常用生物数据分析软件

王　俊　丛丽娟　郑洪坤 著

科学出版社
www.sciencep.com

常用生物数据分析软件

王 俊 丛丽娟 郑洪坤 著

科学出版社
北京

内 容 简 介

本书较为系统全面地介绍了生物信息学分析各个方面的软件用法，结合光盘具体实例，方便使用。全书共分8章，内容包括：Unix/Linux操作系统介绍，介绍了基本的Unix/Linux操作命令；数据的基本处理，介绍了如何处理常用的生物信息学数据；序列的比对，介绍了常用比对软件的用法及其在应用过程中要注意的问题；基因组/基因的注释，介绍了Coding和Non-Codoing基因的预测方法；SNP分析，介绍了常用的从生物学数据中寻找SNP的软件；进化分析专题，介绍了几种分子进化分析软件，内容涉及进化树的构建、Ka/Ks的计算等；基因表达分析专题，介绍了EST及生物芯片分析的流程和方法；蛋白质结构预测，介绍了蛋白质三维结构预测的流程及方法。

本书适合于生物信息学专业本科生及研究生使用。

图书在版编目(CIP)数据

常用生物数据分析软件/王　俊，丛丽娟，郑洪坤著. —北京：科学出版社，2008

ISBN　978-7-03-020622-0

Ⅰ. 常…　Ⅱ. ①王… ②丛… ③郑…　Ⅲ. 生物学－应用软体　Ⅳ. Q-39

中国版本图书馆CIP数据核字(2008)第062938号

责任编辑：李　悦　沈晓晶 / 责任校对：朱光光

责任印制：钱玉芬 / 封面设计：耕者设计工作室

科学出版社 出版

北京东黄城根北街16号

邮政编码：100717

http://www.sciencep.com

双青印刷厂 印刷

科学出版社发行　各地新华书店经销

*

2008年5月第　一　版　　开本：787×1092 1/16

2008年5月第一次印刷　　印张：23 1/4

印数：1—3 000　　字数：539 000

定价：65.00元(含光盘)

(如有印装质量问题，我社负责调换〈环伟〉)

序

生物信息学既是基因组学的孪生兄弟，又是基因组学的核心技术之一，是基因组学研究须臾不可离开的重要工具。基因组学两根重要的理论支柱分别植根于对生命本质的两个理解：一是“生命是序列的”，所有生物的遗传信息都储存在基因组 DNA 的核苷酸序列之中（Watson and Crick，1953）；二是“生命是数据的”（Sulston，2002），不是模拟的，在现代自然科学的前沿——生命科学和信息科学之间建立了实质性的联系。在方法学上，这两根支柱也使测序仪和计算机成为生命科学的最重要的武器。正因如此，生物信息学也可以理解为用信息学的手段来研究生物学，或以计算机为工具来研究生物信息。没有生物信息学，很难设想如今还能在生命科学领域做出具有真正前沿水平的工作成果。

近年来，国内外有关生物信息学的教科书和工具书已有不少，各有侧重，各具千秋。我郑重向相关读者推荐本书，首先是基于它的内容。它并没有试图全面叙述生物信息学的原理和发展历程，而着重于生物信息学方法论的讨论和工具的介绍。本书有着非常明确的目的，也有其明确的读者群，它收集了目前生物信息学研究领域中最常用的工具，涵盖了序列拼接、表达分析、进化分析、比较基因组学等重要方面的工作，提供了精心设计的“手把手”、“一教就会，一学就能”的实例，具有很强的实用性。它将成为基因组学初学者的入门教材，也将成为生命科学其他领域的“行外专家”涉足生物信息学的得力助手。

我推荐本书的另一原因，是由于它出自于研究最前线的科研工作者。这些作者都是生物信息学领域“年轻的老战士”，他们亲身参与了近 10 年来我国几乎所有的大基因组计划，经历了严峻的实战洗礼，积累了丰富的一线经验，并作出了重要贡献。由他们编写的本书，我相信首先会以亲近感和信任感而吸引读者。至于书中存在的不足之处，纵然资深作家也难以避免，应该予以宽容和谅解，只是希望本书的编写者在不久的将来再版时，能在本书的基础上“百尺竿头，更上一步”。

我向广大读者推荐这本书，更是由于它出版于最合适的时机。2007 年被视为“奇迹之年”，甚至可以和奠定物理世纪的 1905 年相媲美。新一代测序技术的应用被视为影响最大的突破之一。正如火如荼进行的“国际千人基因组计划”、“国际癌症基因组计划”，以及 metagenomics 和 epigenomics 等计划，并且由此推动的基因组学的又一轮热潮，都将产生海量的序列及其相关数据，更加突显对生物信息学方法和工具的需求。本书出版，恰逢其时。

