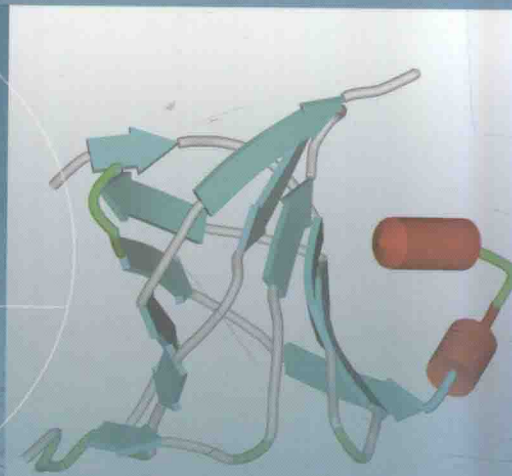
□全国高等学校“十三五”农林规划教材

生物信息学实验教程

主编　吕巍　李滨



高等教育出版社

生物信息学实验教程

□全国高等学校“十三五”农林规划教材

生物信息学实验教程

SHENGWU XINXIXUE SHIYAN JIAOCHENG

主编　吕巍　李滨

编者　（按姓氏拼音排序）

曹雪松　杜娟　李滨　吕巍　宋晓军　张亚南

高等教育出版社·北京

内容简介

本书共设计10个大实验，包括Linux系统入门与操作、美国国家生物技术信息中心（NCBI）网站的相关应用、双序列比对网络应用、双序列比对本地应用、多序列比对分析、进化分析、转录组分析、基因组组装、基因预测、蛋白质分析。每个实验中都介绍了当前较为流行和应用最广的相关软件。其中基础实验部分适合本科生教学，而进阶实验部分适合研究生教学。本教材采用“纸质教材＋数字课程”的出版形式，数字课程是纸质教材的有力补充，主要包括全书插图、教学课件和教学相关视频等，便于教师教学和学生学习使用。

本书可作为理、工、农、医等院校生物学相关专业的教材，也可供从事生物信息学相关研究的师生、科研人员参考使用。

图书在版编目（CIP）数据

生物信息学实验教程 / 吕巍，李滨主编 . -- 北京 : 高等教育出版社，2016.2

ISBN 978-7-04-044429-2

Ⅰ. ①生… Ⅱ. ①吕… ②李… Ⅲ. ①生物信息论－实验－高等学校－教材 Ⅳ. ① Q811.4-44

中国版本图书馆 CIP 数据核字（2016）第 012555 号

策划编辑 孟 丽　　责任编辑 孟 丽　　特约编辑 陈龙飞　　封面设计 姜 磊
责任印制 赵义民

出版发行 高等教育出版社
社　　址 北京市西城区德外大街4号
邮政编码 100120
印　　刷 北京七色印务有限公司

开　　本 787mm×1092mm 1/16
印　　张 6.25
字　　数 150千字
购书热线 010-58581118

咨询电话 400-810-0598
网　　址 http://www.hep.edu.cn
http://www.hep.com.cn
网上订购 http://www.hepmall.com.cn
http://www.hepmall.com
http://www.hepmall.cn
版　　次 2016年2月第1版
印　　次 2016年2月第1次印刷
定　　价 15.00元

物 料 号 44429-00

数字课程（基础版）

生物信息学实验教程

主编　吕巍　李滨

登录方法：

1. 访问http://abook.hep.com.cn/44429，点击页面右侧的“注册”。已注册的用户直接输入用户名和密码，点击“进入课程”。
2. 点击页面右上方“充值”，正确输入教材封底的明码和密码，进行课程充值。
3. 已充值的数字课程会显示在“我的课程”列表中，选择本课程并点击“进入课程”即可进行学习。

自充值之日起一年内为本数字课程的有效期
使用本数字课程如有任何问题
请发邮件至：lifescience@pub.hep.cn

http://abook.hep.com.cn/44429

前　言

生物信息学是一门由生命科学、数学、计算机科学和信息科学相互渗透形成的新型交叉学科。它通过对生物学实验数据的获取、加工、存储、检索和分析，进而揭示这些数据所蕴含的生物学意义，是生命科学在后基因组时代不可或缺的研究工具，成为当今世界生命科学研究的核心领域和前沿领域之一。

