第 4 章　核酸序列分析

课程标题页

4.1　课前甜点

5, -00:05--00:00 (^1.1-)

播放《飞越五千年 -兵圣》中“孙武练兵”的片段。“约束不明，申令不熟，将之罪也；既已明而不如法者，吏士之罪也。”

给大家看这个故事的目的是要大家明白：一是“国有国法，家有家规”，三百六十行都有自己的的职业操守；二是“无规矩不成方圆”，有规矩就要遵守，这样才能把事情做好。

明确课堂纪律，收集学生反馈

在正式上课之前先明确一下作为一个学生应该遵守的课堂纪律：

课堂纪律条目

- 只有正式上课前的请假有效。
- 提前 5 分钟到教室，严禁迟到。
- 上课期间手机关机或调成震动。
- 上课期间离开教室先举手示意。
- 课上有疑问的话先举手后提问。
- 上课期间严禁交头接耳，大声喧哗。
- 随机点名，缺勤扣分如下：1、3、6。
- 缺勤三次或三次以上者，平时成绩为 0。

对上述规范有异议的话，可以以不记名的方式把自己的想法写在纸上，课间或课后交给我。如果对授课内容、方式等有何建议，可以通过各种方式反馈给我。下面正式开始上课。

4.2　自我介绍

1, 00:00--00:01 (-1.1-)

因为是首次上课，我先用一分钟的时间做一下自我介绍。

自我介绍与邮箱网盘

伊现富（Yi Xianfu），1986 年生人，本科毕业于山东大学生命科学学院，学的是生物科学专业；之后保送到中国科学院上海生命科学研究院读研究生，主要从事人类复杂疾病相关的生物信息学研究。

我常用的邮箱有两个：yixfbio@gmail.com，主要用于工作中的科研业务交流；yixf1986@gmail.com，主要用于生活中的闲杂琐事交流。联系电话：15620610763。个人博客：http://yixf.name，内容五花八门，其中有和生物信息学相关的一些资料，感兴趣的可以去看看。网络

上的昵称以“yixf”为主。

密码复杂度，密码管理软件

为了方便进行信息交流与资源共享，我注册了一个 126 的邮箱，账号：bioinfo_TIJMU@126.com，密码：C&563f&nzx!s；申请了一个百度云网盘，账号：bioinfo_TIJMU@126.com，密码：566&Us3Rp6#C。授课讲义、幻灯片、视频等资料都会存储在网盘中，有需要的自行登录下载。

提问：1. 生物信息学的英文读法与拼写。2. 绪论中提到的一种高通量技术。（芯片 ⇒ 第二代测序（Next Generation Sequencing，NGS）技术、单分子实时 (Single Molecule Real Time，SMRT)DNA 测序技术）

4.3　引言

5, 00:01--00:05 (-1.1-)

生物大千世界
中心法则示意图

“龙生龙，凤生凤，老鼠的儿子会打洞！”“种瓜得瓜，种豆得豆。”这些都是大家耳熟能详的谚语。不管是天上飞的、地上跑的、水里游的，还是能动的、不能动的，它们的后代都和它们非常相像，但却也会有少许的差异。这些现象大家都已司空见惯，所以可能没有啥感觉。但仔细想想，你就会发现大自然的奇妙所在。当然，对于生物专业的人来说，这个就没什么奇怪的了，因为我们都知道分子生物学的中心法则（The central dogma of molecular biology）：DNA 转录成 RNA，RNA 翻译成蛋白质。蛋白质执行特定的生物功能从而决定最终的表型，而 DNA 则携带着最原始的决定个体性状的遗传信息，RNA 主要参与遗传信息的表达和调控。在各种生物中，ACGT 都是构成 DNA 和 RNA 核酸序列的基本组分。仅仅这么四种碱基怎么可能构建出缤纷多彩的大千世界呢？其秘诀就在于四种核苷酸的排列顺序。就像搭积木一样，通过不同的排列组合我们可以构建出不同的形状。核酸序列中不同的碱基排列顺序，蕴含着不同的生物信息，包括遗传信息和进化信息等。如何从海量的核酸序列中挖掘信息，并将其与生物性状联系起来，这是生物信息学的主要研究领域之一，也是提取分子生物信息的首要步骤。

01 世界与 ACGT 世界：
序列示例
Matrix 图片

在接下来的四次课中，我将介绍 DNA 和 RNA 序列分析、基因组结构及功能注释的相关内容；而蛋白质序列与结构的内容将由耿鑫和张涛老师分别进行介绍。

基因及调控元件结构示意图

DNA 是主要的遗传物质，是携带遗传信息的载体之一。DNA 序列是指 DNA 的一级结构，基本组份由 ACGT 四种碱基组成，因此又称为碱基序列。不同种属的 DNA 碱基组份存在差异。DNA 主要携带两类遗传信息。一类信息储存在具有功能活性的 DNA 序列中，能够通过转录过程形成 RNA（主要有编码 RNA 和非编码 RNA 两种形式），其中编码 RNA 含有编码蛋白质的氨基酸序列信息，这类 DNA 序列主要是指遗传的基本单位即功能序列。一类信息属于调控信息，主要存在于特定 DNA 的区域，能被各种功能性蛋白分子特异地识别结合，进而完成各种生物过程，例如启动子和增强子调控基因的表达。遗传信息储存于具有特征信息的 DNA 序列中，根据这些特征信息设计不同的算法并开发相应的分析工具，能够在海量的序列数据中挖掘出具有生物学功能的特征信息。

本堂课将介绍 DNA 序列一级结构的基本信息和序列的特征信息分析方法。DNA 基本信息中主要包括序列碱基组份（base composition）分析、序列转换、限制性内切酶位点分析；序列的特征信息主要包括开放阅读框（Open Reading Frame，ORF）、启动子及转录因子结合位点的分析和 CpG 岛（CpG island）的识别。

4.4　DNA 序列转换与组份分析

25, 00:05--00:30 (-1.1-)

美籍奥地利犹太生物学家 Erwin Chargaff 在分析组成 DNA 的碱基时，首先注意到了 DNA 碱基组成的某些规律性。他在 1950 年发现

DNA 中的腺嘌呤与胸腺嘧啶数量几乎完全一样，鸟嘌呤与胞嘧啶的数量也是一样。这项发现后来成为查戈夫第一法则。查戈夫的研究帮助克里克及沃森推断出 DNA 的双螺旋形结构。第二法则则表示不同物种之间的 DNA 组合是不同的，特别是 A、G、T 及 C 之间的相对数量。这两条法则后来都被称为查戈夫法则：

查戈夫法则

- 腺嘌呤和胸腺嘧啶的摩尔数相等，即 A=T；鸟嘌呤和胞嘧啶的摩尔数也相等，即 G=C。由此可推导出含氨基的碱基（腺嘌呤和胞嘧啶）总数等于含酮基的碱基（鸟嘌呤和胸腺嘧啶）总数，即 A+C=T+G；嘌呤的总数等于嘧啶的总数，即 A+G=C+T。
- 不同生物种属的 DNA 碱基组成不同，即 AT/GC 的比值因生物种类不同而异。

GC 含量简介

GC 含量（GC content）是在所研究的对象的全基因组中，鸟嘌呤（Guanine）和胞嘧啶（Cytosine）所占的比例。一种生物的基因组或特定 DNA、RNA 片段有特定的 GC 含量。在 DNA 链中 G 和 C 是以三个氢键相连，而 T 和 A 则是两个氢键相连的。氢键的多少体现连接的能量，氢键多的不容易被打断。因此 GC 含量高的 DNA 比 GC 含量低的 DNA 更加稳定。在双链 DNA 中，腺嘌呤与胸腺嘧啶（A/T）之比，以及鸟嘌呤与胞嘧啶（G/C）之比都是 1。但是，（A+T）/（G+C）之比则随 DNA 的种类不同而异。

