《生物信息数据处理》理论课教学大纲

（适用于四年制生物信息学专业）

## 前 言

《生物信息学数据处理》是生物信息学专业的专业课，根据生物信息学专业本科培养目标开设的一门理工类基础课程。本课程的教学目的是培养学生在具有生物信息学知识的基础上，应用生物信息学的理论和方法，解决医学临床和医学研究中的实际问题。

本课程为72学时，其中理论授课36学时，实验课36学时。授课过程中辅以课堂讨论及课外阅读。理论课程教学共计36学时，主要包括两个模块：linux语言（18学时）、R语言（18学时）。Linux语言的主要内容包括：Linux操作系统的文件系统、常用命令、文本处理、环境配置、shell程序设 计、shell高级命令、linux内核管理。R语言的主要内容：R语言安装、创建数据集、数据管理、图形初级处理、基本图形分类、统计分析模型、广义线性模型、聚类分析、富集分析等。

## 第一章 Linux系统概述

### 一、目的要求

了解Linux的历史和现状，Linux的发行版及应用领域。熟悉Linux相关的基本概念，Linux账户的类型。掌握Linux的登录和退出。

### 二、教学内容

（一）Unix/Linux的发展简史。

（二）Linux简介：概念、操作系统、优点、发行版、应用领域。

（三）Linux入门：从终端登录和退出 ssh、学习与练习、GUI与CLI。

（四）账户基本知识：账户类型、终端提示符、普通用户和root用户。

（五）生物信息学系教学服务器的登录与注意事项。

### 三、教学学时安排

1学时。

### 四、教学方法

课堂讲授：板书，多媒体辅助教学。

## 第二章 Linux的文件系统

### 一、目的要求

（一）了解Linux的文件系统结构：目录文件和普通文件。

（二）掌握文件名命名规则、常用的文件和目录操作命令。

（三）掌握Linux中的路径，文件系统导航的常用命令，权限查看及修改方法。

（四）熟悉Linux中的文件类型，硬链接与软链接的区别和使用。

### 二、教学内容

（一）Linux的目录：文件系统和分区、文件系统结构、目录文件、普通文件、个人目录、系统目录、绝对路径与相对路径

（二）文件名命名规则

（三）目录文件的操作：cd、ls、mkdir、rmdir、pwd、tree

（四）文件系统的操作：常用touch、less、mv、rm命令简介及使用。

（五）文件类型：简介、链接 ln 的使用。

（六）文件和目录权限：简介、查看、访问权限、chmod的使用。

### 三、教学学时安排

2学时。

### 四、教学方法

课堂讲授：板书，多媒体辅助教学。

## 第三章 Linux的常用命令

### 一、目的要求

（一）了解常用的系统导航命令、进程管理命令。

（二）熟悉命令的使用技巧。

（三）掌握常用的文件操作命令、压缩解压命令和获取帮助的命令。

### 二、教学内容

（一）文件操作命令：cat、echo、more/less、head/tail、cp、wc、sort、uniq、cut、du、df。

（二）系统导航命令：which、whereis、locate

（三）进程管理命令：top、ps、kill、killall、jobs、bg、fg、nohup。

（四）压缩解压命令：tar、gzip/gunzip、bzip2/bunzip2、zip/unzip。

（五）获取帮助：man、info、\*\* --help。

（六）命令使用技巧：自动补全、历史、别名、连接符、终端快捷键。

### 三、教学学时安排

4学时。

### 四、教学方法

课堂讲授：板书，操作演示，多媒体辅助教学。

## 第四章 Linux的高级命令

### 一、目的要求

（一）了解find、grep、sed和AWK的基本工作原理。

（二）掌握正则表达式中的字符类型，find、grep、sed和AWK的基本使用方法。

### 二、教学内容

（一）Linux特殊字符：通配符、转义符、输入/输出重定向、管道、命令置换。

（二）正则表达式：简介、元字符类型、扩展正则表达式

（三）Linux中的高级命令：find、grep、sed、AWK简介、原理、使用。

### 三、教学学时安排

3学时。

### 四、教学方法

课堂讲授：板书，操作演示，多媒体辅助教学。

## 第五章 Linux中的文本处理和编辑

### 一、目的要求

（一）了解Linux中的文本处理命令，生物信息学常见格式（FASTQ、FASTA、BED、GFF、VCF，等）

（二）掌握vi编辑器的基本使用方法。