近年来，生物信息学在国内不断发展，其在本科教育领域的重要性逐渐凸显，介绍和讲授生物信息学的书籍也不断涌现。但是，由于编者的学科背景不同，各种专著的侧重点和针对性也不尽相同。基于以上情况，作者在从事生物信息学教学的基础上，充分吸取现有国内外相关教材与著作的长处，结合自己在生物信息学领域的研究，编写了这本以介绍生物信息学领域常用软件，适合理、工、农、医等院校生物学相关专业的本科生及研究生教育的实验教材——《生物信息学实验教程》。

本教程主要包括 10 个实验，其中前 6 个实验属于基础性实验。实验一介绍 Linux 操作系统，主要包括一些与生物信息学软件使用相关的基本命令；实验二介绍美国国家生物技术信息中心网站的相关应用方法；实验三介绍如何通过网络来进行双序列比对；实验四介绍如何在本地服务器通过 BLAST 程序进行序列比对和搜索；实验五介绍多序列比对的方法及使用；实验六介绍进化分析方面的常用软件，包括 MEGA 5.0 和 PHYLIP 软件包的使用方法。而实验七至实验十属于进阶性实验，主要介绍如何分析处理现阶段生物学实验所带来的海量生物数据：实验七和实验八，分别介绍了转录组数据分析和基因组数据分析的常用软件；实验九介绍基因预测软件，包括原核生物基因预测软件 Glimmer 及真核生物基因预测软件 Genscan；实验十简要介绍蛋白质分析方面的实用工具。

生物信息学作为一门新兴学科，正不断向更深更广的方向快速发展。新概念、新理论、新软件、新方法不断涌现，虽然本教程力求反映这些最新的进展，但仍有不少缺失。另外，由于编者水平所限，本教程尚有很多不足及遗漏之处，也敬请同行及专家、读者批评指正。

编　者

2015 年 10 月

目　录

实验一　Linux 系统入门与操作

常见的操作系统包括 Windows、Mac OS X 和 Unix。Linux 是一类 Unix 操作系统，可安装在各种各样的电脑硬件设备，从手机、平板电脑、路由器到超级计算机。Linux 是一个领先的操作系统，世界上运算最快的 10 台超级计算机运行的都是 Linux 操作系统。以其自由度高、安全、强大的内置程序支持，以及灵活的后台管理等特点，Linux 系统受到广大科研工作者的欢迎。目前常用的生物信息学软件的大多数版本只针对于 Linux 系统。Linux 系统是生物信息学工作者常用的操作平台。

【实验原理】

Linux 服务器可以理解为一个超级计算机，它拥有文件管理、数据库管理和应用程序管理功能，以及相对较大的 CPU、存储和内存资源，同时兼有普通计算机的功能。通过 Windows 系统可以登录到 Linux 服务器，对其进行远程控制，从而完成各种复杂的计算。

【实验目的】

掌握 Linux 系统的基本原理，能够熟练使用各种 Linux 命令处理文件。

【设备与软件】

Linux 服务器，Windows 系统个人电脑，Linux 操作系统，SSH Secure Shell Client 服务器登录软件。

【实验方法】

1　登录 Linux 服务器

在 Windows 系统的个人电脑上安装 SSH Secure Shell Client 应用程序。双击打开 SSH Secure Shell Client 程序（图 1－1），点击 Quick Connect 按钮，打开快速连接向导（图 1－2），输入

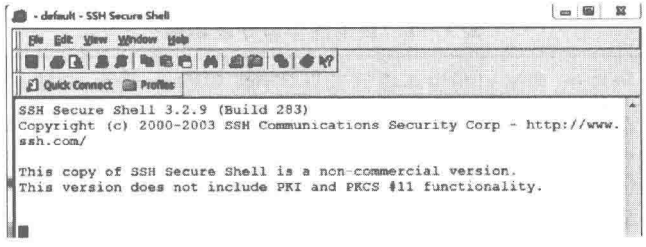


图 1－1

IP 地址和用户名,点 Connect 按钮,弹出输入密码对话框(图 1 -3),输入用户密码,点 OK 按钮,远程登录 Linux 系统,从而远程控制服务器(图 1 -4)。

