<<DNA和蛋白质序列数据分析工具>>

图书基本信息

书名：<<DNA和蛋白质序列数据分析工具>>

13位ISBN编号：9787030345097

10位ISBN编号：7030345096

出版时间：2012-6

出版时间：科学出版社

作者：薛庆中　等编著

页数：356

字数：475250

内容概要

近年来新一代测序技术的研发和应用，极大地推动了基因组科学的发展，也给基因组数据分析带来巨大的新挑战。
第三版对前两版原有内容做了大量更新和补充，《DNA和蛋白质序列数据分析工具（第三版）》17章，分别从基因组学、蛋白质组学、系统生物学三个层次详细介绍了常用的基因数据库和网络工具；为适应Windows7的环境，将BioPerl程序包的数据分析做了重排使其更易操作。
尤其是增添了新一代测序数据分析实例，包括SNVs和Indel识别、小RNA-seq分析、枯草杆菌全基因组序列拼接；并对Bowtie等读序列定位工具和UCSC浏览器的使用做介绍。

《DNA和蛋白质序列数据分析工具（第三版）》内容深入浅出、图文并茂。
书中提及的各种方法均有充实的例证并附上相关数据和图表，供读者理解和参考；书后还附有中英文的专业术语和词汇。
可作为对基因组学、蛋白质组学、生物信息学感兴趣的本科生、研究生和研究人员学习、研究的重要工具手册。

作者简介

无

<<DNA和蛋白质序列数据分析工具>>

书籍目录

第三版前言第二版前言第一版前言第1章 序列比对工具BLAST和ClustalX1.1 BLAST搜索程序1.2 本地运行BLAST(Windows系统)1.3 多序列比对(ClustalX)参考文献第2章 真核生物基因结构的预测2.1 基因可读框的识别2.2 CpG岛、转录终止信号和启动子区域的预测2.3 基因密码子偏好性计算:CodonW的使用2.4 采用mRNA序列预测基因:Spidey的使用2.5 ASTD数据库简介参考文献第3章 电子克隆3.1 种子序列的搜索3.2 序列拼接3.3 在水稻数据库中的电子延伸3.4 电子克隆有关事项的讨论参考文献第4章 分子进化遗传分析工具(MEGA5)4.1 序列数据的获取和比对4.2 进化距离的估计4.3 分子钟假说的检验4.4 系统进化树构建参考文献第5章 蛋白质结构与功能预测5.1 蛋白质信息数据库5.2 蛋白质一级结构分析5.3 蛋白质二级结构预测5.4 蛋白质家族和结构域5.5 蛋白质三级结构预测5.6 蛋白质结构可视化工具参考文献第6章 序列模体的识别和解析6.1 MEME程序包6.2 通过MEME识别DNA或蛋白质序列中模体6.3 通过MAST搜索序列中的已知模体6.4 通过GLAM2识别有空位的模体6.5 通过GLAM2SCAN搜索序列中的已知模体6.6 应用TOMTOM与数据库中的已知模体进行比对6.7 应用GOMO鉴定模体的功能6.8 应用MCAST搜索基因表达调控模块6.9 应用MEME-ChIP发现DNA序列模体6.10 应用SPAMO推测转录因子的结合位点6.11 应用DREME发现短的正则表达模体6.12 应用FIMO寻找数据库已知的模体6.13 应用CentiMo寻找主要的富集模体参考文献第7章 蛋白质谱数据分析7.1 生物质谱技术的基本原理7.2 X!Tandem软件7.3 Mascot软件7.4 Sequest软件7.5 蛋白质组学数据统计分析TPP软件参考文献第8章 基因芯片数据处理和分析8.1 芯片数据的获取和处理8.2 芯片数据聚类分析和差异表达基因筛选8.3 GenMAPP芯片数据的可视化8.4 通过GEO检索和提交芯片数据8.5 应用DAVID工具对芯片数据功能注释和分类参考文献第9章 GO基因本体和KEGG代谢途径分析9.1 Gene Ontology数据库9.2 KEGG数据库参考文献第10章 系统生物学网络结构分析10.1 Cytoscape软件简介10.2 Cytoscape软件安装10.3 Cytoscape基本操作10.4 应用BiNGO插件进行基因注释10.5 应用BioQuali插件进行基因表达分析10.6 应用Agilent Literature Search插件进行文献搜索10.7 链接BOND数据库做网络分析10.8 应用插件Cytoprophet预测潜在蛋白和结构域的相互作用参考文献第11章 Bioperl模块数据分析及其安装11.1 概述11.2 Bioperl重要模块简介和脚本实例11.3 Bioperl安装参考文献第12章 读序列(reads)定位软件Bowtie12.1 Bowtie特性12.2 Burrows-Wheeler(BW)转换程序12.3 不要求精确的比对搜索12.4 回溯过量表达12.5 阶段搜索12.6 Bowtie的输出格式参考文献第13章 UCSC基因组浏览器13.1 基因分类器(Gene sorter)工具13.2 基因组浏览器(Genome Browser)13.3 蛋白质组浏览器(Proteome Browser)13.4 表浏览器(Table Browser)参考文献第14章 SNVs和Indel识别分析方法及工具14.1 Bowtie工具14.2 samtools软件包14.3 识别单核苷酸多态性(SNP)14.4 寻找同义突变和非同义突变14.5 发现读框内插入缺失(in-frame indel)14.6 发现其他类型的突变参考文献第15章 小RNA高通量测序数据分析15.1 数据分析流程15.2 Rfam数据库15.3 miRBase数据库15.4 应用mfold预测RNA二级结构15.5 应用miRAlign搜索miRNA15.6 应用TargetScan预测miRNA的靶基因参考文献第16章 RNA测序(RNA-Seq)分析16.1 TopHat的分析流程16.2 转录组读序列比对16.3 获得基因表达谱及转录物表达谱16.4 差异表达基因鉴定及注释16.5 SNPs/SNVs及InDels鉴定与注释16.6 选择性剪切(alternative splicing)鉴定16.7 TopHat应用实例参考文献第17章 全基因组序列拼接的流程和方法17.1 实例数据的获取17.2 短读序列数据作图到参考基因组17.3 将短读序列数据从头拼接成染色体骨架17.4 大规模染色体骨架拼接17.5 草图和实验物理图谱间的比较参考文献英汉对照词汇英文索引中文索引彩图