我相信所有心态平和的同事们都会赞同我对此书的评价。我期望本书读者和使用者多提出批评与建议，更期待本书对我国基因组学和生物信息学产生的应有影响。

此为序。

杨焕明　于深圳

2008 年 5 月 1 日

前　言

20 世纪末期及 21 世纪，随着测序技术的发展和人类基因组计划的实施完成，我们拥有了海量的生物学数据，这些数据必须经过收集、分析和整理后，才能成为有用的信息与知识，这个过程就是生物信息学分析工作。生物信息学是生物学与计算机科学及应用数学等学科相互交叉而形成的一门新兴学科，它通过对生物学实验数据（当前主要是核苷酸和氨基酸序列）的获取、加工、存储、检索和分析，从而达到揭示数据所蕴含的生物学意义的目的。

关于生物信息学的理论性书籍已经很多，所以本书并没有过多的描述理论方面，而是从实际使用的具体分析工具入手，对信息分析的几个主要方面进行最为细致的讲解，包括软件的安装、输入输出数据的格式说明、常用参数的选取等，并配以实例数据方便大家熟悉及使用。本书附赠的光盘中，每个软件都对应于相应的目录，请参照书中指导进行使用。由于大部分软件使用环境均为 Linux/Unix 系统，读者需要有自己的 Linux/Unix 服务器，或者在自己的电脑上安装虚拟 Linux 环境（如 VMWare）或双系统。

本书第一章主要介绍了 Linux/Unix 常用命令，此为其余软件使用的基础；第二章从 Sanger 法测序数据出发，介绍了如何对测序数据进行初步的峰图转换、去污染及拼接分析；在得到基因组或基因的数据后，如何进行比对分析及注释分析则是本书第三、四章内容，通过此内容，可以了解基因组各种组分的分析方法，如重复序列、编码区、RNA 等，并介绍了如何进行功能分类；第五、六章分别介绍了 SNP 分析及进化分析的几款常用软件，在进化分析方面除了构建系统发育树的几款常用软件外，我们还介绍了计算 Ka/Ks 比值的一个工具及寻找蛋白家族进而进行进化分析的 FGF 工具；第七、八章简单介绍了表达分析及蛋白结构预测的相关工具。

本书作者全部为一线科研人员，在承担各项科研任务之余编写了本书，书中所介绍的软件都是我们工作中反复应用的，我们希望通过本书让更多的生物科研工作者能够尽快熟悉并应用生物信息学工具，更好的为生命科学研究服务。

由于时间和经验等方面原因，本书难免会有一些错误，还请广大读者谅解，也希望大家多提宝贵意见。

作　者

目　录

第 1 章　Unix/Linux 操作系统介绍

1.1　远 程 登 录

进入 Unix/Linux 系统，必须要输入用户的账号和密码，普通用户的账号由系统管理员创建，可以进行有限的操作。登录大型机常用的有三种方式。

1. Telnet 登录

此方式登录大型机，不需要特殊软件。

第一步：打开命令对话框（Windows 系统开始→开始→运行→cmd），输入远程主机 IP，命令为："telnet 192. 168. 1. 120"（图 1-1）。

