第 4 章　核酸序列分析

了解：

1. DNA 携带的两类遗传信息。
2. DNA 与 RNA 序列分析的常见内容及相关数据库和工具。
3. ORF 与 CDS 的区别。
4. 原核基因和真核基因启动子的结构。
5. 原核和真核的基因结构。
6. lncRNA 的研究现状。

熟悉：

1. 限制性核酸内切酶的命名规则，II 型限制酶的特点。
2. 重复序列依重复次数和组织形式的分类。
3. 基因识别的三大类方法。
4. miRNA 及其靶基因预测的方法和工具。

掌握：

1. CpG 岛的概念及其识别依据和判别标准。
2. mRNA 选择性剪接的产生机制。
3. 解决问题的思路。
4. 查找数据库和分析工具的方法。
5. 学习数据库与分析工具使用方法的策略。

4.1　引言

“龙生龙，凤生凤，老鼠的儿子会打洞！”[1]“种瓜得瓜，种豆得豆。”“爹矬矬一个，娘矬矬一窝。”“一母生九子，连母十个样。”“龙生九子各不同。”“天下乌鸦一般黑。”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，所以可能没有什么感觉。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，A、C、G、T/U（Adenine、Cytosine、Guanine、Thymine/Uracil）都是构成 DNA 和 RNA 核酸序列的基本组份。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？其奥秘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可以构建出不同的形状。类似于二进制中运用一连串的 0 和 1 以及英文字母表中运用 26 个不同的字母来表达信息，基因所包含的信息来自于 4 种不同核苷酸沿 DNA 分子的排列顺序。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。

[1] 列举与遗传变异相关的谚语。

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 A、C、G、T 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质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，能够从海量的序列数据中挖掘出具有生物学功能的特征信息。

有关序列的特征，我们可能会提出以下三个基本问题[2]：

1. 总的 GC 含量或者其他核苷酸成分是多少？大多数真核基因组的 GC 含量介于 35%～45% 之间，而细菌则具有一个宽得多的范围。
2. 有哪些重复的 DNA 序列，在什么地方？RepeatMasker 等程序可以识别并标记出重复的成分，比如 Alu 重复。GLIMMER 和 GRAIL 等程序组合了多种可以识别基因组 DNA 中的重复成分的算法。
3. 现在一共有多少个已知基因（编码蛋白质的序列）？基因可以被一定数量的特征所确定，包括：
   - 基因的特殊密码子偏好；
   - 重复 DNA 序列的缺乏；
   - 启动子特异结合位点区域信号的出现。

   这些特征在真核生物和原核生物的 DNA 序列之间是存在很大差异的。

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。

4.2　DNA 组份分析与序列转换

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（A）与胸腺嘧啶（T）数量几乎完全一样，鸟嘌呤（G）与胞嘧啶（C）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：

- 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（GC content）是在所研究对象的全基因组中，鸟嘌呤（Guanine，G）和胞嘧啶（Cytosine，C）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 是以三个氢键相连的，而 T 和 A 则是以两个氢键相连的。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：$\frac{A+T}{G+C}$。

[2]此部分内容摘抄自《生物信息学与功能基因组学》（乔纳森 • 佩夫斯纳 著，孙之荣 译），第 3 篇 基因组分析，第 12 章 全基因组和系统发生树，12.8 基因组分析总览。

对已经完成的基因组进行分析，发现核苷酸组成具有一些特征性质[3]。GC 含量是指鸟嘌呤和胞嘧啶含量的平均百分数，在原核生物的真细菌中从 25% 到 75% 不等，这种组份差异可以用于识别细菌种类。（恶性疟原虫的 GC 含量仅有 20% 左右，这种情况下一般说它富含 AT（AT-rich），而不会说缺少 GC（GC-poor）。）。与细菌相比，真核生物的基因组更大也更为多变，但是他们的 GC 含量却非常一致（40%～45% 左右），与此同时，基因组中不同区域的 GC 含量存在一定的差异。人类基因组的平均 GC 含量是 41%，但是存在一些相对富含 GC 和缺乏 GC 的区域。其中已知基因的 GC 含量较高，当 GC 含量从 30% 增加到 50% 时基因密度增加了 10 倍。此外，GC 含量还与密码子使用偏好性、DNA 双链的熔解温度（Tm，melting temperature）等有关，是进行核酸杂交的重要参数。

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。[4]

首先要分析任务的属性：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太长时间
- 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）[5]；请别人帮忙（术业有专攻）；自己写程序（自己动手，丰衣足食）；……

此处仅提供一个解决问题的思路：遇到问题，先思考后动手[6]；办法总会有的，而且绝不止一种[7]；没有完美的方法，只有合适的方法[8]。

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[9]，这与序列的书写惯例有关：

- 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]

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[10]，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。[11]

4.3 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。[12]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。

[3]此部分内容摘抄自《生物信息学与功能基因组学》（乔纳森 • 佩夫斯纳 著，孙之荣 译），第 3 篇 基因组分析，第 14 章 已完成测序的基因组：细菌和古细菌基因组，14.8 原核基因组分析。

[4]接到一个进行碱基组成和 GC 含量分析的任务后，该如何入手解决？

[5]Don’t Reinvent the Wheel.（不要重复发明轮子。）

[6]三思而后行。

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

[8]卤水点豆腐，一物降一物。

[9]为什么常说“反向互补”，而不是“反向”、“互补”呢？

[10]思考如下操作是否正确：先将 A 全部替换成 T，再将 C 全部替换成 G，再将 G 全部替换成 C，最后将 T 全部替换成 A。如果操作不正确，请思考解决办法（提示：参考 C 语言中交换两个变量的值的方法）。

[11]模仿计算 GC 含量的思路，思考如何获取 DNA 序列的反向、互补及反向互补序列？

[12]限制酶如何实现只切割外源 DNA 而不损伤内源 DNA？

限制酶的命名[13]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[14]。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[15]；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。[16]

核酸序列中的限制性核酸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[17]，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酶切位点分析的功能。

✍ 课外阅读

1. Loenen, W. a M., Dryden, D. T. F., Raleigh, E. a, Wilson, G. G. & Murray, N. E. Highlights of the DNA cutters: a short history of the restriction enzymes. Nucleic Acids Res. 42, 3–19 (2014).
2. Loenen, W. a M., Dryden, D. T. F., Raleigh, E. a & Wilson, G. G. Type I restriction enzymes and their relatives. Nucleic Acids Res. 42, 20–44 (2014).
3. Rao, D. N., Dryden, D. T. F. & Bheemanaik, S. Type III restriction-modification enzymes: a historical perspective. Nucleic Acids Res. 42, 45–55 (2014).
4. Loenen, W. a M. & Raleigh, E. a. The other face of restriction: modification-dependent enzymes. Nucleic Acids Res. 42, 56–69 (2014).
5. Mruk, I. & Kobayashi, I. To be or not to be: regulation of restriction-modification systems and other toxin-antitoxin systems. Nucleic Acids Res. 42, 70–86 (2014).
6. Pingoud, A. & Jeltsch, A. Structure and function of type II restriction endonucleases. Nucleic Acids Res. 29, 3705–3727 (2001).
7. Nikolajewa, S., Beyer, A., Friedel, M., Hollunder, J. & Wilhelm, T. Common patterns in type II restriction enzyme binding sites. Nucleic Acids Res. 33, 2726–33 (2005).

4.4　开放阅读框分析

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（AUG）到终止密码子（UAA、UAG、UGA）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

[13] *Eco*RI 的名称是如何确定的？

[14] 一说分为 I、II、III 和 IV 四大类。

[15] 什么是回文结构？列举回文的例子。（上海自来水来自海上；山东落花生花落东山；人人为我，我为人人；画上荷花和尚画；123321；……）

[16] I、III 型有什么特点，和 II 型有什么区别？

[17] 如何找到 NEBCutter V2.0？

同的开放阅读框架。

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起始密码子和终止密码子为界限的 DNA 序列而其内部不包含终止密码子，识别出蛋白质序列最长的 ORF。长的 ORF 往往不是偶然发生的，所以在任何阅读框中存在 300bp 或更长的未中断的编码序列是判断一个基因的重要依据。

一个 ORF 存在一个潜在的编码序列（Coding DNA Sequence，CDS）[18]，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。

ORF 的预测常与第一个 AUG 和终止密码子的确定相关，但仅凭第一个 AUG 和终止密码子是不足以确定 ORF 的。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 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。

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。

4.5 启动子分析

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的表达调控。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）。在描述碱基的位置时，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，分别位于从转录起始位点起计的-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 TATAAT 6 个核苷酸；位于-35 的序列通常包含 TTGACA 6 个核苷酸。在真核基因中，类似原核基因启动子 Pribnow 区的 Hogness 区（Hogness box），是位于转录起始位点上游-25～-30bp 处的共同序列 TATAAA，也称为 TATA 区。另外，在转录起始位点上游-70～-78bp 处还有另一段共同序列 CCAAT，这是与原核基因中-35bp 相对应的序列，称为 CAAT 区（CAAT box）。[19]

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

[18]ORF 与 CDS 的区别？

[19]各特征序列区有何独特的作用？

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。

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[20]。

随着基因表达调控研究的深入，不断产生越来越多的结合位点的调控区域序列信息。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 数据库[21]是一个真核生物顺式调控元件和反式作用因子数据库，包括转录因子、转录因子结合位点及转录调控关系等信息，数据搜集的对象从酵母一直到人类跨多个物种，而且全部是来自实验证实的数据。TRANSFAC 目前由 BIOBASE GmbH 维护和管理。TRANSFAC 数据库中的数据资源被分为六大数据表，可根据自己的研究需要进行使用和搜索。这六大数据表包括：

1. FACTOR 描述了结合到转录因子结合位点的蛋白质信息；
2. SITE 给出了转录因子结合位点的相关信息；
3. MATRIX 给出了转录因子结合位点的核苷酸矩阵信息；
4. GENE 给出了转录因子的基因信息；
5. CLASS 包含了转录因子背景介绍及其所属类别；
6. CELL 给出了转录因子与其结合位点发生相互作用时的细胞信息。

值得注意的是，这六个数据表是相互独立的（除部分 GENE 和 SITE 有交叉），所以存在冗余现象；与此同时，同一个转录因子在不同数据表中又可以获得不同的信息描述，起到相互补充的作用。TRANSFAC 数据库分为学术（免费）和专业（付费）两个版本。自 1996 年发布第一版以来已作了多次更新，目前学术版已更新至 TRANSFAC 7.0（2006 年）；专业版本已升级至 TRANSFAC 2008.3（2008 年）。与学术版相比，专业版的信息量更多，不仅包含了全部学术版的内容，而且还增加了 ChIP-chip 的数据，2009 年 12 月还增加了 ChIP-seq 的数据。