（三）掌握head/tail、sort、uniq、wc等常用命令的使用方法，sed、AWK的高级使用方法。

### 二、教学内容

（一）vi工作方式、进入和退出、文本输入、修改、ex命令

（二）vi编辑器的工作模式：普通模式、插入模式、命令行模式。

（三）vi编辑器的基本使用：移动定位、剪切删除、搜索替换、退出保存。

（四）Linux文本处理命令：汇总、高频命令（cut、sort、uniq、wc、join、tr、split、paste，等）的介绍与使用。

（五）生物信息学常见格式：FASTQ、FASTA、BED、GFF、VCF

### 三、教学学时安排

2学时。

### 四、教学方法

课堂讲授：板书，操作演示，多媒体辅助教学。

## 第六章 Linux程序安装与环境配置

### 一、目的要求

（一）了解软件包管理常见开源许可证。

（二）掌握linux程序安装的流程及常用命令。

（三）熟悉shell环境文件配置。

### 二、教学内容

（一）软件包管理：软件包类型、管理系统、常用命令、开源许可证。

（二）安装方法：源码安装、yum安装（root权限）、anaconda安装

（三）anaconda与bioconda：安装、配置、使用。

（四）环境变量和软件的配置：环境变量、PATH、LIBRARY\_PATH、bashrc、export。

### 三、教学学时安排

2学时。

### 四、教学方法

课堂讲授：板书，操作演示，多媒体辅助教学。

## 第七章 shell程序设计

### 一、目的要求

（一）对了解shell的类型、配置文件等基本概念。

（二）掌握shell的基本语法，条件测试，流程控制

### 二、教学内容

（一）shell简介：类型、配置文件、环境变量、别名。

（二）shell特殊字符、变量引用

（三）shell的基本语法：脚本结构、变量、注释、输入输出、特殊变量、退出状态、脚本调试。

（四）shell的流程控制结构：if语句、case语句、while语句、for语句、break命令和continue命令、exit命令、select命令。

（五）高级shell脚本编程：输入输出重定向、并行、命令替换、函数、数组。

### 三、教学学时安排

4学时。

### 四、教学方法

课堂讲授：板书，操作演示，多媒体辅助教学。

## 第八章 R语言基础

### 一、目的要求

（一）了解R语言和RStudio。

（二）熟悉R包的来源及其安装与加载。

（三）掌握R语言的基本语法，获取帮助的常用方法。

### 二、教学内容

（一）R基础：R语言简介、RStudio简介。

（二）R包：CRAN/Bioconductor简介、安装、加载、常用R包、环境信息。

（三）获取帮助：help、？、vignette。

（四）基本使用：工作路径、数学运算符、变量赋值、特殊数字、逻辑向量、检查变量。

（五）在R语言中存储、操作DNA序列。

### 三、教学学时安排

1学时。

### 四、教学方法

课堂讲授：板书，操作演示，多媒体辅助教学。

## 第九章 R语言中的数据结构与类型

### 一、目的要求

（一）了解R语言中的日期和时间。熟悉R语言中的函数。

（二）掌握R语言的主要数据结构及其基本操作，R语言中的因子。

### 二、教学内容

（一）数据结构简介：向量、矩阵、数组、列表、数据框。

（二）各种数据结构的使用：创建、操作、输出、索引、转换。字符串与因子：创建、操作、转换、特殊字符。

（三）日期和时间：创建、解析、格式化、算术运算。

（四）函数：简介、创建、调用、作用域。

（五）数据结构在生物信息学数据中的体现：以BED格式为例。

### 三、教学学时安排

3学时。

### 四、教学方法

课堂讲授：板书，操作演示，多媒体辅助教学。

## 第十章 R语言的数据管理

### 一、目的要求

（一）了解数据分析的基本流程。

（二）熟悉探索性数据分析的主要内容和常用R包，数据的长宽格式。

（三）掌握数据读取和存储的方法，数据清理、转换和探索性分析的常用操作。

### 二、教学内容

（一）数据的获取：内置数据集、读取/保存文本文件、读取二进制文件、抓取网页、访问数据库。

（二）长宽格式：简介、利弊、互转。

（三）数据清理和转换：列操作、行筛选、排序、字符串处理、因子处理、日期时间处理。 （四）探索性数据分析：汇总统计、分布的可视化、缺失值和异常值的处理、基本的可视化、常用R包。 基于基因表达数据的分析工作流：读取、变换、探索性数据分析、结果存储。