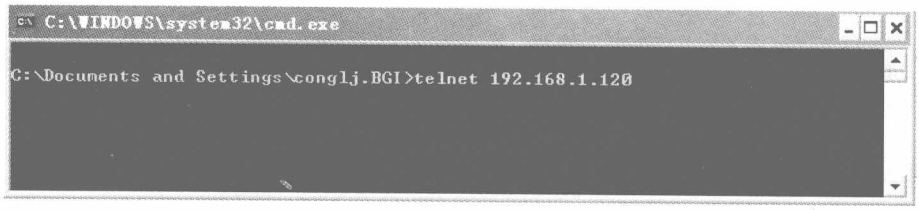

图 1-1　telnet 登录大型机，命令行

第二步：连接成功，系统会提示输入用户名（图 1-2）。

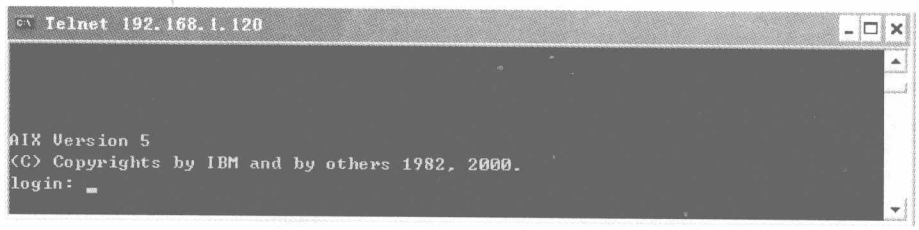

图 1-2　telnet 登录大型机，输入用户名

第三步，如果用户名存在，则提示输入密码（图 1-3）。

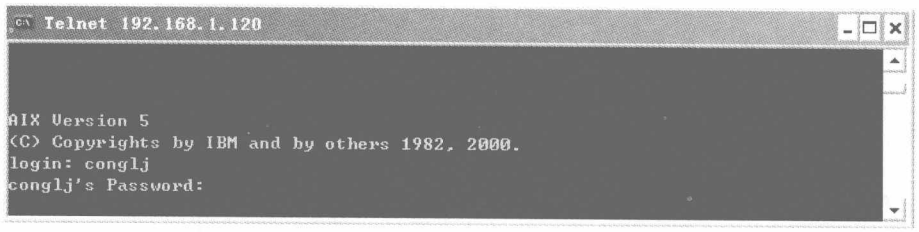

图 1-3　telnet 登录大型机，输入密码

第四步，密码输入正确，成功登录（图 1-4）。

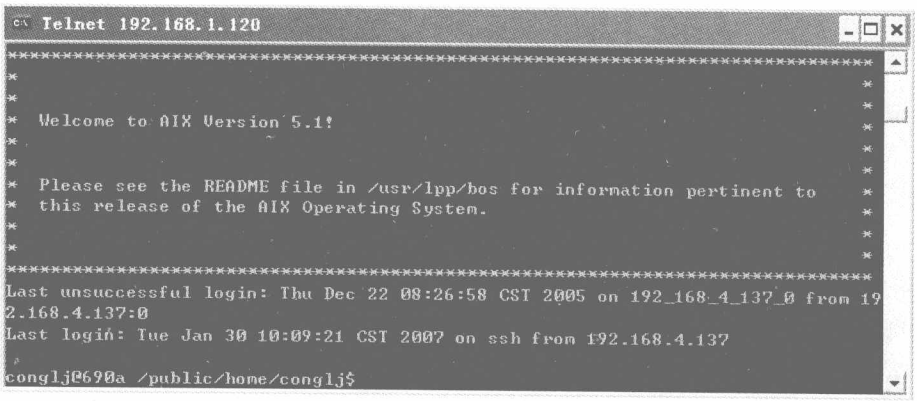

图 1-4　telnet 登录大型机，登录成功

完成任务执行需要退出系统时，只需在 shell 提示符下，键入 exit 命令即可。

2. SSH 登录

需要软件辅助，常用的软件有 SecureCRT、PuTTY 等，下面以 PuTTY 为例讲解 SSH 登录大型机（图 1-5、图 1-6）。