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。[22]

4.6　CpG 岛识别

CpG[23]二核苷酸[24]在人类基因组 DNA 中的含量非常低，出现率仅为预计值的 1/5，占哺乳动物基因组的 5%～10%，其中，70%～80%CpG 的胞嘧啶脱去氨基，呈甲基化状态，称为甲基化的 CpG（mCpG）。然而 CpG 的分布并不均一，在基因组的某些区段，CpG 保持或高于正常概率，这

[20]假阳性、真阳性、假阴性、真阴性的含义及其之间的关系。

[21]此部分内容摘抄自《生物信息学》（李霞，人民卫生出版社），第十一章（转录调控的信息学分析），第四节（转录调控相关数据库）。

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

[23]CpG 中的 p 代表什么？（phosphate，磷酸盐；连接 C 和 G 两个核苷酸的磷酸二酯键。）

[24]此部分内容摘抄自《生物信息学与功能基因组学》（乔纳森 • 佩夫斯纳 著，孙之荣 译），第 3 篇 基因组分析，第 17 章 人类基因组，17.4 人类基因组计划。

些区段被称作 CpG 岛（CpG island），它们通常与看家基因的启动子和外显子区相联系，参与基因沉默、基因组印迹以及 X 染色体失活等过程。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，未甲基化 CpG 岛可能说明基因具有潜在活性；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。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。

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[25]：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：$\frac{Num\ of\ CpG}{Num\ of\ C \times Num\ of\ G} \times Total\ number\ of\ nucleotides\ in\ the\ sequence$。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。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。

EMBL 提供的 CpG 岛的计算工具是 EMBOSS 的 CpGPlot/CpGReport/Isochore，基于传统的滑动窗口法，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。

4.7 操作演示

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

EMBOSS（European Molecular Biology Open Software Suite）[26]软件包是一个开源的序列分析软件包，该软件包源于 1988 年开始开发的 EGCG 系统，是一组为分子生物学家所设计的公开且免费的软件。它整合了目前可以获得的大部分序列分析软件并有一套专门设计的 C 语言函数。该软件包含 160 多个小型程序，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，能很好地进行序列模体挖掘、关键词同源性数据库搜索、序列比较、进化分析、序列二级结构分析、限制性酶切图谱分析、引物设计、序列模式识别与翻译、片段拼接等工作，可以满足一般实验室的各种各样的序列分析要求。同时它提供了一个扩展库，以方便科学家依据自由软件精神编制、发布软件。使用 EMBOSS，可以将系列分析工作进行无缝整合，弥补了很多其他软件功能分散、分析效率底下的缺陷。EMBOSS 遵照 GPL 协议，打破了商业软件包发展的传统模式，使科研工作者自由、免费的使用功能强大的分析工具。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。EMBOSS 的主页网址为 http://emboss.sourceforge.net。

EMBOSS 软件包主要运行于 Linux 操作系统和 Mac 操作系统。现在基于 Windows 操作系统

[25]知识具有时效性，现在的结论未必是正确的！

[26]此部分内容摘抄自《生物信息学》（李霞，人民卫生出版社），第四章（序列特征分析），第四节（序列综合分析）。

的 EMBOSS 也能自由免费使用。需要说明的是基于 Windows 操作系统时，主要采用 Staden 进入 EMBOSS，在使用时，需要安装 Embosswin 软件。Embosswin 的下载网址是：ftp://emboss.open-bio.org/pub/EMBOSS/wEMBOSS Explorerindows。

EMBOSS 程序源码完全公开，其核心程序的基本设计与各种开发平台相兼容，可以供研究人员作为开发应用程序的平台，不同的机构也因此开发了各种 EMBOSS 使用界面，包括 Jemboss、Other GUIs、Web interfacaes、Workfolws、Ports and packages 等，下面介绍两种常用的使用界面。

1. JEMBOSS：JEMBOSS 全名为 Java-EMBOSS，即 Java 界面的 EMBOSS 程序图形使用界面，由英国 HGMP-RC（Human Genome Mapping Program Resource Center）开发，采用友好的 Java 窗口执行 EMBOSS 程序，并加入档案管理功能，使用者下载 Java 执行程序后，可在自己的电脑上执行 EMBOSS 程序并储存分析结果，使用方便。JEMBOSS 的主界面主要分成两块区域，左边为程序列表区，右边为程序执行窗口。左边程序列表区上方是按程序类别进行的分类，中间可输入程序名或前几个字母来快速定位程序，下边是详细的程序列表。
2. EMBOSS Explorer：EMBOSS Explorer 是 Web interfaces 的一种界面，利用 EMBOSS Explorer 可将 EMBOSS 软件从单机版变为网络版，可以让更多的人享用已有的分析平台。这个版本适合安装在服务器上，提供在线分析。EMBOSS Explorer 的主界面，左边的区域是程序列表，右边的区域是执行程序区域。

EMBOSS 使用的主要程序有：

- 最重要的程序。wossname：根据关键字查找程序；showdb：显示所有整合的数据库。
- 序列编辑。revseq：将序列反转并互补；seqret：序列格式转换。
- 两个序列相似性图形表达。dottup：精确匹配；dotmatcher：近似匹配。
- 双序列比对。needle：全局比对；water：局部比对。
- 多序列比对。emma：clustalW。
- 寻找 SNP。deffseq：仅限于双序列比对中。
- 其他。plotorf，getorf：翻译；iep：等电点预测；tmap：跨膜区预测；pepinfo：蛋白质性质；patmatmotifs：Motif 搜索。

4.8　序列分析中的算法【此部分内容还有待整理完善！】

4.8.1　滑动窗口（Sliding Window，SW）

Count occurrence statistics typically use sliding overlapping windows to analyze a sequence.

Sliding window analysis is a commonly used method for studying the properties of molecular sequences: data are plotted as moving averages of a particular criterion, such as the number of nucleotide changes, for a window of a certain length slid along a sequence or sequence alignment.

A sliding window of length N is a common algorithm in sequence analysis.[27] Let's imagine that we have a sequence of length L. For instance if L=5, our sequence might be "ABCDE". How many windows of size N can we make? Well, if N=1 we have 5 (A, B, C, D, E). N=2 gives 4 (AB, BC, CD and DE). N=3 gives 3 (ABC, BCD, CDE) and so on, up to N=5 which gives 1 (ABCDE). You might see a pattern there. The number of windows with size N for a sequence of length L is (L-N)+1.

The sliding window technique is a widely used algorithmic primitive. For example, the sliding window approach has been used to improve the spatial resolution of predicted binding sites using ChIP-Seq data, as well as to analyze sequence polymorphisms that can aid in understanding evolutionary forces and chromosomal functional significance.

The sliding window (SW) method has been extensively used for exploratory DNA polymorphism data analysis. Unfortunately, the SW approach has a number of limitations, such as the determination of the

[27] 此部分内容摘抄自 Sliding windows with (Bio)perl。

appropriate window size or the problem of multiple comparisons, that are critical in genome-wide based analysis.

A sliding window gathers information about properties of nucleotides or amino acids. A simple example is to calculate the %(G+C) content within a window. Then move the window one nucleotide and repeat the calculation.

If the window is too small it is difficult to detect the trend of the measurement. If too large you could miss meaningful data.

课外阅读

1. 滑动窗口的 Perl/Bioperl 实现：Sliding windows with (Bio)perl。
2. 滑动窗口的 R 实现：A sliding window analysis of GC content。

4.8.2 动态规划

动态规划（Dynamic programming，DP）是一种在数学、计算机科学和经济学中使用的，通过把原问题分解为相对简单的子问题的方式求解复杂问题的方法。动态规划常常适用于有重叠子问题和最优子结构性质的问题，动态规划方法所耗时间往往远少于朴素解法。

动态规划背后的基本思想非常简单。大致上，若要解一个给定问题，我们需要解其不同部分(即子问题），再合并子问题的解以得出原问题的解。通常许多子问题非常相似，为此动态规划法试图仅仅解决每个子问题一次，从而减少计算量：一旦某个给定子问题的解已经算出，则将其记忆化存储，以便下次需要同一个子问题解之时直接查表。这种做法在重复子问题的数目关于输入的规模呈指数增长时特别有用。

动态规划在查找有很多重叠子问题的情况的最优解时有效。它将问题重新组合成子问题。为了避免多次解决这些子问题，它们的结果都逐渐被计算并被保存，从简单的问题直到整个问题都被解决。因此，动态规划保存递归时的结果，因而不会在解决同样的问题时花费时间。

动态规划只能应用于有最优子结构的问题。最优子结构的意思是局部最优解能决定全局最优解（对有些问题这个要求并不能完全满足，故有时需要引入一定的近似）。简单地说，问题能够分解成子问题来解决。

适用情况

- 最优子结构性质。如果问题的最优解所包含的子问题的解也是最优的，我们就称该问题具有最优子结构性质（即满足最优化原理）。最优子结构性质为动态规划算法解决问题提供了重要线索。
- 无后效性。即子问题的解一旦确定，就不再改变，不受在这之后、包含它的更大的问题的求解决策影响。
- 子问题重叠性质。子问题重叠性质是指在用递归算法自顶向下对问题进行求解时，每次产生的子问题并不总是新问题，有些子问题会被重复计算多次。动态规划算法正是利用了这种子问题的重叠性质，对每一个子问题只计算一次，然后将其计算结果保存在一个表格中，当再次需要计算已经计算过的子问题时，只是在表格中简单地查看一下结果，从而获得较高的效率。

动态规划的应用实例：计算斐波那契数列（Fibonacci polynomial）。斐波那契数列由 0 和 1 开始，之后的斐波那契系数就由之前的两数相加：0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, ……。

4.8.3　贪心算法

贪心法，又称贪心算法，是一种在每一步选择中都采取在当前状态下最好或最优（即最有利）的选择，从而希望导致结果是最好或最优的算法。

贪心算法在有最优子结构的问题中尤为有效。最优子结构的意思是局部最优解能决定全局最优解。简单地说，问题能够分解成子问题来解决，子问题的最优解能递推到最终问题的最优解。

贪心算法与动态规划的不同在于它每对每个子问题的解决方案都做出选择，不能回退。动态规划则会保存以前的运算结果，并根据以前的结果对当前进行选择，有回退功能。

贪心法可以解决一些最优化问题，如：求图中的最小生成树、求哈夫曼编码……对于其他问题，贪心法一般不能得到我们所要求的答案。一旦一个问题可以通过贪心法来解决，那么贪心法一般是解决这个问题的最好办法。由于贪心法的高效性以及其所求得的答案比较接近最优结果，贪心法也可以用作辅助算法或者直接解决一些要求结果不特别精确的问题。

