

# 北京大学 2018 级研究生教学大纲

2019 年 2 月 20 日

## 实用生物信息技术

Applied Bioinformatics Course

课程编号: 01111590

课程属性: 选修课

学时学分: 64/4

预修课程: 生物化学、分子生物学

授课对象: 从事生命科学实验研究的博士或硕博连读研究生

上课时间: 2019/2/23-2019/6/15, 每周六下午 1:30-5:00, 共 16 周

上课地点: 计算中心三楼 8 号机房, 可自带笔记本电脑

小组讨论: 课外组织小组讨论, 每周 3 学时, 撰写讨论报告

课外作业: 结合本人研究课题, 共 5 次作业

期末交流: 每组推选一位同学, 结合课题进行总结交流

期末总结: 将课上介绍的思路、方法、平台、工具用于自己的研究课题

教学网站: <http://abc.cbi.pku.edu.cn/>

专用邮箱: abc.cbi@pku.edu.cn

### 教学目的

本课程是一门上机操作实验课, 课程教学在多媒体教室。课上将结合具体实例, 介绍常用生物信息技术。欢迎从事实验研究的研究生选修本课程, 针对正在进行或即将开始的研究课题, 利用丰富的生物信息数据库资源和分析工具, 解决研究过程遇到的实际问题。

### 任课教师

罗静初, 北京大学生命科学学院教授, 曾从事生物信息研究开发和教学, 2012 年退休。2000 年起, 为北京大学和中国农业科学院研究生开设“实用生物信息技术”研究生学期课; 2014 年起, 应聘为中国科学院大学硕士研究生开设“实用生物信息技术”暑期课。

### 内容提要

#### 第一讲 序列比对

##### 1. 序列比对基本概念

- 1.1. 序列相似性 (Similarity) 和序列同源性 (Homology)
- 1.2. 直系同源 (Ortholog) 和旁系同源 (Paralog)
- 1.3. 动态规划算法 (Dynamic Programming) 和启发式算法 (Heuristic Programming)
- 1.4. 计分矩阵 (Scoring Matrix) 和空位罚分 (Gap Penalty)
- 1.5. PAM (Point Accept Mutation) 计分矩阵和 BLOSUM (Block Substitution) 计分矩阵

- 1.6. 全局比对 (Global Alignment) 和局部比对 (Local Alignment)
- 1.7. 双序列比对 (Pairwise Sequence Alignment) 和多序列比对 (Multiple Sequence Alignment)
- 1.8. 序列比对点阵图方法 (Dot Plot)
- 1.9. 多序列比对序列图标 (Sequence Logo)

## **2. 序列比对常用工具和网站**

- 2.1. 北京大学生物信息中心 (CBI) 生物信息网上实验室 (WebLab)
- 2.2. 中国科学院网络中心 (CNIC) 微生物模型云服务 (MicroCloud)
- 2.3. 荷兰瓦格宁根大学生物信息中心 (WUR) 序列分析在线服务 (EMBOSS Explorer)
- 2.4. 欧洲生物信息学研究所 (EBI) 在线序列比对工具
- 2.5. 美国国家生物信息中心 (NCBI) 在线序列比对工具

## **3. 简例**

- 3.1. 双序列全局比对
- 3.2. 双序列局部比对
- 3.3. 多序列全局比对
- 3.4. 多序列局部比对
- 3.5. 利用点阵图寻找重复序列
- 3.6. 利用序列图标寻找保守位点

## **4. 序列比对应用实例：人、小鼠和大鼠血红蛋白及其编码区序列比对**

- 4.1. 研究背景
- 4.2. 血红蛋白 alpha 亚基氨基酸序列比对
- 4.3. Alpha 珠蛋白编码区核苷酸序列比对
- 4.4. 结果和讨论

## **第二讲 数据库高级检索**

### **1. 蛋白质序列和功能数据库 UniProt**

- 1.1. UniProt 数据库简介
- 1.2. UniProt 数据库统计报表
- 1.3. UniProt 数据库中序列条目注释信息
- 1.4. UniProt 数据库中序列位点和结构域注释信息
- 1.5. UniProt 数据库序列条目与其它数据库的链接
- 1.6. UniProt 数据库高级检索应用实例
- 1.7. UniProt 数据库系统中的实用程序
- 1.8. UniProt 数据库用户界面
- 1.9. UniProt 数据库帮助文档

