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（4.2）基因组结构注释分析

教师姓名：伊现富　　**职称：**讲师　　**教学日期：**2016 年 9 月 30 日 10:00-12:00

授课对象：生物医学工程与技术学院 2014 级生信班（本）　　**听课人数：**30

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）：

- 掌握重复序列的概念及分类；基因识别的方法与策略。
- 熟悉原核基因和真核基因的结构特点。
- 了解重复序列相关的数据库和分析工具；基因识别的分析工具。
- 自学重复序列相关数据库和分析工具的使用方法；基因识别分析工具的使用方法。

授课内容及学时分配：

- （5’）回顾与导入：回顾序列基本信息和特征信息分析的主要内容，引出重复序列和基因识别的内容。
- （30’）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。
- （50’）基因识别：介绍基因和基因识别的基本概念，回顾原核基因和真核基因的结构特点并进行比较，讲解基因识别的主要方法与策略，介绍基因识别的常用工具。
- （10’）查找数据库与分析工具：讨论查找数据库和分析工具时可以采取的策略。
- （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：重复序列的分类，基因识别的方法和策略。
- 难点：基因识别中“信号”特征和“内容”特征的区别。
- 解决策略：通过示意图和实例帮助学生理解，通过对比加深记忆。

专业外语词汇或术语：

重复序列（repetitive/repeated sequence）
串联重复（tandem repeat）
散在重复（interspersed repeat）
基因识别（gene prediction/finding）
间接识别法（extrinsic approach）
从头计算法（*ab initio* approach）

辅助教学情况：

- 多媒体：原核基因和真核基因的结构，基因识别的策略。
- 板书：查找数据库和分析工具的主要策略。

复习思考题：

- 简述重复序列依重复次数和组织形式的分类。
- 简述原核基因和真核基因结构的异同。
- 简述基因识别的三大类方法和主要策略。
- 论述查找所需数据库和分析工具的策略。

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。
- 李霞，李亦学，廖飞。生物信息学，人民卫生出版社，2010。
- 王明怡，杨益，吴平。生物信息学（中译本，第 2 版），科学出版社，2004。
- 维基百科。

一、回顾与导入（5 分钟）

1. 序列分析（简要回顾上次课的主要知识点）
   - 基本信息：碱基比例、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、……
2. 结构注释（基因识别前要先屏蔽重复序列）
   - 重复序列
   - 基因识别

二、重复序列分析（30 分钟）

对于真核生物的核酸序列而言，已知在重复序列中几乎没有对基因识别或序列比较有用的信息，因此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。

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. 【**重点**】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

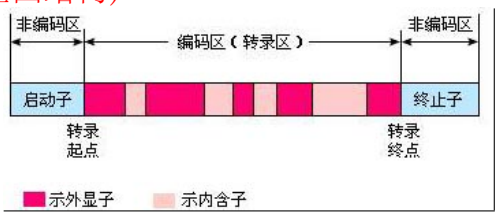
Sequence types	Repeat size(bp)	Array size (kb)	Copy number[a]
Satellites — large tandem arrays		**10–25% of total DNA**	
Microsatellite	2–5	0.2–0.5	3 x 10^5
Minisatellite	~15	0.5–3	10^5
Satellite	5–100	100,000	10^7
Megasatellite	4–10 kb	30–100	30–100
Interpersed elements		**35-40% of total DNA**	
Retrotransposons			
LTR-containing elements			
copia[2], *gypsy*[2]	~5 kb	NA	20–60
Yeast Ty	6.3kb	NA	40
Poly-A elements			
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