对于大部分的问题，贪心法通常都不能找出最佳解（不过也有例外），因为他们一般没有测试所有可能的解。贪心法容易过早做决定，因而没法达到最佳解。

实现过程

1. 建立数学模型来描述问题。
2. 把求解的问题分成若干个子问题。
3. 对每一子问题求解，得到子问题的局部最优解。
4. 把子问题的解局部最优解合成原来解问题的一个解。

4.8.4　聚类分析

聚类分析（Cluster analysis，亦称为群集分析）是对于统计数据分析的一门技术，在许多领域受到广泛应用，包括机器学习、数据挖掘、模式识别、图像分析以及生物信息。聚类是把相似的对象通过静态分类的方法分成不同的组别或者更多的子集（subset），这样让在同一个子集中的成员对象都有相似的一些属性，常见的包括在坐标系中更加短的空间距离等。一般把数据聚类归纳为一种非监督式学习。

聚类类型

数据聚类算法可以分为结构性或者分散性。结构性算法利用以前成功使用过的聚类器进行分类，而分散型算法则是一次确定所有分类。结构性算法可以从上至下或者从下至上双向进行计算。从下至上算法从每个对象作为单独分类开始，不断融合其中相近的对象。而从上至下算法则是把所有对象作为一个整体分类，然后逐渐分小。

分割式聚类算法，是一次性确定要产生的类别，这种算法也已应用于从下至上聚类算法。

基于密度的聚类算法，是为了挖掘有任意形状特性的类别而发明的。此算法把一个类别视为数据集中大于某阈值的一个区域。DBSCAN 和 OPTICS 是两个典型的算法。

许多聚类算法在执行之前，需要指定从输入数据集中产生的分类个数。除非事先准备好一个合适的值，否则必须决定一个大概值，关于这个问题已经有一些现成的技术。

距离测量

在结构性聚类中，关键性的一步就是要选择测量的距离。一个简单的测量就是使用曼哈顿距离，它相当于每个变量的绝对差值之和。该名字的由来起源于在纽约市区测量街道之间的距离就是由人步行的步数来确定的。

一个更为常见的测量是欧式空间距离，他的算法是找到一个空间，来计算每个空间中点到原点的距离，然后对所有距离进行换算。

常用的几个距离计算方法：欧式距离（2-norm 距离）、曼哈顿距离（Manhattan distance, 1-norm 距离）、infinity norm、马氏距离、余弦相似性、汉明距离。

结构性聚类

在已经得到距离值之后，元素间可以被联系起来。通过分离和融合可以构建一个结构。传统上，表示的方法是树形数据结构，然后对该结构进行修剪。树的根节点表示一个包含所有项目的类别，树叶表示与个别的项目相关的类别。

层次聚类算法，要么是自底向上聚集型的，即从叶子节点开始，最终汇聚到根节点；要么是自顶向下分裂型的，即从根节点开始，递归的向下分裂。

任意非负值的函数都可以用于衡量一对观测值之间的相似度。决定一个类别是否分裂或者合并的是一个连动的标准，它是两两观测值之间距离的函数。

在一个指定高度上切割此树，可以得到一个相应精度的分类。

分散性聚类

K-均值算法表示以空间中 k 个点为中心进行聚类，对最靠近他们的对象归类。算法归纳为：

1. 选择聚类的个数 k。
2. 任意产生 k 个聚类，然后确定聚类中心，或者直接生成 k 个中心。
3. 对每个点确定其聚类中心点。
4. 再计算其聚类新中心。
5. 重复以上步骤直到满足收敛要求。（通常就是确定的中心点不再改变。）

该算法的最大优势在于简洁和快速。劣势在于对于一些结果并不能够满足需要，因为结果往往需要随机点的选择非常巧合。

4.8.5 遗传算法

4.8.6 傅里叶变换

4.8.7 机器学习

机器学习是近 20 多年兴起的一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。因为学习算法中涉及了大量的统计学理论，机器学习与统计推断学联系尤为密切，也被称为统计学习理论。算法设计方面，机器学习理论关注可以实现的，行之有效的学习算法。很多推论问题属于无程序可循难度，所以部分的机器学习研究是开发容易处理的近似算法。

机器学习已经有了十分广泛的应用，例如：数据挖掘、计算机视觉、自然语言处理、生物特征识别、搜索引擎、医学诊断、检测信用卡欺诈、证券市场分析、DNA 序列测序、语音和手写识别、战略游戏和机器人运用。

机器学习有下面几种定义：“机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能”。“机器学习是对能通过经验自动改进的计算机算法的研究”。“机器学习是用数据或以往的经验，以此优化计算机程序的性能标准。”一种经常引用的英文定义是：A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

机器学习可以分成下面几种类别：

- 监督学习：从给定的训练数据集中学习出一个函数，当新的数据到来时，可以根据这个函数预测结果。监督学习的训练集要求是包括输入和输出，也可以说是特征和目标。训练集中的目标是由人标注的。常见的监督学习算法包括回归分析和统计分类。
- 无监督学习：与监督学习相比，训练集没有人为标注的结果。常见的无监督学习算法有聚类。
- 半监督学习：介于监督学习与无监督学习之间。
- 增强学习：通过观察来学习做成如何的动作。每个动作都会对环境有所影响，学习对象根据观察到的周围环境的反馈来做出判断。

具体的机器学习算法有：

- 构造条件概率：回归分析和统计分类
  - 人工神经网络
  - 决策树
  - 高斯过程回归
  - 线性判别分析
  - 最近邻居法
  - 感知器
  - 径向基函数核
  - 支持向量机
- 通过再生模型构造概率密度函数：
  - 最大期望算法
  - graphical model：包括贝叶斯网和 Markov 随机场
  - Generative Topographic Mapping
- 近似推断技术：
  - 马尔可夫链
  - 蒙特卡罗方法
  - 变分法
- 最优化：大多数以上方法，直接或者间接使用最优化算法。

4.8.8　判别分析

线性判别分析（Linear Discriminant Analysis），简称判别分析，是统计学上的一种分析方法，用于在已知的分类之下遇到有新的样本时，选定一个判别标准，以判定如何将新样本放置于哪一个类别之中。这种方法主要应用于医学的患者疾病分级，以及人脸识别、经济学的市场定位、产品管理及市场研究等范畴。

关于线性鉴别分析的研究应追溯到 Fisher 在 1936 年发表的经典论文（Fisher R A. The use of multiple measurements in taxonomic problems），其基本思想是选择使得 Fisher 准则函数达到极值的向量作为最佳投影方向，从而使得样本在该方向上投影后，达到最大的类间离散度和最小的类内离散度。在 Fisher 思想的基础上，Wilks 和 Duda 分别提出了鉴别矢量集的概念，即寻找一组鉴别矢量构成子空间，以原始样本在该子空间内的投影矢量作为鉴别特征用于识别。

1970 年 Sammon 提出了基于 Fisher 鉴别准则的最佳鉴别平面的概念。随后，Foley 和 Sammon 进一步提出了采用一组满足正交条件的最佳鉴别矢量集进行特征抽取的方法。

1988 年 Duchene 和 Leclercq 给出了多类情况下最佳鉴别矢量集的计算公式。

2001 年 Jin 和 Yang 从统计不相关的角度，提出了具有统计不相关性的最优鉴别矢量集的概念。与 F-S 鉴别矢量集不同的是，具有统计不相关性的最优鉴别矢量是满足共轭正交条件的，该方法被称为不相关的鉴别分析或 Jin-Yang 线性鉴别法。

以上提到的各种方法仅适用于类内散布矩阵非奇异 (可逆) 的情形，但实际应用中存在着大量的典型的小样本问题，比如在人脸图像识别问题中，类内散布矩阵经常是奇异的。这是因为待识别的图像矢量的维数一般较高，而在实际问题中难以找到或根本不可能找到足够多的训练样本来保证类内散布矩阵的可逆性。因此，在小样本情况下，如何抽取 Fisher 最优鉴别特征成为一个公认的难题。

4.8.9 人工神经网络

人工神经网络（artificial neural network，缩写 ANN），简称神经网络（neural network，缩写 NN），是一种模仿生物神经网络的结构和功能的数学模型或计算模型。神经网络由大量的人工神经元联结进行计算。大多数情况下人工神经网络能在外界信息的基础上改变内部结构，是一种自适应系统。现代神经网络是一种非线性统计性数据建模工具，常用来对输入和输出间复杂的关系进行建模，或用来探索数据的模式。

神经网络是一种运算模型，由大量的节点（或称“神经元”，或“单元”）和之间相互联接构成。每个节点代表一种特定的输出函数，称为激励函数（activation function）。每两个节点间的连接都代表一个对于通过该连接信号的加权值，称之为权重（weight），这相当于人工神经网络的记忆。网络的输出则依网络的连接方式，权重值和激励函数的不同而不同。而网络自身通常都是对自然界某种算法或者函数的逼近，也可能是对一种逻辑策略的表达。

它的构筑理念是受到生物（人或其他动物）神经网络功能的运作启发而产生的。人工神经网络通常是通过一个基于数学统计学类型的学习方法（Learning Method）得以优化，所以人工神经网络也是数学统计学方法的一种实际应用，通过统计学的标准数学方法我们能够得到大量的可以用函数来表达的局部结构空间，另一方面在人工智能学的人工感知领域，我们通过数学统计学的应用可以来做人工感知方面的决定问题（也就是说通过统计学的方法，人工神经网络能够类似人一样具有简单的决定能力和简单的判断能力），这种方法比起正式的逻辑学推理演算更具有优势。

人工神经网络是一个能够学习，能够总结归纳的系统，也就是说它能够通过已知数据的实验运用来学习和归纳总结。人工神经网络通过对局部情况的对照比较（而这些比较是基于不同情况下的自动学习和要实际解决问题的复杂性所决定的），它能够推理产生一个可以自动识别的系统。与之不同的基于符号系统下的学习方法，它们也具有推理功能，只是它们是建立在逻辑算法的基础上，也就是说它们之所以能够推理，基础是需要有一个推理算法则的集合。

