植物细胞与染色体工程国家重点实验室

数据分析服务器用户指南

2013-04-02

Version-1

目　录

第一章 服务器基本信息

植物细胞与染色体工程国家重点实验室数据分析服务器基于生物信息分析中广泛使用的 Linux 操作系统，将常用的数据分析公共软件和数据库集成到系统中，搭建了一套适于进行生物信息和数量遗传学研究的平台，可以适应用户多种个性化分析的需求，该平台由专人进行维护、咨询和培训服务，协助本实验室人员深度挖掘海量数据。

Figure 1. 服务器照片（存放于遗传发育所 2 号楼 S2-117 房间）

服务器配置信息：

品　　牌：IBM

操作系统：Red Hat Enterprise Linux 6.1

处 理 器：8 颗 Intel Xeon 10C E7-8850

内　　存：1024GB DDR3 RDIMM

存　　储：48TB

第二章 服务器的使用

植物细胞与染色体工程国家重点实验室职工和学生均可向服务器管理员申请用户名和初始密码，获得上述信息后即可访问和使用服务器上的计算资源。为了展开服务器上的科学计算工作，需要按以下步骤进行。

1. 账号申请:

填好下表发送至：zhkzhou@genetics.ac.cn，即可获得账号和初始密码。

申请人	课题组	磁盘空间	邮箱

2. 本地准备工作

为了登录服务器,在本地的 PC 机上需要安装链接服务器 SHELL 的客户端（client)程序。对于 windows 用户,建议使用的软件 SSH Secure Shell Client,最新版本的软件和说明书可从以下网站下载:

http://www.onlinedown.net/softdown/20089_2.htm

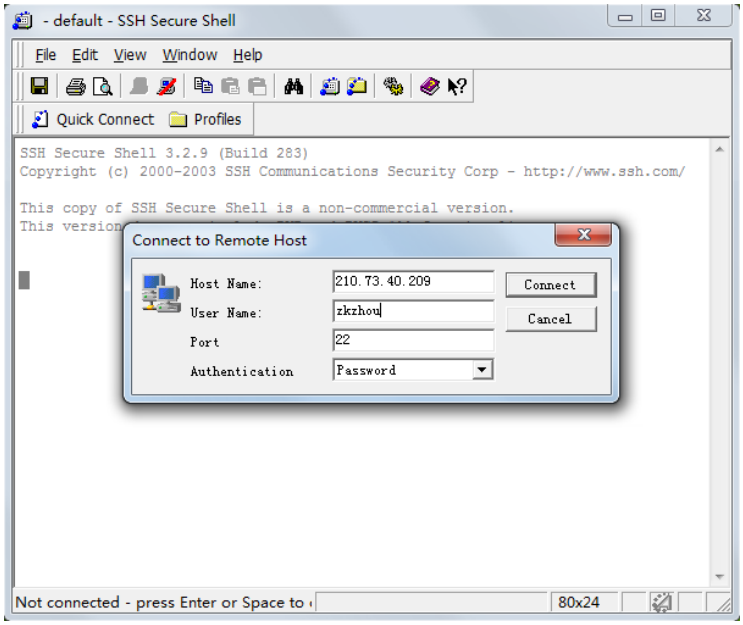

Figure 2. SSH Secure Shell Client 登录设置

本服务器的安全策略是仅允许遗传发育所内 IP 地址登陆，该策略可以有效防止网络攻击，最大限度保证数据安全。

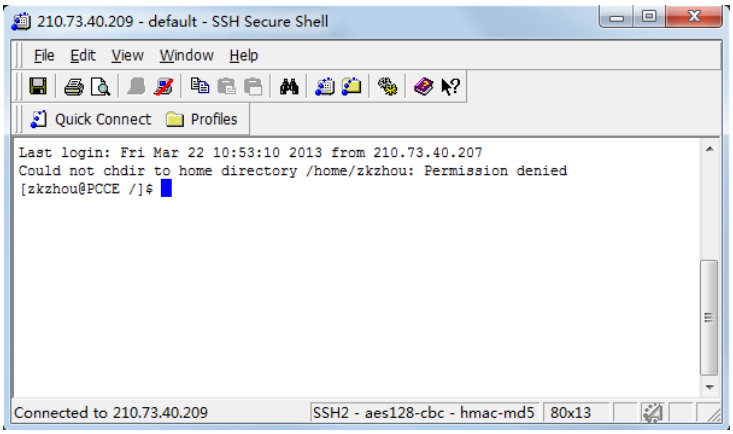

Figure 3. SSH Secure Shell Client 登录后界面

纯字符的 SHELL 命令行通常已经能够满足大多数用户。登陆之后，首先要用 passwd 命令更改自己的密码，从而保证账号和资源安全。当密码遗忘或遇到问题需要管理员解决时，可请管理员修改自己的密码。

修改密码的方法如下:

[zkzhou@PCCE ~]$passwd

Changing password for user zkzhou.

(current) UNIX password:

Linux 系统中的密码在键入时不会显示。修改完密码后，可以管理一下自己的家目录， Linux 系统中，例如当登录用户为 zkzhou 时，家目录路径是/home/zkzhou。在家目录中，一般会有一个 .bashrc 文件存放 bash 的环境变量。bash 是 Linux SHELL 的一种，其语法比较接近 C 语言， Linux 的 SHELL 还有 sh、csh、tcsh 等多种。目前服务器上的用户默认 SHELL 是 bash。关于.bashrc 文件的详细介绍请参考下一部分内容。

至此，可以在家目录中建立目录，存放要运行的软件和数据。在平时的使用中，维护自己的家目录是非常重要的，对于软件的编译和运行，一个有条有理的

目录能够帮助你顺利完成自己的计算任务。所以我们必须熟悉 Linux 系统和操作命令。

3. Linux 操作系统

Linux 是开源的操作系统，由内核和外部模块构成核心功能， Linux 上的软件运行以后台进程的方式进行。软件源代码由编译器编译成可执行文件（bin）存放在文件系统中，供用户调用执行。对于维持系统基本功能的服务（service），比如 httpd，通常以守护进程（daemon）的方式开机后自动在后台执行，等待用户的调用。用户同系统的交互由 SHELL 来完成，类似 Windows DOS 系统的命令行。用户登录服务器后，通过在 SHELL 中输入命令来进行操作。 因此，用户要使用 Linux 系统，必须掌握基本的 Linux 命令。

Linux 的用户分为两种：超级用户（root）和普通用户。 root 用户拥有所有的权限，普通用户的权限在账号被创建的时候可以进行相应的设置。

Linux 系统中的所有文件都被赋予一定的属性，这些属性包括拥有这个文件的用户（user）、组（group）、读写运行的访问权限、最近修改的时间等。其中访问权限的功能非常强大。众所周知 Linux 系统的安全性很高，可以说 Linux 系统的安全就依赖于这样一套严备的访问权限体系。

访问权限在 Linux 系统中由一个 10 位的字符串表示，第一位表示文件的类别： -表示普通文件（file）； d 表示文件目录（directory）； l 表示链接（symbolic link）。后面的 9 位分为 3 组 rwx，第一组为文件所有者的访问权限，第二组为文件所有者所在群组的访问权限，第三组为其他用户的访问权限。每组的 3 个字母： r 代表可读权限（readable）； w 代表可写权限（writable）；x 代表可执行权限（executable）。

