吉康医学
GENECOME MEDICAL

吉康医学常用生物信息学技术教程

GENECOME BIOINFORMATIONAL TECHNICAL MANUL

技术引领医学转化 专业创造行业口碑

北京吉康医学科技有限公司

www.genecome.cn

北京吉康医学科技有限公司

公司总部：北京市北京经济技术开发区荣华南路1号院6号楼

技术服务邮箱：genecomeservice@163.com

公司官网：www.genecome.cn

目录

第 1 章 Unix/Linux 操作系统介绍

1.1 文件和目录相关

mkdir dirname　　建立子目录. 注意:用户不能在一个不存在的目录中建立子目录。

　　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。

rmdir dirname　　删除空目录，目录里面如有文件或目录则无法删除。

pwd　　显示用户目前所在目录

cd dirname　　切换目录。

　　cd .　　"."表示当前目录

　　cd ..　　".."表示上一层目录

　　cd /　　"/"表示根目录

　　cd ~　　"~"表示宿主目录(用户登录时所在的目录)

　　cd /usr/bin　　切换到/usr/bin 目录下

ls　查看文件信息，这是最基本的档案指令。 ls 的意义为 "list"，也就是将某一个目录或是某一个档案的内容显示出来。 ls 命令可加参数很多，我们这里不一一列出，只给出较常用的几个，各参数可以混合使用。

　　ls　不加任何信息，显示目前目录中所有文件。

　　ls [file]　显示特定的文件。如：% ls /home2/X11R5

　　ls –a　　显示所有的文件和目录，若无此参数，句点开始的文件和目录不会显示出来，即以”.”开头的文件，如 tcsh 的初设档 .tcshrc；如果我们要察看这类档案，则必须加上参数 -a

　　ls -l　　这个参数代表使用 ls 的长(long)格式，可以显示更多的信息，包括文件的权限、所有者、大小、最后更改日期等。如：

　　ls -l file1

　　-rwx--x--x 1 soft bgi Aug 8 05:08 file1

　　第一列表示文件得属性，linux 下文件分三个属性：可读 r，可写 w，可执行 x

　　第一个字符表示是目录(d)或链接文件(l)或单纯的文件(-)等

　　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”

　　第二列表示文件个数，如果是文件则为 1，如果是目录则表示里面的文件个数。

　　第三列别是此文件或目录的拥有者。

　　第四列表示文件所有者所属的组

第五列表示文件大小，用 byte 表示

第六列表示文件的修改日期

第七列表示文件或目录名称

ls –t 按文件最后更改时间排序文件

ls –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”@”

more [file] 显示文件，按屏显示，空格键翻页，回车键每次只翻一行，敲入 q/Q/:q/:Q/ZZ 等都可提前退出 more 命令。

less [file] 基本同 more 命令，可以使用方向键随意滚动文件。

less –S 分列显示

less –help 显示详细说明文档

cat [file] 显示文件内容，所有内容全部显示。

cat –n [file] 在显示内容前加上行号

cp 拷贝文件，可以将文件拷贝成另一个文件，或是拷贝到另一个目录中。可以使用通配符拷贝具有同一特征的所有文件。

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 –i 覆盖相同名称文件前先询问用户

cp –R 递归拷贝，即拷贝时将所有目录一并拷贝

cp --help 查阅命令详细使用信息

mv 移走目录或者改文件名

mv file1 file2 将 file1 改名为 file2

mv filename dirname/ 将文件移至某一目录下

mv –help 查阅命令详细使用信息

rm 删除文件或目录

rm file1 file2 file3 ….

rm * 删除当前目录下所有文件

rm –f 强制删除文件，删除时，不提出任何警告讯息。

rm –i 删除文件之前均会询问是否真要删除，y/n 指示下一步。

rm –r 递回式的删除，即逐级删除目录下的子目录。

rm –help 查阅命令详细使用信息

chmod 更改文件或目录权限

chmod –r file 更改所有的权限，包括子目录及其内文件。

chmod nnn file(s) n 从 0 到 7，权限可相加。依次代表用户、组成员、其他人的权限。

0 无任何权限

1 可执行权限

2 可写权限

4 可读权限

chmod a operator b file(s)　a 代表用户 u、组 g 或其他 o，operator 代表＋－＝：权限的更改方式，b 代表权限类型：r 可读 w 可写 x 可执行

chmod g+rw file 增加文件组内可读写的权限

chmod o=rx file 更改文件的权限，使其他用户可读可执行

chown　更改文件或目录所有者，自己不能再改回来。

chown UID:GID files

grep 是一过滤器，它可搜索文件并过滤出有某个特征的行

grep [-nv] match_pattern file1 file2 ....

-n 把所找到的行在行前加上行号列出

-v 把不包含 match_pattern 的行列出

ln [-参数] [源文件或目录][目标文件或目录] 指令用在链接文件或目录。连结又可分为两种:硬连结(hard link)与软连结(symbolic link)，硬连结的意思是一个文件可以有多个名称，而软连结的方式则是产生一个特殊的文件，该文件的内容是指向另一个文件的位置。硬连结是存在同一个文件系统中，而软连结却可以跨越不同的文件系统。常用的参数如下：

-b　删除，覆盖目标文件之前的备份。

-d 或-F　建立目录的硬连接。

-s　对源文件建立符号连接，而非硬连接。

-f　强行建立文件或目录的连接，不论文件或目录是否存在。

-i　覆盖既有文件之前先询问用户。

split [OPTION] [INPUT [PREFIX]] 将一个文件分割成数个，输出依次为 PREFIXaa, PREFIXab…… PREFIX 默认为 x。

-b, --bytes=SIZE SIZE 值为每一输出档案的大小，单位为 byte。SIZE 可加入单位: b 代表 512， k 代表 1K， m 代表 1 Meg。

-l NUMBER　NUMBER 值为每一输出文件的行数。

cut 截取文件中的某字段。

-c m-n 表示显示每一行的第 m 个字元到第 n 个字元。

-f m-n 表示显示第 m 栏到第 n 栏（使用 tab 分隔)。

-d’分隔符’ 用来定义分隔符（单个字符)，默认为 tab 键，和-f 配合使用。

sort 命令的功能是对文件中的各行进行排序，默认为以整行为关键字按 ASCII 字符顺序进行排序。

-u 对排序后认为相同的行只留其中一行。

-f 将小写字母与大写字母同等对待。

-r 按逆序输出排序结果。

uniq 处理文件中重复的行

-d 只显示重复行。

-u 只显示文件中不重复的各行。

find 查找文件，基本用法 find [路径] [参数]，可以使用 find -help 查看详细说明。

find bin/ -name run.sh 查找 bin 目录下名字为 run.sh 的文件

-amin n 查找系统中最后 n 分钟访问的文件

-atime n 查找系统中最后 n 天访问的文件

-cmin n 查找系统中最后 n 分钟被改变状态的文件

-ctime n 查找系统中最后 n 天被改变状态的文件

-empty 查找系统中空白的文件，或空白的文件目录

wc 该命令用来统计给定文件中的字节数、字数、行数。

-c 统计字节数。

-l 统计行数。

-w统计字数。

du [options] [file or dir]统计文件大小

-s 所有文件大小总和

-k 以 kbytes 为单位输出

awk 对文件进行信息提取等处理，基本模式为：awk ‘{操作代码}’ 输入文件

$ awk ’{ print }’ /etc/passwd 此命令输出/etc/passwd 文件的内容。/etc/passwd 为输入文件。花括号用于将几块代码组合到一起，这一点类似于 C 语言。