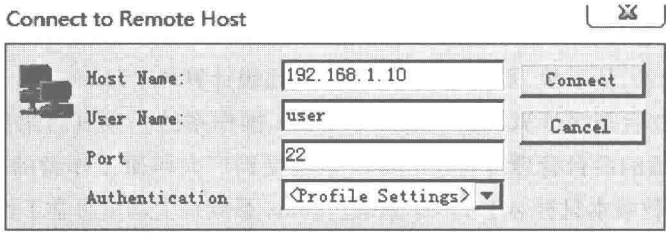


图 1 -2

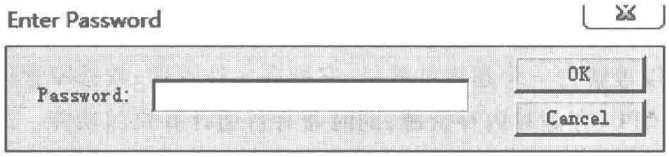


图 1 -3

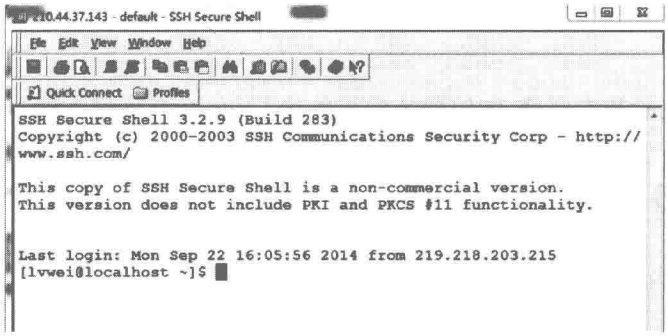


图 1 -4

2 Linux 操作基本命令

Linux 的操作界面对大小写和空格敏感,所以在书写时应注意。每一行命令输入结束后,按 Enter 键运行,得到结果。

2.1 目录操作(表 1 -1)

表 1 -1 目录操作命令及功能

命令	功能
pwd	查看当前所在目录,显示全路径
cd	切换到指定目录 语法:cd [目录名] 目录名可以是全路径,也可以是相对路径

续表

命令	功能
ls	列出指定目录下的文件和目录 语法:ls [参数] [目录或文件] 参数: -a 显示指定目录下的子目录和文件,包括隐藏文件 -c 按照文件的修改时间排序 -L 若指定的名称是一个符号链接文件,则显示链接指向的文件 -R 递归显示指定目录的子目录和子目录中的文件 -l 以长格式显示文件和目录的详细信息,每行列出的信息依次是:文件类型和权限,链接数,文件属主,文件属组,文件大小,建立或最近修改的时间,文件名。用此参数命令显示的信息中,开头是由 10 个字符构成的字符串,其中第一个字符表示文件类型,详细如下:-普通文件;d 目录;l 符号链接;b 块设备文件;c 字符设备文件。后面的 9 位表示文件的访问权限,分为 3 组,每组 3 位。对于目录,表示进入权限。第一组表示文件的属主权限,第二组表示同组用户的权限,第三组表示其他用户的权限。每一组的 3 个字符表示对文件的读(r),写(w)和执行权限(x)
mkdir	创建新目录,要求创建目录的用户在该父目录路径下有可写权限,并且在该父目录下没有同名的文件或目录 语法:mkdir[参数] [目录名或是父目录路径/目录名] -p 递归创建多层目录,即此参数后可以是一个路径,若路径中的某些目录不存在,系统将自动建立那些尚不存在的目录,直至最后一级
rmdir	删除空目录 语法:rmdir [参数] [目录名或是父目录路径/目录名] -p 递归删除目录,当子目录删除后其父目录为空时,也一同被删。如果整个路径被删除或者由于某种原因保留部分路径,则系统在标准输出上显示相应的信息
rm	删除目录,该命令可以删除一个目录下的一个或是多个文件或是目录,它也可以将某个目录下及其下面的所有文件和子目录均删除。对于链接文件,只是断开链接,源文件保持不变。删除目录,必须增加-r 参数 语法:rm [参数] 文件或是目录 参数: -f 强制删除,不给提示 -r 删除目录 -i 进行交互式删除,由于 rm 删除文件和目录后是不能够恢复的,所以可以使用 i 参数在删除前进行系统确认

