天津医科大学理论课教案首页

课程名称：生物信息学　　**课程内容/章节：**第四章（4.1）DNA 序列信息分析

教师姓名：伊现富　　**职称：**讲师　　**教学日期：**2015 年 10 月 9 日 10:00-12:00

授课对象：生物医学工程与技术学院 2013 级生信班（本）　　**听课人数：**28

授课方式：理论讲授　　**学时数：**2　　**教材版本：**生物信息学：基础及应用

教学目的与要求（分掌握、熟悉、了解、自学四个层次）：

- 掌握限制性核酸内切酶的命名规则及 II 型限制酶特点；CpG 岛的概念及其识别依据和判别标准。
- 熟悉 DNA 序列分析的常见内容；ORF 分析中相位的概念；原核和真核基因启动子的结构。
- 了解 DNA 携带的两类遗传信息；DNA 序列分析相关的数据库和工具；ORF 和 CDS 的定义与区别。
- 自学 DNA 序列分析数据库和工具的使用方法。

授课内容及学时分配：

- (5’) 引言与导入：回顾中心法则，阐释核酸序列携带的两类遗传信息。
- (30’) DNA 组份分析与序列转换：回顾 Chargaff 法则，讲解 GC 含量的定义与计算，介绍组份分析和序列转换的原理和思路，讨论解决问题的基本策略。
- (15’) 限制性核酸内切酶位点分析：讲解限制性核酸内切酶的概念、命名规则和 II 型限制酶的特征，介绍常用的数据库与分析工具。
- (10’) 开放阅读框分析：讲解相位的概念以及 ORF 与 CDS 的定义和区别，介绍常用的 ORF 分析工具。
- (10’) 启动子分析：讲解启动子与转录因子的基本概念，回顾原核基因和真核基因启动子的结构，介绍相关数据库与工具。
- (15’) CpG 岛识别：讲解 CpG 岛的概念、识别依据和判别标准，介绍识别 CpG 岛的计算工具。
- (10’) EMBOSS 简介：介绍 EMBOSS 软件包及其中常用的程序。
- (5’) 总结与答疑：总结授课内容中的知识点与技能，解答学生疑问。

教学重点、难点及解决策略：

- 重点：限制酶的命名规则，CpG 岛的识别依据和判别标准。
- 难点：开放阅读框中相位的概念。
- 解决策略：通过示意图和实例帮助学生理解、记忆。

专业外语词汇或术语：

中心法则（central dogma）
GC 含量（GC content）
限制性核酸内切酶（restriction endonuclease）
开放阅读框（Open Reading Frame，ORF）
编码序列（Coding Sequence，CDS）
启动子（promoter）
转录因子结合位点（TFBS）
CpG 岛（CpG island）

辅助教学情况：

- 多媒体：展示中心法则、开放阅读框相位、启动子结构等的示意图。
- 板书：序列的书写惯例，限制酶的命名规则，CpG 岛的识别依据和判别标准。

复习思考题：

- 简述 DNA 携带的两类遗传信息及常见的分析内容。
- 简述限制酶的命名规则及 II 型的主要特点。
- 简述 ORF 与 CDS 的定义和区别。
- 简述 CpG 岛的概念、识别依据和判别标准。
- 论述分析任务属性和解决问题的基本策略。

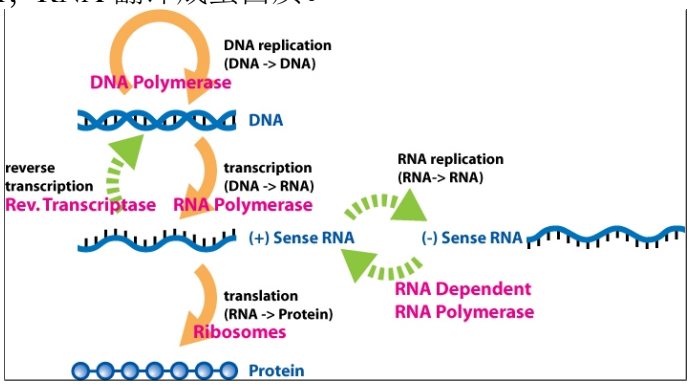
参考资料：

- 朱玉贤，李毅，郑晓峰。现代分子生物学（第 3 版），高等教育出版社，2007。
- 维基百科。

主任签字：　　**年　月　日**　　**教务处制**

一、引言与导入（5 分钟）

1. 分子生物学的中心法则：DNA 转录成 RNA，RNA 翻译成蛋白质。
   - DNA：携带最原始的决定个体性状的遗传信息
   - RNA：参与遗传信息的表达和调控
   - 蛋白质：执行特定的生物功能从而决定最终的表型
   - 排列顺序蕴含生物信息：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（通过类比进行说明）



