天津医科大学实验课教案首页

课程名称：生物信息学　　**实验名称：**实验七 原核生物和真核生物的基因识别

教师姓名：伊现富　　**职称：**讲师　　**教学日期：**2017 年 5 月 19 日 13:30-16:30

授课对象：基础医学院 2014 级基础班（本）　　**实验人数：**18

实验类型（验证型、综合型、设计型、创新型）：验证型　　**实验分组：**一人一机

学时数：3　　**教材版本：**生物信息学实验讲义（自编教材）

实验目的与要求：

- 了解隐马尔科夫模型在基因识别中的应用。
- 掌握原核基因和真核基因的结构特征。
- 掌握 GeneMarkS 和 GENSCAN 的使用方法。

实验内容及学时分配：

- (10') 基因与基因识别：回顾基因和基因识别的基本概念，总结比较原核生物和真核生物基因的结构特点。
- (10') 基因识别的方法：回顾基因识别的三大类方法，总结比较原核和真核基因预测中“信号”和“内容”的异同。
- (10') 基因识别的工具：简单介绍 GeneMarkS 和 GENSCAN，重点强调两者的适用范围。
- (120') 实验操作：对大肠杆菌基因组序列进行基因识别，对人类 CD9 基因进行结构分析。

主要仪器和实验材料：

- 实验材料：大肠杆菌基因组，人类 CD9 基因。
- 主要仪器：联网的计算机。
- 分析工具：GeneMarkS，GENSCAN。

实验重点、难点及解决策略：

- 难点：FASTA 格式与纯序列的区别；解决策略：通过实例进行讲解。
- 重点：GeneMarkS 和 GENSCAN 的使用；解决策略：通过练习熟练掌握。

思考题：

- 原核基因和真核基因的结构有何异同？
- 基因识别的方法主要有哪三大类？
- 原核和真核基因预测中的“信号”和“内容”有何异同？
- GeneMarkS 和 GENSCAN 的适用范围分别是什么？
- GeneMarkS 和 GENSCAN 对输入格式的要求有何差别？

参考资料：

- NCBI
- GeneMarkS
- GENSCAN

一、基因与基因识别（10 分钟）

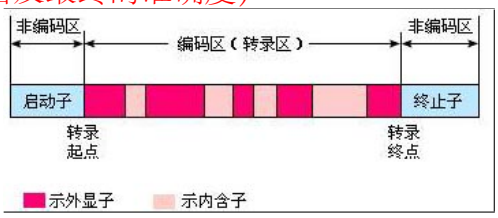
1. 基本概念

- 基因：产生一条多肽链或功能 RNA 所需的全部核苷酸序列（强调既包括编码区，也包括非编码区）

- 基因识别：识别 DNA 序列上具有生物学特征的片段

2. 基因结构（基因结构的复杂性直接影响着基因预测的策略及最终的准确度）

- 共同点：都包括编码区和非编码区
- 原核基因：连续基因
- 真核基因：不连续性



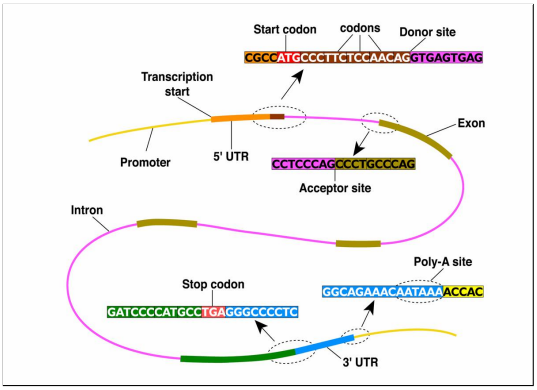
二、基因识别的方法（10 分钟）

1. 识别方法

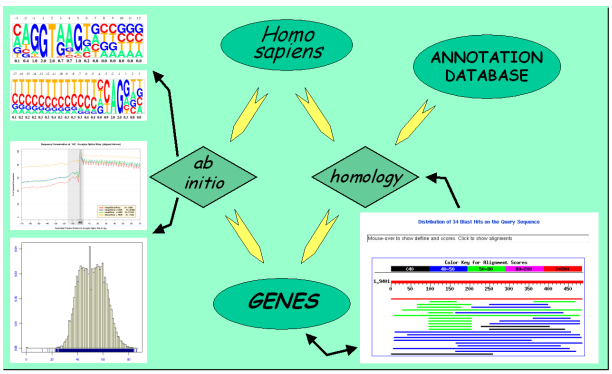
- 间接识别法：mRNA/蛋白质序列 ⇒ DNA 序列
- 从头预测法：基因预测，基于“信号”和“内容”两类特征
- 比较基因组学的方法：比较相关物种的 DNA 序列

