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基础

教师姓名：伊现富　　**职称：**讲师　　**教学日期：**2014 年 4 月 2 日 10 时 -12 时

授课对象：生物医学工程学院 2011 级生信班（本）　　**听课人数：**27

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）：

- 掌握基因组的两种坐标系统；FASTA 和 BED 格式；基因组坐标的常见逻辑运算。
- 熟悉 GFF 格式；不同操作系统中换行符的差异。
- 了解基因组的组装版本及常见版本间的对应关系；纯文本与格式化文本的区别；常见的文本编辑器。
- 自学 VCF 格式；基因组坐标的其他逻辑运算。

授课内容及学时分配：

- (5') 回顾与导入：回顾基因组注释中结构注释的知识，介绍功能注释的主要内容。
- (10') 基因组组装版本：介绍基因组的组装版本、不同数据库中组装版本的命名及其对应关系。
- (15') 基因组坐标系统：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 和 BED 格式，简单介绍 GFF 格式，学生自学 VCF 格式。
- (10') 文本文件与文本编辑器：介绍纯文本和格式化文本的区别，讲解不同操作系统中换行符的差异，推荐几个常用的文本编辑器。
- (35') 基因组坐标的逻辑运算：通过图示和实例讲解交集、减法、补集等常用的基因组坐标逻辑运算。
- (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：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。
- 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。
- 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，通过实例演算帮助学生理解记忆。

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）
集合运算（set operation）

辅助教学情况：

- 多媒体：两种坐标系统、注释常用格式、逻辑运算等的示意图。
- 板书：两种坐标系统、FASTA 和 BED 格式、逻辑运算的简单示例。

复习思考题：

- hg19 和 mm10 分别代表什么含义？
- 举例说明基因组的两种坐标系统。
- BED 格式前 6 列分别代表什么含义？
- 图示常见的基因组坐标逻辑运算。
- 不同操作系统的换行符有何区别？如何进行查看和转换？

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）

- 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、……
- 功能注释
  - 基础知识：组装版本、坐标系统、常用格式、坐标的逻辑运算、……
  - 注释内容：变异位点的注释、富集分析、互作网络分析、进化分析、……

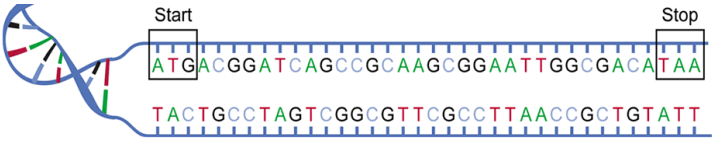
二、基因组组装版本（10 分钟）

1. 基因组也有组装版本（与操作系统、软件的版本变化进行类比）
   - We used DNA sequences from the human and mouse genome assemblies hg18 and mm9.
   - The genome assemblies from which the sequences obtained were Dec 2011 (GRCm38/mm10), Feb 2009 (GRCh37/hg19) and Nov 2004 (Baylor3.4/rn4) for mouse, human and rat respectively.
2. 不同数据库采用不同的命名规则（解释 hg、mm、GRC 等缩写的含义，帮助学生理解其命名）
   - NCBI：Build X
   - Ensembl：NCBIX
   - UCSC：hgX、mmX、……
3. 不同版本间存在对应关系
   - hg19 = GRCh37; hg18 = NCBI Build 36.1
   - mm10 = GRCm38; mm9 = NCBI Build 37

SPECIES	UCSC	DATE	NCBI
Human	hg19	Feb. 2009	Genome Reference Consortium
			GRCh37
	hg18	Mar. 2006	NCBI Build 36.1
	hg17	May 2004	NCBI Build 35
	hg16	Jul. 2003	NCBI Build 34
Mouse	mm10	Dec. 2011	Genome Reference Consortium
			GRCm38
	mm9	Jul. 2007	NCBI Build 37
	mm8	Feb. 2006	NCBI Build 36
	mm7	Aug. 2005	NCBI Build 35

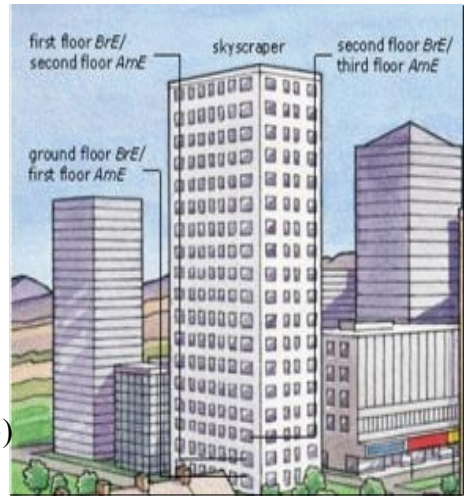
三、基因组坐标系统（15 分钟）

1. 基因组的坐标系统（以数学中的坐标引出基因组的坐标）



   - SNP, rs1800468, hg19: "chr19:41860587"
   - gene, SAMD11, hg19: "chr1:861121-879961"
2. 坐标系统实例（通过实例讲解两种坐标系统；类比英式和美式英语的 first floor 以及编程语言中的计数方式）