三、基因识别（50 分钟）

在介绍基本概念的基础上，通过比较原核基因和真核基因的异同，讲解基因识别的主要策略及各种方法在原核和真核基因识别中的具体应用。

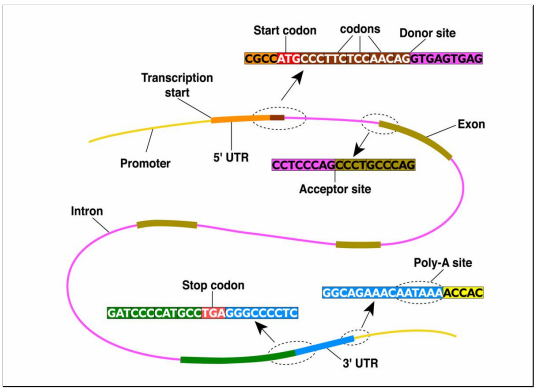
1. 基本概念
   - 基因：产生一条多肽链或功能 RNA 所需的全部核苷酸序列（强调既包括编码区，也包括非编码区）
   - 基因识别：识别 DNA 序列上具有生物学特征的片段

2. 基因结构（通过示意图形象化展示、比较原核和真核的基因结构）
   - 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

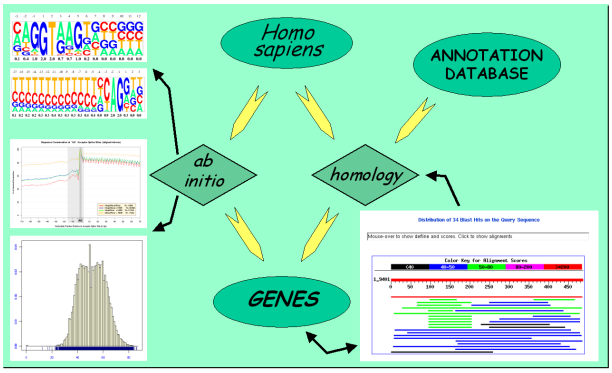


3. 【重点】识别方法
   - 间接识别法：mRNA/蛋白质序列 ⇒ DNA 序列
   - 从头预测法：基因预测，基于“信号”和“内容”两类特征
   - 比较基因组学的方法：比较相关物种的 DNA 序列