GC 含量计算公式

GC 含量通常以百分数的形式进行表示，计算公式如下：$\frac{G+C}{A+T+G+C} \times 100$。但有时也以比值的形式表示，叫做 GC 比（GC-ratio）。AT/GC 比值的计算公式如下：$\frac{A+T}{G+C}$。

物种 GC 含量比较
不同区域 GC 含量比较
基因与基因组 GC 含量比较

原核生物中不同种属的 GC 含量从 25% 到 75% 不等，这种组分差异可用于识别细菌种类。（恶性疟原虫的 GC 含量仅有 20% 左右，这种情况下一半说它富含 AT（AT-rich），而不会说缺少 GC（GC-poor）。）真核生物物种间 GC 含量的差异不如原核生物明显，但真核基因组中不同区域的 GC 含量存在差异，基因区的 GC 含量比整个基因组背景上的 GC 含量要高。GC 含量与密码子使用偏性、DNA 双链的溶解温度有关，是进行核酸杂交的重要参数。

DNA 序列具有双链性与双链互补性，因此进行序列分析时，经常需要针对 DNA 序列进行各种转换，例如：反向序列、互补序列、反向互补序列、显示 DNA 双链、转换为 RNA 序列等。其中最常用的是获取反向互补序列。

提问：问什么常说“反向互补”，而不是“反向”、“互补”呢？

序列书写惯例

序列的书写惯例：

- DNA/RNA：[左] 5' $\Longrightarrow$ 3' [右]
- 多肽/蛋白质：[左] N 端（氨基端）$\Longrightarrow$ C 端（羧基端）[右]

短序列互补实例（板书）

教材中演示了使用 EditSeq 获取 DNA 反向互补序列的操作。本质上就是 A 转换成 T、C 转换成 G、G 转换成 C、T 转换成 A，获得互补序列，然后将互补序列反向输出得到最终的反向互补序列。当然，先反向后互补结果是一样的的。

而对于序列碱基组成和 GC 含量的分析，无非就是进行简单的计数与运算。根据具体任务的不同，采用的策略和工具也会有所不同。

提问：你接到一个进行碱基组成和 GC 含量的分析的任务，该如何入手解决？

实例与策略（板书）

首先分析任务的属性：序列长短、序列数目多少、处理同样任务的频率，然后再决定采用那种策略：

- 序列短、数目少：一个一个查也用不了多少时间
- 序列长、数目少：另辟蹊径，办法绝不止一种（如：巧用 word 替换功能进行计数）
- 序列数目多：找现成工具（他山之石，可以攻玉）；请别人帮忙（术业有专攻）；自己写程序（自己动手，丰衣足食）；……

There's More Than One Way To Do It.（不只一种方法来做一件事。）

Don't Reinvent the Wheel.（不要重复发明轮子）

10, 00:30--00:40 (-1.1-)

此处仅给大家提供一个解决问题的思路：遇到问题，先思考后动手；办法总会有的，而且不止一种；没有完美的方法，只有合适的方法。

提问：[课后思考] 如何获取 DNA 序列的反向、互补及反向互补序列？

4.5　限制性核酸内切酶位点分析

限制酶（restriction enzyme）又称限制内切酶或限制性内切酶（restriction endonuclease），全称限制性核酸内切酶，是可以识别 DNA 的特异序列、并在识别位点或其周围切割双链 DNA 的一类内切酶。限制酶可以将异源性 DNA 切断并使之失活，限制异源 DNA 的侵入。但对自身 DNA 无损伤作用，从而维持细胞原有遗传信息的完整性。限制酶的切割形式有两种，分别是可产生具有突出单股 DNA 的黏状末端，以及末端平整无凸起的平滑末端。染色体或 DNA 上断开的不同限制片段可由 DNA 连接酶黏合，因此限制酶在分子生物学与遗传工程领域有着广泛的应用。

限制酶定义及别名

提问：[课后思考] 限制酶如何实现只切割外源 DNA 而不损伤内源 DNA？

提问：*Eco*RI 的名称是如何确定的？

*Eco*RI 的命名示例

限制酶的命名根据细菌种类而定，一般由微生物属名的第一个字母和种名的前两个字母组成，第四个字母表示菌株（品系）。在同一品系细菌中得到的识别不同碱基顺序的几种不同特异性的酶，可以编成不同的号。以 *Eco*RI 为例：*E* 源于属名 *Escherichia*，*co* 源于种名 *coli*，R 代表 RY13 品系，I 表明在此类细菌中发现的顺序——首次发现。

限制酶分为 I、II、III 型三大类。其中，II 型限制酶能识别专一的、短的 DNA 序列，并在识别位点或附近切割双链 DNA。这类限制酶具有专一的识别和切割位点，是基因工程中实用性较高的限制酶种类。限制酶识别的序列长度一般为 4-8 个碱基，常见的是 6 个碱基，且多数为回文对称结构；切割的序列通常就是其识别的序列，切割位点在 DNA 两条链相对称的位置。切割位点在回文的一侧时，可形成黏性末端，如：*Eco*RI、*Bam*HI、*Hin*dIII 等；另一些酶如 *Alu*I、*Sma*I 等，切割位点在回文序列的中间，形成平滑末端。

回文结构

提问：什么是回文结构？

提问：[课后思考]I、III 型有什么特点，和 II 型有什么区别？

II 型限制酶识别及切割序列示例

核酸序列中的限制性内切酶位点识别依据限制酶所识别的序列结构信息进行预测分析。常用的限制酶资源是限制酶数据库（The Restriction Enzyme Database，REBASE），它收录了限制酶的所有信息，包括限制酶识别序列和作用位点、甲基化酶、甲基化特异性、酶类产品的商业来源和参考文献等。REBASE 提供了限制酶的查询工具、识别位点序列及限制酶酶切双链 DNA 的三维结构等信息；分析工具提供理论酶切消化图谱、序列比对、酶切位点分析等功能。

常用的限制性核酸内切酶位点分析工具是 NEBCutter V2.0，可以产生 DNA 序列的酶切位点分析结果。它使用的限制酶来源于 REBASE 数据库，识别位点列表每天根据 REBASE 数据库数据同步更新。NEBCutter V2.0 可提供单一酶切或多选酶切位点识别和模拟消化图谱。此外，很多 DNA 分析的软件也都含有没切位点分析的功能。

提问：如何找到 NEBCutter V2.0？

4.6 开放阅读框分析

10, 00:40--00:50 (-1.1$)

开放阅读框（Open Reading Frame，ORF，开放阅读框架、开放读架等）是指在给定的阅读框架中，不包含终止密码子的一串序列。这段序列是生物个体的基因组中可能作为蛋白质编码序列的部分，包含从 5' 端翻译起始密码子（ATG）到终止密码子（TAA、TAG、TGA）之间的一段编码蛋白质的碱基序列。由于一段 DNA 或 RNA 序列有多种不同读取方式，因此可能同时存在许多不同的开放阅读框架。

六种读框示例

对于任何给定的核酸序列，根据密码子的起始位置，可以按照三种方式进行解释，其反向互补序列又含有三种，因此，一条 DNA 序列可以按六种框架阅读和翻译。ORF 的识别需要检测这六个阅读框架并决定哪一个包含以启动子和终止子为界限的 DNA 序列而其内部不包含终止子。

一个 ORF 存在一个潜在的编码序列（Coding Sequence，CDS），不同的 ORF 翻译成氨基酸可以得到不同的蛋白质编码。一个 ORF 对应一个候选的 CDS，分析 DNA 序列中的 ORF 是对该序列是否为 CDS 的初步判断，是研究 DNA 序列片段的方法之一。

提问：[课后思考]ORF 与 CDS 的区别?