### **2. 核酸参考序列数据库 RefSeq**

- 2.1. RefSeq 数据库特点简介
- 2.2. RefSeq 数据库与 GenBank 数据库的区别
- 2.3. RefSeq 数据库中序列条目登录号规则和含义

- 2.4. RefSeq 数据库高级检索应用实例
- 2.5. RefSeq 数据库序列条目与 NCBI 其它数据库的链接

### **3. 基因组数据库**

- 3.1. 常用基因组数据库和基因组浏览器
- 3.2. Ensembl 基因组数据库简介
- 3.3. Ensembl 基因组数据库用户界面
- 3.4. Ensembl 基因组数据库应用实例
- 3.5. 模式生物基因组数据库
- 3.6. 植物基因组数据库
- 3.7. 微生物基因组数据库
- 3.8. 病毒分类和生物信息资源数据库

### **4. 生物大分子专业数据库简介**

- 4.1. 生物信息数据库资源导航 Database Common
- 4.2. 生物大分子家族数据库
- 4.3. 蛋白质相互作用网络数据库
- 4.4. 代谢通路数据库
- 4.5. 直系同源序列数据库
- 4.6. 基因本体数据库
- 4.7. 蛋白质表达数据库
- 4.8. microRNA 数据库

## **第三讲 数据库序列相似性搜索**

### **1. 蛋白质搜索程序**

- 1.1. 蛋白质搜索常规程序 BlastP 应用实例
- 1.2. 蛋白质位点特异性迭代搜索程序 PSI-Blast 应用实例
- 1.3. 蛋白质序列模体特异搜索程序 PHI-Blast 应用实例
- 1.4. 蛋白质结构域数据库搜索程序 DELTA-Blast 应用实例
- 1.5. 蛋白质快速搜索程序 Quick Blast 应用实例
- 1.6. 核酸序列搜索蛋白质数据库程序 BlastX 应用实例
- 1.7. 蛋白质序列搜索核酸序列数据库 tBlastN 应用实例

### **2. Blast 算法简介和参数设置**

- 2.1. Blast 算法简介
- 2.2. 如何选择计分矩阵
- 2.3. 如何设置空位罚分
- 2.4. 如何设置错误率 E 值
- 2.5. 如何选择搜索字长 WORD

### **3. NCBI Blast 平台其它工具**

- 3.1. NCBI Blast 系统用户界面和常用程序

- 3.2. 模式生物蛋白质序列数据库搜索程序 SmartBlast
- 3.3. 测序载体接头序列检测程序 VecScreen
- 3.4. 引物设计程序 Primer-Blast
- 3.5. 免疫球蛋白数据库搜索专用程序 IgBlast

#### **4. Blast 使用经验点滴**

- 4.1. 如何选择数据库
- 4.2. 如何选择物种
- 4.3. 如何选择不同程序
- 4.4. 如何改变输出格式
- 4.5. 如何分析输出结果
- 4.6. 如何下载输出结果
- 4.7. 如何保存搜索策略
- 4.8. 如何搜索短序列
- 4.9. 如何屏蔽重复序列

#### **5. Linux 系统本地 Blast 简介**

- 5.1 Linux 系统简介
- 5.2 Linux 系统常用命令
- 5.3 SBP 转录因子本地 Blast 运行实例
- 5.4 SBP 转录因子搜索结果分析

### **第四讲 系统发生树构建**

#### **1. 分子演化和系统发生树基本概念**

- 1.1. 物种分化和分子演化
- 1.2. 分支图和系统树
- 1.3. 物种树和基因树
- 1.4. 有根数和无根树
- 1.5. 二叉树与多歧树
- 1.6. 外部节点和内部节点
- 1.7. 内部节点和根节点
- 1.8. 系统发生树稳定性检验