### 三、教学学时安排

3学时。

### 四、教学方法

课堂讲授：板书，操作演示，多媒体辅助教学。

## 第十一章 图形初级处理

### 一、目的要求

（一）了解R语言的三种绘图系统。

（二）熟悉常用的汇总统计函数，ggplot2的扩展包。掌握ggplot2的原理和使用。

### 二、教学内容

（一）汇总统计：min/max、var/sd、mean/median、table、summary/fivenum，等。

（二）三种绘图系统：base绘图法、lattice图形系统、ggplot2图形系统。

（三）ggplot2绘图系统：基本原理、分层语法、属性映射、统计变换、常见图形的绘制。

（四）ggplot2的扩展包：查询、使用、常用包的简介。

（五）使用ggplot2绘制常见的生物信息学统计图形：以基因表达数据为例。

### 三、教学学时安排

3学时。

### 四、教学方法

课堂讲授：板书，操作演示，多媒体辅助教学。

## 第十二章 R语言的统计分布

### 一、目的要求

（一）了解数据分布在R语言中的实现。

（二）熟悉R语言中的随机与抽样。

（三）掌握R语言中的回归模型构建和评估。

### 二、教学内容

（一）随机数：设置随机种子、从分布中抽样。

（二）统计分布：各种分布的简介、实现及相关函数。

（三）回归模型：线性回归、逻辑回归、模型评估。线性回归的基本应用：以基因表达数据为例。

### 三、教学学时安排

2学时。

### 四、教学方法

课堂讲授：板书，操作演示，多媒体辅助教学。

## 第十三章 常用函数和自定义函数

### 一、目的要求

（一）了解R语言的常用函数的使用场景和目的。

（二）掌握常用函数的功能，输入输出。

（三）掌握自定义函数的组成、内变量的创建

### 二、教学内容

（一）类型及类型转换函数、数学函数、统计函数、概率函数、字符处理函数、向量操作函数其它实用函数

（二）自定义函数的基本组成、函数内变量的创建、函数的输出、函数的保存与调用函数的注释

### 三、教学学时安排

3学时。

### 四、教学方法

课堂讲授：板书，操作演示，多媒体辅助教学。

## 第十四章 循环和条件控制

### 一、目的要求

（一）了解R语言的条件控制和循环控制，R语言中的并行方法。熟悉R语言的高级循环。

（二）掌握流程控制的基本语法。

### 二、教学内容

（一）条件流程控制：if、else、ifelse、矢量化的if、switch、代码风格。

（二）循环流程控制：repeat、while、for。

（三）高级循环：replication、apply家族函数、plyr包、拆分-应用-合并。

（四）流程控制在生物信息学数据分析中的应用：以染色体数据的批量处理为例。

### 三、教学学时安排

3学时。

### 四、教学方法

课堂讲授：板书，操作演示，多媒体辅助教学。

《生物信息数据处理》实验课教学大纲

（适用于四年制生物信息学专业）

## 前言

《生物信息数据处理》课程是根据生物信息学专业本科培养目标开设的一门理工类基础课程。生物信息学数据的处理与分析通常比较繁杂，涉及的数据类型和文件格式多种多样，分析内容更是五花八门。要想胜任生物信息学数据分析的重任，离不开高效的工作平台和操作环境。Linux作为稳定、安全、免费、开放源代码的操作系统，已经成为生物信息学分析的主要工作平台，而R作为统计分析和可视化的语言，则因其语法的简洁、使用的便捷和代码的开源，已经成为生物信息学分析的主要操作环境之一。本课程主要讲述Linux操作系统的基本概念和常用命令，以及R语言的基本语法和数据处理与分析流程，并结合生物信息学数据处理和分析的实际案例讲解它们的具体应用。

《生物信息数据处理》总共72学时，其中，实验教学36学时。实验教学的目的是在学习理论知识的基础上，培养学生使用Linux和R的基本技能，为使用Linux和R进行生物信息学数据的处理与分析打下理论和实践基础。