ORF 的预测常与第一个 ATG 和终止密码子的确定相关，但仅凭第一个 ATG 和终止密码子是不足以确定 ORF 的。ORF 的预测程序主要是对编码区进行特征统计、相关模式的识别或利用同源比对的方法识别。原核生物编码区通常只含有一个单独的 ORF，识别方法相对简单，即最长 ORF 法。而真核生物的编码区被内含子分割成数个不连续的外显子，其编码区序列分析更加复杂。

ORF 分析常用的程序是 NCBI 的在线分析工具 ORF Finder。在其他一些集成化的软件中也有 ORF 分析相关的工具。

4.7 启动子分析

10, 00:00--00:10 (^1.2-)

DNA 序列中储存着调控信息，其中转录调控控制基因的转录活性。基因的转录表达是生命体的基本生物过程，生物个体在不同发育阶段、不同组织、不同生理状态下，基因的转录调控也不同。真核基因调控主要是在转录水平上进行的，受大量特定的顺式作用元件（cis-acting element）和反式作用因子（trans-acting factor，又称跨域作用因子）的调控，真核生物的转录调控大多数是通过顺式作用元件和反式作用因子复杂的相互作用来实现的。顺式作用元件位于基因的旁侧，是能够影响基因表达的核酸序列，包括启动子（promoter）、增强子（enhancer）、应答元件（responsive elements）等，其活性只影响与其自身同处于一个 DNA 分子上的基因。顺式作用元件本身并不编码蛋白质，仅仅提供一个作用位点，与反式作用因子相互作用参与基因表达调控。反式作用因子是参与调控靶基因转录效率的蛋白质，可以直接或间接地识别或结合在各类顺式作用元件核心序列上，可对基因表达产生激活或阻遏的作用。

启动子是一段位于转录起始位点 5' 端上游区的 DNA 序列，能活化 RNA 聚合酶，使之与模板 DNA 准确地结合并具有转录起始的特异性。转录起始位点（Transcription Start Site，TSS）是指与新生 RNA 链第一个核苷酸相对应 DNA 链上的碱基，研究证实通常为一个嘌呤。常把起点前面，即 5' 端的序列称为上游序列（upstream），而把其后面即 3' 端的序列称为下游序列（downstream）。在描述剪辑的位置时，一般用数字表示，起点为 +1, 下游方向依次为 +2,+3,……，上游方向依次为 -1,-2,-3……序列的书写方向通常是固定的，使转录从左（上游）向右（下游）进行，mRNA 同样按照 5'→3' 方向书写。原核基因启动子具有明显共同一致的序列，包含两个短序列，分别位于从转录起结点起计的 -10 及 -35 上游位置，是 RNA 聚合酶与启动子的结合位点。位于 -10 的序列称为 -10 元件或 -10 区（Pribnow 区，Pribnow box），通常包含 TATAAT6 个核苷；位于 -35 的序列通常包含 TTGACA6 个核苷。在真核生物基因中，类似原核基因启动子 Pribnow 区的 Hogness 区（Hogness box），是位于转录起始点上游 -25 ～ -30bp 处的共同序列 TATAAA，也称为 TATA 区。另外，在起始位点上游 -70 ～ -78bp 处

原核基因启动子结构

真核基因启动子结构

还有另一段共同序列 CCAAT，这是与原核生物中 -35bp 相对应的序列，称为 CAAT 区（CAAT box）。

提问：[课后思考] 各特征序列区的作用。

转录因子（transcription factor）是指能够结合在某基因上游特异核苷酸序列上的蛋白质，这些蛋白质能调控其基因的转录。转录因子可以调控 RNA 聚合酶与 DNA 模板的结合。转录因子一般有不同的功能区域，如 DNA 结合结构域与效应结构域。转录因子不单与基因上游的启动子区域结合，也可以和其它转录因子形成转录因子复合体来影响基因的转录。转录因子结合位点（Transcription Factor Binding Site，TFBS）是与转录因子结合的 DNA 序列，长度约为 5 ～ 20bp，它们与转录因子相互作用进行基因的转录调控。同一转录因子能够同时调控多个基因，虽然与不同基因序列的结合位点具有一定的保守性，但又存在一定的可变性。

TFBS 的序列标识图（留悬念）

识别基因的调控区序列特征信息是研究基因功能、基因转录调控规律、识别新基因及解析基因组结构的途径之一。挖掘调控区序列特征信息的方法主要有同源匹配法和模式识别法。TFBS 是较短的 DNA 片段，在整个基因组中会存在大量的重复序列，这些特征给正确识别 TFBS 带来一定的难度，也使得预测方法普遍存在较高的假阳性率。

随着基因表达调控研究的深入，越来越多的结合位点的调控区域序列信息不断产生。启动子和转录因子结合位点信息存储在相关的数据库中。EPD（Eukaryotic Promoter Database）是一个有注释的非冗余的真核生物 RNA 聚合酶 II（Pol II）启动子数据集，其中的转录起始位点（TSS）都是通过实验获得的。TRANSFAC 是真核生物转录调控信息的数据库，包括转录因子、转录因子结合位点及转录调控关系等信息，收录的数据都经过实验验证。

启动子、转录因子结合位点等保守的功能区可通过序列分析获得相应的序列特征信息。分析工具能直接搜索目的 DNA 序列中是否含有已知位点的序列模式。Promoter Scan 根据转录因子结合序列同源性分析预测 DNA 中的启动子区域；Promoter 2.0 基于遗传算法的人工神经网络技术预测脊椎动物启动子区 Pol II 和其他调控因子结合位点的信息。Tfblast（TRANSFAC BLAST）可以根据比对算法找出目标 DNA 序列中可能存在的转录因子结合位点。

TESS（Transcription Element Search System）因资源限制已被移除。

10, 00:10--00:20 (-1.2-)

4.8　CpG 岛识别

CpG 二核苷酸占哺乳动物基因组的 5% ～ 10%，其中，70% ～ 80% 呈甲基化状态，称为甲基化的 CpG（mCpG）。但 CpG 的分布很不均一，在基因组的某些区段，CpG 保持或高于正常概率，这些区段被称作 CpG 岛（CpG island）。CpG 岛主要位于脊椎动物基因，尤其是看家基因（housekeeping gene）的转录起始位点附近，长度约 300 ～ 3000bp。对于哺乳动物的基因来说，约 40% 的启动子（人类约 70%）含有 CpG 岛。几乎看家基因都含有 CpG 岛；一般位于基因的 5' 端区域；大多数 CpG 岛是未甲基化的；CpG 岛中的核小体中 H1 含量低，其他组蛋白被广泛乙酰化，并具有超敏感位点；未甲基化 CpG 岛可能说明基因具有潜在活性。对人类 21 和 22 号染色体全序列进行的分析表明，GC 含量超过 55%、CpG 二核苷酸的出现率（观测值与期望值的比率）达到 65% 且长度超过 500bp 的 DNA 区域更可能是分布在基因 5' 端区域的真的 CpG 岛。CpG 岛是表观遗传学中重要的作用区域，CpG 岛甲基化是基因转录活性的调控因素之一，CpG 岛甲基化异常常伴随着疾病的发生。