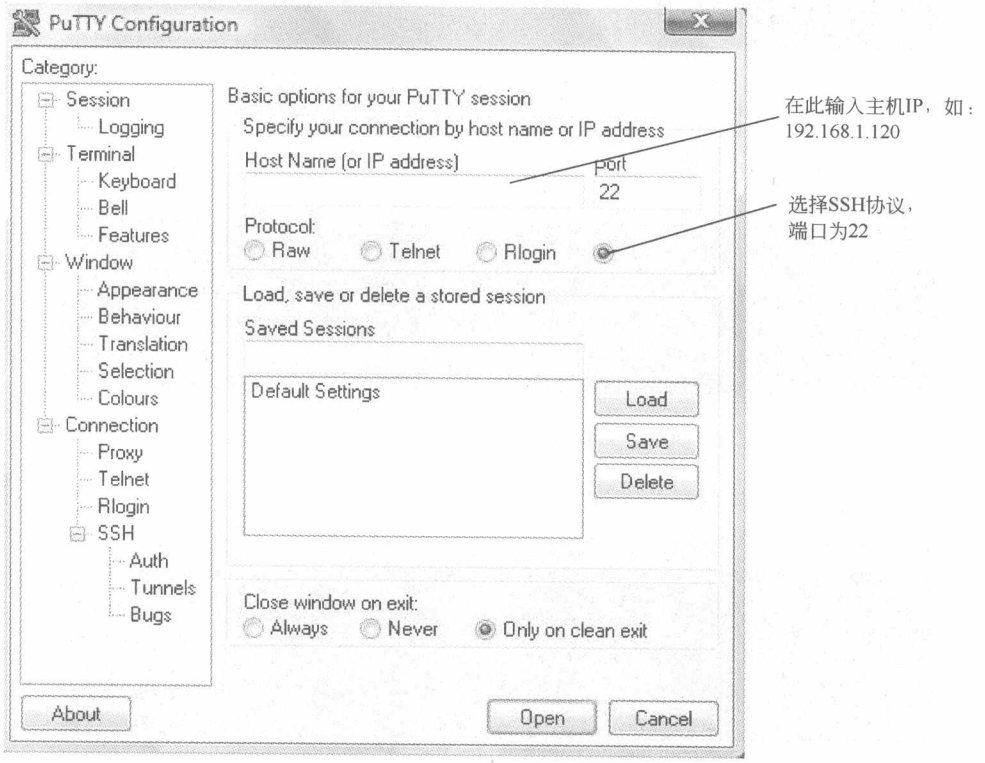

图 1-5　SSH 登录大型机，配置主机 IP 及登录方式

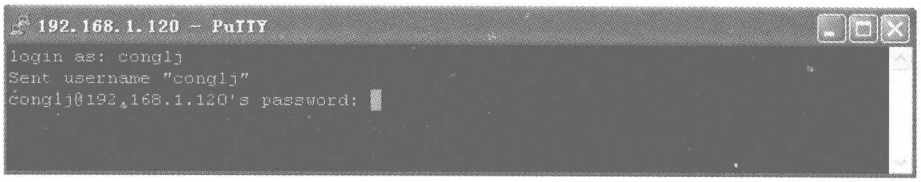

图 1-6　SSH 登录大型机，输入用户名及密码

3. X-Win 登录

X-Win 图形界面登录大型机，可以调用图形窗口，对于 consed 等软件的使用提供了一个很好的平台，可用软件：Xming（http://freedesktop. org/wiki/Xming）、X-Win32 等。以下 Xming 登录步骤仅供参考（图 1-7 ~ 图 1-13）。

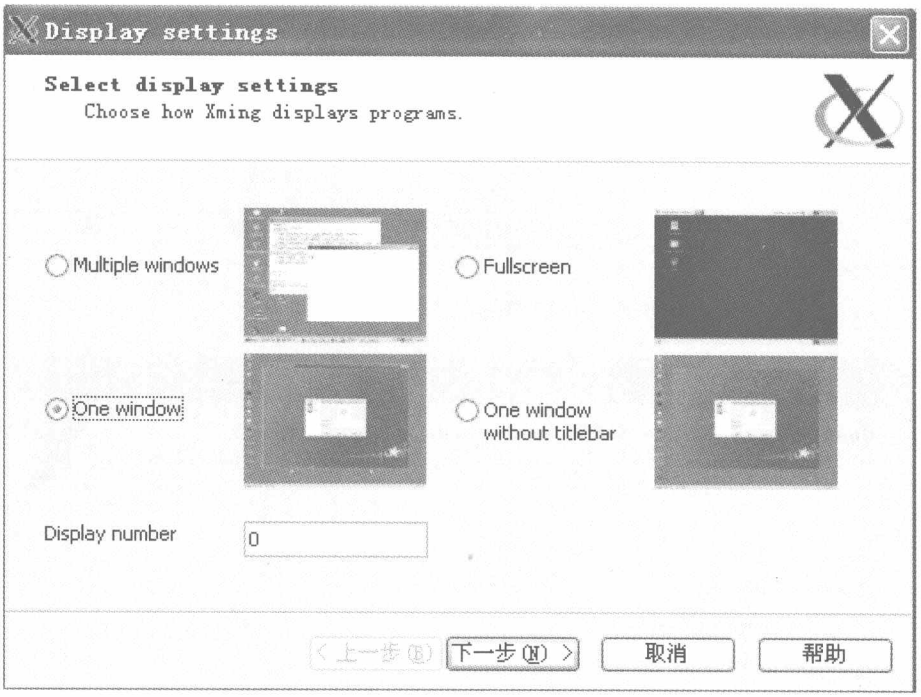

图 1-7　选择界面模式

1）修改密码 passwd 命令

为了更好地保护用户账号的安全，Unix/Linux 允许用户随时修改自己的密码，修改密码的命令是 passwd，键入此命令后，系统将提示用户输入旧密码和新口令，并再次确认新密码，以避免用户无意中按错键。如果用户忘记了密码，可以向系统管理员申请为自己重新设置。