2. DNA 携带两类遗传信息
   - 功能序列：具有功能活性的 DNA 序列，遗传的基本单位
   - 调控信息：特定的 DNA 区域，能被功能性蛋白质分子特异地识别结合
3. DNA 序列分析
   - 基本信息：碱基组份，GC 含量，序列转换，限制性核酸内切酶位点，……
   - 特征信息：开放阅读框，启动子，转录因子结合位点，CpG 岛，……

二、DNA 组份分析与序列转换（30 分钟）

以 Chargaff 法则引申出序列组份分析、序列转换的内容与原理。

1. Chargaff 法则
   - $A = T, G = C \Rightarrow$ 序列长度，碱基数目及比例，序列转换
   - AT/GC 的比值因生物种类不同而异 $\Rightarrow$ GC 含量
2. GC 含量
   - 鸟嘌呤（G）和胞嘧啶（C）所占的比例
   - GC content: $\frac{G+C}{A+T+G+C} \times 100$
   - GC ratio: $\frac{A+T}{G+C}$
3. 序列转换
   - 反向序列，互补序列
   - 反向互补序列 $\Rightarrow$ 序列书写惯例
   - 显示 DNA 双链
   - 转换为 RNA 序列
4. 序列书写惯例
   - 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]
5. 分析解决问题的策略
   - 以计算 GC 含量为例（使用简单例子易于学生理解）
   - 任务属性决定解决策略（使用序列长短、数目多少的实例进行讲解）

三、限制性核酸内切酶位点分析（15 分钟）

1. 限制性核酸内切酶
   - 定义：识别 DNA 特异序列、并在识别位点或其周围切割双链 DNA 的内切酶
   - 【重点】命名规则（以 *Eco*RI 为例）
     - 属名的第一个字母
     - 种名的前两个字母
     - 细菌的菌株/品系
     - 同一品系中的发现顺序

Derivation of the EcoRI name		
Abbreviation	**Meaning**	**Description**
E	*Escherichia*	genus
co	*coli*	species
R	RY13	strain
I	First identified	order of identification in the bacterium

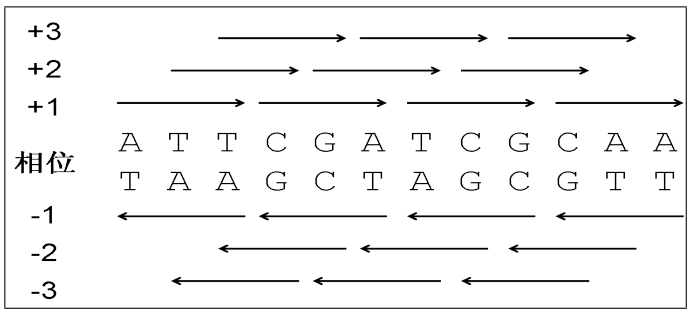
- II 型限制酶的特点（以 *Eco*RI、*Alu*I 等实例加深学生的印象）
  - 识别、切割位点专一
  - 识别序列：4-8 个碱基，回文对称结构
  - 切割序列：识别序列，切割位点对称
  - 切割末端：黏性末端，平滑末端
  - 黏性末端：切割位点在回文序列的一侧
  - 平滑末端：切割位点在回文序列的中间

酵素名称	来源	辩识序列	切法
EcoRI	*Escherichia coli*	5'GAATTC 3'CTTAAG	5'---G AATTC---3' 3'---CTTAA G---5'
*AluI**	*Arthrobacter luteus*	5'AGCT 3'TCGA	5'---AG CT---3' 3'---TC GA---5'