第三章 Linux 常用命令

Linux 操作系统自带的系统命令有很多，然而常用的只有不超过 30 个。这些命令大致分为文件操作和进程管理两大类，具体用法参考相关书籍或网页。

1. 常用命令：

1) man [command]查看 command 命令的说明文档（ manual page）

2) ll 或 ls -[options] [directory]列出目录里的文件，常用的 ls 的选项有 -l -a -t 等

3) cd [directory]进入文件夹（不加目录名则默认进入你的家目录）

4) pwd 显示当前所在目录

5) rm [files]删除文件（删除目录需要加 -r 选项，强制删除用 -f）

6) cp [source] [target]复制文件

7) mv [source] [target]移动文件（也可以理解为改名）

8) touch [filename]新建名为 filename 的文本文档

9) mkdir [-p] [directory]新建文件夹（ -p 为建立整个路径）

10) ln [-s] [path] [link]建立链接（ -s 为建立软链接）

11) cat [textfile]显示文本文档的内容

12) grep 'content' [file]在 file 中查找有 content 的行

13) sed, awk, cut...字符串处理程序

14) chown [user.group] [file]修改 file 的所有人和群组

15) chmod 755 [-R] [file]改变 file 的访问权限， 755 三个数字为三组访问权限的加权值。 r=4，w=2，x=1。755 代表的意思是 -rwxr-xr-x。又比如 644 的意思是-r-xr-r-等。

16) tar zxvf [*.tar.gz]解压缩文件包， z/j=gunzip/bz2 格式， c/x=压缩 /解压缩

17) find -name [filename]在当前文件夹搜索名为 filename 的文件 ,有比较多的高级选项

18) locate [file]快速查找定位文件，只能搜文件名

19) file [file]查看 file 的文件类型

20) vi 功能强大的文本编辑工具 : i 进入编辑模式 Esc 退出编辑模式 r 修改单个字符 :w 保存 :q[!]（放弃修改）退出 :h 帮助 /string 搜索 string :2 co 4 将第 2 行拷贝到第 4 行

21) [command] > outfile 将 command 命令的执行结果写入到 outfile 文本文件中

22) &在后台执行程序

23) [command1] | [command2] 把 command1 执行的结果作为输入送到 command2 中执行

2. 环境变量

Linux 系统的环境变量的作用在于他们定义了应用程序需要多次调用的值，比如：系统文件的路径、软件安装的路径等。定义系统变量可以方便程序获得所需的值，而不必每次都重新定义。PATH 是最重要的一个环境变量，它的作用是存放可执行命令路径，当你在 shell 提示符后键入一个命令后， Linux 会到 PATH 指定的路径去查找相对应的可执行文件，找到后执行它。所以如果你要调用的命令路径不在 PATH 中，就得每次都在命令前加上绝对路径才能正常调用。

在 bash 中，查看系统变量的值，可以用 echo $NAME。定义系统变量的方法是： export NAME=value，这个变量在 bash 被关闭之前有效。为了让环境

变量永久被记住，则需要将它写入.bashrc 文件。系统在打开一个 bash 的同时，会自动加载 .bashrc 中定义的变量。改动.bashrc 内容后，需执行 source .bashrc 或重新登录 bash 才能生效。

例如将编译好的或下载的二进制软件放到/home/zkzhou/bin 目录下，则需要修改.bashrc 文件以便于让软件运行：

#vi .bashrc

在里面加入:

export PATH="$PATH:/home/zkzhou/bin"

3.脚本

概括地说，脚本是 shell 中的一个命令集合，可以将多个命令作为一个单一文件执行，类似于 DOS 里面的批处理文件。使用脚本，可以使繁琐的工作变得简单，也方便管理自己的程序。在日常的工作中我们经常会遇到这样的情况，完成一项工作需要执行一连串的命令，并且在整个过程中需要根据结果的不同做相应的判断，脚本的出现使得我们不必自己一次次重复复杂的操作，而是将规则记录下来让计算机去为我们操作。

掌握脚本的使用，最重要的是理解变量、赋值和条件判断。由于脚本是基于命令行输入输出的编程语言，所有的操作基本以字符串为基础，所以变量的类型只能是字符串。以 bash 为例，

```
#!/bin/sh
for i in {1..10}                    ##   i 为变量，从 1 到 10 循环
do
bwa index ‘myfile’$i’.fasta’      ## 分别对 myfile1.fasta~~myfile10.fasta 建索引
done                                ## 将以上代码保存为 bwaindex.sh
```