## 实验一 熟悉系统的操作界面和登录退出

### 一、目的要求

（一）了解项目分析的基本流程和注意事项。

（二）熟悉项目空间的规划，Linux命令手册的查找与使用。

（三）掌握Linux服务器的登录和退出方法，Linux中目录和文件操作的基本命令。

### 二、教学内容

（一）规划数据分析项目空间，创建项目空间基本目录，修改目录和文件的权限。

（二）通过查找命令手册学习帮助手册的基本结构。

（三）撰写数据分析项目的日志说明。

### 三、教学学时安排

2学时。

### 四、教学方法

实验操作。

## 实验二 文件系统常用操作

### 一、目的要求

（一）了解文件系统的构架。

（二）熟悉Linux中常用的文件处理命令：mkdir、chmod、cut、du等。

（三）掌握文件处理命令的基本用法。

### 二、教学内容

（一）创建自己工作目录

（二）进行个人目录分类：课程分类，实验内容分类，程序分类

（三）熟悉常用的生信分析文件格式，比如gtf文件

（四）更改gtf的访问权限，仅自己可写，其他人仅可读

（五）扩展内容：统计自己目录下各个文件占用空间大小

（六）扩展内容：统计染色体文件中染色体的数目

### 三、教学学时安排

2学时。

### 四、教学方法

实验操作。

## 实验三 常用命令的组合使用

### 一、目的要求

（一）熟悉sort、mv、cp、cut、head、tail等命令的操作。

（二）掌握命令的组合使用方法，统计文件信息。

### 二、教学内容

（一）在自己的课程目录下创建database目录，将gencode.v38.annotation-test.gtf存放在该目录下，并且修改名字为test.gtf

（二）列出gtf文件中染色体上的后 10 个 CDS （按照每个CDS终止位置的基因组坐标进行排序）。

（三）利用 sort 命令依照第4和第5列数字大小排序，将最终结果输出到result.gtf文件中。

（四）统计染色体上各类 feature （gtf文件的第3列，有些注释文件中还应同时考虑第2列） 的数目，并按升序排列。

（五）根据染色体号和feature这两列信息，统计这两列信息的数目

（六）对于示例文件（test.gtf），尝试使用相关命令或命令组合分别统计文件的行数以及字符数。

（七）统计gtf文件中有多少个不同的gene\_id。

### 三、教学学时安排

2学时。

### 四、教学方法

实验操作。

## 实验四 生物信息学常见数据格式的进阶分析

### 一、目的要求

（一）了解BED、GFF和VCF格式。

（二）熟悉sed、AWK、datamash等工具。

（三）掌握sed和AWK的基本用法。

### 二、教学内容

（一）sed的应用：BED、GFF和VCF格式的统计、汇总、处理。

（二）AWK的应用：BED和GFF格式的变换与统计。

（三）其他常用数据分析工具的应用。

### 三、教学学时安排

2学时。

### 四、教学方法

实验操作。

## 实验五 生物信息学分析环境的配置

### 一、目的要求

（一）了解环境变量和配置文件。

（二）熟悉Vim编辑器的高级用法。

（三）掌握conda的使用，Vim编辑器的基本用法。

### 二、教学内容

（一）conda的使用：conda和bioconda的安装、环境的创建、软件的安装。

（二）Vim编辑器的使用：移动定位、插入删除、复制粘贴、搜索替换、保存退出。

（三）使用环境变量和配置文件对conda和Vim进行个性化配置。

### 三、教学学时安排

2学时。

### 四、教学方法

实验操作。

## 实验六 shell编程之条件控制

### 一、目的要求

（一）了解shell中的并行。

（二）熟悉shell中的条件测试和流程控制的基本用法。

（三）掌握shell流程控制在多条染色体数据处理中的应用。

### 二、教学内容

（一）循环读取目录中的多条染色体数据。根据条件筛选染色体数据。

（二）并行处理多条染色体数据。

### 三、教学学时安排

2学时。

### 四、教学方法

实验操作。

## 实验七 shell编程之循环控制

### 一、目的要求

（一）熟悉shell中的循环控制的基本用法。

（二）掌握shell循环控制在修改测序数据处理中的应用。

### 二、教学内容

（一）修改test.fastq文件的序列名称，只保留空格前面的唯一ID, 其余内容保持不变输出。

（二）每四行读取一组序列

（三）处理需要改变的行，其他行保持不变输出

### 三、教学学时安排

4学时。

### 四、教学方法

实验操作。

## 实验八 R语言环境的搭建

### 一、目的要求

（一）了解R包的安装与载入。熟悉RStudio的基本用法。

