天津医科大学实验课教案首页

课程名称：生物信息学　　**实验名称：**实验六 DNA 序列的基本信息及特征分析

教师姓名：伊现富　　**职称：**讲师　　**教学日期：**2018 年 11 月 30 日 8:00-10:45

授课对象：生物医学工程与技术学院 2016 级生信班（本）　　**实验人数：**28

实验类型（验证型、综合型、设计型、创新型）：验证型　　**实验分组：**一人一机

学时数：3　　**教材版本：**生物信息学实验讲义（自编教材）

实验目的与要求：

- 学习并掌握使用 NCBI 查询核酸序列的方法。
- 学习并掌握 EMBOSS 的基本使用方法。
- 掌握 ORF 的性质及其分析方法。

实验内容及学时分配：

- (10') 查询核酸序列：简单介绍 NCBI 数据库，讲解并演示 Nucleotide 数据库的使用。
- (10') EMBOSS 简介：介绍 EMBOSS 软件软件包，讲解并演示 compseq 等工具的使用方法。
- (10') 序列组份分析：回顾序列组份分析的主要内容。
- (10') 开放阅读框分析：回顾 ORF 的定义、相位的概念和最长 ORF 法。
- (95') 实验操作：对人类 CD9 基因序列进行组分分析，对大肠杆菌基因组序列进行 ORF 分析。

主要仪器和实验材料：

- 实验材料：人类 CD9 基因，大肠杆菌基因组。
- 主要仪器：联网的计算机。
- 分析工具：NCBI，EMBOSS。

实验重点、难点及解决策略：

- 重点难点：NCBI 数据库和 EMBOSS 软件包的使用。
- 解决策略：通过演示进行学习，通过练习熟练掌握。

思考题：

- 如何使用 NCBI 查询并下载核酸序列？
- EMBOSS 中进行序列组份分析的程序有哪些？
- getorf 和 ORF Finder 的分析结果有何异同？

参考资料：

- NCBI
- EMBOSS

主任签字：　　**年　月　日**　　**教务处制**

一、查询核酸序列（10 分钟）

1. NCBI 数据库（包含多个子数据库）
   - GenBank: The NIH genetic sequence database, an annotated collection of all publicly available DNA sequences.
   - Gene: A searchable database of genes, focusing on genomes that have been completely sequenced and that have an active research community to contribute gene-specific data. Information includes nomenclature, chromosomal localization, gene products and their attributes (e.g., protein interactions), associated markers, phenotypes, interactions, and links to citations, sequences, variation details, maps, expression reports, homologs, protein domain content, and external databases.
   - Genome: Contains sequence and map data from the whole genomes of over 1000 organisms.
   - Nucleotide Database: A collection of nucleotide sequences from several sources, including GenBank, RefSeq, the Third Party Annotation (TPA) database, and PDB. Searching the Nucleotide Database will yield available results from each of its component databases.
   - Protein Database: A database that includes protein sequence records from a variety of sources, including GenPept, RefSeq, Swiss-Prot, PIR, PRF, and PDB.
   - PubMed: A database of citations and abstracts for biomedical literature from MEDLINE and additional life science journals.
   - Reference Sequence (RefSeq): A collection of curated, non-redundant genomic DNA, transcript (RNA), and protein sequences produced by NCBI.
2. Nucleotide Database
   - 查询：ID，如 AY422198；基因名，如 CD9
   - 下载：选择需要的格式，如FASTA

二、EMBOSS 简介（10 分钟）

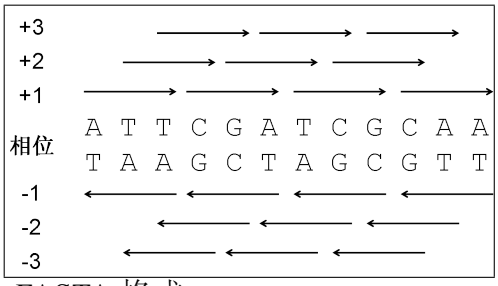
1. EMBOSS 简介
   - 开源、免费的序列分析软件包，整合了目前可以获得的大部分序列分析软件
   - 可以将系列分析工作进行无缝整合，弥补了许多软件功能分散、分析效率低下的缺陷
2. 使用界面
   - 操作系统：Linux，Mac，Windows
   - JEMBOSS：java 界面
   - EMBOSS Explorer：web 界面
3. 主要程序
   - 最重要的程序。wossname：根据关键字查找程序；showdb：显示所有整合的数据库。
   - 序列编辑。revseq：将序列反转并互补；seqret：序列格式转换。
   - 两个序列相似性图形表达。dottup：精确匹配；dotmatcher：近似匹配。
   - 双序列比对。needle：全局比对；water：局部比对。
   - 多序列比对。emma：clustalW。
   - 寻找 SNP。deffseq：仅限于双序列比对中。
   - 其他。plotorf，getorf：翻译；iep：等电点预测；tmap：跨膜区预测；pepinfo：蛋白质性质；patmatmotifs：Motif 搜索。

三、序列组分分析（10 分钟）

1. 碱基组成分析：长度，碱基数目及其比例，GC 含量
2. 序列转换：反向序列，互补序列，反向互补序列

四、开放阅读框分析（10 分钟）

1. ORF：给定的阅读框架中不包含终止密码子的一串序列
2. 相位：六相位（+1, +2, +3, -1, -2, -3）
3. 预测方法：最长 ORF 法（适用于原核生物）



五、实验操作（95 分钟）

1. 人类 CD9 基因的序列组分分析
   - 获取序列：NCBI 的 Nucleotide 数据库，AY422198，FASTA 格式
   - 打开 EMBOSS：EMBOSS Explorer
   - 碱基组分分析：compseq（注意修改参数）
   - 计算 GC 含量：geecee
   - 序列转换：revseq（调整参数即可分别获得反向序列、互补序列和反向互补序列）
2. 大肠杆菌基因组序列的 ORF 分析
   - 获取序列：NCBI 的 Nucleotide 数据库，U00096，FASTA 格式
   - 截取序列：EMBOSS，extractseq，1-3000bp（仅使用部分序列进行练习）
   - ORF 预测：EMBOSS，getorf（注意选择合适的参数）
   - 结果分析：和 NCBI 的 ORF Finder 的结果进行比较