#### **2. 常用系统发生树构建方法**

- 2.1. 距离法
- 2.2. 最大简约法
- 2.3. 最大似然法
- 2.4. 贝叶斯推断

#### **3. 常用系统发生分析软件**

- 3.1. MEGA
- 3.2. DAMBE

3.3. Phylip

3.4. PAML

#### **4. 系统发生树构建和分析实例**

4.1. 人珠蛋白家族 12 个成员基因树

4.2. 人、小鼠、大鼠珠蛋白家族 37 个成员物种和基因树

4.3. 拟南芥和水稻 SBP 转录因子家族

4.4. 典型系统发生树案例分析

#### **第五讲 蛋白质结构比较和分析**

##### **1. 蛋白质结构基本概念**

1.1. 蛋白质结构层次

1.2. 氨基酸种类和性质

1.3. 多肽链构象特征

1.4. 蛋白质分子内部作用力

1.5. 蛋白质中的水分子、金属和有机小分子

##### **2. 蛋白质结构数据库**

2.1. 蛋白质结构数据库 PDB 简介

2.2. 蛋白质结构数据库统计报表

2.3. 蛋白质结构数据库高级检索

2.4. 蛋白质结构在线显示

2.5. 蛋白质序列和结构分析

##### **3. 蛋白质结构分析软件 Swiss-PDBViewer**

3.1. 用户界面

3.2. 基本操作

3.3. 高级操作

##### **4. 蛋白质结构显示分析简例**

4.1. 猪胰岛素 (Insulin, 4INS)

4.2. 人免疫球蛋白 (Immunoglobulin, 1IGT)

4.3. 绿色荧光蛋白 (Green Fluorescent Protein, 1GFL)

4.4. 非洲爪蟾锌指蛋白 (Zinc Finger Protein, 1TF6)

4.5. 非洲爪蟾核小体 (Nucleosome, 1AOI)

##### **5. 斑头雁和灰雁血红蛋白结构分析实例**

5.1. 研究背景

5.2. 序列、结构比较

5.3. 结果讨论

##### **6. 蛋白质结构预测**

6.1. 蛋白质结构预测常用方法

6.2. 蛋白质结构预测常用网站

- 6.3. 同源建模基本原理和步骤
- 6.4. 癌胚抗原 (Carcinoembryonic Antigen, CEA) CEAM5 结构预测实例
- 6.5. 癌胚抗原 CEA21 结构预测实例

## **第六讲 分析实例**

### **1. 豌豆内膜蛋白 PPF1 分析实例**

- 1.1 研究背景
- 1.2 读码框分析工具 PlotORF/ShowORF/GetORF
- 1.3 蛋白质一级结构分析工具 ProtScale
- 1.4 跨膜螺旋预测程序 TMHMM
- 1.5 蛋白质螺旋轮显示程序 PepWheel
- 1.6 亚细胞定位预测程序 TargetP
- 1.7 叶绿体定位程序 ChloroP
- 1.8 蛋白质功能预测网站 PredictProtein
- 1.9 密码子分析程序 CUSP
- 1.10 限制性内切酶分析程序 ReMap

### **2 河豚鱼多药耐药基因 MDR 分析实例**

- 1.1 研究背景
- 1.2 重复序列鉴定方法
- 1.3 基因结构预测方法
- 1.4 如何用 Blast 搜索基因组序列

### **2. 植物转录因子家族 SBP 分析实例**

- 2.1 SBP 转录因子家族研究背景
- 2.2 植物转录因子数据库 PlantTFDB
- 2.3 保守结构域预测网站 MEME
- 2.4 结构域搜索网站 SMART
- 2.5 序列谱分析工具 HMMER

### **3. 抗菌肽和蜘蛛毒素分析实例**

- 3.1 研究背景
- 3.2 低相似度多肽序列比对和手工调整
- 3.3 结构预测结果分析

## **参考文献**

- 1. 罗静初, 实用生物信息技术课程教学实例, 《生物技术通报》, 2015 年, 第 31 卷, 第 11 期, 第 102-111 页.
- 2. Luo J. Teaching the ABCs of bioinformatics: a brief introduction to the Applied Bioinformatics Course. Brief Bioinform. 2014 Nov;15(6):1004-13.

## 参考教材

1. Jin Xiong, Essential Bioinformatics. 2006, Cambridge University Press.
2. David Mount, Bioinformatics: Sequence and Genome Analysis. 2<sup>nd</sup> edition, 2004, CSHL Press.