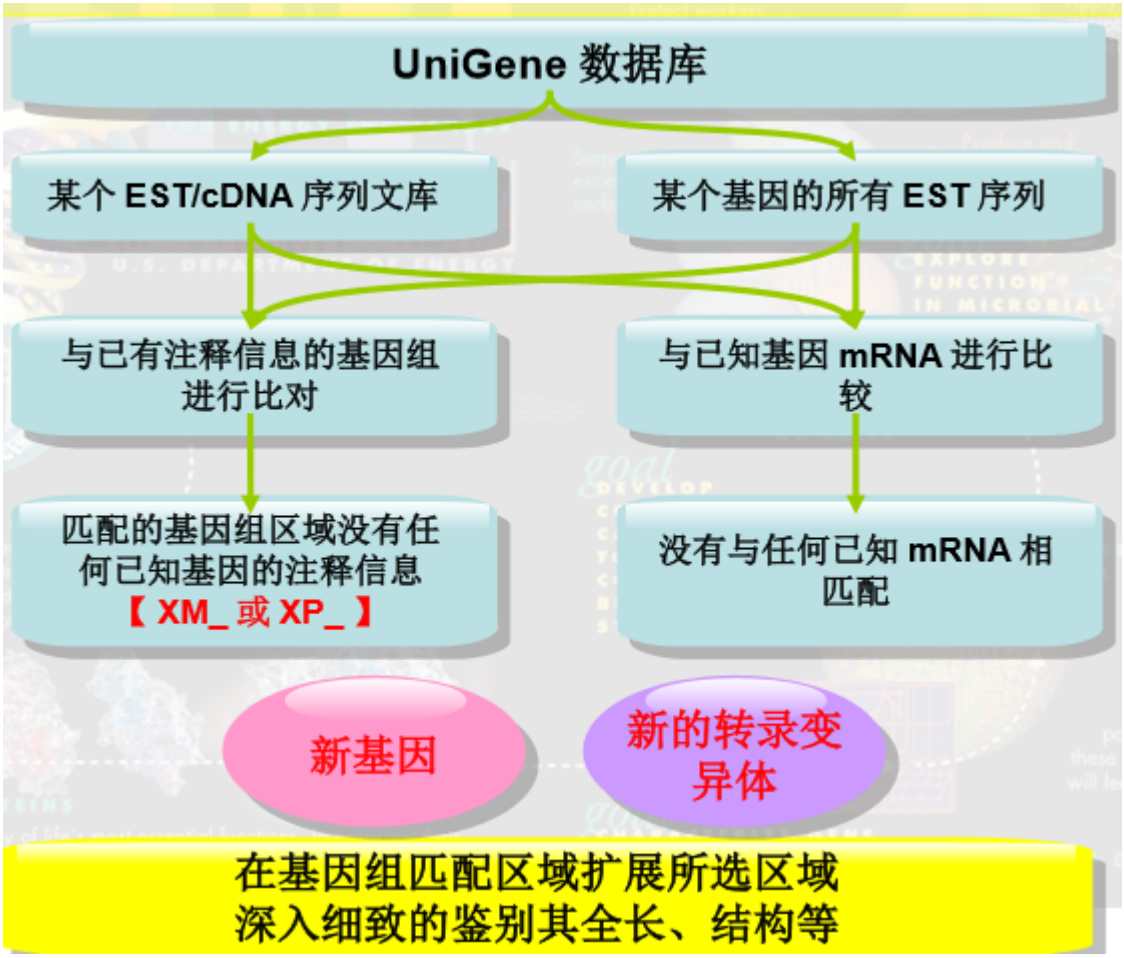
1. 新基因范畴：
   1.1. 该物种没有报道 =》 同源基因搜索
   1.2. 所有物种都没有报道 =》EST/cDNA 序列文库、RNA-seq、从头计算鉴别新基因
2. AP 为例 blast：s