2. 基因预测

- “信号”和“内容”
  - 共同点：都包括编码区和非编码区
  - 信号：不连续的局部序列模体，一般都有一致性序列；如启动子，剪接供体和受体位点，起始和终止密码子，polyA 位点
  - 内容：不同长度的扩展序列，没有一致性序列，但具有把自己与周围 DNA 区分开来的保守特征；如密码子使用偏好性，双联密码子出现频率，基因组等值区



- 原核基因
  - 信号：启动子序列，转录因子结合位点
  - 内容：连续的开放阅读框，统计学特征
  - 总结：信号容易识别，内容容易判别，预测能达到相对较高的精度
- 真核基因
  - 信号：启动子区特征序列，供体和受体位点，起始和终止密码子，polyA 序列；确定外显子的边界，识别编码区域
  - 内容：密码子使用偏好性，双联密码子出现频率，基因组等值区；区分外显子、内含子和基因间区域
  - 总结：信号复杂，内容难判别，预测相当有挑战性；联合信号和内容检测以及同源性搜索，提高识别效率



三、基因识别的工具（10 分钟）

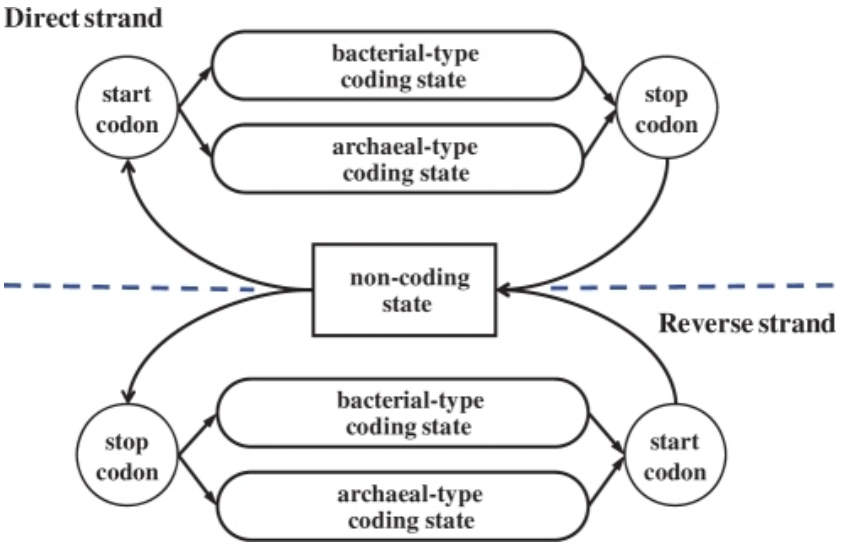
分析工具都有自己的适用范围。

1. GeneMarkS：迭代隐马尔科夫模型，适用于原核生物的基因预测
2. GENSCAN：广义隐马尔科夫模型，人类及脊椎动物基因预测软件

四、实验操作（120 分钟）

基因结构的复杂性直接影响着基因预测的准确度。

1. 大肠杆菌基因组序列的基因识别
   - 获取序列：NCBI 中的 Nucleotide 数据库，U00096，FASTA 格式和 GenBank 格式（复习 GenBank 格式）
   - 截取序列：EMBOSS，extractseq，1-10000bp
   - 基因预测：GeneMarkS，FASTA 格式
   - 结果分析：和 GenBank 格式中的信息进行比较



2. 人类 CD9 基因的结构分析
   - 获取序列：NCBI 中的 Nucleotide 数据库，AY422198，FASTA 格式和 GenBank 格式
   - 基因预测：GENSCAN，纯序列（注意不是 FASTA 格式）
   - 结果解析
     - Type
       * Init: initial exon
       * Intr: internal exon
       * Term: terminal exon
       * Sngl: single-exon gene
       * Prom: promoter region
       * PlyA: polyA signal
     - P
       * P>0.99：可能性极高的外显子
       * 0.50<P<0.99：中等或高可能性的外显子
       * P<0.50：低可能性的外显子
   - 结果分析：和 GenBank 格式中的信息进行比较