0-based index	0	1	2	3	4	5	6	7
Sequence	A	A	T	T	G	G	C	C
1-based index	1	2	3	4	5	6	7	8

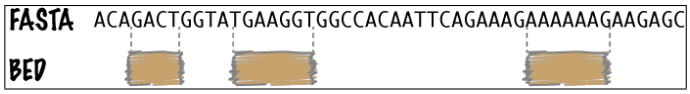


   - 1-based（one-based, fully-closed）：TG = [3,5]
   - 0-based（zero-based, half-closed-half-open）：TG = [4,5)
3. 坐标系统的适用范围（引导学生思考两种坐标系统的优缺点）
   - 1-based：主要给研究人员肉眼查看的数据，如：GFF、VCF、SAM 和 Wiggle 等格式以及 DAS 和 UCSC 的 Genome Browser 等工具
   - 0-based：主要用于计算机程序处理的数据，如：BED、BAM 和 PSL 等格式以及 NCBI 的 dbSNP 和 UCSC 的 Table Browser 等数据库与工具

四、基因组注释常用格式（20 分钟）

类似于多样的计算机数据需要多样的文件格式来存储一样，基因组数据类型的多样性导致了数据格式的多样性。（通过计算机文件格式及格式间的转换引申出基因组数据格式）

- 序列 ⇒ FASTA
- 展示特征 ⇒ BED（Browser Extensible Data）
- 特征 ⇒ GFF（General Feature Format）
- 变异 ⇒ VCF（Variant Call Format）



通过展示实例详细介绍 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。同时强调它们都以纯文本形式进行存储的，可以使用常见的文本编辑器（Notepad++，Vim，Emacs）查看、编辑。

1. FASTA 格式（强调 FASTA 与 sequence 之间的区别）
   - 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），0-based。引申出 BED3，BED4，BED5 和 BED6。
   - BED3：chrom, start, end
     chr1 11873 14409
   - BED4：chrom, start, end, name
     chr1 11873 14409 uc001aaa.3
   - BED5：chrom, start, end, name, score
     chr1 11873 14409 uc001aaa.3 0
   - BED6：chrom, start, end, name, score, strand
     chr1 11873 14409 uc001aaa.3 0 +

```
chr7  127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7  127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7  127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7  127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7  127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7  127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7  127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7  127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7  127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

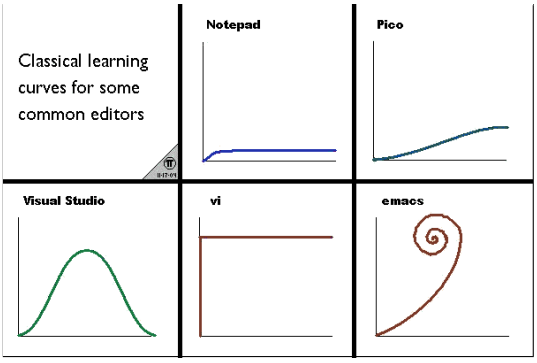
3. GFF 格式：1-based
   - 注释信息：以“##”开头
   - 特征信息：9 列