通常来说，一个人工神经元网络是由一个多层神经元结构组成，每一层神经元拥有输入（它的输入是前一层神经元的输出）和输出，每一层（我们用符号记做）Layer(i) 是由 Ni(Ni 代表在第 i 层上的 N) 个网络神经元组成，每个 Ni 上的网络神经元把对应在 Ni-1 上的神经元输出做为它的输入，我们把神经元和与之对应的神经元之间的连线用生物学的名称，叫做突触（Synapse），在数学模型中每个突触有一个加权数值，我们称做权重，那么要计算第 i 层上的某个神经元所得到的势能等于每一个权重乘以第 i-1 层上对应的神经元的输出，然后全体求和得到了第 i 层上的某个神经元所得到的势能，然后势能数值通过该神经元上的激活函数（activation function，常是 S 型函数（Sigmoid function）以控制输出大小，因为其可微分且连续，方便差量规则（Delta rule）处理。求出该神经元的输出，注意的是该输出是一个非线性的数值，也就是说通过激励函数求的数值根据极限值来判断是否要激活该神经元，换句话说我们对一个神经元网络的输出是否线性不感兴趣。

一种常见的多层结构的前馈网络（Multilayer Feedforward Network）由三部分组成：

- 输入层（Input layer），众多神经元（Neuron）接受大量非线形输入信息。输入的信息称为输入向量。
- 输出层（Output layer），信息在神经元链接中传输、分析、权衡，形成输出结果。输出的信息称为输出向量。
- 隐藏层（Hidden layer），简称“隐层”，是输入层和输出层之间众多神经元和链接组成的各

个层面。隐层可以有多层，习惯上会用一层。隐层的节点（神经元）数目不定，但数目越多神经网络的非线性越显著，从而神经网络的强健性（robustness）（控制系统在一定结构、大小等的参数摄动下，维持某些性能的特性。）更显著。习惯上会选输入节点 1.2 至 1.5 倍的节点。

人工神经网络分类为以下两种：

1. 依学习策略（Algorithm）分类主要有：
   - 监督式学习网络（Supervised Learning Network）为主
   - 无监督式学习网络（Unsupervised Learning Network）
   - 混合式学习网络（Hybrid Learning Network）
   - 联想式学习网络（Associate Learning Network）
   - 最适化学习网络（Optimization Application Network）
2. 依网络架构（Connectionism）分类主要有：
   - 前向式架构（Feed Forward Network）
   - 回馈式架构（Recurrent Network）
   - 强化式架构（Reinforcement Network）

4.8.10　支持向量机

4.8.11　决策树

4.8.12　最近邻居法

4.8.13　隐马尔可夫模型

马尔可夫性质

马尔可夫性质是概率论中的一个概念。当一个随机过程在给定现在状态及所有过去状态情况下，其未来状态的条件概率分布仅依赖于当前状态；换句话说，在给定现在状态时，它与过去状态(即该过程的历史路径）是条件独立的，那么此随机过程即具有马尔可夫性质。具有马尔可夫性质的过程通常称之为马尔可夫过程。

最有名的马尔可夫过程为马尔可夫链，但不少其他的过程，包括布朗运动也是马尔可夫过程。

马尔可夫过程

在概率论及统计学中，马尔可夫过程（Markov process）是一个具备了马尔可夫性质的随机过程，因为俄国数学家安德雷 · 马尔可夫得名。马尔可夫过程是不具备记忆特质的（memorylessness）。换言之，马尔可夫过程的条件概率仅仅与系统的当前状态相关，而与它的过去历史或未来状态，都是独立、不相关的。

具备离散状态的马尔可夫过程，通常被称为马尔可夫链。马尔可夫链通常使用离散的时间集合定义，又称离散时间马尔可夫链。有些学者虽然采用这个术语，但允许时间可以取连续的值。

马尔可夫链

马尔可夫链（Markov chain），又称离散时间马尔可夫链（discrete-time Markov chain，缩写为 DTMC），因安德雷 · 马尔可夫得名，为状态空间中经过从一个状态到另一个状态的转换的随机过程。该过程要求具备“无记忆”的性质：下一状态的概率分布只能由当前状态决定，在时间序列中它前面的事件均与之无关。这种特定类型的“无记忆性”称作马尔可夫性质。马尔科夫链作为实际过程的统计模型具有许多应用。

马尔可夫链是马尔可夫过程中的一个特例，为具备马尔可夫性质与离散时间状态的随机过程。该过程中，在给定当前知识或信息的情况下，只有当前的状态用来预测将来，过去（即当前以前的历史状态）对于预测将来（即当前以后的未来状态）是无关的。

在马尔可夫链的每一步，系统根据概率分布，可以从一个状态变到另一个状态, 也可以保持当前状态。状态的改变叫做过渡，与不同的状态改变相关的概率叫做过渡概率。随机漫步就是马尔可夫链的例子。随机漫步中每一步的状态是在图形中的点，每一步可以移动到任何一个相邻的点，在这里移动到每一个点的概率都是相同的（无论之前漫步路径是如何的）。

隐马尔可夫模型

隐马尔可夫模型（Hidden Markov Model，HMM）是统计模型，它用来描述一个含有隐含未知参数的马尔可夫过程。其难点是从可观察的参数中确定该过程的隐含参数。然后利用这些参数来作进一步的分析，例如模式识别。

在正常的马尔可夫模型中，状态对于观察者来说是直接可见的。这样状态的转换概率便是全部的参数。而在隐马尔可夫模型中, 状态并不是直接可见的，但受状态影响的某些变量则是可见的。每一个状态在可能输出的符号上都有一概率分布。因此输出符号的序列能够透露出状态序列的一些信息。

隐马尔可夫模型最初是在 20 世纪 60 年代后半期 Leonard E. Baum 和其它一些作者在一系列的统计学论文中描述的。HMM 最初的应用之一是开始于 20 世纪 70 年代中期的语音识别。在 1980 年代后半期，HMM 开始应用到生物序列尤其是 DNA 的分析中。此后，在生物信息学领域 HMM 逐渐成为一项不可或缺的技术。

✍ 扩展阅读

假设你有一个住得很远的朋友，他每天跟你打电话告诉你他那天做了什么。你的朋友仅仅对三种活动感兴趣：公园散步，购物以及清理房间。他选择做什么事情只凭天气。你对于他所住的地方的天气情况并不了解，但是你知道总的趋势。在他告诉你每天所做的事情基础上，你想要猜测他所在地的天气情况。

你认为天气的运行就像一个马尔可夫链。其有两个状态——“雨”和“晴”，但是你无法直接观察它们，也就是说，它们对于你是隐藏的。每天，你的朋友有一定的概率进行下列活动：“散步”、“购物”或“清理”。因为你朋友告诉你他的活动，所以这些活动就是你的观察数据。这整个系统就是一个隐马尔可夫模型 HMM。

你知道这个地区的总的天气趋势，并且平时知道你朋友会做的事情。也就是说这个隐马尔可夫模型的参数是已知的。你可以用程序语言（Python）写下来：

```
states = ('Rainy', 'Sunny')

observations = ('walk', 'shop', 'clean')

start_probability = {'Rainy': 0.6, 'Sunny': 0.4}

transition_probability = {
    'Rainy' : {'Rainy': 0.7, 'Sunny': 0.3},
    'Sunny' : {'Rainy': 0.4, 'Sunny': 0.6},
}

emission_probability = {
    'Rainy' : {'walk': 0.1, 'shop': 0.4, 'clean': 0.5},
    'Sunny' : {'walk': 0.6, 'shop': 0.3, 'clean': 0.1},
```

```
}
```

在这些代码中，start_probability 代表了你对于你朋友第一次给你打电话时的天气情况的不确定性（你知道的只是那个地方平均起来下雨多些）。在这里，这个特定的概率分布并非平衡的，平衡概率应该接近（在给定变迁概率的情况下）`{'Rainy': 0.571, 'Sunny': 0.429}`。transition_probability 表示基于马尔可夫链模型的天气变迁。在这个例子中，如果今天下雨，那么明天天晴的概率只有 30%。emission_probability 表示了你朋友每天做某件事的概率。如果下雨，有 50% 的概率他在清理房间；如果天晴，则有 60% 的概率他在外头散步。

HMM 有三个典型（canonical）问题：

- 已知模型参数，计算某一特定输出序列的概率。通常使用 forward 算法解决。
- 已知模型参数，寻找最可能的能产生某一特定输出序列的隐含状态的序列。通常使用 Viterbi 算法解决。
- 已知输出序列，寻找最可能的状态转移以及输出概率。通常使用 Baum-Welch 算法以及 Reversed Viterbi 算法解决。

另外，最近的一些方法使用 Junction tree 算法来解决这三个问题。

✍ 扩展阅读

想象一个乡村诊所。村民有着非常理想化的特性，要么健康要么发烧。他们只有问诊所的医生的才能知道是否发烧。聪明的医生通过询问病人的感觉诊断他们是否发烧。村民只回答他们感觉正常、头晕或冷。

假设一个病人每天来到诊所并告诉医生他的感觉。医生相信病人的健康状况如同一个离散马尔可夫链。病人的状态有两种“健康”和“发烧”，但医生不能直接观察到，这意味着状态对他是“隐含”的。每天病人会告诉医生自己有以下几种由他的健康状态决定的感觉的一种：正常、冷或头晕。这些是观察结果。整个系统为一个隐马尔可夫模型 (HMM)。

医生知道村民的总体健康状况，还知道发烧和没发烧的病人通常会抱怨什么症状。换句话说，医生知道隐马尔可夫模型的参数。这可以用 Python 语言表示如下:

```
states = ('Healthy', 'Fever')

observations = ('normal', 'cold', 'dizzy')

start_probability = {'Healthy': 0.6, 'Fever': 0.4}

transition_probability = {
    'Healthy' : {'Healthy': 0.7, 'Fever': 0.3},
    'Fever' : {'Healthy': 0.4, 'Fever': 0.6},
}

emission_probability = {
    'Healthy' : {'normal': 0.5, 'cold': 0.4, 'dizzy': 0.1},
    'Fever' : {'normal': 0.1, 'cold': 0.3, 'dizzy': 0.6},
}
```

在这段代码中，起始状态概率 start_probability 表示病人第一次到访时医生认为其所处的 HMM 状态，他唯一知道的是病人倾向于是健康的。这里用到的特定概率分布不是均衡的，如平衡概率大约是 `{'Healthy': 0.57, 'Fever': 0.43}`。状态转换概率 transition_probability 表示潜在的马尔可夫链中健康状态的变化。在这个例子中，当天健康的病人仅有 30% 的机会第二天会发烧。输出概率 emission_probability 表示每天病人感觉的可能性。假如他是健康的，50% 会感觉正常。如果他发烧了，有 60% 的可能感觉到头晕。

病人连续三天看医生，医生发现第一天他感觉正常，第二天感觉冷，第三天感觉头晕。于是医生产生了一个问题：怎样的健康状态序列最能够解释这些观察结果。维特比算法解答了这个问题。

函数 viterbi 具有以下参数: obs 为观察结果序列, 例如 `['normal', 'cold', 'dizzy']`; states 为一组隐含状态; start_p 为起始状态概率; trans_p 为状态转换概率; 而 emit_p 为输出概率。