执行脚本前，需要将文件 chmod 为可执行文件：

chmod +x bwaindex.sh

对于字符串处理， Linux 系统提供的 awk 和 sed 命令具有更加完备和强悍的功能。有兴趣的读者可以自行查阅它们的说明文档。这两个强大的工具使得复杂的字符串操作成为可能，所以在脚本中经常被用到。

4. screen 命令的用法

很多用户会遇到这样的情况，一个任务需要运行很长时间，例如 Genome mapping，这个时候，我们被迫一直开启一个 SSH 客户端的链接，以便观察任务执行的状态、进行下一步操作等等。如果这个时候你的 PC 机需要关机或者重启，那么对于我们来说可能是个很痛苦的选择，因为一旦 SSH 连接断开后，任务即中断！因此向大家推荐 screen 命令，这个命令是一个虚拟 shell 环境工具。你可以在你登录到服务器之后，用这个工具创建一个虚拟的 shell 环境，在这个环境中工作，可以完全不用考虑断开连接对你的影响。以下是 screen 的基本使用方法。

语 法：screen [-AmRvx -ls -wipe][-d <作业名称>][-h <行数>][-r <作业名称>][-s <shell>][-S <作业名称>]

Step 1. > screen －S myjob ##指定 screen 作业的名称;

Step 2. > input you job then ctrl+a d ##离开 screen;

Step 3. > screen －ls ##显示目前所有的 screen 作业;

Setp 4. > screen －r myjob ##恢复离线的 screen 作业

第四章 软件编译和安装

Linux 系统下的软件安装与 Windows 系统中不同，Windows 系统中我们习惯于运行 setup.exe 文件，其实它所做的工作是将自身压缩的可执行文件解压并拷贝到系统中，并在注册表中留下相关的记录，使得软件能够正常运行。在 Linux 系统中没有注册表，另外由于开源软件的流通，软件经常以代码包的形式被下载使用。这样做的一个好处是软件包所占空间非常小，但在安装前需要编译。

1.编译器

软件的代码通常由 C 语言、 Fortran 语言等写成，Linux 也提供了相应的编译器，例如 GNU 的 gcc，gfortran，Intel 的 f90 等。本服务器上使用的是 GNU 的编译器（gcc 等）。

在 Linux 系统中编译源文件非常简单，只需要执行命令

[compiler] -flag [sourcefile] -o [executive]。

compiler 为编译器名， flag 为编译过程中的参数设置， sourcefile 为源代码文件， executive 为编译成的可执行文件，默认的文件是 a.out。例如：

gcc test.f -o test.exe

该命令用 gcc 编译器编译 test.f 文件，将生成的可执行文件命名为 test.exe。编译器的选项设置是否正确，影响到编译是否能顺利完成。这些选项包括系统类型的选择、内存的使用、需要用到的链接库等。

2.库

熟悉编程的读者应该对库的概念有所了解， Linux 系统中编译源代码指定链接库文件的方法是（以 gcc 编译器为例）：

gcc -L/usr/lib –libdemo.so test.f

这样指定了在编译 test.f 的过程中链接 demo 库文件 libdemo.so，该文件位于 /usr/lib 中。

3. Makefile

由于软件包往往是由大量的源码文件组成，它们之间又有着复杂的依赖关系，如果依次单个进行编译的话，会非常的耗时耗力，所以 Linux 提供了 make 机制来处理复杂的软件编译过程。在软件包的各目录中，都有一个 Makefile 文件，记载了该文件夹中的源码按什么样的规则来编译。在软件的根目录中，同样有一个 Makefile 记录软件的作者提供的可能的编译方式。 Makefile 中记录了编译器名、编译器使用的选项以及源代码被编译的先后顺序等。因此，我们在软件的编译过程中只需要修改 Makefile 中的相应项就可以了。