```
##gff-version 3
ctg123 . operon        1300 15000  .  +  .  ID=operon001;Name=superOperon
ctg123 . mRNA          1300  9000  .  +  .  ID=mrna0001;Parent=operon001;Name=sonichedgehog
ctg123 . exon          1300  1500  .  +  .  Parent=mrna0001
ctg123 . exon          1050  1500  .  +  .  Parent=mrna0001
ctg123 . exon          3000  3902  .  +  .  Parent=mrna0001
ctg123 . exon          5000  5500  .  +  .  Parent=mrna0001
ctg123 . exon          7000  9000  .  +  .  Parent=mrna0001
ctg123 . mRNA         10000 15000  .  +  .  ID=mrna0002;Parent=operon001;Name=subsonicsquirrel
ctg123 . exon         10000 12000  .  +  .  Parent=mrna0002
ctg123 . exon         14000 15000  .  +  .  Parent=mrna0002
```

4. VCF 格式：1-based（自学）
   - 元信息：以“##”起始
   - 标题行：以“#”起始
   - 数据行：8+1+N 列

五、文本文件与文本编辑器（10 分钟）

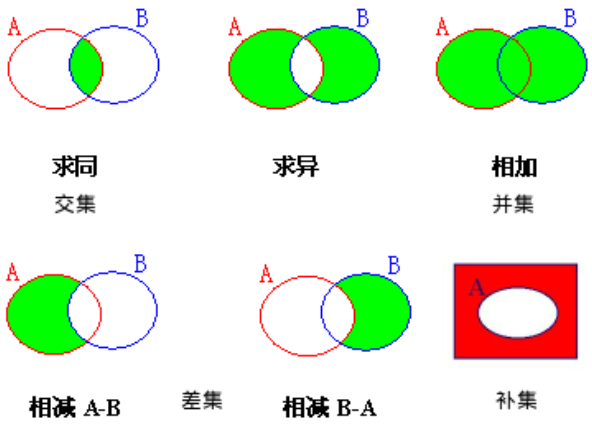
1. 纯文本 vs. 格式化文本（通过比较 txt 和 doc 这两种常见的文本文件格式，帮助学生理解纯文本和格式化文本的区别）
2. 换行符及其识别与转换
   - Windows：`\r\n`（CR+LF，回车 + 换行），文件尾部直接 EOF（文件结束标志）
   - Unix：`\n`（LF，仅有换行），文件最后一行也会增加该字符，然后才是 EOF
   - Mac：`\r`（CR，仅有回车）
   - Windows 中的识别与转换：文本编辑器，如 Notepad++
   - Unix 中的识别与转换：file 识别，fromdos & todos 转换

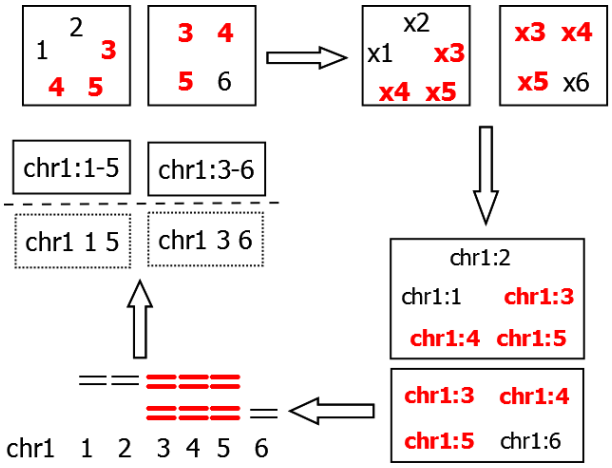


3. 文本编辑器：Notepad++，Vim，Emacs，Sublime Text

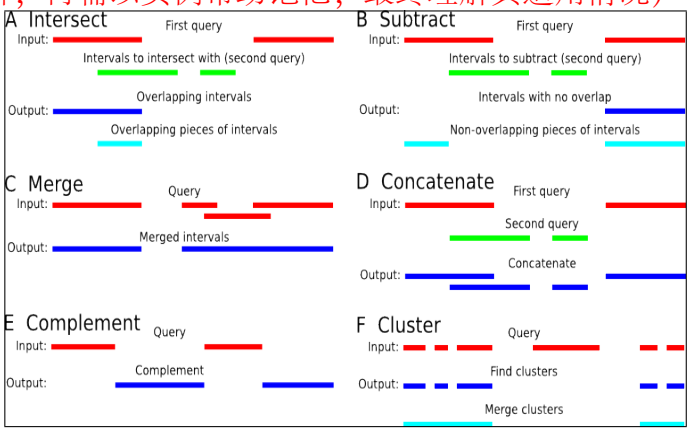
六、基因组坐标的逻辑运算（35 分钟）

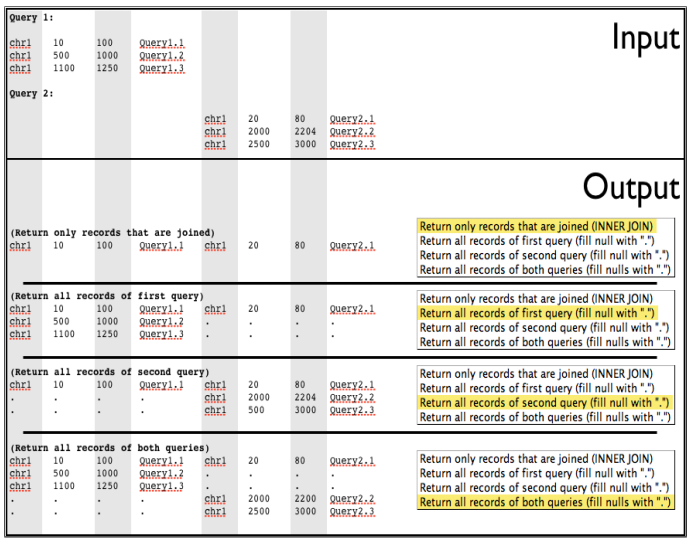
1. 集合运算与逻辑运算
   - 类比数学中的集合运算
   - 以交集为例，讲解由集合运算转换到逻辑运算的过程





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
   - coverage, flank, closest, slop, window, …（学生自学）



七、总结与答疑（5 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
   - 坐标的逻辑运算：intersect，subtract，merge，concatenate，complement，cluster，join
2. 技能
   - 纯文本与格式化文本
   - 不同操作系统中的换行符
   - 文本编辑器：Notepad++, Vim, Emacs