《生物信息学算法》理论课教学大纲

(适用于四年制生物信息学专业)

前 言

《生物信息学算法》是生物信息学基础课程，讲授典型计算机算法设计方法和算法复杂性概念及分析方法相关的知识，培养学生利用典型计算机算法解决生物学问题的能力，使学生掌握算法设计的基本方法，熟悉算法分析的基本技术，并能熟练运用一些常用算法解决生物学问题。

本教学大纲所安排的教学内容是为四年制生物信息学专业的学生而制定。 《生物信息学算法》课程为专业必修课，总学时数为54，其中理论课学时为36。

# 绪论

## 目标要求

* 1. 掌握算法的定义、特性，算法的设计步骤等内容
  2. 掌握伪代码基本命令
  3. 掌握算法的时间和空间复杂度估计方法

## 教学内容

* 1. 算法的基本概念：
     1. 定义
        1. 算法是对特定问题求解步骤的一种描述，它是指令的有限序列，其中的每条指令表示一个或多个操作。
        2. 算法就是为了解决一个适当的公式化表示的问题而必须执行的一系列指令。
        3. 算法是任何定义好了的计算程式，它取某些值或值的集合作为输入，并产生某些值或值的集合作为输出。因此，算法是将输入转化为输出的一系列计算步骤。
     2. 算法的5个特性：有穷性、确定性、可行性、输入、输出
     3. 好算法的特质：正确性、可读性、健壮性、高效率与低储存量需求（时间复杂度和空间复杂度）
     4. 算法设计步骤：
        1. 识别、分析遇到的问题
        2. 设计算法
        3. 算法评估（正确性，运行效率）
  2. 算法效率的度量
     1. 时间复杂度
     2. 空间复杂度
  3. 伪代码基本命令

## 教学学时安排

2学时

## 教学方法

板书、多媒体辅助教学

# 枚举算法

## 目标要求

* 1. 理解枚举搜索算法的原理
  2. 掌握枚举搜索算法的设计步骤

## 教学内容

* 1. 枚举算法的基本概念：
     1. 导入：水仙数
     2. 定义
        1. 按照问题本身的性质，一一列举出该问题所有可能得解，并在列举的过程中，逐一检验每个可能解是否是问题的真正解。若是则采纳这个解，否则抛弃它。
     3. 重点
        1. 不能遗漏，否则可能导致结果不正确
        2. 不要重复，否则可能导致效率比较低
     4. 特点
        1. 穷举算法的解准确且全面
        2. 实现简单，通过循环/递归实现
        3. 执行效率提升空间往往比较大
     5. 穷举算法设计步骤
        1. 确定穷举对象
        2. 逐一列举可能解
        3. 逐一验证可能解
  2. 生物学问题：
     1. 限制酶切作图（重点）
     2. 基序发现问题
        1. 生物学背景：生物钟 → 基因调控
  3. 算法优化：分支定界法

## 教学学时安排

2学时

## 教学方法

板书、多媒体辅助教学

# 贪婪算法

## 目的要求

* + 1. 理解贪婪算法的原理
    2. 掌握贪婪算法的设计步骤

## 教学内容

* 1. 贪婪算法的基本概念
     1. 导入：找零钱问题
     2. 定义：贪婪算法是一个分阶段决策过程，在每个局部阶段，贪婪算法都做出一个当前最优的局部决策，并期望通过每次所做的局部最优决策产生一个全局最优解。
     3. 特点：以当前情况为基础作出最优选择，而不考虑各种可能得整体情况；一般可以快速得到满意的解，但可能会返回次优解而不是最优解。
     4. 贪婪算法设计步骤：
        1. 分解：将原问题求解过程划分为连续的若干个决策阶段；
        2. 决策：在每一个阶段依据贪婪策略进行贪婪决策，得到局部的最优解，并缩小待求解问题的规模
        3. 合并：将各个阶段的局部解合并为原问题的一个全局最优解
     5. 近似算法
        1. 近似率
  2. 生物学问题举例：
     1. 基因组重排
        1. 反序排序法
        2. 断点优化
     2. 基序发现问题