维特比算法揭示了观察结果 `['normal', 'cold', 'dizzy']` 最有可能由状态序列 `['Healthy', 'Health` 产生。换句话说，对于观察到的活动，病人第一天感到正常，第二天感到冷时都是健康的，而第三天发烧了。

✍ 课外阅读

1. Hidden Markov Models
2. 致老婆的那些模型与算法之韩梅梅
3. 【经典】数据科学家教你用数据模型来恋爱
4. 隐马尔可夫模型（Hidden Markov Model，HMM）理解
5. 史上最详细最容易理解的 HMM 文章
6. 隐马尔科夫模型 HMM 攻略
7. HMM 相关文章索引
8. 隐马尔科夫模型 HMM 自学
9. 隐马尔科夫过程在生物信息学中的应用
10. Eddy, S. R. What is a hidden Markov model? *Nature Biotechnology*, 2004, 22(10), 1315–6.
11. Ghahramani, Z. An Introduction to Hidden Markov Models and Bayesian Networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 2001, 15(1), 9-42.
12. Brent, M. R. How does eukaryotic gene prediction work? *Nature Biotechnology*, 2007, 25(8), 883–5.
13. Mukherjee, S. & Mitra, S. Hidden markov models, grammars, and biology: a tutorial. *Journal of Bioinformatics and Computational Biology* 2005, 3(2), 491–526.
14. Saeys, Y., Inza, I. & Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 2007, 23(19), 2507–2517.
15. Eddy, S. R. What is Bayesian statistics? *Nature Biotechnology*, 2004, 22(9), 1177–1178.

4.9 总结与答疑

本次课涉及的主要知识点与技能:

- 知识点——DNA 序列的基本信息与特征信息分析:
  - DNA 序列基本信息分析——查戈夫法则，GC 含量，序列转换。
  - 限制酶位点分析——命名，II 型的特点。
  - 开放阅读框分析——相位，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——概念、判别依据及标准。
- 技能——解决问题的思路:
  - 首先分析任务的属性。
  - 寻找可能的解决方案。
  - 确定最合适的方法。
  - 先易后难，由浅入深。

4.10 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：计算碱基比例和 GC 含量、序列转换、寻找限制性核酸内切酶位点；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；之后就是对其进行注释。

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。

4.11 重复序列分析

细菌和古细菌的基因组内包含了基因和基因间的小片段[28]。通常原核基因组是环形的，每 1000 个碱基对的长度中含有一个基因。相反地，真核生物基因组包含了小部分的蛋白质编码基因和大量的非编码 DNA 序列。这些非编码序列包括重复 DNA、编码具有调控功能的 RNA 的基因、外显子之间的将被切除出成熟 mRNA 的内含子。真核生物基因组中的非编码序列的很大一部分由重复性 DNA 序列组成。许多不同长度的重复的核苷酸序列遍布整个基因组。在哺乳动物中，大约 60% 的基因组 DNA 是重复的，而在酵母中也达到了 20%。重复性 DNA 有时候也被称为“垃圾 DNA”或者“自私的 DNA”，因为它们不断向整个基因组扩张。但是，重复性 DNA 可能在染色体结构、重组事件和某些基因的功能中起重要作用。

根据生物信息学的观点，许多基因组分析算法（包括数据库搜索程序）都是从“屏蔽”已知重复序列如 SINE 和 LINE 开始的，因为已知在这些重复序列中几乎没有对基因识别或序列比较有用的信息。

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[29]：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。

串联重复序列（tandem repeat）根据重复序列的重复单位的长度可分为卫星 DNA（satellite DNA）、小卫星 DNA（minisatellite DNA）和微卫星 DNA（microsatellite DNA）。卫星 DNA 是一类高度重复序列，在不同真核生物中长度从 5bp 到 200bp 不等，通常存在着几百万个拷贝。虽然有一些卫星 DNA 分散在整个真核基因组中，但大部分集中在着丝粒部位。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 DNA 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。虽然微卫星相对较短但其中的许多序列是非常均匀地分布在整个复杂的真核基因组中的。STR 是存在于人类基因组 DNA 中的一类具有长度多态性的 DNA 序列，不同数目的核心序列呈串联重复排列而呈现出长度多态性，通常多态性片段长度在 100～300bp。一般认为，人类基因组 DNA 中平均每 6～10kb 就有一个 STR 位点，其多态性成为法医物证检验个人识别和亲子鉴定的丰富来源。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

[28]此部分内容摘抄自《生物信息学与功能基因组学》（乔纳森 • 佩夫斯纳 著，孙之荣 译），第 3 篇 基因组分析，第 16 章 真核基因组：从寄生生物到灵长类，16.2 真核生物的普遍特性。

[29]不同分类标准之间的关系。

Element，LINE）。前者长度在 500bp 以下，在人类基因组中的重复拷贝数达 10 万以上；后者长度在 1000bp 以上，在人类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。LINE 不同成员之间的序列有较大差异，但是同一个物种中的 LINE 的不同成员仍有较大的同源性。在灵长类中，主要的 SINE 和 LINE 分别为 Alu 和 L1。Alu 元件的长度约为 300bp，以 100 万个拷贝的形式分布在人类基因组中，是人类基因组中丰度最高的转座元件。人类 L1 重复序列是 LINE 的一个典型例子，该重复序列全长 6100bp，约有 3500 个拷贝，分散在人类基因组中。完整的 L1 重复和几十万个被截短了的片段总共构成了每个哺乳动物基因组的约 5%。

✍ 扩展阅读

重复 DNA 序列约占据人类基因组的 50% 以上。真核基因组中的重复性 DNA 主要分为 5 种[30]：

1. 散布的重复性 DNA（转座子导致的重复）。散在的重复 DNA 组成了人类基因组的大约 45%。这些重复 DNA 可以由拷贝 RNA 中间产物（反转录元件）或者 DNA 中间产物（DNA 转座子）生成。散布性的重复序列可以分成 4 个类：
   (a) 长末端重复（LTR）转座子，是由 RNA 反转录而成的元件。它们也称为类反转录病毒元件。LTR 转座子在两端有长达数百个碱基对的 LTR。占人类基因组 8%，拷贝数 450000，其中自主 LTR 长 6～11kb，非自主 LTR 长 1.3～3kb。
   (b) 长散布元件（LINEs），编码了具有反转录活性的酶，还可能编码了其他蛋白质。在哺乳动物中，LINE1 和 LINE2 是最常见的。占人类基因组 21%，拷贝数 850000，长 6～8kb。
   (c) 短散布元件（SINEs），也是由 RNA 产生的元件。在灵长类动物中发现的 Alu 重复序列，是 SINEs 的著名例子，占人类基因组 13%，拷贝数 1500000，长 100～300bp。
   (d) DNA 转座子，占人类基因组的大约 3%，拷贝数 300000，其中自主 DNA 转座子长 2～3kb，非自主 DNA 转座子长 80～3000bp。

   LINE、SINE 和 LTR 转座子都是具有编码反转录酶活性的反转录转座子，它们通过产生 RNA 中间体整合入基因组。与它们不同，DNA 转座子具有反向末端重复序列，并编码类似细菌转座子的转座酶。反转录转座子可以更细地分为自主的反转录转座子（含有它们转座所需的酶的编码序列）和非自主反转录转座子（依靠外源的酶，如宿主细胞的 DNA 修复酶）。最常见的非自主反转录转座子是 Alu 元件。散在重复序列占人类基因组的比例比在其他真核生物基因组中大得多，拷贝总数估计有 300 万，它们的存在为分子进化的研究提供了有力的工具。
2. 被修饰的假基因。这些基因并不被转录或者翻译。它们曾经是有功能的基因，但是它们缺乏蛋白质产物。因为它们有一个终止密码子或者由阅读框移位导致的中断，所以它们可以被识别出来。假基因可以由反转录转座事件产生，或者由基因的复制产生然后失去其基因功能。
3. 简单重复序列。这些微卫星序列（通常长度为 1～6bp）和小卫星序列（通常长度为 12～500bp 的重复）包含了短的序列，例如 (A)n、(CA)n 或者 (CGG)n。复制过程中的滑移是产生简单重复序列的一个可能原因。简单重复序列在不同的物种中具有不同的长度和位置偏好性。在人类基因组中，简单重复序列引起了人们的极大兴趣，因为它们通常在不同个体间体现了多态性，可以用来作为遗传学标记。
   简单重复序列是完美的或略不完美的特殊 k-mer（k 个碱基）的串联重复。具有短的重复单位（1～12bp）的简单序列重复称为微卫星，具有较长的重复单位的序列（13～500bp）通常称为小卫星。微卫星和小卫星占人类基因组的 3%。最常见的重复单位是二核苷酸 AC、AT 和 AG。
4. 片段复制。这些复制的片段长度为 1～200kb 或 300kb，被从基因组的一个位置拷贝到另一个位置，出现在同一个染色体上或者不同染色体之间。人类基因组包含了大约 3.6% 的复制区域（长度至少 1kb，并具有 90%～98% 的序列相似性），一般长 10～50kb。着丝粒包含了大量的染色体间片段性重复。较小范围的片段性重复也发生在接近端粒的区域。
5. 串联重复序列块，例如在端粒、着丝粒和核糖体基因群附近的串联重复序列块。端粒是在真核生物染色体末端发现的重复性序列。它们维持染色体的稳定性，防止被降解，同时阻止染色体末端之间的融合。着丝点是染色体上一个重要的位置，它为纺锤体的微管丝提供了附着点，使得在细胞进行有丝分裂和减数分裂时染色体能分离开来。所有的真核生物染色体都有有功能的着丝点，虽然其序列可能因

[30]此部分内容摘抄自《生物信息学与功能基因组学》（乔纳森 • 佩夫斯纳 著，孙之荣 译），第 3 篇 基因组分析，第 16 章 真核基因组：从寄生生物到灵长类，16.2 真核生物的普遍特性。

物种的不同而不同。在人类基因组中，着丝点的 DNA 包含了长为 171bp 的 α 卫星 DNA 序列重复，共 1～4Mb。着丝粒附近的卫星序列是所有真核生物的共同特征，只有两个已知的例外。一个例外是，在酵母基因组中，整个着丝粒序列一共只有几百个碱基对的长度。第二个例外是新着丝粒——一种异常的着丝粒，它有功能正常的着丝点，在有丝分裂中稳定，但是缺少 α 卫星 DNA 序列。

