

实验七 原核生物和真核生物的基因识别

一、实验目的

原核生物的基因结构简单，属连续基因，识别过程相对简单；而真核基因的编码区是不连续的，结构比较复杂，需要根据基因的整体结构进行基因预测。应用GeneMarkS识别大肠杆菌基因组序列（U00096）中的基因，并用GENSCAN分析人类CD9抗原基因（序列号为AY422198）的基因结构。

1. 学习和掌握原核基因和真核基因的结构特征。
2. 学习和了解隐马尔可夫模型在基因识别中的应用。
3. 学习和掌握应用GeneMarkS、GENSCAN识别基因。

二、实验内容

1. 大肠杆菌基因组序列的基因识别

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。同时保存一份GenBank格式的基因信息，用于后续比较。
2. 截取序列。利用EMBOSS中的程序extractseq截取大肠杆菌基因组序列的前10000bp。
3. 打开GeneMarkS网站。通过搜索引擎搜索GeneMarkS，或者直接在浏览器的地址栏中输入网址，打开<http://exon.gatech.edu/GeneMark/genemarks.cgi>即可。
4. 基因识别。在Inputsequence输入框中粘贴截取的前10000bp序列片段，其他参数默认即可。点击Start GeneMarkS后，在新的页面中点击gms.out即可看到基因的识别结果。
5. 结果分析。打开保存的GenBank格式文件，找到并查看前10000bp中基因的位置信息。将其与GeneMarkS的预测结果进行比较分析。

2. 人类CD9抗原基因的结构分析

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。同时保存一份GenBank格式的基因信息，用于后续比较。
2. 打开GENSCAN网站。通过搜索引擎搜索GENSCAN，或者直接在浏览器的地址栏中输入网址，打开<http://genes.mit.edu/GENSCAN.html>即可。
3. 基因识别。上传人类CD9抗原基因的序列，其他参数默认即可。点击Run GENSCAN得到基因结构的预测结果。
4. 结果分析。打开保存的GenBank格式文件，找到并查看该基因的CDS区域信息。将其与GENSCAN的预测结果进行比较分析。