CpG 岛的特征

提问：现在学习的知识一定正确吗？（知识的时效性）

提问：[课后思考] 假阳性、真阳性、假阴性、真阴性的含义及其之间的关系。

传统的 CpG 岛识别方法主要依据三个序列特征：GC 含量、CpG 岛长度、CpG 二核苷酸的出现频率。确定一个区域为 CpG 岛的常用标准为：至少长 200bp，GC 含量超过 50%，CpG 的观察值与预测值的比率高于 60%。其中，CpG 的观察值与预测值的比率计算公式为：$\frac{Num\ of\ CpG}{Num\ of\ C \times Num\ of\ G} \times Total\ number\ of\ nucleotides\ in\ the\ sequence$。另一类主要方法是基于统计学特征的识别方法，如使用马尔科夫链和隐马尔科夫链识别 CpG 岛。

CpG 岛预测标准

EMBL 提供的 CpG 岛的计算工具是 EMBOSS 的 CpGPlot/CpGReport/Isochore，基于传统的窗口滑动法，参数设置默认 CpG 岛跨度至少为 200bp，GC 含量 >50%，CpG 出现频率 >0.6，满足这些条件的区域都预测为 CpG 岛。其他的类似工具还有 CpG Island Searcher、CpGcluster2 等。

演示使用 cpgplot 预测 CpG 的操作

4.9 重复序列分析

10, 00:20--00:30 (-1.2-)

基因组注释包括结构注释和功能注释，结构注释的核心是基因识别，而为了提高基因识别的效率，首先要寻找并屏蔽重复的、低复杂性的序列。

重复序列的分类

重复序列（repetitive sequence, repeated sequence）是指真核生物基因组中重复出现的核苷酸序列。这些序列一般不编码多肽，在基因组内可成簇排布，也可散布于基因组。

根据重复次数的多少，可以分成三大类：（1）低度重复序列（lowly repetitive sequence），在整个基因组中只含有 2 ～ 10 个拷贝，如酵母 tRNA 基因、人和小鼠的珠蛋白基因等；（2）中度重复序列（moderately repetitive sequence），重复次数为几十次到几千次，重复单元的平均长度约 300bp，如 rRNA 和 tRNA 基因；（3）高度重复序列（highly repetitive sequence），重复几百万次，一般是少于 10 个核苷酸残基组成的短片段，如异染色质上的卫星 DNA。

按照重复序列的组织形式可以分成两大类：串联重复序列和散在重复序列。前一种成簇存在于染色体的特定区域，后一种分散于染色体的各位点上。串联重复（tandem repeat）包括卫星 DNA（satellite DNA）、小卫星（minisatellite）和微卫星（microsatellite）。卫星 DNA 是一类高度重复序列。小卫星 DNA (minisatellite DNA) 又称可变数目串联重复（variable number tandem repeat，VNTR），由 10 100bp 的基本单位串联而成，总长通常不超过 20kb，重复次数在群体中是高度变异的。在人类基因组中，约 90% 的小卫星序列出现在靠近端粒的位置。微卫星又称为简单重复序列（Simple Sequence Repeats，SSRs）或短串联重复序列（Short Tandem Repeats，STRs），指两个或多个核苷酸重复排列、且不同的重复序列相邻的形式，只有 2 ～ 10bp，串联成簇，长度 50 ～ 100bp，常见于非编码的内含子中。微卫星是多型性的一种类型，由于重复单位及重复次数不同，使其在不同种族、不同人群之间的分布具有很大差异性，构成了 STR 遗传多态性。散在重复（Interspersed repeat）一般都是中度重复序列。根据重复序列的长度可以分为短散在重复序列（Short Interspersed Nuclear Element，SINE）和长散在重复序列（Long Interspersed Nuclear Element，LINE）。前者长度在 500bp 以下，在人基因组中的重复拷贝数达 10 万以上；后者长度在 1000bp 以上，在人基因组中有上万份拷贝。LINE 是可以自主转座的一类反转录转座子，来源于 RNA 聚合酶 II 的转录产物；SINE 则是非自主转座的反转录转座子，来源于 RNA 聚合酶 III 的转录产物。在灵长类中，主要的 LINE 和 SINE 分别为 L1 和 Alu。

提问：[课后思考] 不同分类标准之间的关系。

对于真核生物的核酸序列而言，在进行基因识别之前首先应该把简单的大量的重复序列标记出来并去除，目的是为了避免重复序列对预测程序产生干扰，尤其是涉及数据库搜索的程序。

不同重复序列数据库储存了不同类型重复序列的信息：Repbase 是常用的真核生物 DNA 重复序列数据库；L1Base 是 LINE-1 的数据库；STRBase（Short Tandem Repeat DNA Internet DataBase）是存储短串联重复序列的数据库。RepeatMasker 是比较常用的重复序列片段分析程序，应用于识别、分类和屏蔽重复元件，包括低复杂性序列和散在重复，通过与已知重复序列数据库比对搜索基因组序列中的相似序列进行识别。RepeatMasker 一共提供了四个搜索引擎：Cross_match 速度慢但比其他引擎的精度高；ABBlast（以前叫做 WUBlast）速度快精度略低；RMBlast 是 NCBI Blast 工具的兼容版；HMMER 使用 nhmmer 程序搜索 Dfam 数据库，但它只适用于人类基因组序列。

4.10　操作演示

10, 00:30--00:40 (-1.2-)

EMBOSS（The European Molecular Biology Open Software Suite）是一个开放源代码的序列分析软件包，它是一组为分子生物学家所设计的公开且免费的软件。该软件能够自动识别处理以不同格式存储的数据，甚至可以通过互联网提取数据，此外同软件包一同提供的还包括大量的程序库，软件包整合了 100 多个的序列分析程序，可以满足一般实验室的各种各样的序列分析要求。并且，因为该软件包同时提供了一个扩展库，它也是允许其他科学家依据自由软件精神编制、发布软件的一个平台。EMBOSS 同时将现在可以得到的一系列序列分析工具整合成一个无缝的整体。EMBOSS 遵照 GPL 协议，打破了向商业软件包发展的传统模式。使用者可以通过三种不同的方式使用 EMBOSS 软件：第一种是通过命令行的方式；第二种是通过 X-Windows 的方式使用 EMBOSS 软件的图形界面；第三种是内联网的方式。使用者可以免费获得这些软件以及相关界面程序。

1. 使用 EMBOSS 中的相关程序对人类 CD9 基因序列（序列号：AY422198.1）的组份进行分析。
    - compseq: Calculate the composition of unique words in sequences
    - geecee: Calculate fractional GC content of nucleic acid sequences
    - revseq: Reverse and complement a nucleotide sequence
2. 使用 EMBOSS 中的相关程序对人类 TERT 基因序列（序列号：NG_009265.1）中 4000-5300bp 区域进行 CpG 岛的分析。
    - extractseq: Extract regions from a sequence
    - cpgplot: Identify and plot CpG islands in nucleotide sequence(s)
    - cpgreport: Identify and report CpG-rich regions in nucleotide sequence(s)
    - isochore: Plot isochores in DNA sequences

4.11　总结与答疑

10, 00:40--00:50 (-1.2$)

本次课需要掌握的知识点与技能：

- 知识点：
    - DNA 序列基本信息分析——查戈夫法则，序列转换，GC 含量。
    - 限制酶位点分析——命名，II 型。
    - 开放阅读框分析——ORF 与 CDS。
    - 启动子与转录因子结合位点分析——启动子结构。
    - CpG 岛识别——判别依据及标准。