使用以上命令进行操作,参见图 1-5。

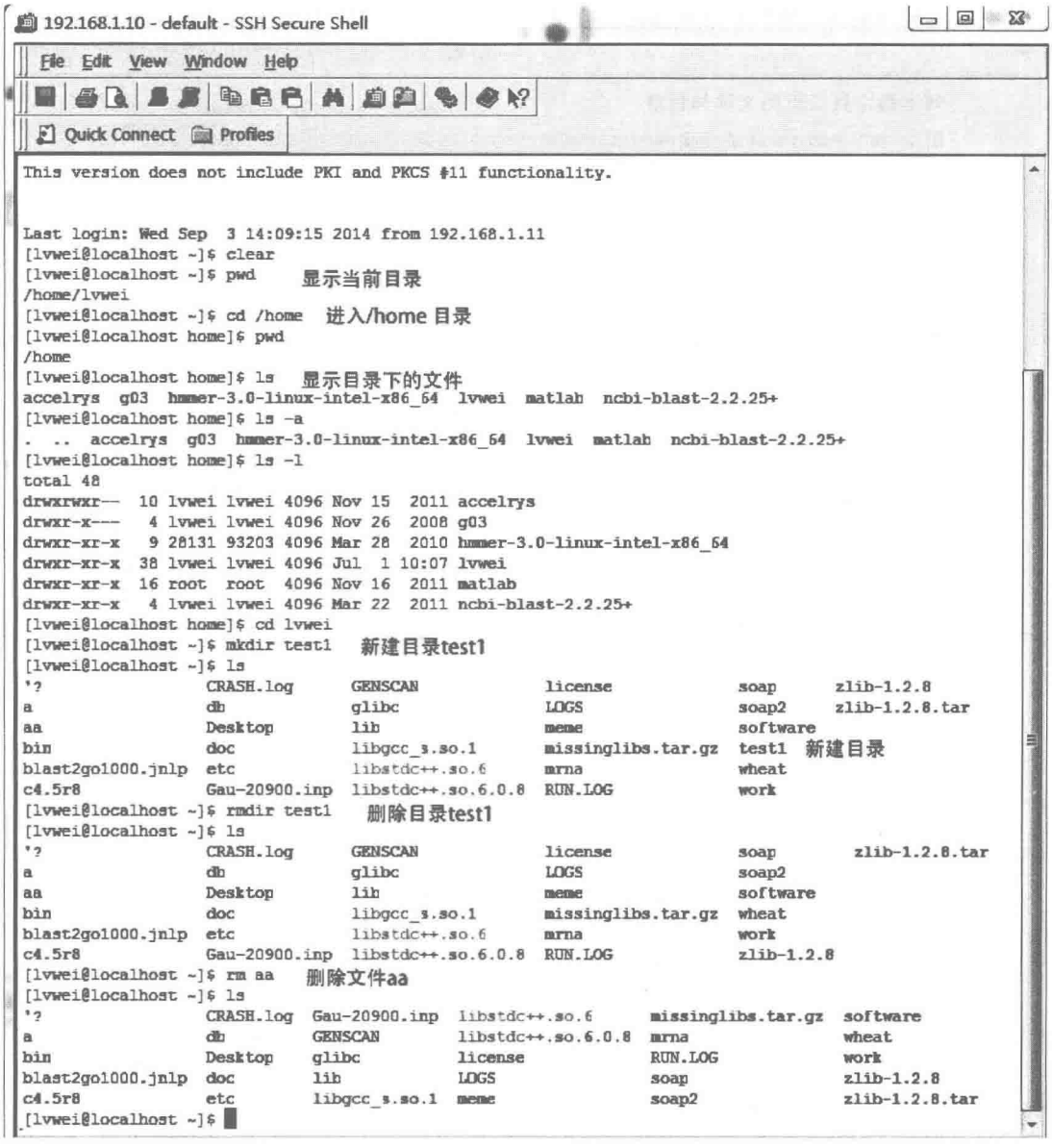


图 1-5 目录操作命令运行结果

2.2 文件操作

表 1-2 文件操作命令及功能

命令	功能
touch	创建一个新的空文件,无内容 语法:touch [参数] [文件名]