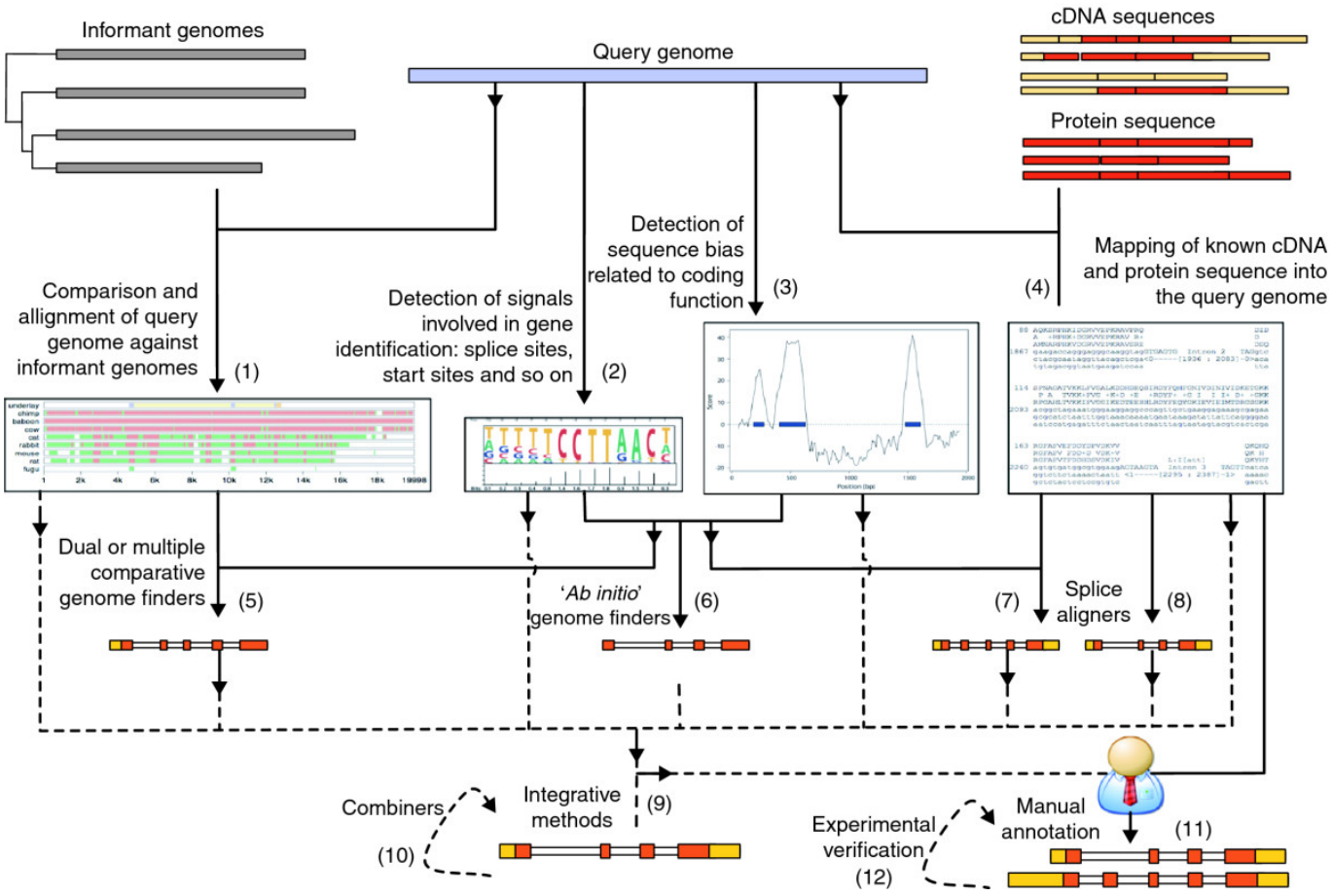
4. 基因预测
   - 【难点】“信号”和“内容”
     - 共同点：都包括编码区和非编码区
     - 信号：不连续的局部序列模体，一般都有一致性序列；如启动子，剪接供体和受体位点，起始和终止密码子，polyA 位点
     - 内容：不同长度的扩展序列，没有一致性序列，但具有把自己与周围 DNA 区分开来的保守特征；如密码子使用偏好性，双联密码子出现频率，基因组等值区



   - 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

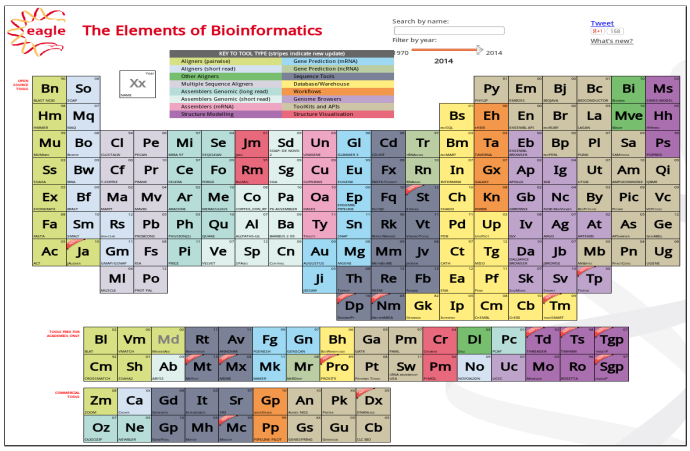


5. 【重点】识别策略
6. 识别工具（强调分析工具的适用范围）
   - 识别原核基因：GeneMarkS, Glimmer
   - 识别真核基因：GENSCAN



四、查找数据库与分析工具（10 分钟）

- 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



五、总结与答疑（5 分钟）

1. 知识点
   - 重复序列：分类
   - 基因识别：原核和真核的基因结构，基因识别的方法和策略
2. 技能
   - 查找数据库：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