4.编译软件的步骤

软件编译的第一步，始终是阅读 readme 文件，因为软件的作者会在里面详细地介绍软件编译安装的过程以及可能遇到的问题。

为了方便用户的编译，许多软件的作者提供了 configuration 这一步，有点类似于 windows 软件“下一步”的风格。运行 configure 脚本进行用户交互，根据得到的选择生成合适的 Makefile，使得用户不用亲自研究 Makefile 的语法。configure 脚本在执行完后，会生成一个配置文件， Makefile 会调用这个文件，使得设置生效。

正确修改 Makefile 之后，在软件的根目录下执行 make 命令，开始编译。如果在编译中遇到了错误，会在输出文本中体现出来。常见的错误有：使用了错误的编译模式、链接的库文件没有找到、系统兼容性等。

第五章 文件传输

本服务器开通了 sftp 服务，可以供用户进行 sftp 协议的文件传输。使用账号登录 sftp，可以访问自己家目录下的文件，并可以执行读写操作。sftp client 软件有许多种，对于 windows 用户，可以使用专门的 sftp 软件，比如 FileZilla 等。该软件可以在 http://filezilla-project.org 网站上自由下载。 sftp 登录的方法是： 主机：210.73.40.209；用户名：zkzhou；密码：xxxxx；端口：22.

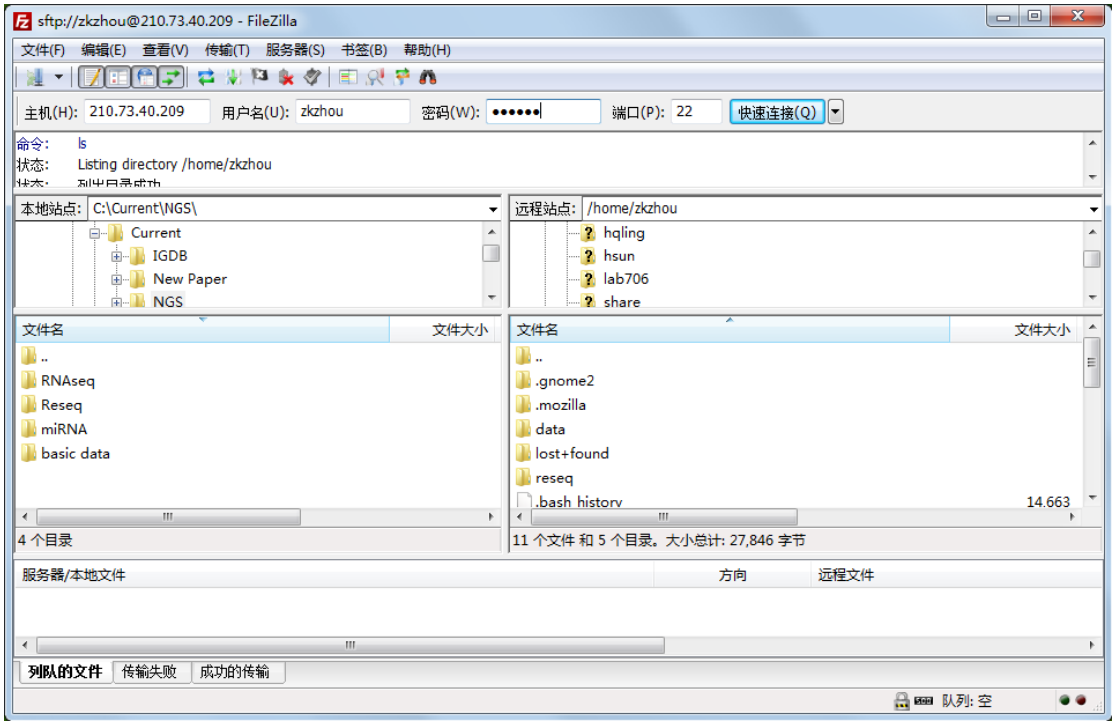

Figure 4. FileZilla 登陆界面

SSH Secure Shell Client 软件也具备文件传输功能，在 SSH 登陆状态下，只需点击 New file transfer window 按钮即可出现该功能界面。