2. 相关资源
   - 数据库：REBASE 收录了限制酶的所有信息
   - 分析工具：NEBCutter V2.0 产生 DNA 序列的酶切位点分析结果

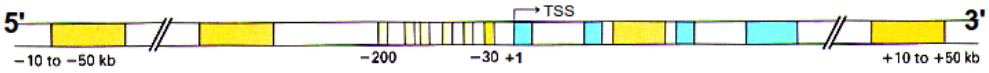
四、开放阅读框分析（10 分钟）

1. ORF：开放阅读框
2. 【难点】frame：相位（通过示意图加深理解）
3. CDS：编码序列
4. ORF vs. CDS：理论预测 vs. 实验证实
5. 分析工具：ORF Finder

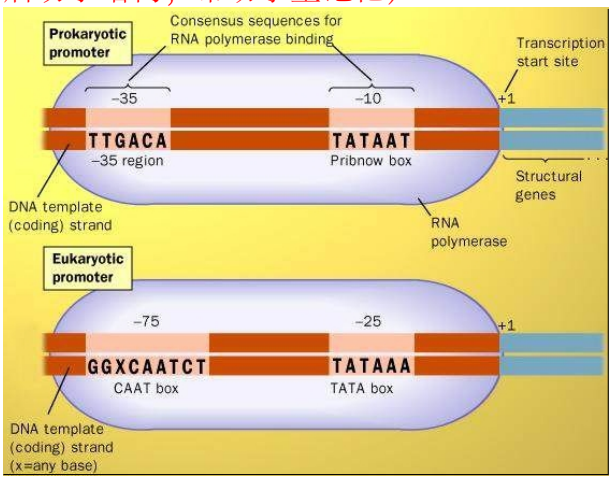


五、启动子分析（10 分钟）

1. 转录调控
   - 顺式作用元件：核酸序列 ⇒ 启动子
   - 反式作用因子：蛋白质
   - 两者相互作用实现转录调控
2. 启动子
   - 基本概念
     - 启动子：一段位于转录起始位点 5’ 端上游区的 DNA 序列
     - 转录起始位点：与新生 RNA 链第一个核苷酸相对应 DNA 链上的碱基（图示 TSS 附近的坐标）



   - 启动子结构（图示、对比原核和真核基因的启动子结构，帮助学生记忆）
     - 原核基因
       - * -10 区，-10，TATAAT
       - * -35 区，-35，TTGACA
     - 真核基因
       - * TATA box，-25 ～-30，TATAAA
       - * CAAT box，-70 ～-80，CCAAT



3. 转录因子
   - 转录因子：蛋白质
   - 转录因子结合位点：DNA 序列，5 ～ 20bp
4. 相关资源
   - 数据库：EPD；TRANSFAC
   - 分析工具：Promoter Scan，Promoter 2.0；Tfblast

六、CpG 岛识别（15 分钟）

1. CpG 岛简介
   - CpG 保持或高于正常概率的基因组区段
   - 一般位于基因（尤其是看家基因）的 5’ 端区域，长度约 300 ～ 3000bp；大多数未甲基化
2. 【重点】识别依据与判别标准（提醒学生判别标准不是唯一的）
   - GC 含量：50% → 55%
   - CpG 岛的长度：200bp → 500bp
   - CpG 二核苷酸的出现频率：60% → 65%
     （计算公式：$\frac{Num\ of\ CpG}{Num\ of\ C \times Num\ of\ G} \times Total\ number\ of\ nucleotides\ in\ the\ sequence$）
3. 分析工具：EMBOSS（CpGPlot/CpGReport/Isochore）

七、EMBOSS 简介（10 分钟）

1. EMBOSS 简介
   - 开源、免费的序列分析软件包，整合了目前可以获得的大部分序列分析软件
   - 可以将系列分析工作进行无缝整合，弥补了许多软件功能分散、分析效率低下的缺陷
2. 使用界面
   - 操作系统：Linux，Mac，Windows
   - JEMBOSS：java 界面
   - EMBOSS Explorer：web 界面
3. 主要程序
   - 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. 使用实例：以使用 EMBOSS 识别 CpG 岛的实例操作加深学生对 CpG 岛识别依据和标准的理解，同时熟悉 EMBOSS 的使用方法

八、总结与答疑（5 分钟）

1. 知识点
   - DNA 序列基本信息分析：Chargaff 法则，GC 含量，序列转换
   - 限制性核酸内切酶位点分析：命名规则，II 型核酸酶的特点
   - 开放阅读框分析：相位，ORF 和 CDS 的区别
   - 启动子分析：原核基因和真核基因的启动子结构
   - CpG 岛识别：概念、识别依据及判别标准
2. 技能
   - 任务属性决定解决方案
   - 寻找最合适的方法
   - 先易后难，由浅入深