2）更改用户命令 su

su 命令可以将用户更改为其他用户，运行此命令时系统会提示输入另一用户的密码。默认更改为管理员 root。

su - user 更改为其他用户并使用其环境变量设置。

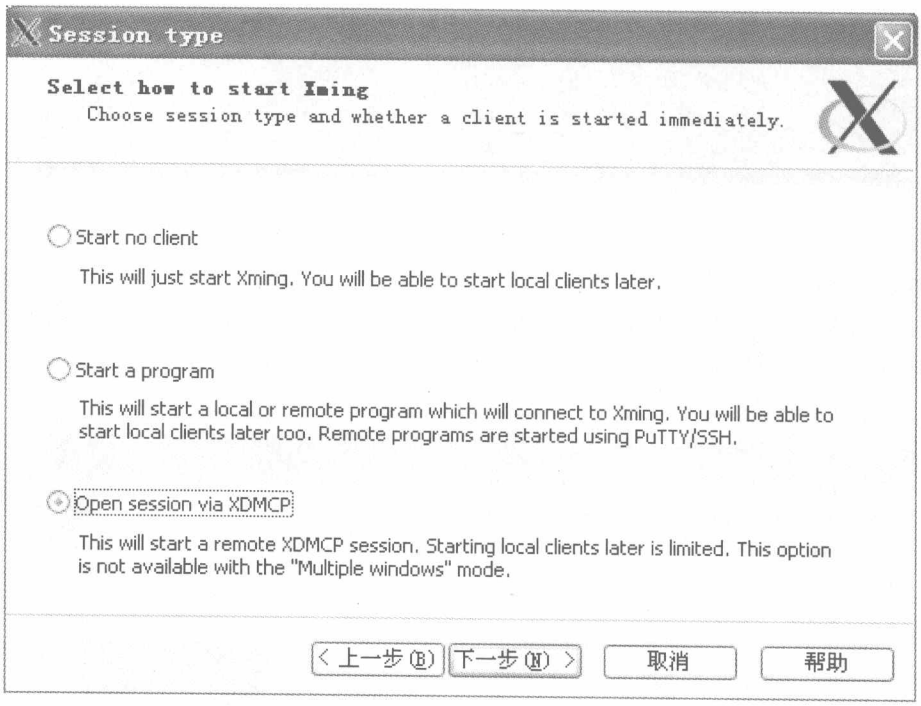

图 1-8　Xming 登录大型机，选择登录方式

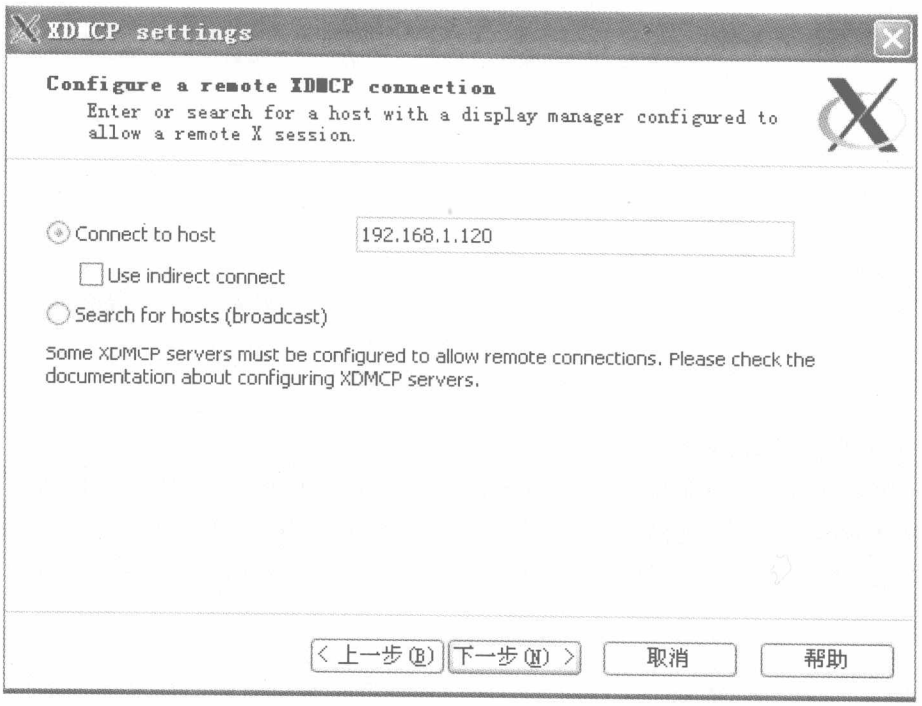

图 1-9　填写登录主机 IP

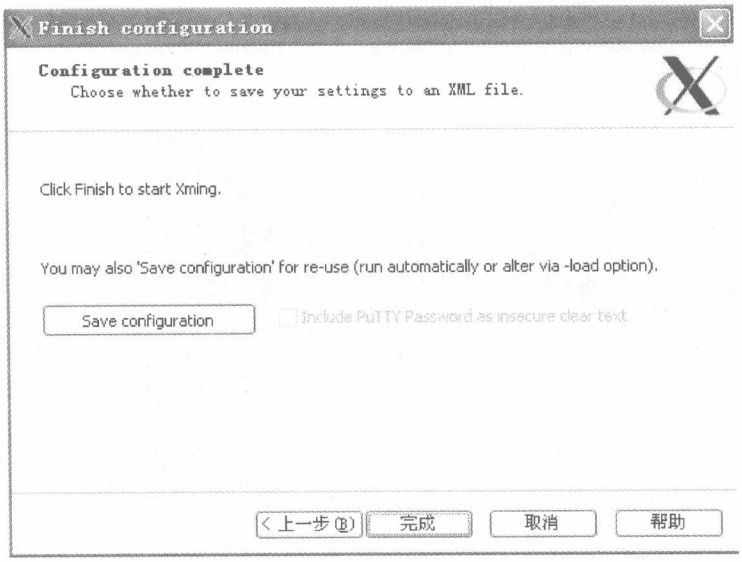

图 1-10　完成配置

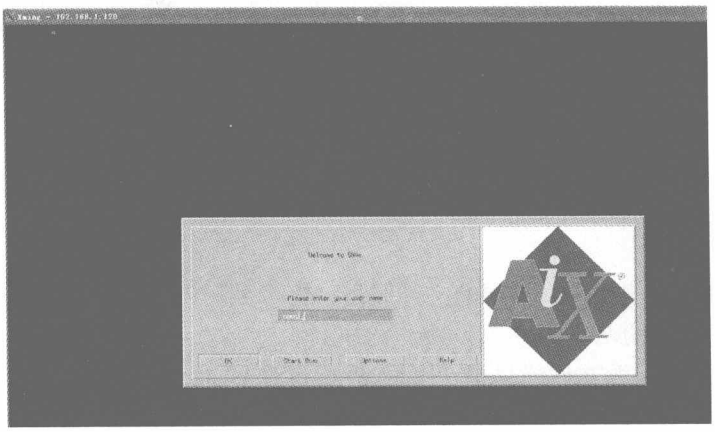