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。在真核基因组 DNA 中寻找重复序列（如 SINEs 和 LINEs），对于基因组分析来说是非常必要的。知道重复性序列的位置可以帮助鉴定非编码区域，而重复序列本身也是“分子化石”，可以用在比较基因组学研究中。寻找重复性序列的一个实用方法是对数据库进行搜索，寻找那些已知的重复序列和低复杂度区域。不同重复序列数据库储存了不同类型重复序列的信息：Repbase Update（RU）[31]是常用的真核基因组重复序列数据库，由美国遗传信息研究所（Genetic Information Research Institute，GIRI）创建并维护，收录转座子及其他重复序列的序列和注释信息；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。

为了鉴定和屏蔽重复性序列，可以使用 RepeatMasker 网站服务器。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 由 Arian Smit 和 Robert Hubley 开发，它将把输入序列中已知的重复序列都屏蔽为 N 或 X，并给出相应的重复序列统计表。一般情况下，一条人的基因组数据的大约 50% 会被该程序屏蔽。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。RepeatMasker 的输出信息包含了用 Smith-Waterman 算法得出的一个按照分值排列的列表、重复序列的位置和关于重复序列类型的信息（如 SINE/Alu、LTR 或者简单重复序列）。两两的序列比对结果也被显示出来。输入的序列也用 FASTA 的格式输出，其中的重复性序列用 X 或者 N 屏蔽。这个版本的序列对于后续的数据库搜索尤其有用。最后，RepeatMasker 程序对找到的重复序列提供了一个总结。

4.12　基因识别

理解真核生物基因组的两个最大挑战[32]就是定义什么是基因和鉴定基因组 DNA 中包含的基因。我们将首先定义基因，然后给出鉴定基因的准则。

1. 编码蛋白质的基因组成了基因的大多数。许多准则被用来判断一条 DNA 序列是否是一个编码蛋白质的基因。最原则性的要求是它必须含有一个长度至少为 90bp（对应 30 个编码氨基酸的密码子，或者大约分子量 3ku 的蛋白质）的开放阅读框。
2. 假基因并不编码有功能的基因产物。
3. 除了假基因外，还有其他许多种类的基因，它们也不编码蛋白质，而是编码许多有功能的 RNA 产物。这些基因包括转运 RNA（tRNA）基因。它们将信息从 mRNA 的三联密码子翻译成为氨基酸。
4. 核糖体 RNA（rRNAs）基因也在翻译中执行重要功能。
5. 小核仁 RNA（snoRNAs）在核内执行功能。
6. 小核 RNA 在剪切体中起作用，它把初生的 RNA 中的内含子移除。
7. 微小 RNA（miRNAs）是大约 21～25 个碱基长度的 RNA，在各个物种中保守，可能作为其他 RNA 的反义调节者。

在基因组 DNA 的注释过程中，人们通常注重那些编码蛋白质的基因。但是，现在知道编码各

[31]此部分内容摘抄自《常用生物数据分析软件》（王俊，科学出版社），第 4 章 基因组/基因的注释，4.1 重复序列分析。

[32]此部分内容摘抄自《生物信息学与功能基因组学》（乔纳森 • 佩夫斯纳 著，孙之荣 译），第 3 篇 基因组分析，第 16 章 真核基因组：从寄生生物到灵长类，16.2 真核生物的普遍特性。

种 RNA 的基因也有多种多样的重要功能。

基因（gene）[33]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。[34]

课外阅读

1. 《生命是什么？》，薛定谔。
2. 《双螺旋》、《基因 • 女郎 • 伽莫夫》，沃森。
3. 《创世纪的第八天》，贾德森。
4. 《物理世界奇遇记》、《从一到无穷大》，伽莫夫。

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。

原核与真核生物的基因[35]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过程的调节区，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始位点的一段长度为 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位点和两个区（-35 区和-10 区）。转录起始位点是 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[36]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，与原核基因的启动子相似，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 盒与原核基因启动子的-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[37]，所以使用的基因识别方法也不同[38]。

---

[33] 你对基因概念的理解。

[34] 基因概念的提出与发展。（维基百科）

[35] 原核基因和真核基因最主要的区别是什么？

[36] 区分模板链和编码链：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。

[37] 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）

[38] 井蛙不可以语于海者，拘于虚也；夏虫不可以语于冰者，笃于时也；曲士不可以语于道者，束于教也。——《庄子 •

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质编码基因。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，通过比较相关物种的 DNA 序列，我们能够取得预测基因的新线索。

最有效的程序合并了信号元件和内容自动检测以及同源性搜索，从而产生了寻找基因的集成软件包[39]。

- 信号是不连续的局部序列模体，如启动子，剪切供体和受体位点，起始和终止密码子，以及 poly(A) 尾位点（polyadenylation site）。这类模体一般都有一致性序列（consensus sequence），可开发算法来搜索这些单独信号或结合信号周围背景。这类算法叫作信号传感器（signal sensor）。
- 内容是不同长度的扩展序列。基因最重要的内容通常是编码区，当然其他序列也是有一定关系的，如脊椎动物基因中的 CpG 岛（CpG island）。内容没有一致性序列，但是它们具有把自己与周围 DNA 区分开来的保守特征。例如，核苷酸频率和核苷酸相关性（两个特定的核苷酸同时在两个特定位置出现的可能性）对于外显子和内含子或基因之间的 DNA 来说是不同的。这些可以通过采用统计模型的内容传感器算法来检测。
- 同源性是已知基因之间的匹配。

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，编码蛋白质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

秋水》

[39]此部分内容摘抄自《生物信息学（中译本，第二版）》（王明怡，科学出版社），H 序列注释，H1 基因组注释原理。

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[40]；3’ 端的外显子下游包含终止密码子和 poly(A) 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[41]。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。

✍ 扩展阅读

细菌和太古细菌的基因组都包含附加的、相对较小的基因间区域[42]。一般来说，这些基因组都是环状的，并且基因在基因组组序列中的分布密度大约为每 1000 个碱基对含有一个有效基因。对于原核生物来说，基因可以通过简单确定长开放阅读框（ORF）的出现来确定，这些长开放阅读框的长度一般都远远大于一些系统设定的阈值，如 90 个核苷酸的长度（30 个氨基酸的长度大约为 3ku 蛋白质序列的长度）。这样的程序有 GLIMMER 和 GenMark，能够有效确定基因在细菌基因组序列上的位置。Glimmer 是用来确定原核生物 DNA 序列上基因的一个程序。该程序需要两个输入元素：一个是 DNA 基因序列文件（FASTA 格式的文件），另一个是有关基因识别的马尔可夫模型。

细菌和古细菌[43]具有较高的基因密度（大约每 1000 个碱基 1 个基因），没有内含子和很少有重复 DNA。这样比起在真核生物 DNA 中寻找基因就要相对简单一些了。对于基因识别有用的基因组 DNA 有下面 4 个主要的特点。

1. 开放阅读框长度。一个开放阅读框并不一定就是一个基因。一个开放阅读框被一个起始密码子（如 ATG，编码甲硫氨酸）和一个终止密码子（TAA、TAG、TGA）所定义。
2. 在起始密码子的附近有结合核糖体的一致性序列存在。找到核糖体结合位点是判断起始密码子的重要方法。在细菌中，核糖体结合位点被称为 Shine-Dalgarno 序列。这是一段和 16S rRNA 的 3’ 端互补的多嘌呤核苷酸片段，从-20（起始密码子的 5’ 方向）到 +13（起始密码子的 3’ 方向下游 13 个核苷酸）。Shine-Dalgarno 序列和高基因表达水平呈正相关。
3. 存在和基因一致的密码子使用模型。
4. 推测的基因和其他已知基因有同源性。基因组 DNA 序列（包括推测的基因）能够在蛋白质数据库中用 blastx 进行查询。这个方法对在真核生物中查找基因有帮助。

前 3 个特征被研究并用作搜寻基因的基本特征。它们被称为基本特征是因为这些特征并不决定于其他物种的基因序列的比较。第 4 个特征（和其他基因的联系）被称为非基本特征。一些诸如 Critica 的原核生物基因搜寻程序同时结合了基本特征和非基本特征。GLIMMER 系统是最重要的基因搜寻算法之一。它能够鉴别出细菌基因组中所有基因的 97%～98%。这个算法采用插入马尔可夫模型（IMMs）。

对原核基因的预测也有一些困难。

1. 一个基因组 DNA 片段可能有多个基因在相同方向或者相反方向的可选阅读框中。
2. 很难判断一个短开放阅读框是否真正被转录了。
3. 一个开放阅读框编码一个基因，并有一个终止密码子。如果基因组 DNA 中同一条链上发生连续编码其他阅读框，就可能发生阅读框移动。
4. 有一些基因属于操纵子，经常在原核生物中起着相关作用。操纵子有启动子和终止子序列模体，但是对这些并没有很好地研究过。
5. 细菌和古细菌中经常会发生水平基因转移。

在真核生物基因组中寻找基因是比在原核生物中寻找基因复杂得多的一个问题[44]。因为细菌的基因一般是与长的 ORF 联系在一起的，而真核生物的基因包含了内含子和外显子。外显子有多种：

1. 非编码的外显子即 DNA 序列的 5’ 非编码区和 3’ 非编码区。

[40] 供体位点和受体位点在外显子上还是在内含子上？

[41] 什么是密码子使用偏好性，什么是基因组等值区（isochore）？

[42] 此部分内容摘抄自《生物信息学与功能基因组学》（乔纳森 • 佩夫斯纳 著，孙之荣 译），第 3 篇 基因组分析，第 12 章 全基因组和系统发生树，12.8 基因组分析总览。

[43] 此部分内容摘抄自《生物信息学与功能基因组学》（乔纳森 • 佩夫斯纳 著，孙之荣 译），第 3 篇 基因组分析，第 14 章 已完成测序的基因组：细菌和古细菌基因组，14.8 原核基因组分析。

[44] 此部分内容摘抄自《生物信息学与功能基因组学》（乔纳森 • 佩夫斯纳 著，孙之荣 译），第 3 篇 基因组分析，第 16 章 真核基因组：从寄生生物到灵长类，16.2 真核生物的普遍特性。

2. 包括启动位点的甲硫氨酸在内的起始外显子，直到第一个 5’ 剪切位点。
3. 内在的外显子，从 3’ 剪切位点开始，到 5’ 剪切位点结束。
4. 最后一个外显子，从 3’ 剪切位点到终止子。
5. 单外显子基因是没有内含子的，从起始密码子开始，到终止密码子结束。

除了具有内含子外，真核生物的基因占基因组的比例远远小于原核生物基因占基因组的比例。真核生物编码蛋白质的基因在线虫和昆虫基因组中只占 25%，而在人类和小鼠中只占不到 3%。

