实验六 DNA序列的基本信息及特征分析

一、实验目的

人类CD9抗原基因属于4跨膜蛋白超家族，与癌细胞黏附、迁移及增生有关，具有肿瘤转移抑制作用。应用集成工具EMBOSS的网页版EMBOSS Explorer，对人类CD9抗原基因（序列号为AY422198）进行核酸序列组分分析，同时对大肠杆菌基因组序列（U00096）进行ORF的分析，从中识别真实的ORF。

1. 学习和掌握应用NCBI查询核酸序列信息。
2. 学习和掌握EMBOSS的基本使用方法。
3. 学习和掌握核酸序列ORF的性质。

二、实验内容——图形化操作

1. 人类CD9基因的序列组分分析
    1. 获取序列。打开NCBI主页（http://www.ncbi.nlm.nih.gov/），在核酸（Nucleotide）数据库查询人类CD9抗原基因（AY422198），将其序列以FASTA格式保存到本地计算机上。
    2. 打开EMBOSS。有 http://emboss.bioinformatics.nl/ 和 http://genome.csdb.cn/emboss/ 多个EMBOSS Explorer可供选用，任选其一打开链接即可使用。
    3. 碱基组成分析。找到程序compseq，在Input section项目中，使用upload上传CD9的序列;在Required section项目中，把Word size to consider修改为“1”;其他参数默认即可。最后，点击Run compseq获得ATGC四种碱基的数目及百分比含量。
    4. GC含量分析。找到程序geecee，以上传文件的方式提交DNA序列，之后点击Run geecee得到CD9基因的GC含量。
    5. 序列转换。找到程序revseq，如前所述上传CD9基因的DNA序列，点击Run revseq后，得到其反向互补序列。通过调整Advanced section项目中的参数，可以仅获得反向序列或互补序列。
    6. 比较AY422198及其互补序列的组成成分。
2. 大肠杆菌基因组序列的ORF分析
    1. 获取序列。登录NCBI主页（http://www.ncbi.nlm.nih.gov/），在核酸（Nucleotide）数据库查询大肠杆菌基因组序列（U00096），将其以FASTA格式保存到本地计算机上。
    2. 截取序列。找到EMBOSS中的程序extractseq，在Required section项目中，通过设定Regions to extract为1-3000截取大肠杆菌基因组序列的前3000bp。
    3. ORF预测。找到程序getorf，在序列输入框中粘贴上一步截取的前3000bp序列片段，Code to use中选择“Bacterial”，Type of output中选择“Nucleic sequences between START and STOP codons”，其他默认即可。之后点击Run getorf得到ORF的预测结果。
    4. 和NCBI的ORF Finder预测结果进行比较分析。

三、实验内容——命令行操作

1. 配置环境。安装conda、bioconda，新建环境（略）。
2. 安装软件。

```
# NCBI Entrez Direct UNIX E-utilities for sequence and data retrieval from NCBI
conda install entrez-direct
# Seqtk is a fast and lightweight tool for processing sequences in the FASTA or FASTQ format
conda install seqtk
# SeqKit - a cross-platform and ultrafast toolkit for FASTA/Q file manipulation
conda install seqkit
# Bioawk is an extension to Brian Kernighan's awk, adding the support of several common biological
data formats, including optionally gzip'ed BED, GFF, SAM, VCF, FASTA/Q and TAB-delimited formats
with column names.
conda install bioawk
# OrfM: A simple and not slow open reading frame (ORF) caller.
conda install orfm
# EMBOSS
conda install emboss
```

3. 下载序列。

```
# AY422198，人类CD9基因
esearch -db nucleotide -query "AY422198" | efetch -format fasta > AY422198.fasta
esearch -db nucleotide -query "AY422198" | efetch -format gb > AY422198.gb
# U00096，大肠杆菌基因组
esearch -db nucleotide -query "U00096" | efetch -format fasta > U00096.fasta
esearch -db nucleotide -query "U00096" | efetch -format gb > U00096.gb
```

4. 组分分析与ORF预测。

```
# ATGC四种碱基的数目/百分比含量，GC含量，序列长度
seqtk comp AY422198.fasta
seqkit fx2tab AY422198.fasta  -l -g -n -i -H -BA -BC -BG -BT
compseq -sequence AY422198.fasta -word 1 -outfile AY422198_composition.txt
bioawk -c fastx '{ print $name, length($seq) }' AY422198.fasta
bioawk -c fastx '{ print $name, gc($seq) }' AY422198.fasta
# 序列转换：反向序列，互补序列，反向互补序列
seqtk seq -r AY422198.fasta > AY422198_rev_com_seqtk.fasta
seqkit seq -r AY422198.fasta > AY422198_rev_seqkit.fasta
seqkit seq -p AY422198.fasta > AY422198_com_seqkit.fasta
seqkit seq -r -p AY422198.fasta > AY422198_rev_com_seqkit.fasta
bioawk -c fastx '{ print ">"$name; print revcomp($seq) }' AY422198.fasta >
AY422198_rev_com_bioawk.fasta
# 截取序列
seqkit subseq -r 1:3000 U00096.fasta > U00096_1-3000bp_seqkit.fasta
bioawk -c fastx '{ print ">"$name; print substr($seq,1,3000) }' U00096.fasta > U00096_1-
3000bp_bioawk.fasta
# ORF预测
orfm -t orfm_dna.fasta -c 11 U00096_1-3000bp.fasta > orfm_protein.fasta
getorf -sequence U00096_1-3000bp.fasta -outseq getorf_dna.fa -table 11 -minsize 90 -find 3
```

5. 参考资料

- Bioconda Homepage
- Available packages in Bioconda
- NCBI Entrez Direct UNIX E-utilities
- Seqtk at GitHub

- SeqKit Homepage
- Usage and Examples for SeqKit
- bioawk at GitHub
- OrfM at GitHub
- EMBOSS