注意事项

Linux 系统具有高度可靠性，但并不能保证你的数据万无一失，因此所有的数据资料要在本地和服务器之间互为备份，以便万一数据丢失时可以恢复。

用户（尤其是初学者）在使用 Linux 系统和一些生物信息学软件时可能会遇到各种各样的问题，这时建议优先考虑在 Google 和 Baidu 百度 中搜索答案，大多数情况下会获得解决方案。

附录：已安装软件列表

Function	Name	Version	Webpage
Basic tools	BioPerl	1.6.0	www.bioperl.org
	picard	1.87	picard.sourceforge.net
	fastx_toolkit	0.13	hannonlab.cshl.edu/fastx_toolkit
	samtools	0.1.18	samtools.sourceforge.net
	R	2.11.1	www.r-project.org
	tabix	0.2.6	samtools.sourceforge.net/tabix.shtml
	BEDTools	v2.17	code.google.com/p/bedtools
	Python-2.7	2.7	www.python.org
	blast		blast.ncbi.nlm.nih.gov
Mapping	bwa	0.6.1	bio-bwa.sourceforge.net
	bowtie	2.0.2	bowtie-bio.sourceforge.net/bowtie2
Resequence	GATK	2.4-7	www.broadinstitute.org/gatk/
	samtools	0.1.18	samtools.sourceforge.net
	vcftools	0.1.10	vcftools.sourceforge.net
RNA-seq	tophat	2.0.6	tophat.cbcb.umd.edu
	cufflinks	2.0.2	cufflinks.cbcb.umd.edu
	AStalavista	v2.2	genome.crg.es/astalavista
miRNA-seq	mireap	0.2	mireap.sourceforge.net
	miRDeep		www.mdc-berlin.de/rajewsky/miRDeep
	randfold	2.0	bioinformatics.psb.ugent.be/software /details/Randfold
	ViennaRNA	2.0.7h	www.tbi.univie.ac.at/RNA/
Other	cluster	1.50	http://rana.lbl.gov/EisenSoftwareSource.htm
	EVER-seq	1.0.7	code.google.com/p/ever-seq
	hmmSplicer	0.9.5	http://derisilab.ucsf.edu/index.php?software=105
	libpng	1.2.7	www.libpng.org/pub/png/libpng.html
	libsequence	1.7.5	molpopgen.org/software/libsequence.html
	meme	4.9.0	meme.nbcr.net
	nawk		gnuwin32.sourceforge.net/packages/nawk.htm
	numpy	1.7.0b2	www.numpy.org
	passion	1.2.1	https://trac.nbic.nl/passion/
	promoter	2.0	www.cbs.dtu.dk/services/Promoter
	RetroSeq	master	www.sanger.ac.uk/resources/software/retroseq
	RSeQC	2.3.3	code.google.com/p/rseqc
	smalt	0.5.7	www.sanger.ac.uk/resources/software/smalt
	snappy	1.0.3	code.google.com/p/snappy
	fastqc	0.10.0	www.bioinformatics.babraham.ac.uk/projects /fastqc/

注：1. 以上软件均安装在共享目录/home/share/bin/，详细功能请参照各自主页

2. 生物信息学软件开发市场一片繁荣，推陈出新速度惊人，在此所列软件必不能满足用户的所有需求，绝大多数情况下仍需要自行安装。

3. 软件更新信息将在植物细胞与染色体工程国家重点实验室网站/平台建设/数据分析平台/(http://pcce.genetics.cas.cn/ptjs/sjfx/)上实时发布。