$ awk -F:″:″ ’{ print $1 }’ /etc/passwd 使用 -F 选项来指定 ″:″ 作为字段分隔符，打印出在输入文件中每一行中出现的第一个字段。

1.2 压缩和解压缩

gzip (gunzip) 压缩（解压缩）文件，产生后缀为.gz 的压缩文件。

gzip –d file 解压缩文件

gzip –f file 如果压缩的文件重名，则强制覆盖

gzip –h 显示此命令的帮助信息

zip 压缩文件

unzip 解压缩文件，该命令用于解扩展名为.zip的压缩文件。

-t 检测压缩的档案文件

-d 解压缩文件到 exdir

tar 打包多个文件到一个压缩包或反之

tar –cf bin.tar /usr/bin 将/usr/bin 目录下所有文件打包成 bin.tar

tar –xf bin.tar 提出 bin.tar 包里所有文件

tar –tvf bin.tar 给出 bin 包里的文件列表，并不解压缩

tar –help 显示此命令的帮助信息

compress 压缩文件，压缩后的文件会加上一个 .Z 后缀以区别未压缩的文件，可以用 uncompress 解压缩或使用参数－d 解压缩

1.3 进程及其他

man [命令] man 是手册（manual）的意思。用来让使用者查询某一命令的具体使用帮助。

Ctrl+f 或空格键 后翻一页

Ctrl+b 或 b 前翻一页

Ctrl+c 或 q 离开

重定向，可将某命令的结果输出到文件中

>file将结果输出到文件 file 中，如果该文件原本就存在，则该文件原有的内容会被删除

>>file 将结果输出到文件 file 中，如果原文件存在，则附加在原文件后面，原文件的内容不会被清除

管道符|，可将某命令的结果输出给另一命令

su 更改为其他用户，默认为 su 到 root，会提示输入另一用户的密码

su – user 更改为其他用户并使用其环境变量设置

passwd 更改用户密码，会提示输入旧密码，并两次输入新密码以确认

top 即时显示进程动态，进入 top 命令后可以使用如下几个命令进行操作：

h：显示帮助信息

q：离开此命令

s：更新速度，每几秒更新一次，也可使用空格键手动更新。

n：只显示最上面运行的几个进程

i： 不显示任何闲置（idle）或无用（zombie）的行程

u：单独显示某一用户的进程，“+”为显示所有用户的进程，

history 查询历史命令记录

history number 显示前面几个命令

history –c 从下一个命令开始记录

history –h 只显示命令历史记录，不显示命令编号、时间等信息

History –r 反向显示命令的历史纪录，即从最近的一个命令开始显示

ps 显示用户的运行程序或系统程序

ps –e 列出所有用户的进程

ps –u [user]　列出用户 user 的进程

ps –f 给出详细列表

kill 杀掉某一进程

kill [-signal] pid　signal 为 0 到 31 的数字，也可以是特定字符串。如数字 9 代表 KILL，可以杀掉一般无法终止的程序。

kill -l 查看 signal 代表的意思。常用的 signal 有 HUP、STOP、CONT 等。

1.4 远程登陆

登陆大型机的三种方式：

1. Telnet 登陆大型机，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

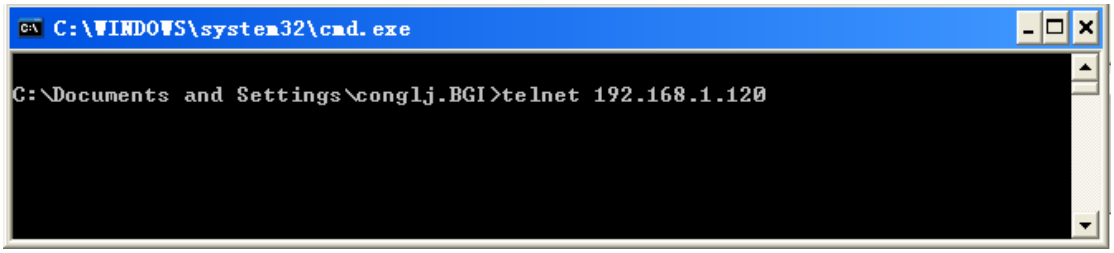

图 1-1 telnet 登陆大型机，命令行

第二步：连接成功，系统会提示输入用户名：

图 1-2 telnet 登陆大型机，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 telnet 登陆大型机，输入密码