  - 重复序列分析——分类。
- 技能：
  - 解决问题的思路。
  - 寻找最合适的方法。
  - 搜索及学习新软件。
  - 先易后难，由浅入深。

4.12 课前甜点

播放“中心法则”的动画视频。

10, -00:10--00:00 (^2.1-)

4.13 回顾与导入

对于一条未知的核苷酸序列，我们首先进行基本信息的分析，包括：序列转换、计算碱基比例和 GC 含量、寻找限制酶切位点；接下来进一步进行序列特征的分析，包括：开放阅读框的预测、启动子和转录因子结合位点的分析、CpG 岛的识别；之后就是比较重要的编码区基因的预测，而在基因识别之前，为了提高识别效率，需要先屏蔽掉重复序列。这些都是上次课介绍的内容，本次课将承接以上内容，首先讲解基因识别的相关内容，之后把分析的对象转换到转录产物 mRNA 上，介绍选择性剪接的分析，最后对 miRNA 及其靶基因的预测进行讲解，并用几分钟的时间介绍一下近几年备受瞩目的长链非编码 RNA。

5, 00:00--00:05 (-2.1-)

4.14 基因识别

25, 00:05--00:30 (-2.1-)

提问：你对基因概念的理解。

基因（gene）指的是产生一条多肽链或功能 RNA 所需的全部核苷酸序列。基因，一段具有特定功能和结构的连续的 DNA 片段，携带着遗传信息，是编码蛋白质或 RNA 分子遗传信息、控制性状的基本遗传单位。基因通过指导蛋白质的合成来表达自己所携带的遗传信息，从而控制生物个体的性状表现。一个完整的基因，不仅包括编码区，还包括 5' 末端和 3' 末端长度不等的特异性序列，它们虽然不编码氨基酸，却在基因的转录过程中起着重要的调节作用。基因识别（gene prediction，gene finding），是生物信息学的一个重要分支，使用生物学实验或计算机等手段识别 DNA 序列上的具有生物学特征的片段。基因识别的对象主要是蛋白质编码基因，也包括其他具有一定生物学功能的因子，如 RNA 基因和调控因子。基因识别是基因组研究的基础。

课外读物：
薛定谔的《生命是什么？》
伽莫夫的《从一到无穷大》
沃森的《双螺旋》和《基因 • 女郎 • 伽莫夫》
贾德森的《创世纪的第八天》

提问：[课后思考] 基因概念的提出与发展。（维基百科）

提问：原核基因和真核基因最主要的区别是什么？

原核基因结构

原核与真核生物的基因都包括编码区和非编码区，但两者的结构有着很大的差别。原核基因的结构比较简单，为连续基因，其编码区是一个完整的 DNA 片段，非编码区位于编码区的上游及下游。所有原核基因都有一个编码区，依基因类型的不同，或是编码一种蛋白质多肽或是编码一种 RNA 结构，如 tRNA 和 rRNA。在原核基因编码区两侧，还存在着用于控制转录作用的调节区，即启动子和终止子。在

DNA 链上，由起始密码子开始到终止密码子为止的一个连续编码序列，叫做开放阅读框（Open Reading Frame，ORF），也就是所谓的编码区。启动子（promoter）是位于基因 5' 末端上游外侧紧挨转录起点的一段长度为 20 ～ 200bp 的非编码的核苷酸序列，其功能是与 RNA 聚合酶结合形成转录起始复合物。原核生物的启动子大约 40 ～ 50bp，其中包含有转录的起始点和两个区（-35 区和 -10 区）。起始点是 DNA 模板链上开始进行转录作用的位点，通常在其互补的编码链对应位点（碱基）标以“+1”。-10 区是 RNA 聚合酶核心酶与 DNA 分子紧密结合的部位，大多包含有 6bp 的共有序列，即：TATAAT。-35 区是 RNA 聚合酶因子识别 DNA 分子的部位，其共有序列为：TTGACA。终止子（terminator）是位于一个基因或一个操纵子的末端，提供转录停止信号的 DNA 区段。与启动子不同的是终止子仍能被 RNA 聚合酶转录成 mRNA。与原核基因一样，一个完整的真核基因，不仅包括编码区，还包括编码区两侧的调节序列。但真核基因和原核基因在结构上存在着许多基本的区别，其中最重要的一点是其不连续性，许多真核生物的蛋白质编码基因以及某些 tRNA 基因的编码序列，都被一种叫做内含子（intron）的非编码序列所间断。在基因的表达过程中，内含子便从初级 mRNA 分子中被剪接掉，形成成熟的功能 mRNA。真核基因的非编码序列包括非编码区的所有序列以及编码区里面的内含子。真核生物编码蛋白质的基因启动子，与原核生物的启动子相似，也具有两个高度保守的共有序列。其一是在 -25 ～ -35 区含有 TATAA 序列，称为 TATA 盒（TATA box）。TATA 盒与原核生物启动子的 -10 区相似，是转录因子与 DNA 分子的结合部位。其二是在多数启动子中，-70 ～ -80 区含有 CAAT 序列，称为 CAAT 盒。另外，还有一部分 DNA 序列能增强或减弱真核基因转录起始的频率，这些区域称为增强子（enhancer）和沉默子（sliencer）。原核生物和真核生物的基因结构不同，所以使用的基因识别方法也不同。

真核基因结构

提问：模板链和编码链。（具有转录功能、合成 RNA 的模板的那条链是模板链、反义链；无转录功能、编码蛋白质的那条链是编码链、有义链，它与 RNA 的序列相同）

提问：基因结构不同，识别方法是否也不同？（扩展：任务性质不同，处理策略不同）

基因识别的方法

基因识别的方法主要包括三大类：间接识别法（Extrinsic Approach）、从头计算法（*Ab Initio* Approach）和比较基因组学的方法。在基因的间接识别法中，人们利用已知的 mRNA 或蛋白质序列为线索在 DNA 序列中搜寻所对应的片段。由给定的 mRNA 序列确定唯一的作为转录源的 DNA 序列；而由给定的蛋白质序列，也可以由密码子反转确定一组可能的 DNA 序列。因此，在线索的提示下搜寻工作相对较为容易，搜寻算法的关键在于提高效率，并能够容忍由于测序不完整或者不精确所带来的误差。BLAST 是目前以此为目的最广泛使用的软件之一。若 DNA 序列的某一片段与 mRNA 或蛋白质序列具有高度相似性，这说明该 DNA 片段极有可能是蛋白编码基因。但是，测定 mRNA 或蛋白质序列的成本高昂，而且在复杂的生物体中，任意确定的时刻往往只有一部分基因得到了表达。这意味着从任何单个细胞的 mRNA 和蛋白质上都只能获得一小部分基因的信息；要想得到更为完整的信息，不得不对成百上千个不同状态的细胞中的 mRNA 和蛋白质测序。这是相当困难的。鉴于间接识别法的种种缺陷，仅仅由 DNA 序列信息预测蛋白质编码基因的从头计算法就显得十分重要了。一般意义上基因具有两种类型的特征，一类特征是“信号”，由一些特殊的序列构成，通常预示着其周围存在着一个基因；另一类特征是“内容”，即蛋白质编码基因所具有的某些统计学特征。使用 *Ab Initio* 方法识别基因又称为基因预测。通常仍需借助实验证实预测的 DNA 片段是否具有生物学功能。由于多个物种的基因组序列已完全测出，使得比较基因组学得以发展，并产生了新的基因识别的方法。该方法基于如下原理：自然选择的力量使得基因和 DNA 序列上具有生物学功能的其他片段较其他部分有较慢的变异速率，在前者的变异更有可能对生物体的生存产生负面影响，因而难以得到保存。因此，通过比较相关的物种的 DNA 序列，我们能够取得预测基因的新线索。

