

天津医科大学理论课教案首页

(共 5 页、第 1 页)

课程名称：生物信息学

课程内容/章节：DNA 序列及其特征分析

教师姓名：伊现富

职称：讲师

教学日期：2014 年 10 月 11 日 13:30-15:20

授课对象：生物医学工程与技术学院 2011 级生医班（本）

听课人数：55

授课方式：理论讲授

学时数：2

教材版本：生物信息学（自编教材）

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；重复序列的概念及分类。
- 熟悉 DNA 序列分析的常见内容；ORF 分析中相位概念；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；重复序列相关的数据库和分析工具。
- 自学 DNA 序列分析数据库和工具的使用方法；重复序列相关数据库和分析工具的使用方法。

授课内容及学时分配：

- (5') 引言与导入：介绍中心法则，阐释核酸序列携带的两类遗传信息。
- (20') DNA 组份分析与序列转换：介绍 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。
- (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：讲解相位概念以及 ORF 与 CDS 的定义和区别，介绍常用的 ORF 分析工具。
- (15') 启动子分析：讲解启动子与转录因子的基本概念，介绍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。
- (15') 重复序列分析：讲解重复序列的概念、分类及特点，介绍常用数据库与分析工具。
- (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)

启动子 (promoter)

GC 含量 (GC content)

转录因子结合位点 (TFBS)

限制性核酸内切酶 (restriction endonuclease)

CpG 岛 (CpG island)

开放阅读框 (Open Reading Frame, ORF)

编码序列 (Coding Sequence, CDS)

重复序列 (repetitive/repeated sequence)

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 的两类信息及常见分析内容。
- 简述 CpG 岛的概念、识别依据和判别标准。
- 简述限制酶的命名规则及 II 型的主要特点。
- 简述重复序列依重复次数和组织形式的分类。
- 简述 ORF 与 CDS 的定义和区别。

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：

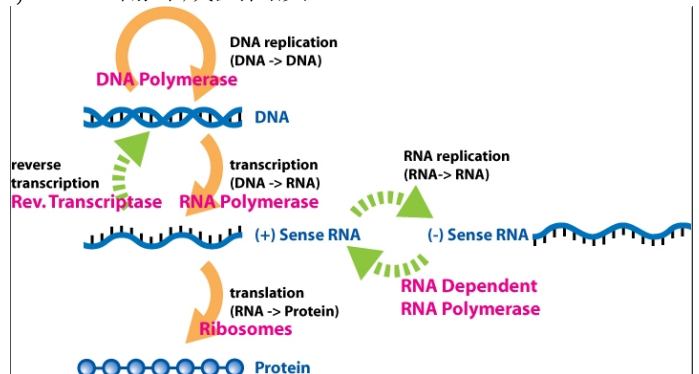
年 月 日

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 (20 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' \Rightarrow 3' [右]
- 多肽/蛋白质: [左] N 端 (氨基端) \Rightarrow C 端 (羧基端) [右]

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

• II 型限制酶的特点 (以 *EcoRI*、*AluI* 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

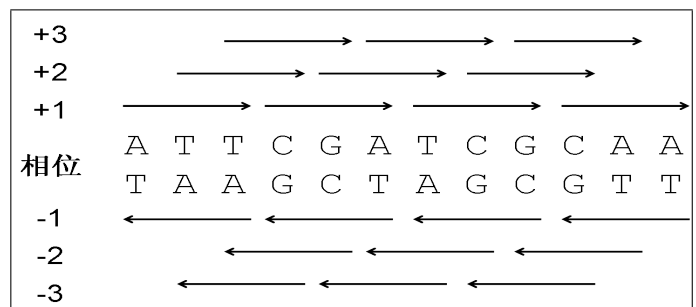
酶名称	来源	识别序列	切法
<i>EcoRI</i>	<i>Escherichia coli</i>	5' GAATTC 3' CTTAAG	5' ---G AATTC---3' 3' ---CTTAA G---5'
<i>AluI</i> *	<i>Arthrobacter luteus</i>	5' AGCT 3' TCGA	5' ---AG CT---3' 3' ---TC GA---5'

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



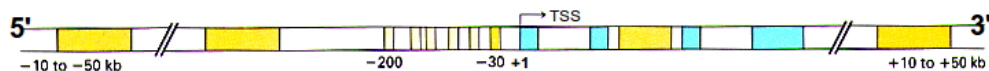
五、启动子分析 (15 分钟)