第四步，密码输入正确，成功登陆：

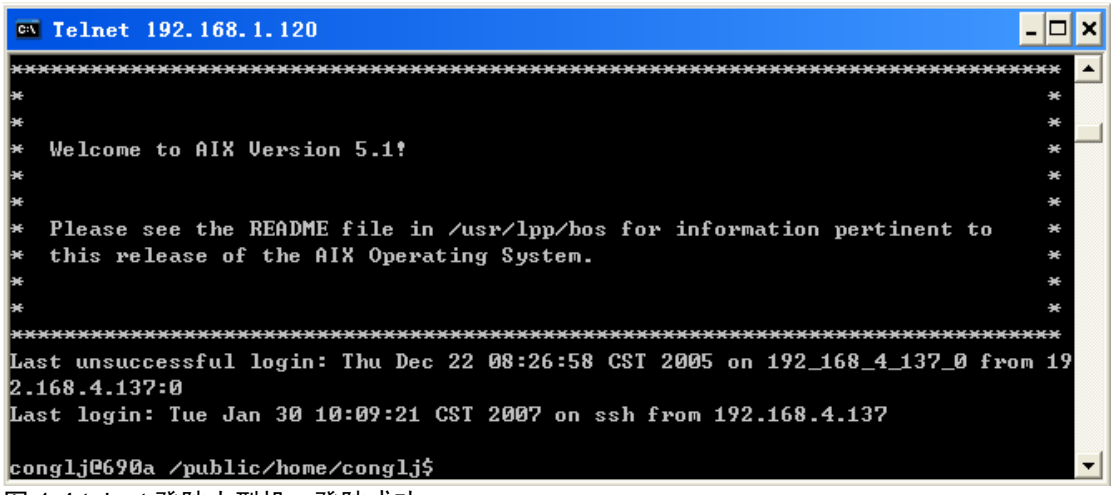

图 1-4 telnet 登陆大型机，登陆成功

2. SSH 登陆大型机，需要软件辅助，常用的软件有：SecureCRT，等，下面以 SecureCRT 为例讲解 SSh 登陆大型机：

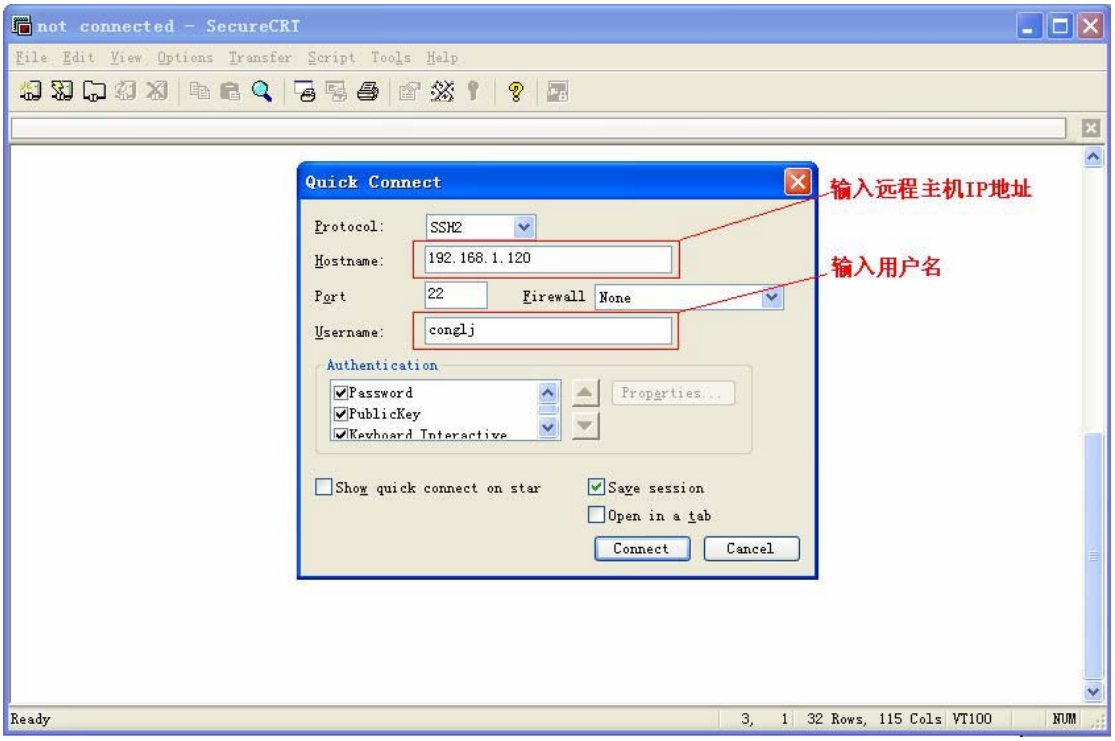

图 1-5 SSH 登陆大型机

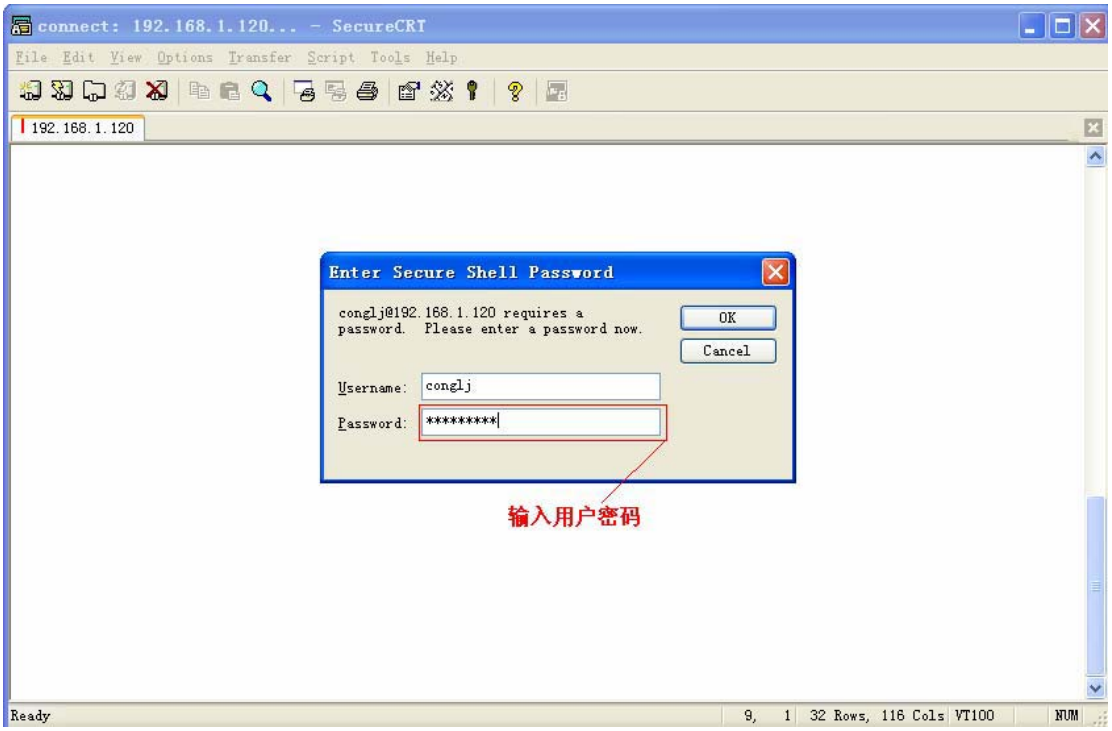

图 1-6

3． X-Win 登陆大型机

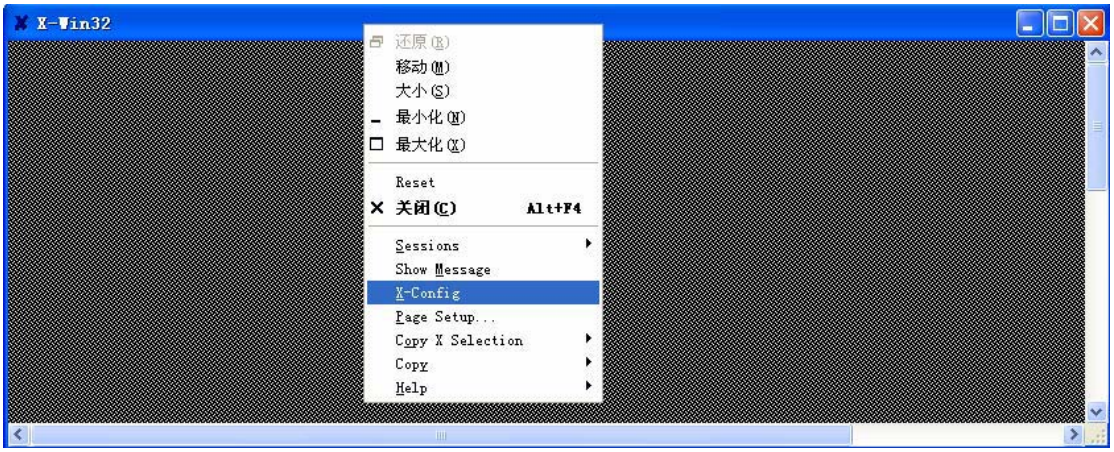

图 1- 7

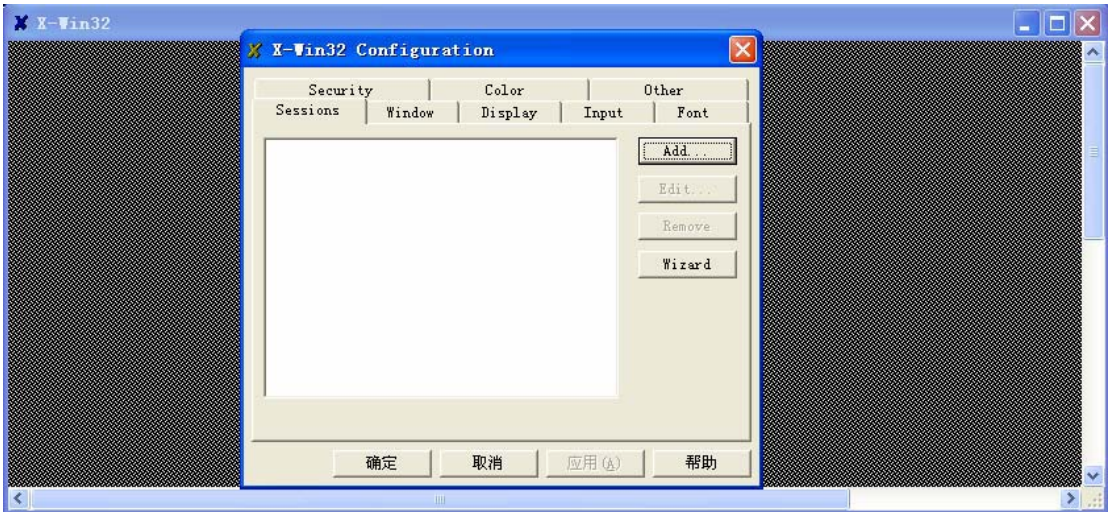

图 1- 8

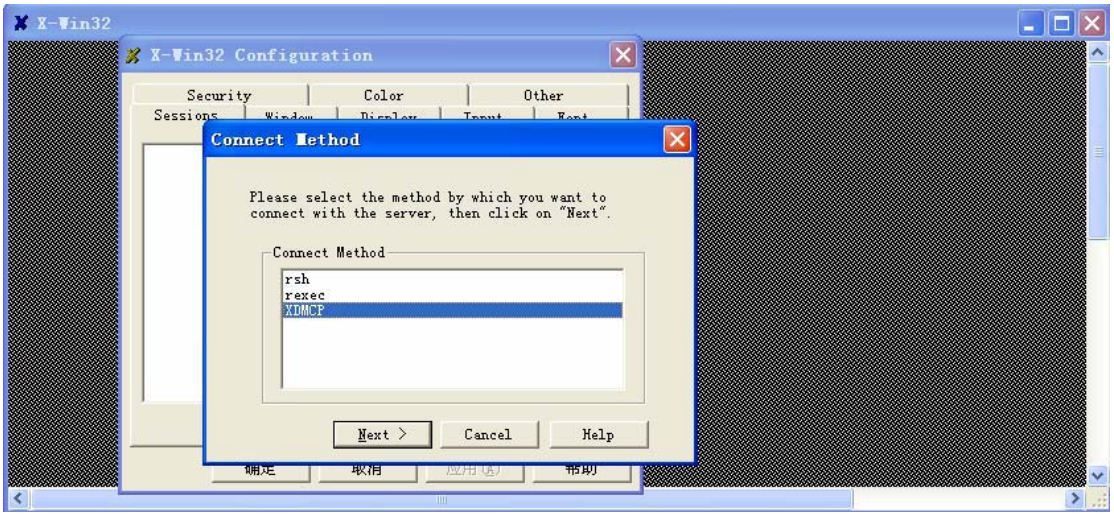
X-Win32
X-Win32 Configuration
Security
Color
Other
Sessions
Connect Method
Please select the method by which you want to connect with the server, then click on "Next".
Connect Method
rsh
rexec
XDMCP
Next >
Cancel
Help

图 1- 9

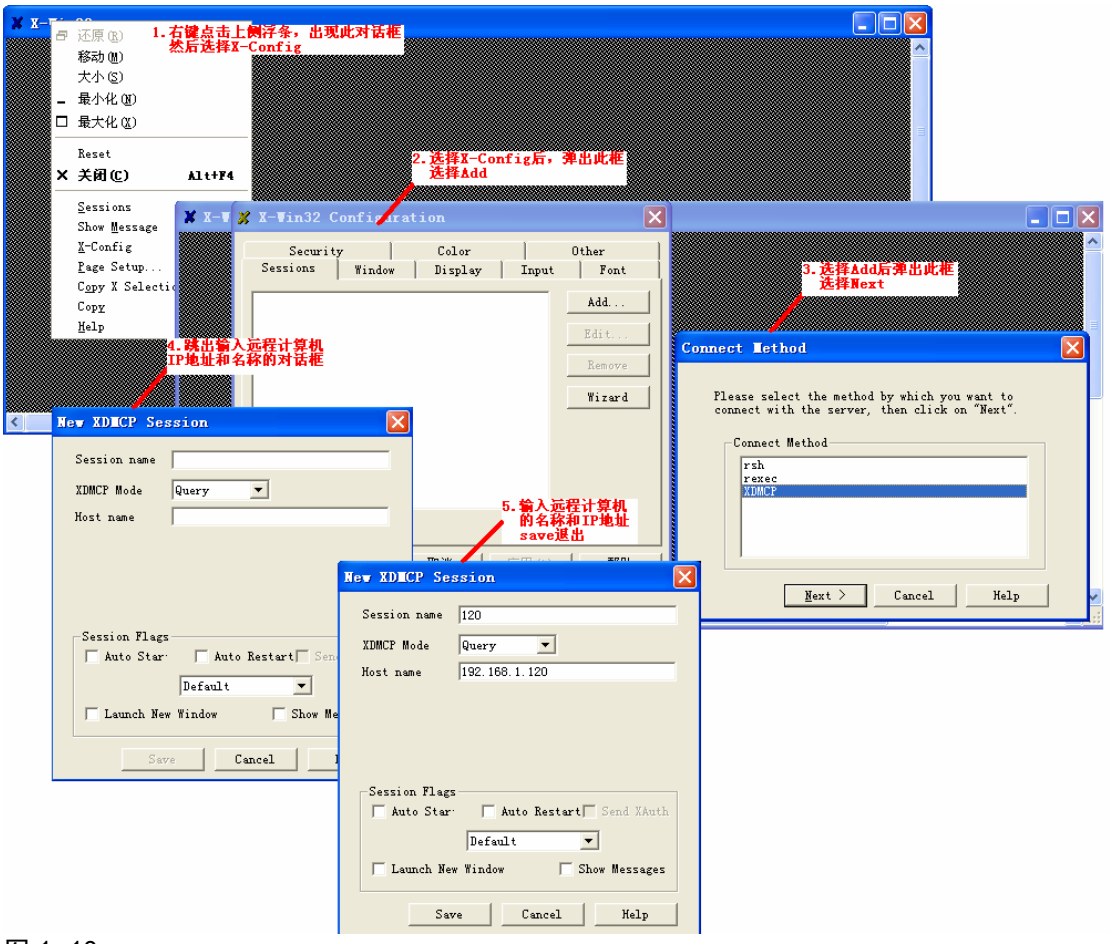
1.右键点击上侧浮条，出现此对话框
然后选择X-Config
还原(R)
移动(M)
大小(S)
最小化(N)
最大化(X)
Reset
关闭(C) Alt+F4
Sessions
Show Message
X-Config
Page Setup...
Copy
Help
2.选择X-Config后，弹出此框
选择Add
X-Win32 Configuration
Security
Color
Other
Sessions
Window
Display
Input
Font
Add...
Edit...
Remove
Wizard
3.选择Add后弹出此框
选择Next
Connect Method
Please select the method by which you want to connect with the server, then click on "Next".
Connect Method
rsh
rexec
XDMCP
Next >
Cancel
Help
4.跳出输入远程计算机
IP地址和名称的对话框
New XDMCP Session
Session name
XDMCP Mode
Query
Host name
Session Flags
Default
Launch New Window
Save
Cancel
5.输入远程计算机
的名称和IP地址
save退出
New XDMCP Session
Session name
120
XDMCP Mode
Query
Host name
192.168.1.120
Session Flags
Auto Restart
Send XAuth
Default
Launch New Window
Show Messages
Save
Cancel
Help

图 1- 10


X-Win32
还原(R)
移动(M)
大小(S)
最小化(N)
最大化(X)
Reset
关闭(C) Alt+F4
Sessions
120
Show Message
X-Config
Page Setup...
Copy X Selection
Copy
Help