在原核生物中，基因往往具有特定且容易识别的启动子序列（信号），如 Pribnow 盒和转录因子结合位点。与此同时，构成蛋白质编

码的序列构成一个连续的开放阅读框（内容），其长度约为数百个到数千个碱基对。除此之外，原核生物的蛋白质编码还具有其他一些容易判别的统计学的特征。这使得对原核生物的基因预测能达到相对较高的精度。对真核生物（尤其是复杂的生物如人类）的基因预测则相当有挑战性。一方面，真核生物中的启动子和其他控制信号更为复杂，还未被很好的了解。两个被真核生物基因搜寻器识别到的讯号例子有 CpG islands 及 poly(A) tail 的结合点。另一方面，由于真核生物所具有的剪接机制，基因中一个蛋白质编码序列被分为了若干段（外显子），中间由非编码序列连接（内含子）。人类的一个普通蛋白质编码基因可能被分为了十几个外显子，其中每个外显子的长度少于 200 个碱基对，而某些外显子更可能只有二三十个碱基对长。因而蛋白质编码的一些统计学特征变得难于判别。真核基因预测中可利用的信号有：上游启动子区特征序列（TATA box、CAAT box、GC box）；5' 端外显子位于核心启动子 TATA 盒的下游，含有起始密码子；内部的外显子两端的供体位点和受体位点；3' 端的外显子下游包含终止密码子和 polyA 信号序列。综合多个序列信号信息确定外显子的边界，从而达到识别编码区域的目的。可以利用的内容信息主要包括密码子使用偏好性（codon usage bias）、双联密码子出现频率、核苷酸周期性分析 (即分析同一个核苷酸在 3，6，9，... 位置上周期性出现的规律)、基因组等值区 (isochore) 等。对已知编码区进行统计学分析找出编码规律和特性，通过统计值区分外显子、内含子和基因间区域。在实际应用中常常联合几种方法，以提高识别效率。

提问：供体位点和受体位点在外显子上还是在内含子上？

提问：[课后思考] 什么是密码子使用偏好性，什么是基因组等值区（isochore）？

基因预测的工具列表

高级的基因识别算法常使用更加复杂的概率论模型，如隐马尔科夫模型、人工神经网络、决策树方法等。GeneMarkS 是采用迭代隐马尔科夫模型（iterative Hidden Markov model）的识别工具；Glimmer 是一个广泛应用的高级基因识别程序，基于插入式马尔科夫模型（interpolated Markov models，IMMs），它对原核生物基因的预测已非常精确。相比之下，对真核生物的预测则效果有限。GENSCAN 是脊椎动物基因预测软件，使用广义隐马尔科夫模型（GHMM）根据基因的整体结构进行基因预测，包括外显子、内含子、基因间区域、转录信号、翻译信号、剪接信号等信息，能对基因组 DNA 序列识别完整的外显子 -内含子结构，能识别多个基因，具有同时处理正、反两条链的功能。其他常用识别工具还有利用神经网络技术同时组合各种编码度量的 GRAIL。

4.15 mRNA 选择性剪接

25, 00:30--00:55 (-2.1$)

提问：RNA 包括哪些种类？

DNA 转录生成 RNA。RNA 即是携带遗传信息的主要生物大分子，也是重要的功能单位。RNA 包括 mRNA、tRNA、rRNA 三种主要形式，参与蛋白质的生物合成；还包括微小 RNA（miRNA）、干扰小 RNA（siRNA）、长链非编码 RNA（long non-coding RNA，lncRNA）等参与生物调控。mRNA 属于编码 RNA，tRNA、rRNA、miRNA、siRNA、lncRNA 属于非编码 RNA。前面介绍了 DNA 分析的相关主题，接下来重点介绍 RNA 的相关内容，主要是 mRNA 和 miRNA 的生物学特征及其分析方法。

真核细胞的基因序列中，包含了内含子（intron）与外显子（exon），两者交互穿插。其中内含子在基因转录成 mRNA 前体后会被 RNA 剪接体移除，剩下的外显子才是能够存在于成熟 mRNA（之后再进一步翻译成蛋白质）的片段。剪接（splicing），又称拼接，指基因信息在转录后的一种修饰，即将内含子移除及合并外显子，是真核生物的信使 RNA 前体（precursor messenger RNA）变成成熟 mRNA 的过程之一。这些成熟的 mRNA 会接着进行蛋白质生物合成中的翻译，以产生蛋白质，称翻译作用。剪接也是真核生物与原核生物的区别之一。

在很多时候，剪接过程可以通过对同一个基因转录的相同 pre-mRNA 使用不同的剪接选择，产生不同的 mRNA 异构体 (isoform)，最后产生多种相似却又独特的蛋白质，或是产生出稳定性低的 mRNA 产物以达到调节基因表达的目的。RNA 的选择性剪接（alternative splicing），又称可变剪接，就是指这种用不同的剪接方式（选择不同的剪接位点组合）从一个 mRNA 前体产生不同的 mRNA 剪接异构体的过程。大多数真核生物的基因都存在选择性剪接的现象。由于选择性剪接的存在而使基因组可以产生比基因数量还多许多倍的基因产物。

提问：[课后思考] 剪接的分子生物学过程。

选择性剪接利用基因的不连续性，一条未经剪接的 RNA，含有的多种外显子被剪成的不同组合，可翻译出不同的蛋白质。从而将同一基因中的外显子以不同的组合方式来表现，使一个基因在不同时间、不同环境中能够制造出不同的蛋白质（基因表达调控），这可增加生理状况下系统的复杂性或适应性。Pre-mRNA 的剪接也并不是完美的，其中相当一部分的剪接产物 (spliced transcripts) 因为剪接过程的不够精确、或是形成未成熟的终止密码子 (premature termination codon, PTC) 而造成该 RNA 的降解 (RAN degradation)。选择性剪接受时间和空间的限制，在不同的组织中，在相同组织的不同细胞中，在同一组织的不同发育阶段，在对病理过程的不同反应过程中都会产生不同的剪接异构体。有研究表明，人体细胞中有 92% ～ 94% 的基因会进行选择性剪接。多达 50% 的致病突变会影响剪接，选择性剪接的异常改变使得基因在转录后期产生异常的剪接变体，编码出异常的蛋白质，导致人类遗传疾病甚至癌变。

选择性剪接类型的示意图

选择性剪接的形式多样，主要有以下五种产生机制：

1. 外显子跳跃（exon skipping），也叫盒式外显子（cassette exon）、外显子遗漏等，在剪接时外显子会被移除或者保留下来，这是选择性剪接最常见的方式。
2. 互斥外显子（mutually exclusive exons），即相互排斥性剪接，两个外显子只有一个会保留在剪接后的产物中，两者不会同时存在。
3. 5' 选择性剪接（alternative donor site），选择性使用不同的 5' 端的剪接连接点（即供体位点），从而改变上游外显子的 3' 边界。
4. 3' 选择性剪接（alternative acceptor site），选择性使用不同的 3' 端的剪接连接点（即受体位点），从而改变下游外显子的 5' 边界。
5. 内含子保留（intron retention），一段序列在剪接过程中或作为内含子被去除，或作为外显子被保留下来。它和外显子跳跃的区别在于这段序列的两边不是内含子。这是最少见的选择性剪接机制。