## 教学学时安排

2学时

## 教学方法

板书、多媒体辅助教学

# 动态规划算法

## 目的要求

1. 理解动态规划算法的原理
2. 掌握动态规划算法的设计步骤

## 教学内容

* 1. 动态规划的基本概念
     1. 导入：曼哈顿游客问题
     2. 定义：动态规划是运筹学的一个分支，20世纪50年代初美国数学家R.E.Bellman等人在研究多阶段决策过程的优化问题时，提出了著名的最优化原理，把多阶段过程转化为一系列单阶段问题，逐个求解，创立了解决这类过程优化问题的新方法——动态规划。
     3. 基本要素：
        1. 最优子结构性质
        2. 子问题重叠性质
     4. 动态规划算法设计步骤
        1. 分析最优子结构性质
        2. 确定状态表示和状态递推方程，递归地定义最优值
        3. 确定状态转移顺序，以自底向上的方式计算出最优值
        4. 根据计算最优值时得到的信息，构造最优解
  2. 普通问题：
     1. 矩阵连乘
     2. 背包问题
  3. 生物学问题：序列比对
     1. 最长公共子序列
     2. 全局序列联配
     3. 得分联配
     4. 局部序列联配
     5. 缺口罚分联配
     6. 多重联配

## 教学学时安排

4学时

## 教学方法

板书、多媒体辅助教学

# 分而治之算法

## 目标要求

1. 理解分而治之算法的原理
2. 掌握分而治之算法的设计步骤

## 教学内容

1. 分治的基本概念
   * 1. 导入：称假币
     2. 分治算法设计步骤
        1. 分：将大规模的原问题分割成k个更小规模的子问题，如果子问题的规模仍然不够小，则再划分为k个“子子”问题，如此递归地进行下去，直到子问题规模足够小（基础问题），很容易求出其解为止；
        2. 治：求解规模足够小的基础问题；
        3. 合：将求出的小规模的问题的解合并为一个更大规模的问题的解，自底向上逐步求出原来问题的解。
     3. 重要原则
        1. 平衡子问题原则：分割出的k个子问题其规模最好大致相当；
        2. 独立子问题原则：分割出的k个子问题之间重叠越少越好，最好k个子问题是相互独立，不存在重叠子问题。
2. 普通问题：合并排序
3. 生物学问题：空间效率高的序列联配算法

## 教学学时安排

4学时

## 教学方法

板书、多媒体辅助教学

# 图算法

## 目标要求

1. 理解图的基本概念和术语
2. 掌握图的表示方式，两种遍历的算法思想、步骤
3. 理解图的应用

## 教学内容

1. 基本概念
   * 1. 导入
        1. 一笔画问题：手机锁屏问题，柯尼斯堡七桥问题
        2. 生活中的图问题：计算机网络、社交网络等都可以抽象为图问题
     2. 图的基本概念
        1. 图的定义：顶点、边
2. 有向图
3. 无向图
4. 简单图、多重图
5. 完全图
6. 子图
7. 连通、连通图和连通分量：相邻和关联
8. 强连通图、强连通分量
9. 生成树、生成森林
10. 度：顶点的度、图的度

握手定理：无向图的度是边数的两倍

a.证明柯尼斯堡七桥问题无解

1. 边的权和网
2. 路径、路径长度和回路
3. 简单路径、简单回路
4. 距离
5. 有向树
   * 1. 图的表示方法：
        1. 邻接链表
        2. 邻接矩阵
        3. 十字链表
        4. 邻接多重表
6. 图的遍历
   * 1. 广度优先搜索
     2. 深度优先搜索
7. 图的应用
   * 1. 最小生成树
     2. 最短路径
     3. 拓扑排序
     4. 关键路径
8. 生物学问题
   * 1. DNA序列组装
     2. 蛋白质鉴定