图 1- 11

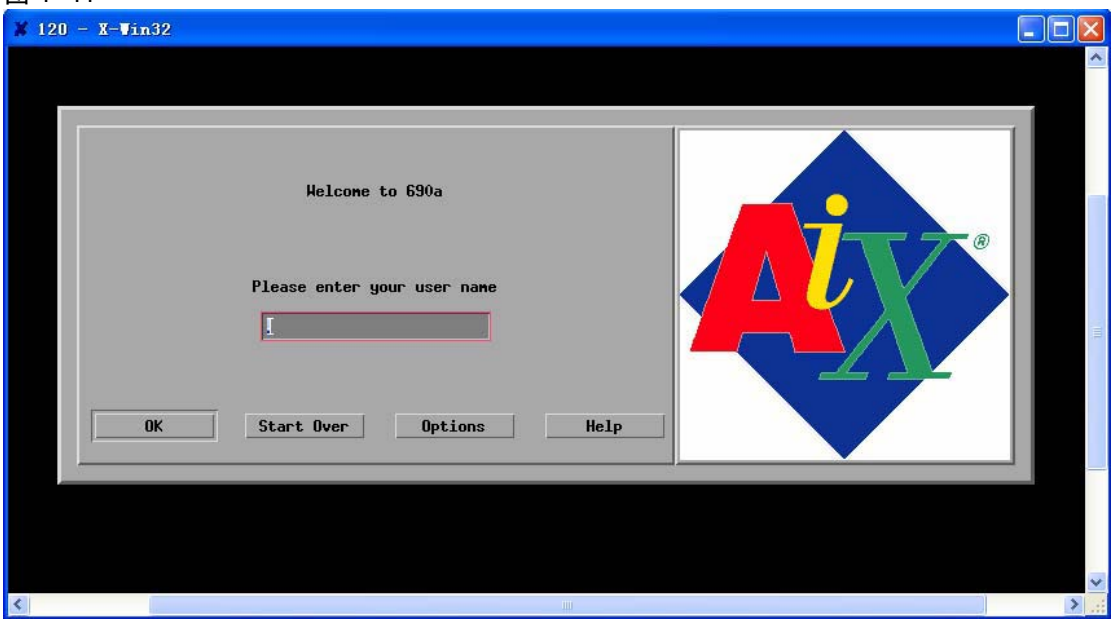
120 - X-Win32
Welcome to 690a
Please enter your user name
OK
Start Over
Options
Help