提问：外显子跳跃和内含子保留的区别。

除了上述五种主要的选择性剪接机制外，还有两种从同一基因生成不同 mRNA 的主要机制：多启动子（multiple promoters）或选择性起始（alternative initiation），多 polyA 位点（multiple polyadenylation sites）或选择性终止（alternative termination）。通过在不同的位点起始转录，可以产生含有不同 5' 端外显子的转录产物；但它常被看做转录调节（transcriptional regulation）的机制而非选择性剪接的方式。与之类似，不同的 polyA 位点会产生含有不同 3' 端的转录本。这两种机制与选择性剪接相结合，为从同一基因产生不同 mRNA 增加了多样性。

在各种选择性剪接的类型中，外显子跳跃最为常见，互斥外显子相对较少见，内含子保留是最少见的。

提问：为什么无法统一各种观点？（生物系统的复杂性）

真实剪接的示例

真实情况远比理论模型复杂

以上介绍的只是选择性剪接机制的基本模型，但真实的剪接事件则更为复杂，比如小鼠 hyaluronidase 3 基因的三个剪接异构体。前两者（黄色和绿色）的比较表明是内含子保留的机制，但后两者（黄色和蓝色）的比较则表明是外显子跳跃。

选择性剪接数据资源根据数据来源分成两大类。一类是基于文献报道的数据库，通过收集、整理已有的实验数据和文献报道而建立。

一类是基于 EST 数据的选择性剪接数据库，主要是通过采用 EST 序列数据与基因组或 DNA、mRNA 序列进行比对的方法，发现新的或已经存在的选择性剪接形式后建立的数据库或数据集。常用的选择性剪接数据库有 ASTD 和 ASAP。ASTD（Alternative Splicing and Transcript Diversity database），选择性剪接和转录多样性数据库，由 ASD（Alternative Splicing Database）和 ATD（Alternate Transcript Diversity Project）合并而来，提供人、小鼠、大鼠、斑马鱼、线虫、果蝇等多个物种的选择性剪接数据，是目前常用的选择性剪接数据库。ASD 包含了多种模式生物的选择性剪接数据。ASD 由三个子数据库组成：AEDB（Alternative Exon Database），从文献中收集的经实验验证的人类选择性外显子（alternative exons）；AltExtron，由 EST 与全基因组序列比对得到选择性剪接数据及选择性外显子和内含子；AltSplice（Alternative Splicing Database），收集了通过计算方法得到的选择性剪接事件及其模式。ATD 试图通过创建人和小鼠的全长选择性剪接转录本，来揭示转录异构体的产生机制。ASAP（Alternative Splicing Annotation Project）是通过全基因组范围内比对 EST 数据得到的人和小鼠的选择性剪接数据库。它提供基因的外显子、内含子结构、选择性剪接、组织特异性选择性剪接、选择性剪接产生的蛋白质异构体等信息。

ASAP 网站失效（注意数据库的更新日期）

提问：英语中 ASAP 缩写的含义。

剪接位点的精确定位是确定真核生物基因结构的关键。生物信息学已开发出从头预测法、基于 EST/cDNA 序列比对法和基于 RNA-seq 数据识别等多种方法。从头预测法主要采用支持向量机、概率模型、隐马尔科夫模型、神经网络和二次判别分析法等技术预测剪接位点。选择性剪接过程的调控机制具有多样性，主要由剪接调节因子（splicing-regulatory element）和调节蛋白相互作用来进行调节。剪接调节因子主要由外显子剪接增强子（Exonic Splicing Enhancer，ESE）、外显子剪接沉默子（Exonic Splicing Silencer，ESS）、内含子剪接增强子（Intronic Spicing Enhancer，ISE）和内含子剪接沉默子（Intronic Splicing Silencer，ISS）。目前常用工具都结合剪接调节因子预测进行选择性剪接的分析。ESEfinder 和 RESCUE-ESE 等用于外显子剪接增强子的预测。ASPicDB（Alternative Splicing Prediction Database）旨在提供人类基因选择性剪接模式的可靠注释和剪接异构体的功能注释，能够在基因、转录本、外显子、蛋白质或剪接位点水平上进行分析，提供两类蛋白质（球状蛋白和跨膜蛋白）及有关定位、PFAM 结构域、信号肽（signal peptides）、跨膜片段和卷曲螺旋片段的信息。

4.16 miRNA 及其靶基因预测

25, 00:05--00:30 (^2.2-)

提问：ncRNA 包括哪些?

人类基因组中，虽然仅有 1.5% 左右的序列可以编码蛋白质，但研究表明，90% 以上的基因组都是可以转录的。这些转录后不编码蛋白质的 RNA 分子统称为非编码 RNA（non-coding RNAs，ncRNA）。ncRNA 主要分成基础结构性 ncRNA（infrastructural non-coding RNAs）和调节性 ncRNA（regulatory non-coding RNAs）大类。基础结构性 ncRNA 即看家 ncRNA（housekeeping non-coding RNAs），主要包括 tRNA、rRNA、snRNA 和 snoRNA。根据转录本的长度，调节性 ncRNA 分为短于 200nt 的小 RNA（small RNAs，sRNA）和长于 200nt 的长链非编码 RNA（long ncRNAs，lncRNA）。sRNA 主要包括已为研究人员所熟知的 miRNA、siRNA 和 piRNA，其产生过程及作用机理都已研究的比较透彻。

微 RNA（microRNAs，miRNA，小分子 RNA）归属小 RNA 范畴，是真核生物中广泛存在的一种长约 20 到 24 个核苷酸的内源性非编码单链 RNA 分子。miRNA 通过 RNA 诱导沉默复合体（RISC）与靶基因的 3' 非翻译区（3' UTR）相结合，导致靶基因 mRNA 降解或者抑制其翻译，从而调节基因转录后的表达。miRNA 在调控基因表达、细胞周期、生物体发育时序等方面起重要作用。

miRNA 的特点

miRNA 不具有开放阅读框，不编码蛋白质，表达具有时序性和组织特异性，进化上具有高度的保守性。在植物、动物和真菌中发现的 miRNAs 只在特定的组织和发育阶段表达。miRNA 的组织特异性和时序性，决定组织和细胞的功能特异性，表明 miRNA 在细胞生长和发育过程的调节过程中起多种作用。

提问：miRNA 的生成过程。

miRNA 生成过程

编码基因在核内产生长度为 300 ～ 1000nt 的初级转录本（primary transcript），即初始 miRNA（pri-miRNA），被双链 RNA 特异的核糖核酸酶 Drosha 切割成长度为 70 ～ 90nt、具有茎环二级结构（发卡结构）的单链前体 miRNA（pre-miRNA）。这些发夹结构的 pre-miRNA 通过核输出蛋白 Exportin 5 机制转运到细胞质，被第二个双链 RNA 特异的核糖核酸酶 Dicer 及其辅因子加工形成 20 ～ 24nt 的成熟 miRNA 及其互补体。miRNA 前体在各个物种间具有高度的进化保守性，茎部保守性最强，环部可以容许更多的突变位点存在。

miRNA 与靶 mRNA 的互补类型

miRNA 与其靶基因间是多对多的关系：一个 miRNA 可以调控多个靶基因，一个基因也可以受多个 miRNA 的调控。这种复杂的调节网络既可以通过一个 miRNA 来调控多个基因的表达，也可以通过几个 miRNAs 的组合来精细调控某个基因的表达。在动物中，一个 miRNA 通常可以调控数十个基因。miRNA 的作用机制取决于 miRNA 与靶 mRNA 的互补程度，包括完全互补型和不完全互补型。miRNA 与靶 mRNA 完全互补（或者几乎完全互补）导致靶基因 mRNA 降解（在植物中比较常见），通过这种机制起作用的 miRNAs 的结合位点通常都在 mRNA 的编码区或开放阅读框中；不完全互补导致靶基因 mRNA 的翻译受到抑制，使用这种机制的 miRNA 结合位点通常在 mRNA 的 3' 端非翻译区。。