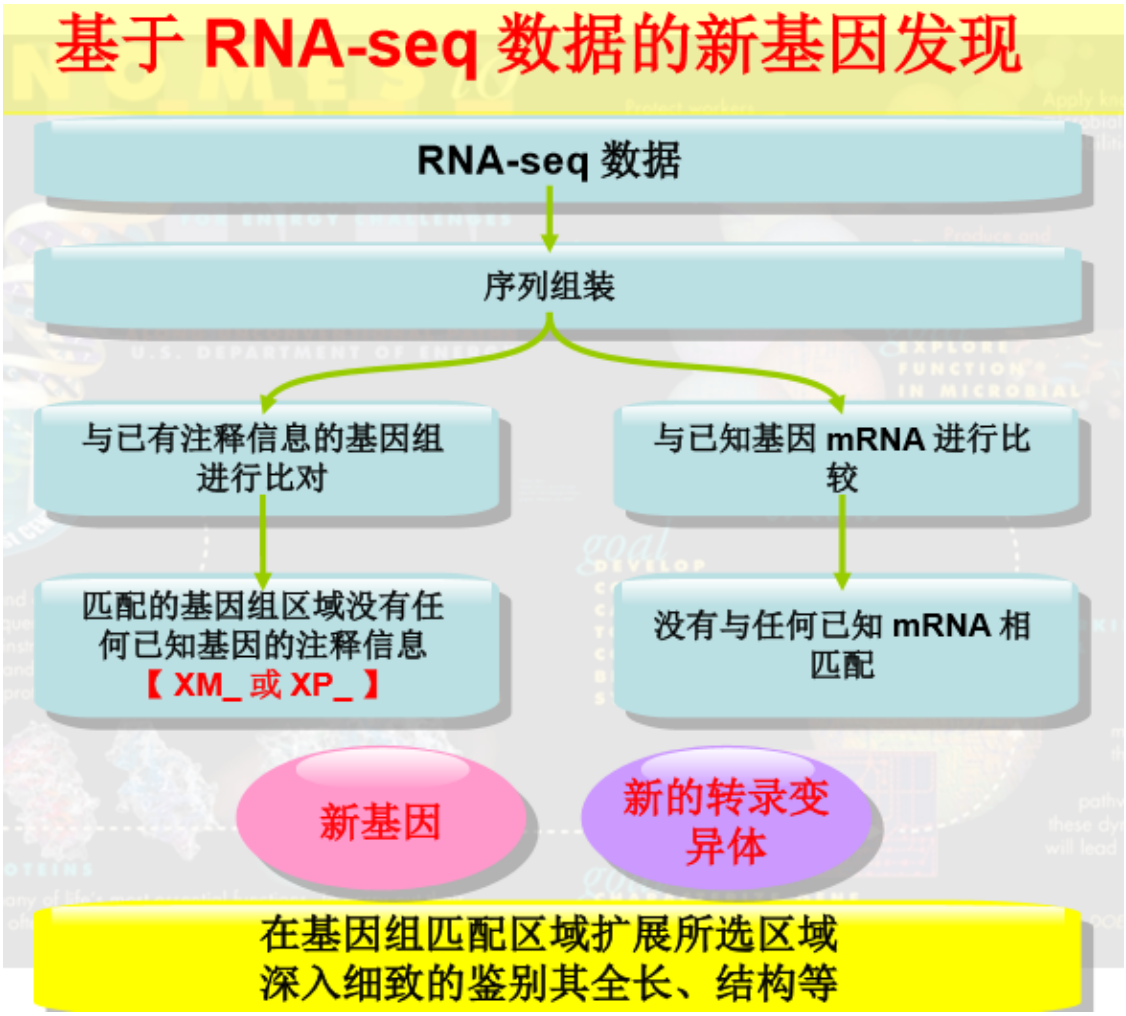


在结果中选择高相似区域，扩展上下游，各延伸 1kb

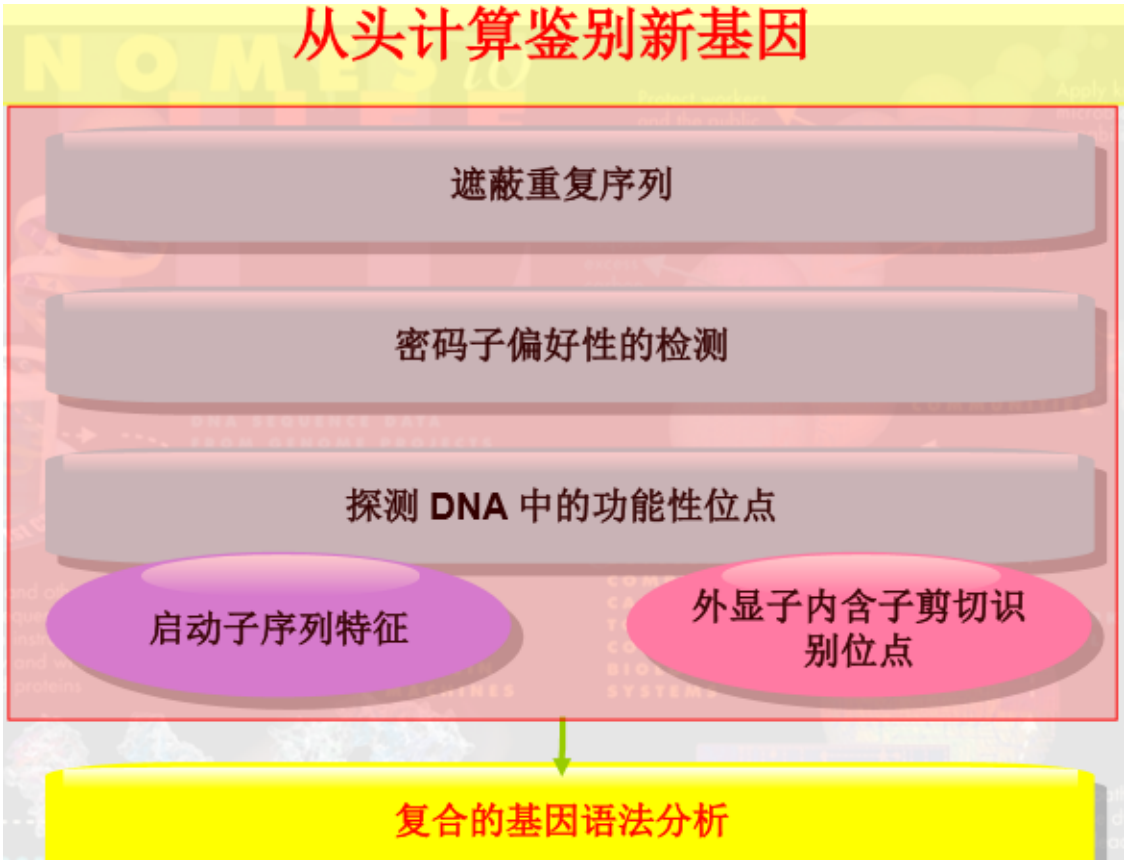
3. 基于 EST/cDNA 序列文库的新基因发现：



4. 基于 RNA-seq 数据的新基因发现：


基于 RNA-seq 数据的新基因发现
RNA-seq 数据
序列组装
与已有注释信息的基因组进行比对
与已知基因 mRNA 进行比较
匹配的基因组区域没有任何已知基因的注释信息【XM_ 或 XP_】
没有与任何已知 mRNA 相匹配
新基因
新的转录变异体
在基因组匹配区域扩展所选区域
深入细致的鉴别其全长、结构等

5. 从头计算鉴别新基因：



从头计算鉴别新基因 >> 相关软件

Software /WebServer	applicability	link
RepeatMasker	various familiar/model species	NHGRI
GENSCAN	Vertebrate/Arabidopsis/Maize	MIT
tRNAscan-SE	any	lowelab Eddy lab bioweb