图 1- 12

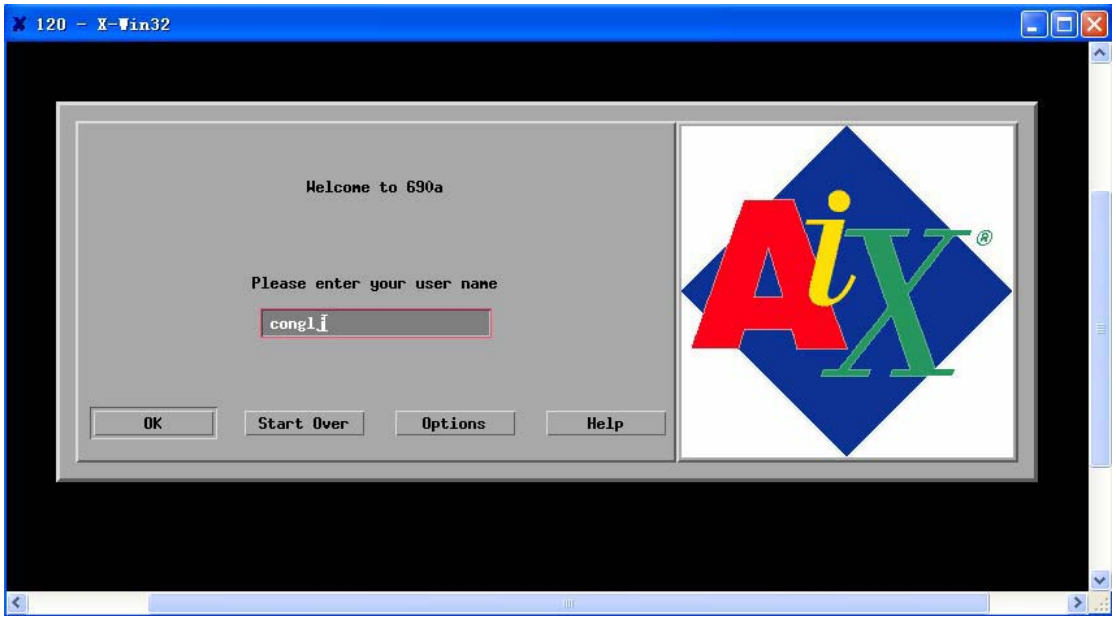

图 1- 13

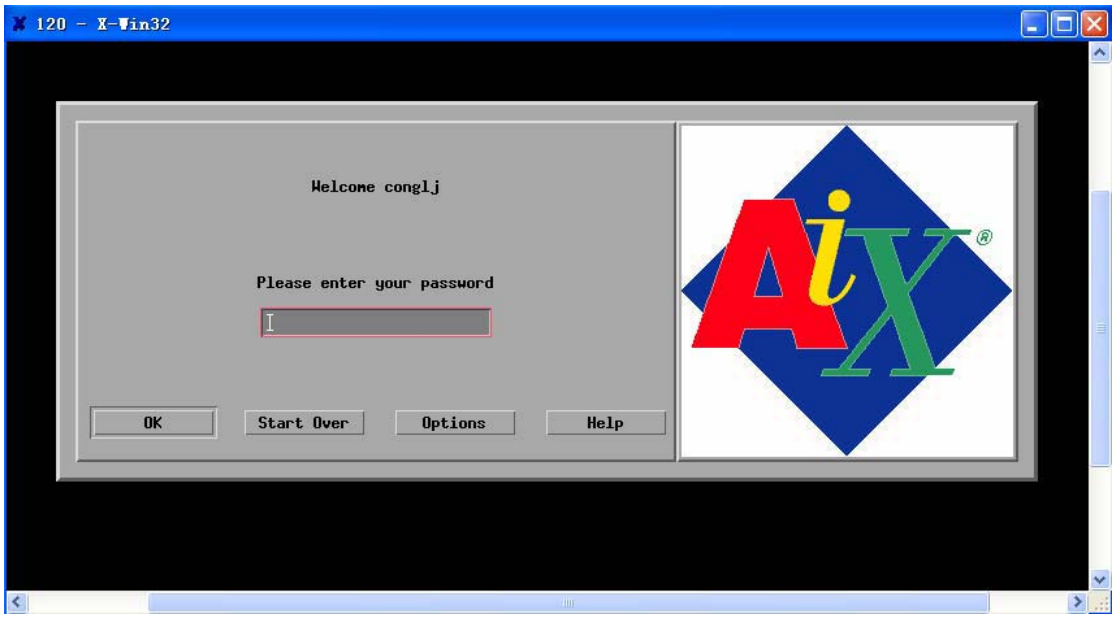

图 1- 14

登陆成功，此处为远程计算机操作界面（不同的计算机有不同的界面）

图 1- 15

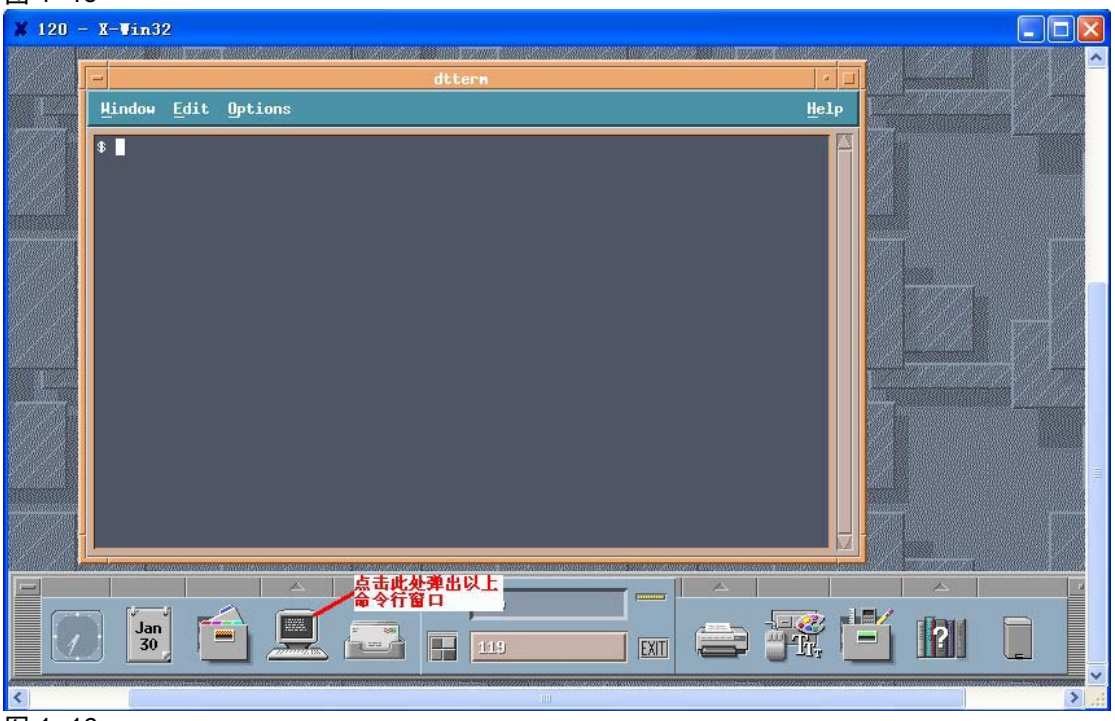

图 1- 16

1.5 软件安装简介

1. 后缀为.rpm 的软件。RPM 全称是 Red Hat Package Manager（Red Hat 包管理器）。;

Rpm 的安装基本命令为：rpm -ivh [software].rpm

RPM 命令主要参数：

-i 安装软件。

-t 测试安装，不是真的安装。

-p 显示安装进度。

-f 忽略任何错误。

-U 升级安装。

-v 检测套件是否正确安装。

卸载软件

rpm -e 软件名

目前 RPM 有两种模式，一种是已经过编码的（i386.rpm），一种是未经编码的（src.rpm）。如果是未经编码的包，需要先运行命令：rpm --rebuild Filename.src.rpm，这时系统会建立一个文件 Filename.rpm，在/usr/src/redflag/RPMS/子目录下，一般是 i386，具体情况和 Linux 发行版本有关。然后执行下面代码即可：rpm -ivh /usr/src/regflag/RPMS/i386/Filename.rpm

2. 后缀为.tar.gz、tar.Z、tar.bz2 或.tgz 是使用 linux/Unix 系统打包工具 tar 打包的解压数据包。首先要解压缩，不同扩展名解压缩命令也不相同，如：

类型 命令
.gz gunzip
.Z uncompress
.zip unzip
.bz2 bunzip2

进入解压缩目录，查看 README/INSTALL，如果有此类文件，安装前阅读，里面会有安装过程。不同的软件安装不尽相同。一般大致过程如下：

./configure 配置

make 调用 make 命令进行编译

make -f file 指定 file 文件为描述文件。如果没有"-f"参数，则系统将默认当前目录下名为 makefile 或者名为 Makefile 的文件为描述文件。

make install 安装可执行程序

make clean 删除安装时产生的临时文件

卸载软件：＃make uninstall

有些软件包的源代码编译安装后可以用 make uninstall 命令卸载。如果不提供此功能，则软件的卸载必须手动删除。

第 2 章 数据的基本处理

2.1 测序原理介绍

2.2 峰图转化 Phred

简介

Phred 是 phred\phrap 软件包的一部分，phred\phrap 软件包由华盛顿大学分子生物技术学院的 Phil Green 和 Brent Ewing 开发，主要用于学术科研活动。Phred 功能是处理测序仪直接生成的色谱图，给出相应的碱基和质量值。不同的测序仪会给出不同的色谱文件，Phred 能够识

别三种格式的色谱文件，SCF，ABI 和预先处理的 ESD 格式。

碱基的测序质量值 Q 和此碱基出错的概率 Pe 相关。公式：

```
Q = -10 log10( Pe )
```

下载

该软件包可以从 phrap 的的网站申请后免费下载，网站链接：http://www.phrap.org/consed/consed.html#howToGet

安装

1、上传 phred 的压缩包到本地 linux/unix 运算服务器；

2、解压缩：

```
gzip -d phred-dist-020425.c-acd.tar.gz
tar -xvf phred-dist-020425.c-acd.tar
```

3、查看解压缩后的文件：

```
bash-2.05b$ ls -l
total 4628
-rw-r--r--    1 bgi   soft      6230 Jul 26  2002 DAEV.DOC
-rw-r--r--    1 bgi   soft      7700 Jul 26  2002 INSTALL
-rw-r--r--    1 bgi   soft      5632 Jul 26  2002 Makefile
-rw-r--r--    1 bgi   soft     60946 Jul 26  2002 PHRED.DOC
-rw-r--r--    1 bgi   soft     84528 Jul 26  2002 qualTableABI3700Prim.h
-rw-r--r--    1 bgi   soft     20834 Jul 26  2002 phred.h
-rw-r--r--    1 bgi   soft      6078 Jul 26  2002 phredData.h
-rw-r--r--    1 bgi   soft      4561 Jul 26  2002 trimPhred.c
-rw-r--r--    1 bgi   soft     21581 Jul 26  2002 trimSeq.c
-rw-r--r--    1 bgi   soft      3976 Jul 26  2002 logFile.c
-rw-r--r--    1 bgi   soft      6987 Jul 26  2002 phred.c
-rw-r--r--    1 bgi   soft      9445 Jul 26  2002 phredpar.dat
… …
```

4、编译源程序：

在命令行键入 make all

敲入“make >& make.log”，完成 phred 的编译。

敲入“make daev”，完成 phred 程序包中 daev 程序的编译。

编译完成后，可将执行文件 phred、daev 拷到/usr/local/genome/bin 目录下面。

默认是用 cc 编译源代码，如果编译报错的话，很可能是 CC 编译器有问题，可以试一下用 gcc 编译，将 Makefile 文件中 CC= cc 改为 CC=gcc 或用命令：make CC=gcc all

5、设置环境变量

为了以后使用方便，可以把 phred 需要的环境变量设置在用户宿主目录下面的.profile 和.bashrc 或.cshrc 文件里面，把配置文件的路径付给 PHRED_PARAMETER_FILE，

例如：

1. C shell ,tcsh:

```
% setenv PHRED_PARAMETER_FILE /usr/local/PhredPar/phredpar.dat
```

2. sh,bash:

```
$ HRED_PARAMETER_FILE=/usr/local/PhredPar/phredpar.dat
```

```
$ export PHRED_PARAMETER_FILE
```

