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：**基因组功能注释分析

教师姓名：伊现富　　**职称：**讲师　　**教学日期：**2013 年 9 月 2 日 14 时 -16 时

授课对象：生物医学工程学院 2010 级生信班（本）　　**听课人数：**23

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）：

- 掌握基因组的组装版本与坐标系统，熟悉组装版本间坐标转换的工具并了解其用法。
- 掌握基因组注释常用格式中的 FASTA 和 BED 格式，熟悉 GFF 格式，了解 VCF 格式，了解基因组注释格式间的转换工具并自学其使用方法。
- 掌握基因组坐标的逻辑运算模式，熟悉坐标逻辑运算的适用范围，了解进行逻辑运算的常用工具并自学其使用方法。

授课内容及学时分配：

- （5'）回顾与导入：回顾序列分析中结构注释的内容，介绍功能注释的基础知识及主要内容。
- （20'）基因组组装版本与坐标系统：介绍基因组的组装版本、不同数据库中基因组组装版本的命名及其对应关系，举例讲解 0-based 和 1-based 两种不同的基因组坐标系统及其各自的适用范围。
- （20'）基因组注释常用格式：通过实例详细讲解 FASTA、BED、GFF 和 VCF 格式，介绍 GFF 和 VCF 格式的解读方法。
- （25'）基因组坐标的逻辑运算：讲解交集、减法、补集等常用的逻辑运算模式，介绍各种逻辑运算的适用范围及实例。
- （20'）操作演示：介绍坐标转换、格式转换、逻辑运算的常用工具，并通过实例演示各种工具的使用方法。
- （10'）总结与答疑：回顾授课内容中的知识点，解答学生疑问。

教学重点、难点及解决策略：

- 重点：基因组的两种坐标系统，基因组注释中常用的 BED 格式，基因组坐标的逻辑运算模式。
- 难点：基因组坐标中的 0-based 坐标系统，基因组坐标的逻辑运算模式。
- 解决策略：通过形象化的图示、与集合运算等的类比解释基本概念，通过实例帮助学生理解记忆。

专业外语词汇或术语：

基因组注释（genome annotation）
功能注释（functional annotation）
基因组组装版本（genome build）
坐标系统（coordinate system）
坐标转换（coordinate transform）
单核苷酸多态性（SNP）

辅助教学情况：

- 多媒体：两种坐标系统、注释常用格式、逻辑运算模式的示意图。
- 板书：两种坐标系统、FASTA 和 BED 格式、逻辑运算模式的简单示例。
- 演示：hg19 和 hg18 间坐标转换、BED 和 GFF 格式转换、减法和联合运算的操作实例。

复习思考题：

- 不同数据库间基因组组装版本的对应关系。
- 基因组坐标的两种表示方法。
- BED 格式中每一列的含义。
- 常见的基因组坐标逻辑运算模式。

参考资料：

- UCSC FAQ（Frequently Asked Questions）
- Galaxy Wiki
- File formats on Wikipedia

主任签字：　　**年　月　日**　　**教务处制**

一、回顾与导入（5 分钟）

基因组注释：（回顾结构注释的相关内容，引出功能注释的主要工作）

- 基因组结构注释：序列基本信息分析、寻找限制酶切位点、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、屏蔽重复序列、基因识别、……
- 基因组功能注释（结合第二代测序数据的生物信息学处理过程逐步引出功能注释的内容）
  - 基础工作：坐标转换、格式转换、坐标的逻辑运算、……
  - 高级注释：变异位点的注释、富集分析、序列标识、……