章节摘录

第1章 序列比对工具BLAST 和ClustalX骆迎峰 丁文超 程尹 陈辰 薛庆中序列比对是基因组学研究的核心手段之一，从测序拼接到基因表达分析都需要将未知序列和数据库中的已知序列进行相似性比较。
序列比对工具很多，其中以基本局部比对搜索工具（BLAST，basic local alignment search tool）最为常用。
生物不同基因的DNA 序列或氨基酸序列通过比对，可以在相应数据库中找到相同或相似序列。
本章主要介绍美国国家生物技术信息中心（The National Center for Biotechnology Information，NCBI）数据库提供的BLAST搜索在线服务及本地运行程序，用户可以通过提交核苷酸或蛋白质序列，并选择所要比较的NCBI 序列数据库，进行序列相似性（Sequence similarity）搜索。
本章还将介绍多序列比对工具ClustalX的使用方法，以便预测基因的功能，探索物种的亲缘关系及其进化。
1.1 BLAST 搜索程序NCBI 的BLAST搜索程序（http://www.ncbi.nlm.nih.gov/blast/Blast.cgi）下设3 个部分（图1.1）：用BLAST 拼接的参考基因组（BLAST Assembled RefSeq Genome）、基础的BLAST（Basic BLAST）、特殊的BLAST（Specialized BLAST）1.1.1 用BLAST拼接的参考基因组在做BLAST搜索前，用户可根据自己的需求，选择与某个特定物种（special genome） 基因组数据库或所有拼接的基因组参考序列数据库BLAST。
如选择后者，点击list all genomic BLAST databases 后，从图1.2 可知目前正在测序或已完成测序的物种及其数量，包括：脊椎动物（Vertebrates）26 种、无脊椎动物（Invertebrates）16 种、原生动物（Protozoa）18 种、植物（Plants）47 种、真菌（Fungi）17 种。
1.1.2 基础的BLAST确定了相应的数据库，接下来是选择搜索方法。
表1.1 列出了BLAST 家族的5 个子程序及其查询序列、数据库、搜索方法。
子程序nucleotide blast（blastn）和protein blast（blastp）最为常用，使用也较简便，可以直接进行比对，搜索时只需将查询序列粘贴到搜索框中，点击BLAST 即可完成。
其中，blastn 用来发现高分值匹配的核酸序列，而blastp 能发现氨基酸残基的相似性和找到其同源蛋白。
与前两个子程序相比，后三个子程序（blastx、tblastn 和tblastx）搜索过程较为复杂，在比对前需要先经过“翻译”。
例如，运行blastx 需先将查询序列翻译成蛋白质序列，tblastn 需将核酸数据库中的序列翻译成蛋白质序列，而tblastx 需对查询序列和数据库中的核酸序列都进行翻译。
现以blastx 为例（图1.3），说明核苷酸序列翻译后可能生成6 种蛋白质序列。
假设目标序列为ATG AGT ACC GCT AAA TTA GTT AAA TCA AAA GCG ACC AAT CTG CTT TAT ACC CGC，理论上此核苷酸序列翻译时，可以分别从查询序列的正向链或反向互补链的1、2、3 相位起始。
正向链（5′→3′端）（1）第一位起始：ATG AGT ACC GCT AAA TTA GTT AAA TCA AAA GCGACC AAT CTG CTT TAT ACC CGC（2）第二位起始：TG AGT ACC GCT AAA TTA GTT AAA TCA AAA GCGACC AAT CTG CTT TAT ACC CGC（3）第三位起始：G AGT ACC GCT AAA TTA GTT AAA TCA AAA GCG ACCAAT CTG CTT TAT ACC CGC反向链（3′→5′端）（4）第一位起始：GCG GGT ATA AAG CAG ATT GGT CGC TTT TGA TTTAAC TAA TTT AGC GGT ACT CAT（5）第二位起始：CG GGT ATA AAG CAG ATT GGT CGC TTT TGA TTTAAC TAA TTT AGC GGT ACT CAT（6）第三位起始：G GGT ATA AAG CAG ATT GGT CGC TTT TGA TTT AACTAA TTT AGC GGT ACT CAT上述目标序列翻译后便会产生相应的6 个不同相位的氨基酸序列：（1）M S T A K L V K S K A T N L L Y T R（2）－V P L N－L N Q K R P I C F I P（3）E Y R－I S－I K S D Q S A L Y P（4）A G I K Q I G R F－F N－F S G T H（5）R V－S R L V A F D L T N L A V L（6）G Y K A D W S L L I－L I－R Y S结果如图1.4 所示（注：“－”为终止子）。
通过blastx 程序比对，将匹配分值最高的序列视为最有可能表达的靶标核苷酸序列。
本例最佳比对为MSTAKLVKSKATNLLYTR（图1.5），暗示该序列是从正向第一位起始翻译，由此说