注意路径要根据不同用户安装目录的不同做相应的修改，不能照抄这个例子。

phredpar.dat 文件内容：

```
########################################################################
#                                                                      #
# phredpar.dat - phred parameter file: 980806                          #
#                                                                      #
#   known chemistries: primer, terminator, unknown                     #
#   known dyes       : rhodamine, d-rhodamine, big-dye                 #
#                      energy-transfer, bodipy, unknown                #
#   known machines   : ABI_373_377, MolDyn_MegaBACE,                   #
#                      ABI_3700, LI-COR_4000                           #
#                                                                      #
# Notes:                                                               #
#       (1) enclose the `dye primer' name in double quotes             #
#           and include spaces in the names.                           #
#       (2) leave one or more spaces between the `dye primer'          #
#           and chemistry names, between the chemistry and             #
#           dye names, and between the dye and machine names.          #
#       (3) add entries between the `begin chem_list' and              #
#           `end chem_list' lines.                                     #
#                                                                      #
########################################################################
#
begin chem_list
"DP6%25Ac{-21M13}"              primer          rhodamine               ABI_373_377
"DP6%Ac{-21M13}"                primer          rhodamine               ABI_373_377
"DP6%25Ac{M13Rev}"              primer          rhodamine               ABI_373_377
"DP6%Ac{M13Rev}"                primer          rhodamine               ABI_373_377
"DyePrimer{-21m13}"             primer          rhodamine               ABI_373_377
"DyePrimer{KS}"                 primer          rhodamine               ABI_373_377
"DyePrimer{M13RP1}"             primer          rhodamine               ABI_373_377
"DyePrimer{SK}"                 primer          rhodamine               ABI_373_377
"DyePrimer{SP6}"                primer          rhodamine               ABI_373_377
"DyePrimer{T3}"                 primer          rhodamine               ABI_373_377
"DyePrimer{T7}"                 primer          rhodamine               ABI_373_377
"DyePrimer{-21M13LR}"           primer          rhodamine               ABI_373_377
"DP5%LR{HS}"                    primer          rhodamine               ABI_373_377
"DP4%25Ac{-21M13}"              primer          rhodamine               ABI_373_377
```

图 2- 1 phredpar.dat 文件内容

最后两行：

```
"__no_matching_string__"            unknown         unknown             unknown
end chem_list
```

如果有如下报错信息，说明环境变量还没有设置成功，需要重新设置环境变量：

```
FATAL_ERROR: PHRED_PARAMETER_FILE environment variable not set. type `phred -doc' for
more information
```

使用

程序运行命令行：

```
phred  -id <chromat-file-dir>  -pd <phd-file-dir> [other options]
```

键入 phred -help（-h）查看帮助信息：

```
bash-2.05b$ phred -help
parameter   argument    default     description
-if     <filename>      none        read input filenames from file
-id     <dirname>       none        read input files from <dirname>
-zd     <dirname>       path        uncompress program path
-zt     <dirname>       /usr/tmp    uncompress temporary directory
-st         <type>      fasta       sequence file type (fasta|xbap)
-s      none            nofile      write *.seq sequence file(s)
-s      <filename>      nofile      write <filename> sequence file
-sa     <filename>      none        append sequence files to <filename>
```

```
-sd     <dirname>       nofile      write *.seq file(s) to <dirname>
-qt         <type>      fasta       quality file type (fasta|xbap|mix)
-q          none        nofile      write *.qual quality file(s)
-q      <filename>      nofile      write <filename> quality file
-qa     <filename>      none        append quality files to <filename>
-qd     <dirname>       nofile      write *.qual file(s) to <dirname>
-qr     <filename>      nofile      write quality report to <filename>
-p          none        nofile      write *.phd.1 file(s)
-p      <filename>      nofile      write <filename> phd file
-pd     <dirname>       nofile      write *.phd.1 file(s) to <dirname>
-cv     <version>       2           SCF format version (2 or 3)
-cp     <precision>     maxval      SCF data precision in bytes (1 or 2)
-cs          none       no scale    always scale traces in SCF files
-c           none       nofile      write * phred SCF file(s)
-c      <filename>      nofile      write <filename> phred SCF file
-cd     <dirname>       nofile      write * SCF file(s) to <dirname>
-d           none       nofile      write *.poly poly file(s)
-d      <filename>      nofile      write <filename> poly file
-dd     <dirname>       nofile      write *.poly file(s) to <dirname>
-raw    <seq name>      NULL        seq name written in output files
-log                    nolog       write phred.log file
-nocall      none       call        disable basecalling
-trim   <enzyme seq>    notrim      enable auto trim
-trim_alt   <enzyme seq>    notrim  enable alternate auto trim
-trim_cutoff    <n>     0.05        trim_alt error probability
-trim_fasta none        none        trim FASTA bases and qual. values
-trim_scf   none        none        trim SCF bases and qual. values
-trim_phd   none        none        trim base call data in phd files
-trim_out   none        none        trim data in most output files
-nonorm     none        normalize   disable trace normalization
-nosplit    none        none        no compressed peak splitting
-nocmpqv    none        none        no compressed peak quality values
-ceilqv     <ceiling qv>    none    quality value ceiling value
-beg_pred   <point>     none        set peak prediction start point
-v          <n>         none        verbose operation <n> = 1 to 63
-tags       none        not tags    label common messages with tags
-V          none        none        show version
-help       none        none        help
-h          none        none        help
-doc        none        none        show phred documentation
 For the warning messages `unable to identify chemistry and dye' and `unknown chemistry
(...) in chromat ...' please read the phred documentation using the command `phred
-doc'.
    no input files specified
```

输入

测序仪产生的峰图文件，可识别：SCF, ABI model 373 and 377 DNA sequencer chromatogram, and MegaBACE ESD chromatograms files

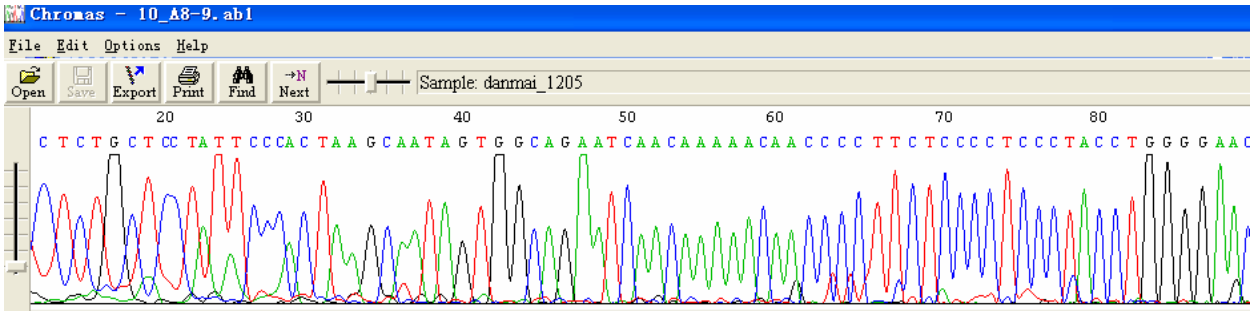

图 2- 2 峰图

输出

运行过程中的屏幕输出：

```
  chromat_dir/10_A8-9.ab1
  chromat_dir/11_A8-9_R.ab1
  chromat_dir/15_A8-9.ab1
  chromat_dir/21_A8-9.ab1
  chromat_dir/22_A8-9.ab1
  chromat_dir/23_A8-9.ab1
```