6. 基因结构建模：

   NCBI Splign

   UCSC BLAT

   EMBL-EBI GeneWise

NCBI >> Sequence Analysis >> Splign >> Results

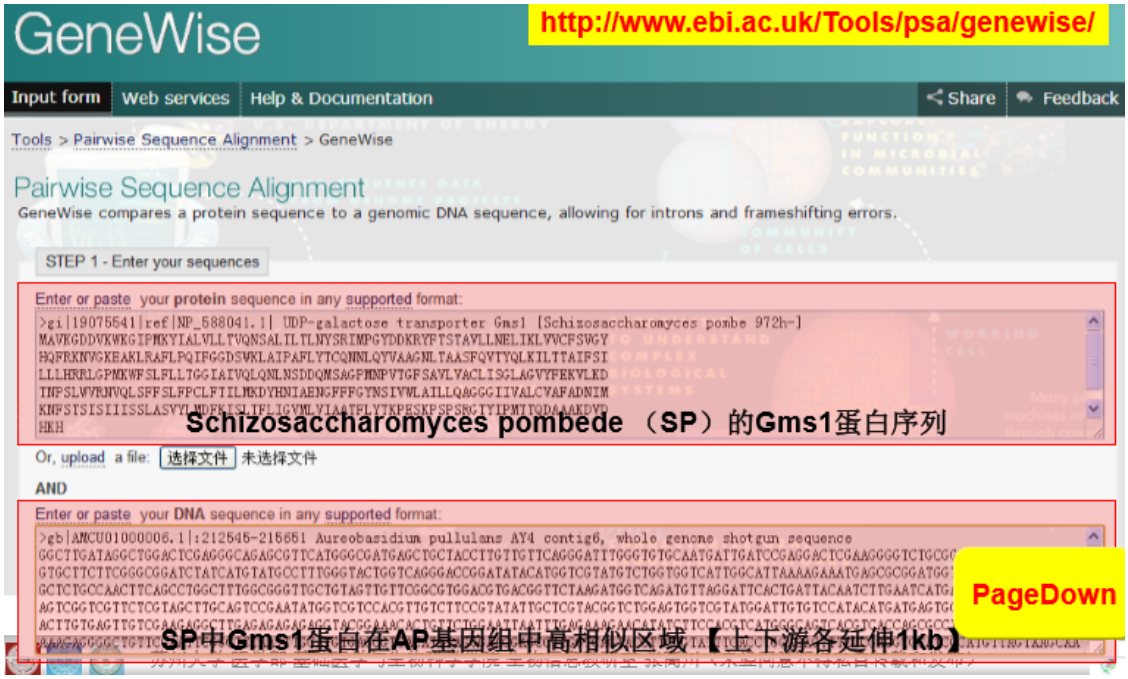
匹配的染色体编号、正负链、基因组区间以及跨越长度

BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	NP_035770.2	592	77	390	390	84.3%	17	+-	7572930	7579449	6520
browser details	NP_035770.2	99	86	255	390	77.8%	1	++	3639926	3643780	3855

一致残基的比例

相似序列得分



7. 对比各个基因结构建模软件：

软件	目标序列	参照序列	物种来源	新基因发现	已知基因的新转录变异体鉴别	基因结构建模效果	
NCBI Blast系列	基因组核酸序列	蛋白质、EST/cDNA、mRNA	亲缘关系越近越好	√	√	较差	
UCSC BLAT				√	√	较差	
NCBI Splign		EST/cDNA、mRNA				一般	只能输入一个目标信息和一个基因组信息，只可以基因结构建模
EMBL-EBI GeneWise		蛋白质				较好	亲缘关系越近，才易得出结果
GENSCAN		从头计算				较好	

8. 原核生物从头预测基因：

8.1. 原核生物的启动子区域信号：

Pribnow box

转录因子结合位点

8.2. ORF 区

有成百上千的碱基对长

终止密码子

9. 真核生物从头预测基因

9.1. 真核生物的启动子区域信号：

CpG 岛

poly(A)尾巴的结合位点

9.2. 外显子和内含子的剪接机制：

剪接位点

外显子

10. 密码子偏好：

RNA 二级结构、转录/基因表达、翻译延伸速度、蛋白质折叠

JAVA Codon Adaptation Tool (JCat) http://www.jcat.de/

Codon Optimization Tool http://sg.idtdna.com/CodonOpt

11. 原核生物从头预测基因：
    GLIMMER
    GeneMark GeneMarkS GeneMarkS+
12. 真核生物从头预测基因：
    GENSCAN
    Geneid 【HMM】
    Augustus 【HMM】
    GeneMark-ES GeneMark-ET
    GlimmerHMM 【HMM】
    mSplicer
    CONTRAST
    mGen
    FGENESH 【HMM】

数据库

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