由于基因在生物学中的重要地位[45]，研究人类基因组的基因含量成为人们关注的焦点。然而，辨别基因是在人类基因组序列识别中最难的部分，是项具有挑战性的工作，原因如下。

1. 外显子的平均长度只有 50 个密码子（150 个核苷酸），这样小的序列难以明确地识别。
2. 外显子被内含子隔开，有些间隔达数万碱基对。
3. 存在许多假基因，它们和编码功能蛋白的真基因难以区分。
4. 我们对非编码基因还很不了解。

在真核生物中找寻编码蛋白质的基因的算法可以被分为两类：基于同源性的（也称为外在的）和基于算法的（也称为内在的）。基于同源的方法通常包括了对已表达的基因的比对（来自 cDNA 文库的 EST）。在这种情况下，ESTs 可以帮助定义基因组 DNA 的外显子和内含子结构。因此，基于同源的方法通常是很成功的。另外一种基于同源鉴定基因的形式是把两个亲缘关系相近的物种基因组 DNA 进行比较。

虽然用 EST 数据注释真核生物基因是非常有帮助的，但是这种方法也有一些限制条件。

1. EST 数据的质量通常很低，只有一条链上的克隆，而且经常有序列上的错误。
2. 高表达量的基因通常过于不成比例地出现，虽然一些 cDNA 文库已经经过归一化。
3. EST 数据并不提供关于基因定位的信息。

内在性的方法也被广泛用于注释基因组 DNA。有相当一部分基因既没有已经鉴定的同源基因，也没有可以使用的 EST 数据，因此必须使用从头预测的方法来鉴定编码蛋白质的基因。

算法利用编码的统计特性（例如编码的和非编码的区域的核酸序列的不同特性）和各种信号（例如，启动子元件、起始和终止密码子、剪切位点和多腺嘌呤位点）。

当前的基因预测并不完美[46]，主要存在假阳性（False Positive，FP）、假阴性（False Negative，FN）、过界预测（Over Prediction，OP）、片段化（Fragmentation）和融合化（Fusion）等问题。评价预测的准确性是用 cDNA 定位或已知基因结构作为基准的。值得注意的是：对于一个基因，只能预测出一种剪接形式，而无法识别可变剪接；只能预测从起始密码子到终止密码子的区域，而不能预测两端的 UTR 区域；对于高等生物，基因组中存在着大量的转座子重复序列，将对基因预测造成严重的影响。

- 假阳性：多预测了假的编码区，即在非编码区预测出编码区。
- 假阴性：漏掉了真实的编码区，即将编码区预测为非编码区。
- 过界预测：由于基因的边界很难准确定位，预测经常会超出实际的边界。
- 片段化：内含子过大的基因，在预测时容易断裂成两个或多个基因。
- 融合化：距离过近的两个或多个基因，在预测时容易被融合成一个很大的基因。

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。虽然不同的

[45]此部分内容摘抄自《生物信息学与功能基因组学》（乔纳森 • 佩夫斯纳 著，孙之荣 译），第 3 篇 基因组分析，第 17 章 人类基因组，17.4 人类基因组计划。

[46]此部分内容摘抄自《常用生物数据分析软件》（王俊，科学出版社），第 4 章 基因组/基因的注释，4.3 基因预测。

软件使用的各种算法都不尽相同，但是它们都有一些共同的缺点，即过高的假阳性率以及不能准确预测出基因的完整结构。

GENSCAN[47]是由美国麻省理工大学的 Burge 和 Karlin 于 1997 年开发的，基于广义隐马尔可夫模型的人类及脊椎动物基因预测软件，它不依赖于已有的蛋白质库，是一种“从头预测”的软件。GENSCAN 是从未注释的基因组序列中寻找基因的经典软件，是基因识别领域的第四代软件，所用的数学方法是构造基因的 HSMM 概率模型（隐半马尔可夫模型）。目标是确定编码外显子的位置，预测序列中的基因个数，既能处理完整的基因，也能处理不完整的基因，而且能够包含在正负两条链上的基因。基因模型中包含的基因特征有：剪接信号模型，外显子长度分布，启动子和 ploy(A) 信号。考虑到不同的 $C+G$ 组份区域在基因密度和结构（例如内含子长度）方面的差别，GENSCAN 还能够对自己的预测的可靠性给出有用的估计。还考虑了亚最优外显子，它们对应于基因的异常剪接。GENSCAN 目前还开发了适用于果蝇、拟南芥和玉米的专用版本，对于其他物种可以先采用相近的物种版本来预测。总体来说，对中间外显子预测的准确性高于起始外显子和末端外显子，外显子的准确性高于 poly(A) 或启动子。继 GENSCAN 之后发展起来的基因预测软件有 Fgenesh、BGF、GlimmerM 等。Glimmer 预测系统先用 build-icm 程序对该物种已知的基因序列生成一个马尔可夫模型参数集合，Glimmer 再应用这个参数集对 DNA 序列进行基因预测。此软件适合对原核生物进行预测。GlimmerM 是 TIGR 最早开发用于预测 *Plasmodium falciparum*（恶性疟原虫）的一个软件，适用于基因密度在 20% 左右的小的真核生物的基因预测。现在此软件已经训练应用于 *Arabidopsis thaliana*（拟南芥）、*Oryza sativa*（水稻）和 *Aspergillus fumigatus*（烟曲霉）等几个物种的基因预测。GlimmerM 应用动态规划算法，基因预测基于以下几个假设：

1. 每个基因的编码区开始于起始密码子 ATG；
2. 除了最后一个密码子外基因没有阅读框内的终止密码子；
3. 每个 exon 都和前一个 exon 在一致的阅读框内。

此软件不能判断移码。GlimmerM 一次只能对一条序列进行预测，如果基因组含多条序列，需拆分成多个文件，每个文件只含有一条 FASTA 序列。

✍ 课外阅读

1. Zhang, M. Q. Computational prediction of eukaryotic protein-coding genes. Nat. Rev. Genet. 3, 698–709 (2002).
2. Brent, M. R. How does eukaryotic gene prediction work? Nat. Biotechnol. 25, 883–5 (2007).
3. Harrow, J. et al. Identifying protein-coding genes in genomic sequences. Genome Biol. 10, 201 (2009).
4. Rose, D. et al. Computational discovery of human coding and non-coding transcripts with conserved splice sites. Bioinformatics 27, 1894–900 (2011).
5. Alioto, T. Chapter 6, Gene prediction. Evolutionary Genomics 855, 175-201 (2012).
6. List of gene prediction software(Wikipedia)
7. Computational prediction of eukaryotic protein-coding genes, Box 2, Useful internet resources

4.13 查找数据库与分析工具

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[48]，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：

- 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。

[47]此部分内容摘抄自《常用生物数据分析软件》（王俊，科学出版社），第 4 章 基因组/基因的注释，4.3 基因预测。
[48]如何查找需要的数据库和工具？

- The Elements of Bioinformatics。
- 使用 Google 等搜索引擎搜索。
- 图书馆。

4.14 总结与答疑

本次课涉及的主要知识点与技能：

- 知识点——重复序列和基因识别：
  - 重复序列——分类。
  - 基因识别——原核和真核的基因结构，基因识别方法与策略。
- 技能——查找数据库与分析工具：
  - 借鉴文献、收集专刊、请教专家、搜索网络。
  - 数据库有其时效性。
  - 分析工具有其适用范围。

4.15 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。这些都是针对 DNA 序列进行的分析。

DNA、RNA 和蛋白质是三种重要的生物大分子[49]。传统的观念认为 DNA 携带着遗传信息，蛋白质是体现生物功能的分子，而 RNA 在这两者间起传递遗传信息的桥梁作用。通常把 RNA 按照是否编码蛋白质分成两大类：（1）编码 RNA，也就是我们经常说的 mRNA。（2）非编码 RNA，如核糖体 RNA、转运 RNA、核仁 RNA、小 RNA 等。DNA 的遗传信息决定生命的主要性状，而 mRNA 在信息传递中起很重要的作用。其他两大类 RNA：rRNA 和 tRNA，同样在蛋白质的生物合成中发挥不可替代的重要功能。

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）[50]。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[51] 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。与之相反，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。现在有关 lncRNA 的研究已经引起国内外的广泛关注，对其认识和研究正逐步深入。

✍ 扩展阅读

DNA 转录生成 RNA[52]。RNA 既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、

[49] 此部分内容摘抄自《常用生物数据分析软件》（王俊，科学出版社），第 4 章 基因组/基因的注释，4.2 RNA 分析。

[50] ncRNA 包括哪些？

[51] nt 和 bp 分别是什么的缩写，代表什么含义？（nt：nucleotide；bp：base pair）[严格地说，碱基对是一对相互匹配的碱基（即 A:T，G:C，A:U 相互作用）被氢键连接起来。然而，它常被用来衡量 DNA 和 RNA 的长度（尽管 RNA 是单链）。它还与核苷酸互换使用，尽管后者是由一个五碳糖、磷酸和一个碱基组成。碱基对通常简写做 bp（base pair），千碱基对为 kbp，或简写作 kb（对于双链核酸；对于单链核酸，kb 指千碱基）。]

[52] RNA 包括哪些种类？

rRNA、miRNA、siRNA、lncRNA 属于非编码 RNA。

人类基因中有许多并不编码蛋白质[53]，而以非编码 RNA 为其产物。非编码 RNA 在基因组 DNA 中难以识别，因为它们较短小、缺乏开放阅读框、也没有 poly(A)，基因寻找算法难以检测，也不存在于 cDNA 文库中。这类非编码 RNA 包括：

1. 转运 RNA（tRNA），用于将 mRNA 翻译成蛋白质的氨基酸序列；
2. 核糖体 RNA（rRNA），用于 mRNA 的翻译；
3. 小核仁 RNA（snoRNA），是核仁中 RNA 加工过程所必需的；
4. 小核 RNA（snRNA），剪接体行使功能所必需的。

前面主要介绍了 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：先介绍转录产物 mRNA 选择性剪接的分析，然后对 miRNA 及其靶基因的预测进行讲解，最后简要介绍一下近几年备受瞩目的长链非编码 RNA。

4.16 mRNA 选择性剪接

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[54]。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。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。

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不同的组合，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

[53]此部分内容摘抄自《生物信息学与功能基因组学》（乔纳森 • 佩夫斯纳 著，孙之荣 译），第 3 篇 基因组分析，第 17 章 人类基因组，17.4 人类基因组计划。

[54]剪接有几种类型？其分子生物学过程是怎样的？。

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[55]。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。

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 poly(A) 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。[56]

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。

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[57]，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 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的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）[58]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[59]。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。

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。剪接调节因子主要有外显子剪接增强子（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