续表

命令	功能
cp	拷贝文件或目录，该命令可以将指定的目录或是文件拷贝到另一个目录或是文件，可以使用通配符拷贝具有同一特征的所有文件。 语法：cp［参数］源文件或是目录目标文件或是目录 参数： -a 通常在拷贝目录的时候使用，将保留链接、文件属性，并递归拷贝目录 -d 拷贝时保留链接 -f 删除已经存在的目标文件并不提示 -i 交互式拷贝，如果目标文件存在，则提示用户，以免覆盖重要文件或是目录 -p 除拷贝源文件的内容外，还将其修改时间和访问权限也复制过来 -r 若源文件是目录，将递归拷贝该目录下的所有子目录和文件
mv	移动或重命名文件或目录 语法：mv［参数］源文件或是目录，目标文件或是目录可以使用该命令来为文件或是目录改名或将文件由一个目录移入另外一个目录中 参数： -i 交互式操作，如果目标文件存在，则提示用户，以免覆盖重要文件或是目录 -f 强制执行，删除已经存在的目标文件并不提示
more	显示文件，在终端按屏显示文件，一次显示一屏内容，显示满之后，停下来，并在底部打印出“—More—”，系统还将显示出已显示文本占全部文本的百分比，若要继续显示，空格键显示下一页，回车键显示下一行，q 为退出 语法：more［参数］文件 参数： -p 显示下一屏之前先清屏 -s 文件中连续的空白行压缩成一个空白行显示
wc	统计文件的字节数，字数，行数 语法：wc［参数］文件 参数： -c 统计字节数 -l 统计行数 -w 统计字数

使用以上命令进行操作，参见图 1－6。

图 1-6 文件操作命令运行结果

2.3 权限控制

表 1-3 权限控制命令及功能

命令	功能
chmod	改变文件或是目录的访问权限,权限类型可以是以下字母的组合,多个权限方式或是多个文件,用逗号隔开。“用户”可以是下述字母中任意一个或是其组合: 语法 a: chmod [用户] [操作] [权限类型] 文件名 u 表示“用户(user)”,即文件或是目录的所有者 g 表示“同组(group)用户”,即与文件属主有相同组 ID 的所有用户 o 表示“其他(other)用户” a 表示“所有(all)用户”,系统默认值

续表

命令	功能
chmod	操作选项可以是： + 添加某个权限 - 取消某个权限 = 赋予给定权限并取消其他所有权限 语法 b： chmod [ddd] 文件名 数字属性的格式是 3 个从 0 到 7 的八进制数字，其代表的用户顺序为 u,g,o。0 表示没有权限，1 表示可执行权限，2 表示可写权限，4 表示可读权限，数字可以相加，表示权限的累加