## 教学学时安排

6学时

## 教学方法

板书、多媒体辅助教学

# 聚类和树

## 目标要求

1. 理解聚类和树的原理
2. 掌握聚类和树的设计步骤

## 教学内容

1. 基本概念
   * 1. 导入：基因表达分析
     2. 距离计算
        1. 闵可夫斯基距离：
2. 欧氏距离
3. 曼哈顿距离
   * 1. 聚类算法分类：
        1. 原型聚类
4. 定义：基于原型的聚类。通常情形下，算法先对原型进行初始化，然后对原型进行迭代更新求解。采用不同的原型表示、不同的求解方式，将产生不同的算法。
5. 分类：
   * + - 1. K均值算法
         2. 学习向量化
         3. 高斯混合聚类
       1. 密度聚类
6. 定义：基于密度的聚类。此类算法假设聚类结构能通过样本分布的紧密程度确定
7. 分类：
   * + - 1. DBSCAN
       1. 层次聚类（hierarchical clustering，分级聚类，系统聚类）
8. 定义：层析聚类试图在不同层次对数据集进行划分，从而形成树形的聚类结构。
9. 分类：
   * + - 1. AGNES算法

单链接算法：最小距离

全链接算法：最大距离

均链接算法：平均距离

1. 生物学问题
   * 1. DNA阵列分析
     2. 基于距离法的建树方法：UPGMA法、Fitch-Margoliash法、邻接法和最小进化法

## 教学学时安排

4学时

## 教学方法

板书、多媒体辅助教学

# 隐马尔科夫模型

## 目标要求

1. 理解隐马尔科夫模型的原理

## 教学内容

1. 背景知识
   * 1. 概率论
2. 基本概念
   * 1. 定义：隐马尔科夫模型是关于时序的概率模型，描述由一个隐藏的马尔科夫链随机生成不可观测的状态随机序列，再由各个状态生成一个观测而产生观测随机序列的过程。隐藏的马尔科夫链随机生成的状态的序列，称为状态序列；每个状态生成一个观测，而由此产生的观测的随机序列，称为观测序列。序列的每一个位置又可以看作是一个时刻。
     2. 数学定义：初始概率分布、状态转移概率分布以及观测概率分布
     3. 隐马尔科夫模型的两个假设：
        1. 齐次马尔科夫性假设：假设隐藏的马尔科夫链在任意时刻t的状态只依赖于其前一时刻的状态，与其他时刻的状态及观测无关，也与时刻t无关
        2. 观测独立性假设：假设任意时刻的观测只依赖于该时刻的马尔科夫链的状态，与其它观测及状态无关
     4. 举例：盒子和球模型
     5. 三个问题：
        1. 评估问题、解码问题和学习问题
3. 生物学问题
   * 1. 基因结构预测
     2. CpG岛识别

## 教学学时安排

4学时

## 教学方法

板书、多媒体辅助教学

# 随机化算法（概率算法）

## 目标要求

1. 理解随机化算法的原理
2. 掌握随机化算法的设计步骤

## 教学内容

1. 基本概念
   * 1. 介绍：随机化算法允许算法在执行过程中随机地选择下一个计算步骤。这种算法把随机性注入到算法中。
     2. 特点：对所求解问题的同一实例用同一随机化算法求解两次可能得到完全不同的结果，这两次求解问题所需的时间甚至所得到的结果可能会有相当大的差别。
     3. 分类：
        1. 数值随机化算法
        2. 蒙特卡罗算法（重点）
        3. 拉斯维加斯算法
        4. 舍伍德算法
2. 普通问题
   * 1. 计算π的值
     2. 计算定积分
3. 生物学问题
   * 1. 基序发现的随机化方法
        1. 吉布斯抽样（重点）
        2. 随机投影