图 1-11　主机连接成功，输入用户名

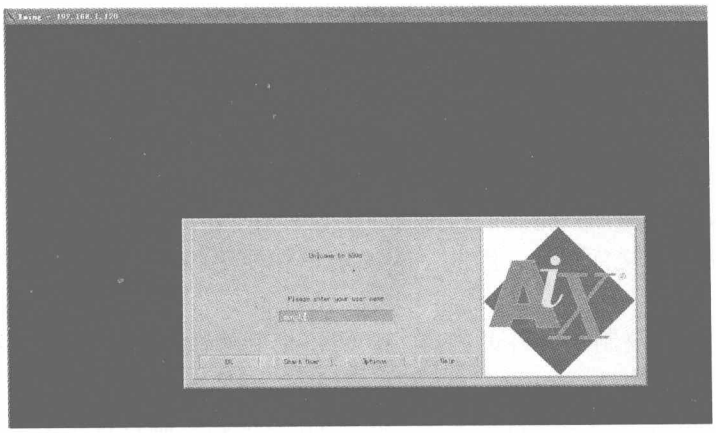

图 1-12　输入用户的密码，点击 OK 登录

登录成功，图 1-13 为远程计算机操作界面（不同的计算机有不同的界面）。

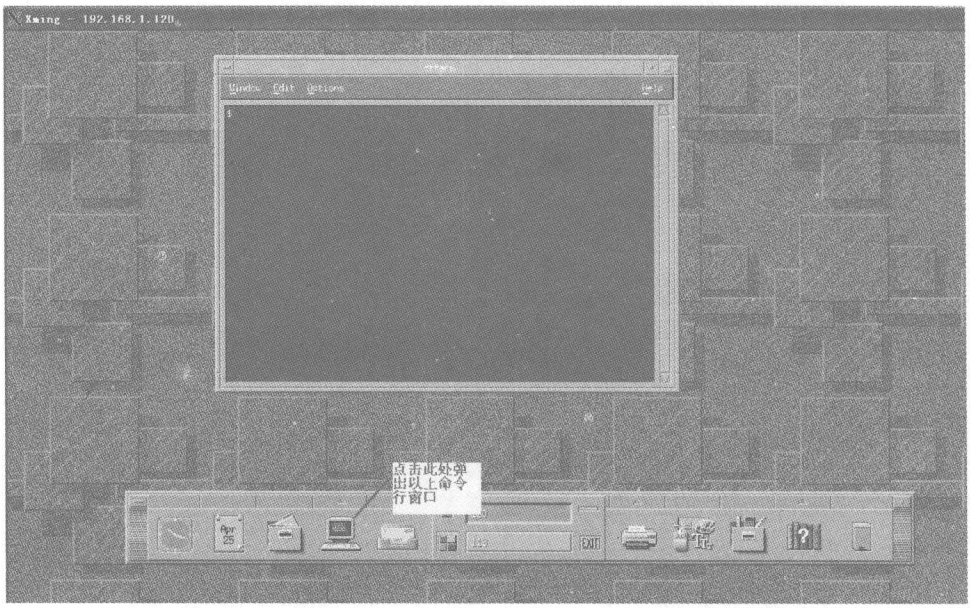

图 1-13 登录成功，打开命令行窗口

1.2 文件的复制、删除和移动命令

1. cp 命令

说明：该命令的功能是将指定的文件或目录拷贝到另一文件或目录中。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］源文件或目录 目标文件或目录