明，blastx 子程序在编码区分析时，可对相位的确定起一定作用。
1.1.2 网上blastx 比对工具在BLAST主界面点击“blastx”（图1.1），进入序列提交界面（图1.6）。
该界面由输入查询序列（Enter Query Sequence）、搜索设置选项（Choose Search Set）和算法参数设置（Algorithm Parameters）（图1.7）三部分组成。
（1） 输入查询序列：用户可以在提交框中直接输入NCBI 数据库GI 号（每行1个号），或粘贴序列；也可以点击“浏览”（Browse）按钮上传保存在本地的fasta格式序列文件。
网上运行BLAST服务允许选择比对两条或多条序列（Align two or more sequences）。
此时，比对序列必须采用fasta 格式。
为方便管理，用户可以为BLAST 搜索任务命名（Job Title）。
在本例中，填入的fasta 格式序列名称是“lesson.seq.screen.Contig34”，相应地，搜索任务名称自动变为“lesson.seq.screen.Contig34”（图1.6）。
若提交的是单条fasta 格式序列，默认搜索任务就是该序列名称。
（2） 搜索设置：本例选择的数据库（Database）为默认的非冗余蛋白库（nr）。
物种（Organism）选择填入“human”；密码子表（Genetic code）采用默认标准密码子。
在“Entrez Query” 中可选择使用布尔表达式（Boolean expression）。
（3）算法参数设置：a） 通用的参数（General Parameters）设置包括：最多靶序列数（Max target）和期望阈值（Expect threshold，简称E值）、搜索词大小（Word size）、查询区域最多匹配数（Max matches in a query range）。
E 值表示在数据库搜索时与期望值随机匹配的可能性，E＝1 表示匹配是随机产生的；反之，E＝0 表示匹配不是随机产生的，由此可见，设置的E 值越小，置信度就越高（图1.7）。
b）记分参数：蛋白质序列相似性通常采用突变数据（mutation data，MD）和BLOSUM 两种矩阵估算。
突变数据基于可接受点突变（point accepted mutation，PAM）值。
PAM1 表示一个进化变异单位，即有1%的氨基酸变异。
常用的矩阵PAM250 相似性记分值相当于两个序列间保留20%匹配。
在测定远距离序列相关性时可采用BLOSUM 矩阵，BLOSUM 值表示相同序列的百分比（如常用的BLOSUM62 表示比对结果中至少有62%的氨基酸相同）。
对于相似性较高序列的比对，一般选择较低的PAM 值或较高的BLOSUM 值，反之亦然。
为补偿插入与缺失对序列相似性的影响，通常采用空位开放罚分（gap opening penalty）和空位延伸罚分（gap extension penalty）。
一个长度为n的空位，罚分数=空位开放罚分+空位延伸罚分×n。
每个记分矩阵都有默认的空位罚分值。
本例采用默认选项。
c） 过滤（Filter）和屏蔽（Mask）：通常需对低复杂性区域序列进行过滤，使其不参与显著性统计。
设置屏蔽查询种子序列只用于扫描数据库，不能用于扩展。
屏蔽fasta 格式中的小写字母。
搜索选项设置完成后，点击“BLAST”即可运行搜索。
blastx运行后，服务器会自动以网页形式返回结果，其中包括比对上的序列、相似性程度及显著性水平等信息，如图1.8所示。
……

版权说明