## 教学学时安排

4学时

## 教学方法

板书、多媒体辅助教学

# 组合模式匹配

## 目标要求

1. 理解组合模式匹配的原理
2. 掌握组合模式匹配的设计步骤

## 教学内容

1. 组合模式匹配的基本概念
   * 1. 定义：组合模式匹配是在一个长文本文件中寻找给定模式的精确或者近似的匹配。
     2. 与基序发现问题对比
     3. 基序发现问题：在不给定任何特定模式的情形下，在DNA序列中查找高度重复的模式（pattern）
     4. 难点：实际应用中，基因组数据量庞大，使得整个问题变得相当困难。尤其是当搜索许多较长模式的近似匹配的时候。
2. 生物学问题：重复序列发现
   * 1. 哈希表
     2. 精确模式匹配
     3. 关键词树
     4. 后缀树
     5. 启发式相似搜索算法
     6. 近似模式匹配

## 教学学时安排

4学时

## 教学方法

板书、多媒体辅助教学

《生物信息学算法》实验课教学大纲

（适用于四年制生物信息学专业）

前 言

#### 《生物信息学算法》的实验课程是与理论课程紧密结合的实践性教学环节，旨在通过实际操作加深学生对生物信息学算法原理和应用的理解。实验课程的主要目的是让学生通过实践操作，掌握生物信息学算法的基本原理，理解算法的理论基础，并能够将所学知识应用于解决实际生物学问题。实验课教学学时为18。

# 实验一 枚举算法

一、实验目的

（一）掌握枚举算法基本原理

（二）能够使用python实现简单枚举算法

二、实验内容

（一）模糊数字问题

（二）Counting Words

（三）The Frequent Words Problem

三、学时安排

2学时

四、实验教学方法

在机房中，教师先进行实验演示，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

# 实验二 贪婪算法

一、实验目的

（一）掌握贪婪算法基本原理

（二）能够使用python实现简单贪婪算法

二、实验内容

（一）活动安排问题

（二）小数背包问题

（三）Bio.motifs

（四）Profile-most Probable k-mer Problem

（五）Greedy Motif Search

三、学时安排

2学时

四、实验教学方法

在机房中，教师先进行实验演示，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

# 实验三 动态规划算法

一、实验目的

（一）掌握动态规划算法基本原理

（二）能够使用python实现简单动态规划算法

二、实验内容

（一）斐波那契数列

（二）0-1背包问题

（三）最长公共子序列问题

三、学时安排

2学时

四、实验教学方法

在机房中，教师先进行实验演示，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

# 实验四 分而治之算法

一、实验目的

（一）掌握分治算法基本原理

（二）能够使用python实现简单分治算法

二、实验内容

（一）合并排序

（二）空间效率高的序列联配算法

三、学时安排

2学时

四、实验教学方法

在机房中，教师先进行实验演示，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

# 实验五 图算法

一、实验目的

（一）掌握图算法基本原理

（二）能够实现图算法操作

二、实验内容

（一）最小生成树

（二）DNA序列组装

三、学时安排

2学时

四、实验教学方法

在机房中，教师先进行实验演示，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

# 实验六 聚类和树

一、实验目的

（一）掌握聚类和树算法基本原理

（二）能够实现聚类和树算法操作

二、实验内容

（一）各种距离计算方法

（二）各种聚类计算方法

（三）基于不同方法的建树方法

三、学时安排

2学时

四、实验教学方法

在机房中，教师先进行实验演示，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

# 实验七 隐马尔科夫模型

一、实验目的

（一）掌握隐马尔科夫模型基本原理

（二）能够实现隐马尔科夫模型操作

二、实验内容

（一）基因结构预测

（二）CpG岛识别

三、学时安排

2学时

四、实验教学方法

在机房中，教师先进行实验演示，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

# 实验八 随机化算法

一、实验目的

（一）掌握随机化算法基本原理

（二）能够实现随机化算法操作

二、实验内容

（一）计算π的值

（二）计算定积分

（三）基序发现问题的随机化方法

三、学时安排

2学时

四、实验教学方法

在机房中，教师先进行实验演示，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

# 实验九 组合模式匹配

一、实验目的

（一）掌握随组合模式匹配基本原理

（二）能够实现组合模式匹配操作

二、实验内容

（一）重复序列发现：哈希表

（二）重复序列发现：精确模式匹配

三、学时安排

2学时

四、实验教学方法

在机房中，教师先进行实验演示，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。