参数：

-a 该选项通常在拷贝目录时使用。它保留链接、文件属性，并递归地拷贝目录，其作用等于 dpR 选项的组合。

-d 拷贝时保留链接。

-f 删除已经存在的目标文件而不提示。

-i 和 f 选项相反，在覆盖目标文件之前将给出提示要求用户确认。回答 y 时，目标文件将被覆盖，是交互式拷贝。为防止用户在不经意的情况下用 cp 命令破坏另一个文件，如用户指定的目标文件名已存在，用 cp 命令拷贝文件后，这个文件就会被新源文件覆盖，因此，建议用户在使用 cp 命令拷贝文件时，最好使用 i 选项。

-p 此时 cp 除复制源文件的内容外，还将把其修改时间和访问权限也复制到新文件中。

-r 若给出的源文件是一目录文件，此时 cp 将递归复制该目录下所有的子目录和文件。此时目标文件必须为一个目录名。

-l　不作拷贝，只是链接文件。

例子：

$ cp file1 file2　将 file1 拷贝成 file2

$ cp /usr/file2. /　将/usr 目录下的文件 file2 拷到当前目录下

$ cp — help　查阅命令详细使用信息

2. mv 命令

说明：用户可以使用 mv 命令来为文件或目录改名，或将文件由一个目录移入另一个目录中。视 mv 命令中第二个参数类型的不同（是目标文件还是目标目录），mv 命令将文件重命名或将其移至一个新的目录中。当第二个参数类型是文件时，mv 命令完成文件重命名，此时，源文件只能有一个（也可以是源目录名），它将所给的源文件或目录重命名为给定的目标文件名。当第二个参数是已存在的目录名称时，源文件或目录参数可以有多个，mv 命令将各参数指定的源文件均移至目标目录中。在跨文件系统移动文件时，mv 先拷贝，再将原有文件删除，而链至该文件的链接也将丢失。

语法：mv［参数］源文件或目录 目标文件或目录

参数：

-i　交互方式操作。如果 mv 操作将导致对已存在的目标文件的覆盖，此时系统询问是否重写，要求用户回答 y 或 n，这样可以避免误覆盖文件。

-f　强制执行，在 mv 操作要覆盖某已有的目标文件时不给任何指示，指定此选项后，i 选项将不再起作用。如果所给目标文件（不是目录）已存在，此时该文件的内容将被新文件覆盖。为防止用户用 mv 命令破坏另一个文件，使用 mv 命令移动文件时，最好使用 i 选项。

例子：

$ mv file1 file2　将 file1 改名为 file2

$ mv. /test /sdb/conglj/　将文件 test 移至/sdb/conglj 目录下

$ mv -help　查阅命令详细使用信息

3. rm 命令

说明：用户可以用 rm 命令删除不需要的文件。该命令可删除一个目录中的一个或多个文件或目录，它也可以将某个目录及其下的所有文件及子目录均删除。对于链接文件，只是断开链接，源文件保持不变。如果没有使用 -r 选项，则 rm 不会删除目录。

语法：rm［参数］文件…

参数：

-f　忽略不存在的文件，不给任何提示。

-r　将列出的全部目录和子目录逐级递归地删除。

-i　进行交互式删除。文件被删除后是不能被恢复的。为防止误删有用文件，可以使用 i 选项来逐个确认要删除的文件。如果用户输入 y，文件将被删除。否则文件不会删除。

例子：

$ rm file1 file2 file3　删除三个文件

$ rm * 删除当前目录下所有文件（目录不删除）

$ rm-help 查阅命令详细使用信息

1.3 目录的创建、删除及更改目录命令

1. mkdir 命令

说明： 创建一个目录（类似 MSDOS 下的 md 命令）。要求创建目录的用户在当前目录中（dir-name 的父目录中）具有写权限，并且 dirname 不能是当前目录中已有的目录或文件名称。

语法： mkdir［参数］目录名

参数：

-m 对新建目录设置存取权限。也可以用 chmod 命令设置。

-p 可以是一个路径名称。此时若路径中的某些目录尚不存在，加上此选项后，系统将自动建立好那些尚不存在的目录，即一次可以建立多个目录。

例子：

$ mkdir data 在当前目录下建立子目录 data

$ mkdir /usr/data 在/usr/目录下建立子目录 data，此时/usr 目录必须已经存在。

2. rmdir 命令

说明： 该命令从一个目录中删除一个或多个子目录项。一个目录被删除之前必须是空的。rm-r dir 命令可代替 rmdir。

语法： rmdir［参数］目录名

参数：

-p 递归删除目录，当子目录删除后其父目录为空时，也一同被删除。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。