（二）掌握R的基本语法和R包中函数的使用。

### 二、教学内容

使用R的基本语法处理DNA序列：生成、统计、变换。安装、载入seqinr包，查找并学习comp函数的用法。使用seqinr包中的函数处理DNA序列。

### 三、教学学时安排

2学时。

### 四、教学方法

实验操作。

## 实验九 网络数据的统计分析

### 一、目的要求

（一）了解网络中点和边的含义。熟悉生物信息中TF和target的统计含义。

（二）掌握R语言中数据的匹配规则。

### 二、教学内容

（一）统计调控网络中节点和边个数（这里可能涉及判断节点和边是否有重复）

（二）输出网络中的TFs和targets（去除冗余情况），并统计个数。

（三）输出靶基因中哪些是TFs,并提取对应的调控关系。

（四）对于TF\_target，给出调控关系对应的基因名字。

### 三、教学学时安排

2学时。

### 四、教学方法

实验操作。

## 实验十 基因表达数据的统计和筛选

### 一、目的要求

（一）了解基因表达数据的含义。

（二）熟悉数据中显著性的使用规则。

（三）掌握差异基因的含义和筛选规则。

### 二、教学内容

（一）加载工作空间” geneExp5000.Rdata” ，查看里面的变量的类型和维度。

（二）统计每个基因在不同样本中的表达均值（行），以及相同样本中不同基因的表达均值（列）

（三）基于基因表达谱数据，计算基因之间的相关系数，以及相关性显著性p值。

（四）提取相关系数最大的前20%，且显著性水平P<0.05的基因名字，将结果存入result.Rdata文件中。

### 三、教学学时安排

2学时。

### 四、教学方法

实验操作。

## 实验十一 拷贝数变异信号的识别

### 一、目的要求

（一）了解R中随机函数的使用。

（二）熟悉对称矩阵的使用规律。

（三）掌握基因组中拷贝数信号的含义和计算方法。

### 二、教学内容

（一）构建一个10\*10的随机矩阵

（二）使其变成对称矩阵，要求左上三角与右下三角对称；记录运行所需时间

（三）假设一条染色体上有1000个基因，现在有每个基因的表达向量geneExp，需要对每个基因根据它上下游x（x默认值设为50）个基因的表达取均值，作为该基因的拷贝数信号。

教学学时安排

### 三、教学学时安排

2学时。

### 四、教学方法

实验操作。

## 实验十二 常用数据格式转化

### 一、目的要求

（一）了解常见的测序文件格式。

（二）熟悉差异分析的常用方法。

（三）掌握R语言中蛋白质序列分析和二级结构预测的方法。

### 二、教学内容

（一） R语言编写一组程序，实现sam文件格式到bed文件格式的转换。

（二）常规基因的差异分析：limma。

（三）使用DECIPHER和bamboo包预测蛋白质的二级结构并进行比较。

### 三、教学学时安排

2学时。

### 四、教学方法

实验操作。

## 实验十三 数据的简单可视化

### 一、目的要求

（一）了解基因集富集分析的目的。

（二）熟悉基因集富集分析的基本原理。

（三）掌握R语言中基因集富集分析的实现方法。

### 二、教学内容

（一）安装clusterProfiler包，查找并理解相关包的作用。使用clusterProfiler包进行基因集富集分析。

（二）对基因集富集分析的结果进行可视化。

### 三、教学学时安排

2学时。

### 四、教学方法

实验操作。

## 实验十四 ggplot2绘制数据分析图

### 

### 一、目的要求

（一）了解生物信息学中的常见数据分析图。熟悉ggplot2的基本语法。

（二）掌握ggplot2的基本用法。

### 二、教学内容

（一）使用ggplot2绘制单变量图：直方图、密度分布图、箱线图、条形图、饼图，等。

（二）使用ggplot2绘制双变量图：散点图、 热图、火山图、韦恩图，等。

（三）ggplot2的属性调整：颜色、文字、图例、注释、主题。

### 三、教学学时安排

4学时。

### 四、教学方法

实验操作。

## 实验十五 数据处理与综合实践

### 一、目的要求

（一）了解DNA、RNA和表观遗传学的常见数据内容。熟悉生物信息学数据分析项目的规划与协作。

（二）掌握生物信息学数据常见分析和可视化的技能。

### 二、教学内容

（一）分小组选题并理解数据内容：DNA变异数据分析、RNA差异表达数据分析、ATAC染色质可及性数据分析、单细胞RNA表达数据分析、自选题目。

（二）小组内部讨论数据分析内容、分析策略，并查找可用工具、学习使用方法。

（三）整理数据分析结果并撰写项目的数据分析报告。

### 三、教学学时安排

4学时。

### 四、教学方法

自主学习，小组讨论，实验操作。