二、基因组组装版本与坐标系统（20 分钟）

1. 基因组组装版本
   - 基因组组装版本在不断变化（与操作系统、软件的版本变化进行类比）
   - 不同数据库采用不同的命名规则（板书对应关系）
     - NCBI：Build X
     - Ensembl：NCBIX
     - UCSC：hgX、mmX、……（解释 hg、mm 等缩写的含义）
2. 基因组的坐标系统（以数学中的坐标引出基因组的坐标，并通过实例予以讲解；类比英式和美式英语的 first floor 以及编程语言中的计数方式；引导学生思考两种坐标系统的优缺点）
   - 1-based（one-based, fully-closed）：[start, end]
   - 0-based（zero-based, half-closed-half-open）：[start, end)
3. 坐标系统的适用范围
   - 1-based：主要给研究人员肉眼查看的数据，如：GFF、VCF、SAM 和 Wiggle 等格式以及 DAS 和 UCSC 的 Genome Browser 等工具
   - 0-based：主要用于计算机程序处理的数据，如：BED、BAM 和 PSL 等格式以及 NCBI 的 dbSNP 和 UCSC 的 Table Browser 等数据库与工具

三、基因组注释常用格式（20 分钟）

基因组数据类型的多样性导致了数据格式的多样性。通过实例详细介绍 FASTA、BED、GFF 和 VCF 四种常用格式，讲解格式中每一列的含义，引导学生学会解析特定格式中的信息。

1. FASTA 格式
   - 首行：起始标识符“>”，ID，描述信息；其余行：具体的序列。
   - IUB/IUPAC 核酸代码：“N”代表任意一种核酸，“-”代表空位。
   - IUB/IUPAC 氨基酸代码：“X”代表任意一种氨基酸，“*”代表翻译终止，“-”代表空位。
2. BED 格式：3+9=12 列（BED12）。引申出简化版的 BED3，BED4，BED5 和 BED6。
   - BED3：chrom, chromStart, chromEnd
   - BED4：chrom, chromStart, chromEnd, name
   - BED5：chrom, chromStart, chromEnd, name, score
   - BED6：chrom, chromStart, chromEnd, name, score, strand
3. GFF 格式
   - 注释信息：以“##”开头
   - 特征信息：9 列
4. VCF 格式
   - 元信息：以“##”起始
   - 标题行：以“#”起始
   - 数据行：8+1+N 列

强调它们都以纯文本形式进行存储，简单介绍常用的文本编辑器（Notepad++，Vim，Emacs）。

四、基因组坐标的逻辑运算（25 分钟）

1. 集合运算与逻辑运算
   - 类比数学中的集合运算
   - 以交集为例，板书由集合运算转换到逻辑运算的过程
2. 基因组坐标逻辑运算（先用示意图予以讲解，再辅以实例帮助记忆，最终理解其适用情况）
   - intersect，交集：保留重叠的坐标
   - subtract，减法：去除重叠的坐标
   - merge，合并：合并重叠的坐标
   - concatenate，串联：合并多组坐标
   - complement，补集：取坐标的补集
   - cluster，聚类：聚合符合要求的坐标
   - join，联合：根据坐标重叠把两组记录对应起来

五、操作演示（20 分钟）

1. 坐标转换
   - 工具：liftOver
   - 实例：把人类的基因坐标从 hg19 转换到 hg18
2. 格式转换
   - 工具：Galaxy
   - 实例：BED 与 GFF 格式的互转
3. 逻辑运算
   - 工具：Galaxy，BEDTools
   - 实例：外显子与 SNP 的比较
4. 处理问题的基本步骤（将“三步走”的思想贯穿在每个操作实例中）
   - 获取输入：数据来源，文件格式，……
   - 数据处理：程序选择，参数调整，……
   - 解析输出：文件格式，数据校验，……

六、总结与答疑（10 分钟）

1. 知识点
   - 基因组组装版本：命名规则，对应关系
   - 两种基因组坐标系统：1-based，0-based
   - 四种注释常用格式：FASTA，BED，GFF，VCF
   - 逻辑运算模式：intersect，subtract，merge，concatenate，complement，cluster，join
   - 坐标转换、格式转换、逻辑运算的工具：liftOver，Galaxy，BEDTools
2. 技能
   - “输入 ⇒ 加工 ⇒ 输出”三步走
   - 获取输入：数据来源，文件格式，……
   - 数据处理：程序选择，参数调整，……
   - 解析输出：文件格式，数据校验，……