3. cd 命令

说明： 该命令将当前目录改变至指定的目录。若没有指定目录，则回到用户的主目录。

语法： cd［目录名］

例子：

$ cd /usr/bin 切换至/usr/bin 目录

$ cd.. 切换至上一层目录

$ cd/ 切换至根目录

$ cd~ 切换至宿主目录（用户登录时所在的目录），效果等同于不加指定目录。

4. pwd 命令

说明： 该命令显示用户当前所在的路径，为全路径。

语法： pwd

5. ls 命令

说明： ls 是英文单词 list 的简写，其功能为列出目录的内容。对于每个目录，该命令将列出其中所有的子目录与文件。对于每个文件，ls 将输出其文件名以及所要求的其他信息。默认情况下，输出条目按字母顺序排序。当未给出目录名或是文件名时，就显示当前目录的信息。

语法： ls［参数］［目录或是文件］

参数：

-a　显示指定目录下所有子目录与文件，包括隐藏文件。

-c　按文件的修改时间排序。

-F　在文件后面加上类型标识：如果是目录，则在后面加“/”；如果是可执行文件，则在后面加“*”；如果是个链接，则在后面加“@”，管道（或 FIFO）后面标记“|”，socket 文件后面标记“=”。

-L　若指定的名称为一个符号链接文件，则显示链接所指向的文件。

-r　按字母逆序或最早优先的顺序显示输出结果。

-R　递归式地显示指定目录的各个子目录中的文件。

-s　给出每个目录项所用的块数，包括间接块。

-t　显示时按修改时间（最近优先）顺序而不是按名字排序。若文件修改时间相同，则按字典顺序。

-l　以长格式来显示文件的详细信息。是最常用的参数之一。每行列出的信息依次是：文件类型与权限、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，它可以是下述类型之一：-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 个字符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。第一组表示文件属主的权限，第二组表示同组用户的权限，第三组表示其他用户的权限。每一组的三个字符分别表示对文件的读（r）、写（w）和执行权限（x）。

例子：

$ ls　显示目前目录中所有文件。

$ ls /usr/bin　显示/usr/bin 目录下的文件

$ ls-a /home/conglj　显示/home/conglj 目录下所有的文件和目录，若无此参数，“.”开始的隐藏文件和目录不会显示。此命令显示结果为：

. . . . bash_history . bash_profile. bashrc. tsshrc

$ ls -l file1

-rwxr-xr- - 1 soft bgi Aug 8 05：08 file1

第一列是文件的属性：

第一个字符（-）表示是单纯的文件

第 2 ~ 4 字符“rwx”　表示此文件属主 soft 对文件 file1 的权利为“可读、可写、可执行”；

第 5 ~ 7 字符 “r - x” 表示此用户组 bgi 内的用户对文件 file1 的权利为：“可读、不可写、可执行”；

第 8 ~ 10 字符 “r - -” 表示其他用户对文件 file1 的权利为 “可读、不可写、不可执行”。

第二列表示文件的链接数为 1。

第三列表示此文件或目录的拥有者是 soft 用户。

第四列表示文件所有者 soft 用户所属的组是 bgi。

第五列表示文件大小为 8byte。

第六列表示文件的修改日期是 8 月 8 日。

第七列表示文件名为 file1。

1.4 文本查看命令

1. more 命令

说明： 在终端屏幕按屏显示文本文件，该命令一次显示一屏文本，显示满之后停下来，并在终端底部打印出 “ - - More - - ”，系统还将同时显示出已显示文本占全部文本的百分比，若要继续显示，按回车或空格键即可。若要退出，按 q 或 Q。

语法： more [参数] 文件

参数：

-p 显示下一屏之前先清屏。

-c 作用同 -p 基本一样

-l 不处理 ctrl +l (换页符)。如果没有给出这个选项，则 more 命令在显示了一个包含有 ctrl +l 字符的行后将暂停显示，并等待接收命令。

-s 文件中连续的空白行压缩成一个空白行显示。

例子：

$ more -c -5 example. txt 执行该命令后，先清屏，然后将以每五行的方式显示文件 example. txt 的内容。

2. less 命令

说明： less 命令的功能几乎和 more 命令一样，也是用来按页显示文件。此命令可以使用方向键滚动文件。用 less 命令显示文件时，若需要在文件中往前移动，按 b 键；要移动到用文件的百分比表示的某位置，则指定一个 0 到 100 之间的数，并按 p 即可。

语法： less [参数] 文件

参数：

less -S 分列显示

less -help 显示详细说明文档

3. cat 命令

说明： 该命令功能之一是用来显示文件，它依次读取其后所指文件的内容并将其输出到标准输出。该命令功能之二是用来将两个或多个文件连接起来。