使用以上命令进行操作，参见图 1 -7。

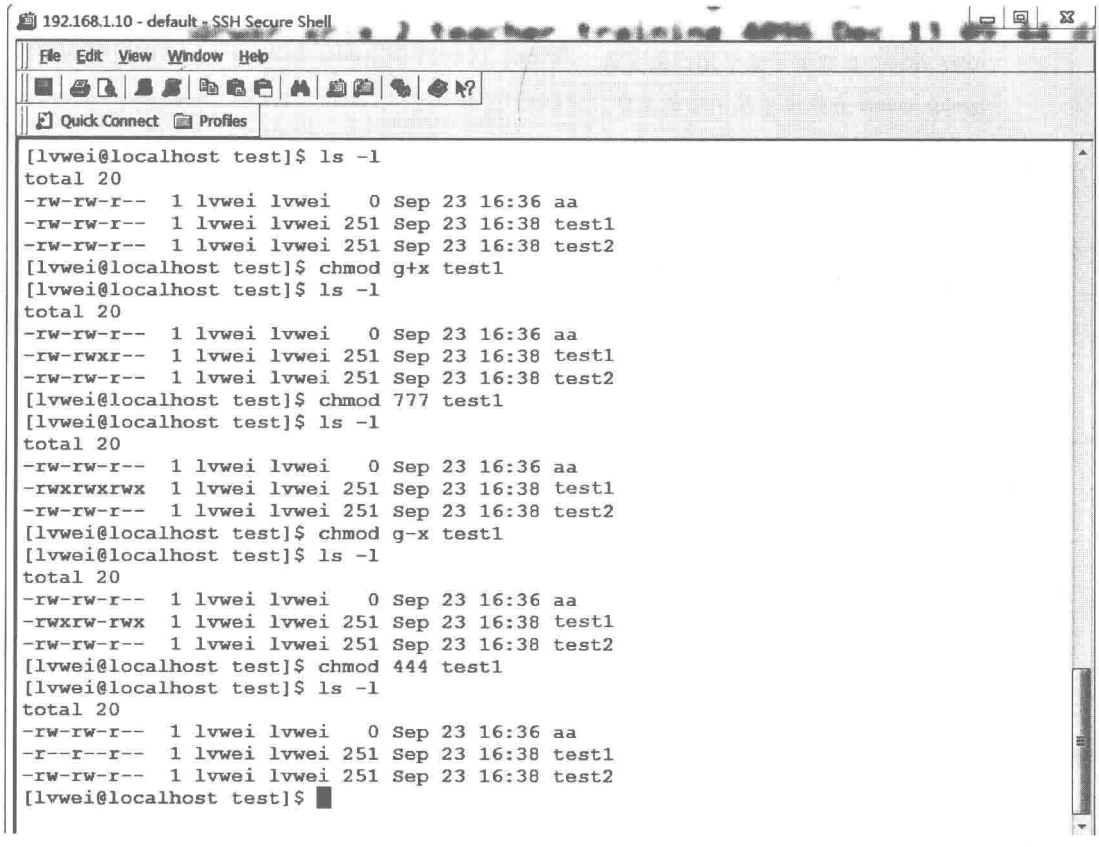


图 1 -7　权限控制命令运行结果

2.4 搜索和查找

表 1 - 4 搜索和查找命令及功能

命令	功能
grep	搜索文件并过滤出有某个特征的行 语法： grep [参数] 匹配模式文件 -n 把所有匹配的行在行前加上行号列出 -v 把不包含匹配模式的行列出
find	在目录结构中搜索文件，并执行指定的操作 语法： find [路径] [参数] [操作] -empty 查找系统中空白的文件或是空白的文件目录 -name 要查找的文件名，可以加通配符 -user uname 查找属于某个用户的文件 -group gname 查找属于某个组的文件 -links n 查找有 n 个链接的文件 -exec 执行操作
locate	查找文件，该命令提供的寻找条件可以是一个用逻辑运算符 not、and、or 组成的负荷条件。locate 会在保存文件与目录名称的数据库中查找合乎范本样式条件的文件或是目录 语法： locate [参数] [操作] -d 设置 locate 指令所使用的数据库

使用以上命令进行操作，参见图 1 - 8。

【实验内容】

1. 利用 SSH Secure Shell Client 软件远程登录 Linux 服务器。

2. 练习“pwd”“cd”“ls”命令的使用，利用“mkdir”“rmdir”“rm”命令建立目录 test，对其进行删除等操作。

3. 进入 test 目录下，用“touch”创建一个新的空文件 aa，用“mv”命令从电脑中剪切一个文件到 test 文件夹下，并命名为 test1，用“cp”命令将 test1 复制为 test2，使用“more”命令查看 test1 文件内容，使用“wc”命令统计 test1 的字节数、字数及行数等内容。