1. 转录调控

- 顺式作用元件：核酸序列 \Rightarrow 启动子
- 反式作用因子：蛋白质
- 两者相互作用实现转录调控

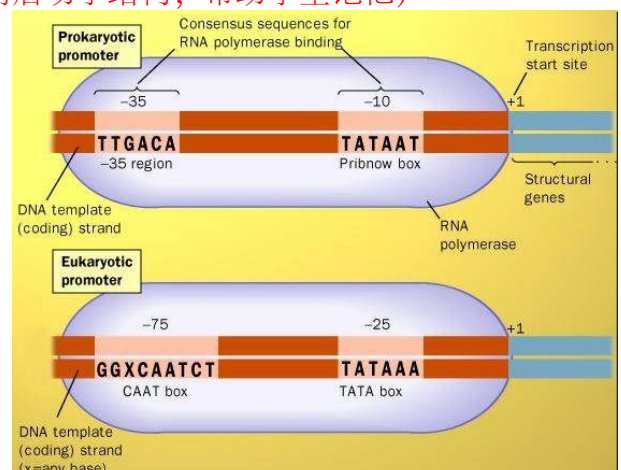
2. 启动子

- 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; Tfbblast

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: $\frac{\text{Num of CpG}}{\text{Num of C} \times \text{Num of G}} \times \text{Total number of nucleotides in the sequence}$)

3. 分析工具: EMBOSS (CpGPlot/CpGReport/Isochore)

七、重复序列分析 (15 分钟)

1. 基本概念: 重复序列是指真核生物基因组中重复出现的核苷酸序列。

2. 分类 (引导学生思考不同分类标准之间的关系)

- 重复次数 (对每一类别都给出实例, 帮助学生理解记忆)
 - 低度重复序列: 2 ~ 10 个拷贝, 如酵母 tRNA 基因
 - 中度重复序列: 重复几十~几千次, 长度约 300bp, 如 rRNA 基因
 - 高度重复序列: 重复几百万次, 少于 10 个核苷酸残基, 如卫星 DNA
- 组织形式
 - 串联重复序列: 成簇存在于染色体的特定区域
 - * 卫星 DNA: 5 ~ 200bp, 几百万个拷贝, 大多集中在着丝粒部位
 - * 小卫星: 10 ~ 100bp 的基本单位, 总长不超过 20kb, 重复次数高度变异, 靠近端粒的位置
 - * 微卫星: 2 ~ 10bp, 长度 50 ~ 100bp, STR 遗传多态性, 常见于内含子
 - 散在重复序列: 分散于染色体的各位点上
 - * 短散在重复序列 (SINE): 500bp 以下, 拷贝数达 10 万以上; 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; 如 Alu——300bp, 100 万个拷贝
 - * 长散在重复序列 (LINE): 1000bp 以上, 上万份拷贝; 可以自主转座的反转录转座子; 来源于 RNA 聚合酶 II 的转录产物; 如 L1——6100bp, 3500 个拷贝

3. 相关资源

- Repbase: 真核生物 DNA 重复序列数据库
- L1Base: L1 数据库
- STRBase: STR 数据库
- RepeatMasker: 通过与已知重复序列数据库进行比对, 识别、分类和屏蔽重复序列; 提供四个搜索引擎
 - Cross_match: 速度慢、精度高
 - ABBlast: 速度快、精度略低
 - RMBlast: NCBI Blast 的兼容版
 - HMMER: 只适用于人类基因组序列

Sequence types	Repeat size(bp)	Array size (kb)	Copy number ^a
Satellites — large tandem arrays		10–25% of total DNA	
Microsatellite	2–5	0.2–0.5	3 x 10 ³
Minisatellite	~15	0.5–3	10 ³
Satellite	5–100	100,000	10 ⁷
Megasatellite	4–10 kb	30–100	30–100
Interpersed elements		35–40% of total DNA	
Retrotransposons			
<i>LTR-containing elements</i>			
<i>copia</i> ² , <i>gypsy</i> ²	~5 kb	NA	20–60
Yeast Ty	6.3kb	NA	40
<i>Poly-A elements</i>			
LINE1 (L1)	1–7 kb	NA	~10 ⁵
HeT-A, TART ²	6–10 kb	5–10	~10 ⁴
SINEs			
Alu	300	NA	~10 ⁶

八、总结与答疑 (5 分钟)

- 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
- 重复序列: 分类