[55] 外显子跳跃和内含子保留的区别是什么？

[56] 生物系统的复杂性导致多种观点并存，无法统一。

[57] 真实情况远比理论模型复杂！

[58] 英语中 ASAP 是什么的缩写，有何含义？其他缩写名比较有趣的数据库或工具：TCGA——The Cancer Genome Atlas；TiGER——Tissue-specific Gene Expression and Regulation；DAVID——The Database for Annotation, Visualization and Integrated Discovery；GOLD——Genome Online Database；COSMIC——Catalogue of somatic mutations in cancer；GREAT——Genomic Regions Enrichment of Annotation Tool；……

[59] ASAP 网站已失效，提醒我们要注意数据库的更新日期。

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。

4.17 miRNA 及其靶基因预测

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。[60]

miRNA 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。[61]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。

研究表明 miRNA 在序列、表达、调控、物理位置等方面主要有如下特征[62]：

1. 在序列特征上主要有两方面特点，即 miRNA 本身不具有开放阅读框，不编码蛋白质；成熟的 miRNA 5’ 端为单一磷酸基团，3’ 端为羟基；
2. miRNA 的表达具有时序性以及组织特异性；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 在细胞生长和发育的调节过程中起多种作用。
3. miRNA 与其靶基因间呈多对多的调控关系，即一个 miRNA 可能调控多个靶基因，而一个基因也可能受多个 miRNA 调控；
4. miRNA 的物理位置倾向于成簇地出现在染色体上；
5. miRNA 还具有在物种间高度保守的特点。

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。[63]

---

[60]比较 miRNA 和 siRNA。（miRNA 与 siRNA 之间有许多相同之处：1. 二者的长度都约在 22nt 左右。2. 二者都依赖 Dicer 酶的加工，是 Dicer 的产物，所以具有 Dicer 产物的特点。3. 二者生成都需要 Argonaute 家族蛋白存在。4. 二者都是 RISC 组分，所以其功能界限变得不清晰，如二者在介导沉默机制上有重叠。5.miRNA 和 siRNA 合成都是由双链的 RNA 或 RNA 前体形成的。miRNA 与 siRNA 的不同点：1. 根本区别是 miRNA 是内源的，是生物体的固有因素；而 siRNA 是人工体外合成的，通过转染进入人体内，是 RNA 干涉的中间产物。2. 结构上，miRNA 是单链 RNA，而 siRNA 是双链 RNA。3.Dicer 酶对二者的加工过程不同，miRNA 是不对称加工，miRNA 仅是剪切 pre-miRNA 的一个侧臂，其他部分降解；而 siRNA 对称地来源于双链 RNA 的前体的两侧臂。4. 在作用位置上，miRNA 主要作用于靶标基因 3′ UTR 区，而 siRNA 可作用于 mRNA 的任何部位。5. 在作用方式上，miRNA 可抑制靶标基因的翻译，也可以导致靶标基因降解，即在翻译水平和转录后水平起作用，而 siRNA 只能导致靶标基因的降解，即为转录后水平调控。6.miRNA 主要在发育过程中起作用，调节内源基因表达，而 siRNA 不参与生物生长，是 RNAi 的产物，原始作用是抑制转座子活性和病毒感染。）

[61]查阅 miRNA 的具体生成过程。

[62]此部分内容摘抄自《生物信息学》（李霞，人民卫生出版社），第十六章（miRNA 与复杂疾病），第二节（miRNA 与其靶基因）。

[63]差之毫厘而谬以千里，为什么仅仅完全、不完全互补就会导致完全不同的结果？

✍ 扩展阅读

miRNA 基因是一类高度保守的基因家族，按其作用模式不同可分为三种[64]：第一种以线虫 lin-4 为代表，作用时与靶标基因不完全互补结合，进而抑制翻译而不影响 mRNA 的稳定性（不改变 mRNA 丰度），这种 miRNA 是目前发现最多的种类；第二种以拟南芥 miR-171 为代表，作用时与靶标基因完全互补结合，作用方式和功能与 siRNA 非常类似，最后切割靶 mRNA；第三种以 let-7 为代表，它具有以上两种作用模式：当与靶标基因完全互补结合时，直接靶向切割 mRNA，如果蝇和 Hela 细胞中 let-7 直接介导 RISC 分裂切割靶 mRNA；当与靶标基因不完全互补结合时，起调节基因表达的作用，如线虫中的 let-7 与靶 mRNA 3’ 端非翻译区不完全配对结合后，抑制靶标基因的翻译。

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 2000 个不同的 miRNAs，调控至少 30% 以上基因的表达。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 作用机理的深入研究，以及利用最新的例如 miRNA 芯片、miRNA-seq 等高通量技术手段对于 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。

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 序列能较好得互补配对结合，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：

[64]此部分内容摘抄自《常用生物数据分析软件》（王俊，科学出版社），第 4 章 基因组/基因的注释，4.2 RNA 分析。

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。

miRBase[65]是一个集 miRNA 序列、注释信息以及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（网址：http://www.mirbase.org）。miRBase 提供便捷的网上查询服务，允许用户使用关键词或序列在线搜索已知的 miRNA 和靶基因信息。该数据库主要包括三部分内容，即 miRBase Registry、miRBase Sequence 以及 miRBase Targets。miRBase Registry 主要是为新发现的 miRNA 命名服务；miRBase Sequence 包含所有已发布的成熟 miRNA 序列，同时提供对应的预测的发卡结构、注释信息以及与其他数据库的链接；miRBase Targets（已更名为 microCosm，网址：http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5）主要采用 miRanda 算法预测 miRNA 靶基因。TarBase[66]是一个目前使用最广泛的存储已被实验验证的真实 miRNA 与靶基因间关系的数据库，其网址为：http://diana.cslab.ece.ntua.gr/tatbase，存储来自大约 200 篇文献、涵盖多种实验方法的 1333 个 miRNA 与靶基因关系对。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）。

4.18 lncRNA 简介

目前已知在人类基因组中，lncRNA（长链非编码 RNA，long non-coding RNAs）基因[67]的数目已经达到了 15779（GENCODE，version 28；蛋白质编码基因的数目为 19901）。lncRNA 作为一类特殊的 RNA 分子，在生命体内发挥着重要的生物功能，因其独特的生物特征和复杂的生物功能引起了国内外的广泛关注，对其研究正逐步深入。

对于长度小于 200nt 的小非编码 RNA 的生物特征已经开展了广泛的研究。小非编码 RNA 在序列和结构上存在很多标志性的特征，例如，miRNA 前体可以形成发卡结构，在酶的作用下，在茎结构处产生成熟的 miRNA。因此不同类型小非编码 RNA 预测、靶位点查找和功能分析等都取得了快速的进展。与之不同，lncRNA 没有明显的、独特的生物特征，一般不能通过单一的方法鉴定和识别，其长度大于 200nt，产生方式类似 mRNA，具有许多与其类似的特征。

lncRNA 的基因结构类似 mRNA，具有 5′ 端帽子结构、3′ 端聚腺苷酸尾巴以及剪接现象等类似于蛋白质编码基因的特性，因此有时也将其称为 mRNA-like 的非编码 RNA。然而，lncRNA 在序列、结构和功能等各个层面都体现了其独特的生物特征：lncRNA 大多数都不存在较长的 ORF；其密码子偏好性与内含子区域相似，与 mRNA 的密码子偏好性不同；lincRNA 的外显子与随机选取的基因间区域相比具有明显的序列保守性，但低于蛋白质编码基因的外显子区域，高于其内含子区域，lincRNA（long intergenic non-coding RNAs）的启动子区域和剪接位置具有极高的保守性，类似蛋白质编码基因；lncRNA 二级结构中丰富的长茎发夹结构或许暗示了其功能特异性。[68]

随着研究不断深入，大量的 lncRNA 序列数据和表达数据陆续产生，其作为有功能的 RNA 分子这一观点正在被不断证实，然而绝大多数 lncRNA 的功能未知。研究表明 lncRNA 生物功能的多样性，同时其生物功能也体现了特异性：（1）lncRNA 与特定的生物过程相关，其表达具有时空特异性；（2）lncRNA 具有复杂的调控功能，其在染色质改变、转录调控及转录后调控中发挥重要作

[65]此部分内容摘抄自《生物信息学》（李霞，人民卫生出版社），第十六章（miRNA 与复杂疾病），第二节（miRNA 与其靶基因）。

[66]此部分内容摘抄自《生物信息学》（李霞，人民卫生出版社），第十六章（miRNA 与复杂疾病），第二节（miRNA 与其靶基因）。

[67]以学校等社会机构进行类比，垃圾 DNA 真的是垃圾吗？

[68]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。（此段是自己的总结）

用；（3）lncRNA 可能具有复杂的代谢机制，大多数 lncRNA 是稳定的，其半衰期的变化范围较大，半衰期的平均值要小于 mRNA，这可能暗示 lncRNA 具有更复杂的代谢机制和更广泛的功能；（4）lncRNA 与疾病存在密切关系[69]。

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）。

4.19　学习数据库与分析工具的使用

查找到所需要的数据库与分析工具后，如何快速掌握它的使用方法与步骤呢？以下经验可供参考：

- 阅读官方的帮助手册。
- 请教有使用经验的专家。
- 查找简单的使用实例，并重复其操作步骤。
- 使用 Google 等搜索引擎搜索相关资料。
- 参考相关的专业书籍。
- 各种 protocols 期刊：*Nature protocols, Current Protocols (in Bioinformatics/Human Genetics/Molecular Biology/...), SpringerProtocols, Methods in Molecular Biology, ...*。

4.20　总结与答疑

本次课涉及的主要知识点与技能：

- 知识点——mRNA 选择性剪接和 miRNA 分析：
  - 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——学习数据库与分析工具的使用：
  - 阅读手册、请教专家、重复实例、搜索网络。
  - 历史资料使用的是历史版本。

课后思考

1. DNA 序列主要携带哪两类遗传信息？基于这两类遗传信息，可以对 DNA 序列进行哪些分析？
2. 简述限制性核酸内切酶的命名规则及 II 型限制酶的主要特点。
3. 简述 CpG 岛的概念及其识别依据和判别标准。
4. 简述重复序列依重复次数和组织形式的分类。
5. 简述基因识别的三大类方法及策略。
6. 简述选择性剪接的产生机制。
7. 简述 miRNA 预测和 miRNA 靶基因预测的方法。
8. 以计算 GC 含量为例，论述解决思路，即如何通过分析问题的属性确定相应的策略从而找到最合适的方法。
9. 在解决生物信息学问题时，论述找到所需数据库和分析工具并掌握其使用方法的策略。

[69] 查阅 lncRNA 与疾病关系的相关资料。