4. 使用“chmod”命令，增加或减少 test1 文件的可读、可写、可执行的权限。

5. 使用“grep”命令，查找 test1 文件中包含某些字符的特定行，使用“find”和“locate”查找 test1 文件所在位置。

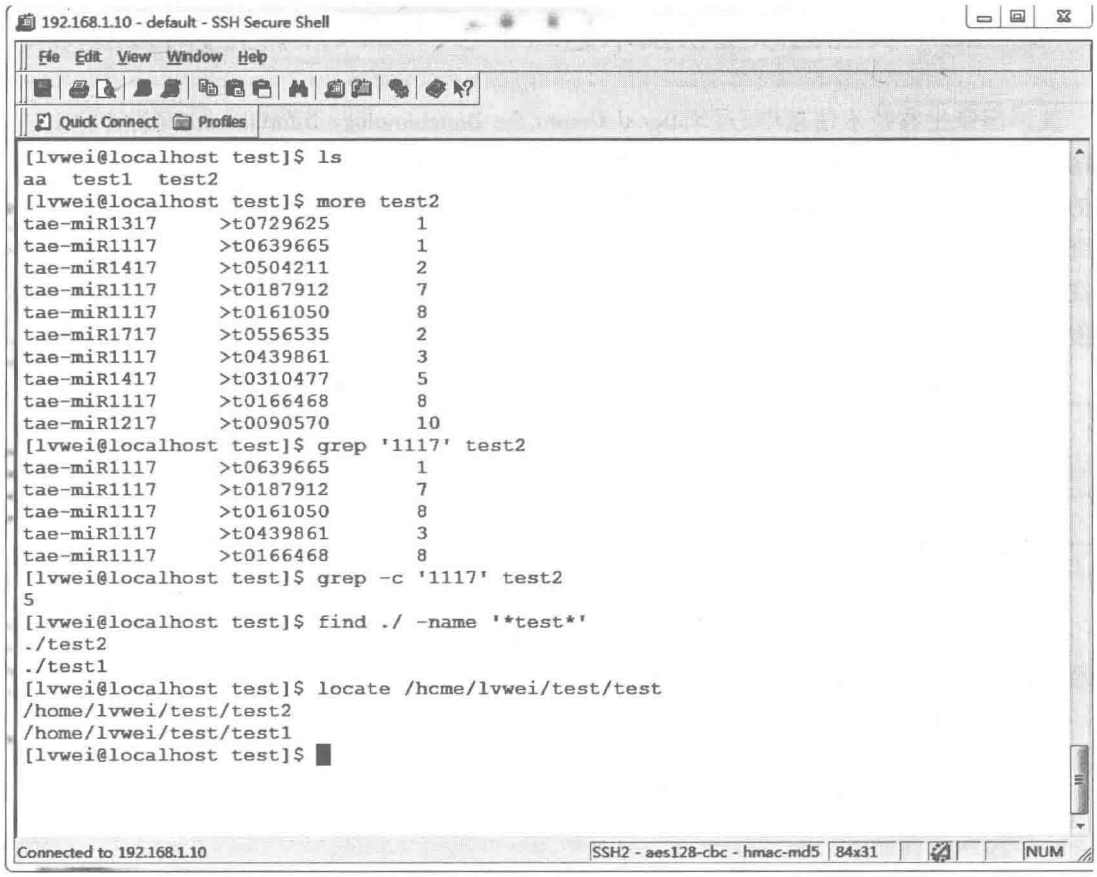


图 1－8　搜索和查找命令运行结果

实验二 美国国家生物技术信息中心(NCBI)网站的相关应用

美国国家生物技术信息中心(National Center for Biotechnology Information,简称 NCBI)是美国国家医学图书馆(NLM)的一部分(该图书馆属于美国国家卫生研究院)。NCBI 位于马里兰州的贝塞斯达,建立于 1988 年。NCBI 保管 GenBank 的基因测序数据和 MEDLINE 的生物医学研究论文索引。所有的这些数据库都可以通过 Entrez 搜索引擎在线访问。许多受尊敬的研究者在 NCBI 工作,如比较基因组学领域的一位多产的科学家 Eugene Koonin 和 BLAST 序列数据库搜索算法的作者 Stephen Altschul。

GenBank 是美国国家卫生研究院维护的基因序列数据库,汇集并注释了所有公开的核酸序列。它与日本 DNA 数据库以及欧洲分子生物学实验室核苷酸数据库一起,都是国际核苷酸序列数据库集团的成员。

【实验原理】

NCBI 网站包括了大量的数据库,像 PubMed 和 GenBank。

【实验目的】

了解 NCBI 网站的构成与包含的内容,熟练掌握其使用方法。

【设备与软件】

Windows 系统个人电脑,网络。

【实验方法】

1 打开 NCBI 网站

在浏览器中输入网址 http://www.ncbi.nlm.nih.gov/,进入 NCBI 主页,如图 2-1。左侧是 NCBI 工具栏,列出了网站的所有工具,右侧是网站中使用率最高的工具列表,最上方是数据搜索栏,可以输入关键词进行搜索。例如,输入关键词“NAC”(一类转录因子的缩写),选择所有数据库“All Database”,点击 Search 按钮,将进入搜索结果页面,如图 2-2。可以看到在所有数据库中搜索出 270 572 条与“NAC”相关的信息,网站根据不同的数据库对其进行了详细的分类,主要包括六大类:文献相关“Literature”、健康相关“Health”、基因组“Genomes”、基因“Genes”、蛋白质“Proteins”和化学“Chemicals”。其中我们常用的有 Literature 中的 PubMed、Genes 中的 Gene 以及 Proteins 中的 Protein 这几个搜索结果。