提问：[课后思考] 为什么仅仅完全、不完全互补就会导致完全不同的结果？

miRNA 通过作用于相应靶基因 mRNA 完成生物学功能，如个体发育的调控、细胞分化和组织发育等。据推测脊椎动物基因组有多达 1000 个不同的 miRNAs，调控至少 30% 以上的基因表达。miRNA 的异常与疾病发生发展具有相关性。最近的研究发现，miRNA 表达与多种癌症相关，大约 50% 得到注释的 miRNAs 在基因组上定位于与肿瘤相关的脆性位点（fragile site）。这说明 miRNAs 在肿瘤发生过程中起至关重要的作用，这些 miRNAs 所起的作用类似于抑癌基因和癌基因的功能。

miRNA 在细胞分化，生物发育及疾病发生发展过程中发挥巨大作用，越来越多地引起研究人员的关注。随着对于 miRNA 作用机理的进一步的深入研究，以及利用最新的例如 miRNA 芯片、miRNA-seq 等高通量的技术手段对于 miRNA 和疾病之间的关系进行研究，将会使人们对于高等真核生物基因表达调控的网络理解提高到一个新的水平。这也将使 miRNA 可能成为疾病诊断的新的生物学标记，还可能使得这一分子成为药靶，或是模拟这一分子进行新药研发，这将可能会给人类疾病的治疗提供一种新的手段。

miRNA 分析主要包括 miRNA 预测和 miRNA 靶基因预测两方面。

miRNA 主要通过 cDNA 克隆测序和计算预测两种方法获得。早期克隆测序直接、可靠，但很难克隆出在不同时期表达或只在特定组织或细胞系中表达的 miRNA，由于它的固有局限性，也很难捕获表达丰度较低的 miRNA。最近几年发展起来的 miRNA-seq 等高通量技术使得通过实验预测 miRNA 得到了一定程度的改观。随着 miRNA 研究的发展，生物信息学预测 miRNA 的方法成为一条重要的辅助途径，其优势是不受 miRNA 表达的时间和组织特异性以及表达水平的影响。

miRNA 预测方法

常用的 miRNA 预测方法主要有 5 种：

1. 同源片段搜索方法。将已知 miRNA 或 pre-miRNA 序列在自身或其他相近基因组中用比对算法搜索同源序列，结合序列二级结构特征进行筛选。

2. 基于比较基因组学的预测方法。依据进化过程中的保守性在多物种中搜索潜在的 miRNA。
3. 基于序列和结构特征打分的预测方法。根据已知 miRNA 序列和结构的特征对全基因组范围中能形成茎环结构的片段进行筛选，是发现非同源、物种特异性 miRNA 的方法。
4. 结合作用靶标的预测方法。依据 miRNA 与其靶基因序列间的碱基互补配对的保守性的特点预测 miRNA。
5. 基于机器学习的预测方法。通过对阳性 miRNA 和阴性 miRNA 数据集的训练来构建区分两者的分类器，根据所得分类器对未知序列进行预测。其中支持向量机（SVM，Support Vector Machine）方法是目前 miRNA 分类和预测最常用的机器学习方法。

常用的 miRNA 预测软件有 MiRscan、MiPred、miRFinder 等。

miRNA 通过与靶基因 mRNA 的 3' UTR 不精确互补配对使靶 mRNA 的翻译受到抑制，二者相互作用以 miRNA:mRNA 二聚体结构形式存在。miRNA 序列 5' 端的 2-8nt 为种子区域，在 miRNA 靶基因预测中起主导作用。种子区域具有保守性，与靶 mRNA 序列能较好地互补配对结合，且在不同物种中靶序列也是保守的，这些特征是靶基因预测方法的重要依据。miRNA 靶基因预测方法主要有 2 类：

种子区域

miRNA 靶基因预测方法与工具

1. 基于种子区域互补和保守性的规则预测，常用软件有 miRanda、TargetScan 等。
2. 基于机器学习方法训练参数进行靶基因预测，常用软件有 PicTar、miTarget 等。

miRNA 数据库资源

miRBase 是集 miRNA 序列、注释信息和预测的靶基因数据为一体的数据库，是目前存储 miRNA 信息最主要的公共数据库之一。TarBase 数据库是存储已被实验验证的 miRNA 与靶基因间关系的数据库。miRGen 是整合了 miRNA 靶基因数据、基因组注释信息以及位置关系的综合数据库。更多数据库可以参看微 RNA 与微 RNA 靶数据库（维基百科）。

4.17 lncRNA 简介

5, 00:30--00:35 (-2.2-)

提问：垃圾 DNA 真的是垃圾吗？（以学校等社会机构进行类比）

目前已知在人类基因组中，lncRNA 基因的数目已经达到了 13249。据估计，人类基因组中 lncRNA 基因的总数可能在 15000 条以上。在基因组范围上，对已知 lncRNA 进行的研究表明：大多数 lncRNA 是被 RNA 聚合酶 II 所转录的，有 5' 帽子和 3' 端的 poly(A) 尾巴，主要富集在细胞核；与蛋白质编码基因相比，lncRNA 的长度偏短、外显子数目偏少，在不同物种间的保守性差，稳定性偏低，表达水平很低，而且表达具有细胞、组织、发育、疾病等时空特异性。lncRNA 以 RNA 分子形式在表观遗传学水平、转录水平和转录后水平上调控基因的表达，参与基因转录、剪接、翻译、修饰和印迹等重要的生物学过程。lncRNA 的异常表达与众多疾病的发生发展相关，如肿瘤、阿尔兹海默病和心血管疾病等。

lncRNA 的特点

虽然在基因表达调控过程中具有重要作用，且与肿瘤、阿尔兹海默病等疾病密切相关，但迄今为止，仅有数量有限的 lncRNA 得到了比较细致的研究，对 lncRNA 的大规模分析更是处于起步阶段。lncRNA 的相关数据库可以参看长链非编码 RNA 数据库（维基百科）。

lncRNA 数据库资源

提问：[课后思考] 查阅 lncRNA 与疾病关系的相关资料。

4.18 查找数据库与分析工具

5, 00:35-00:40 (-2.2-)

每一个研究领域或研究对象都涉及众多的数据库与分析工具，如何快速查找到它们也是生物信息学工作者的必备技能之一。实际工作

中可以采用以下几种途径：

提问：如何查找需要的数据库和工具？

- 借鉴相关文献中使用的数据库与工具。
- 向特定领域的专家请教。
- *Nucleic Acids Research* 每年的第一期为数据库专刊。
- 维基百科等总结性网站。
- The Elements of Bioinformatics。
- 使用 Google 等搜索引擎搜索。

4.19　总结与答疑

10, 00:40-00:50 (-2.2$)

本次课需要掌握的知识点与技能：

- 知识点：
  - 基因识别——原核和真核的基因结构，基因识别方法。
  - mRNA 选择性剪接——选择性剪接的类型，数据资源。
  - miRNA——miRNA 的特点，miRNA 预测方法与工具，miRNA 靶基因预测方法与工具。
- 技能：
  - 查找数据库——时效性。
  - 查找分析工具——适用范围。

提问：[课后思考] 总结自己遇到的文本型问题（如：根据学号整合其姓名与成绩），经典问题将在 Galaxy 操作演示中予以解决。