Warn 输出：

```
Chromat_dir/10_A8-9.ab1
unknown chemistry (KB_3730_POP7_BDTv3.mob) in chromat tmp/10_A8-9.ab1 add a line of
the form
   "KB_3730_POP7_BDTv3.mob"    <chemistry>     <dye type>     <machine type>
   to the file phredpar.dat type `phred -doc' for more information
```

程序的输出结果是文件输出，格式可以是 FASTA 格式, 也可以是 XBAP, PHD 格式或 SCF 格式。

1. Phd 文件，用于组装后 consed 查看编辑，名字为<filename>.phd.1，

```
bash-2.05b$ ls -l phd_dir/
total 44
-rw-r--r--    1 bgi soft      3040 Dec 20 06:58 23_A8-9.ab1.phd.1
-rw-r--r--    1 bgi soft      6996 Dec 20 06:58 22_A8-9.ab1.phd.1
-rw-r--r--    1 bgi soft      7013 Dec 20 06:58 21_A8-9.ab1.phd.1
-rw-r--r--    1 bgi soft      7026 Dec 20 06:58 15_A8-9.ab1.phd.1
-rw-r--r--    1 bgi soft      6908 Dec 20 06:58 11_A8-9_R.ab1.phd.1
-rw-r--r--    1 bgi soft      7041 Dec 20 06:58 10_A8-9.ab1.phd.1
```

2. Fasta 格式的核酸序列文件

FASTA 头注释行包含修饰信息（序列没有影响），此行有如下格式：

a. 序列名称

b. phred 读出的碱基数

c. 序列开始部分被修饰掉的碱基数

d. 修饰后余下的碱基数

e. 描述输入文件类型

```
>23_A8-9.ab1    289     0    289  ABI
GCATGGGATTCCGATCAGGATGATCTTCAGAGACTGTCTCAGATTAGACT
CAAGAGCCCTCAGAGGTACTGTGACTTTTTATGGGGGTGGGGGTGGGGGG
TTATTGCCTTCTCTCCAGGATGAAGATGGGAAGAAGTTGTCCCATCCACT
CCCTCTCAGCGCACCCGGACACCTTTAGGTTTGCCCGGCGAGACGCGCCA
CCTGGTGGCTAGGGTGCGTGGCTAGGGGGACACCGGATCCCAGGACAGAC
CGTGTGCTGCGCCTGTCATGGCCTGGGGGGCAGCCCCGC
```

3. Fasta 格式的质量文件(和序列文件相对应，给出每个碱基的质量值)

```
>23_A8-9.ab1    289     0    289  ABI
4 4 4 4 7 9 8 8 7 7 7 10 13 6 6 6 7 6 6 6 6 8 8 14
18 13 15 8 8 10 9 16 18 29 29 29 37 37 30 27 17 22
24 29 29 18 18 25 27 31 22 22 16 14 19 16 19 30 33
23 11 10 10 19 33 46 42 42 42 42 40 40 40 40 40 40
… …
9 9 9 11 13 6 7 8 8 8 9 10 9 9 7 7 10 11 9 7 7 9 9
9 10 9 11 11 11 8 8 9 10 7 8 8 9 9 10 10 9 8 9 8 10
10 20 10 11 9 9 14 12 14 11 11 11 11 11 11 9 9 8 7
11 9 12 13 10
```

参数

详细的参数列表及说明可以通过键入 phred –doc 查看：

```
bash-2.05b$phred -doc
```

输入选项：

-id　　输入文件目录

运行选项：

-nocall　　关闭 phred 碱基读取而使用 ABI 碱基读取，默认为采用 phred

```
-trim            修饰当前序列
-trim_alt        Perform sequence trimming on the current sequence.
-trim_cutoff     Set trimming error probability for the `-trim_alt' option and the
                 trimming points written in the phd files. 默认值 0.05.
-trim_fasta      修饰序列写入到 FASTA 文件 ，FASTA 注释信息行显示序列的高质量信息
-trim_scf        修饰序列、质量值、碱基位置写入到 SCF 文件。
-trim_phd        修饰序列、质量值、碱基位置写入到 PHD 文件
-trim_out        FASTA、SCF、PHD 的输出， `-trim_fasta', `-trim_scf',`-trim_phd' 参数
                 结合
输出选项
-st fasta        输出 FASTA 文件（默认）
-s               输出文件加后缀".seq"
-sd              输出序列文件到特定目录
-sa              Write a sequence output file in FASTA format with the name .
-qt fasta        Set the output quality file format to FASTA. (Default.) Trimming
                 options affect the FASTA file; see the Notes below for more
                 information.
-qt xbap         Set the output quality file format to XBAP.
-qt mix          Set the output quality file format to FASTA.
-q               Write quality output files with the names obtained by appending
                 ".qual" to the names of the input files, and store them in the directory
                 where phred is running. This option is valid for FASTA format output
                 files only.
-q               输出质量文件，仅当只有一个文件输入时有效
-qd              输出".qual"质量文件，并存储在目录中
-qa              Write a quality output file in FASTA format with the name.
-qr              Write a histogram of the number of high quality bases per read.  This
                 is meaning-ful when phred processes more than one read.
-c               输出 SCF 文件，包含序列、碱基位置。
-cd              输出 SCF 文件到特定目录下
-cp              Store SCF trace data as 1 or 2 byte values.
-cv              以 Version2 或 Version3 的格式输出 SCF 文件，默认为 2。
-cs              Always scale traces before writing them to an SCF output file.
-p               输出 PHD 文件。
-pd              输出 PHD 文件到特定目录。
-d               Write a data file that is used for detecting polymorphic bases.
-dd              Write polymorphism data files in directory.
-raw             Write  in the header of the sequence output file and the quality output
                 file.
-log             程序运行日志"phred.log"。
Miscellaneous
-h, -help        显示命令行基本参数。
-doc             显示 phred 文档。
-V               显示 phred 版本。
```

在线帮助文档： http://www.phrap.org/phredphrap/phred.html

DAEV 简介：

```
bash-2.05b$ daev -h
option      argument      default    description
------     --------      -------    -----------
-cutoff     <cutoff QV>     20      set high quality value cutoff
-phd_hq     none            none    print HQ base count for each file
-no_stats   none            none    print only HQ base count for each file
-V          none            none    show version
-help       none            none    help
-h          none            none    help
missing phd directory name

usage: daev <option(s)> <phd_dirname>
```

```
(use -h for help summary)
For example:
bash-2.05b$ daev -cutoff 20 phd_dir
```

结果为标准屏幕输出：

```
daev version: 0.020426.c
command line: ./daev -cutoff 20 phd_dir
time:        061220:091443

High Quality (QV >= 20) Bases Per Read
======================================
num HQ base    num read  (%)    rcum read (%)
------------      -------------       -------------
    50-99     1  (16.7)       6 (100.0)   |XXXXXXXX
   550-599    2  (33.3)       5  (83.3)   |XXXXXXXXXXXXXXXX
   600-649    3  (50.0)       3  (50.0)   |XXXXXXXXXXXXXXXXXXXXXXXX

mean number of high quality bases per read:  522.3

Trimmed Read Length Distribution
================================
    length      num read  (%)    rcum read (%)
 ------------   -------------    -------------
   100-149      1  (16.7)    6 (100.0)   |XXXXXXXX
   550-599      1  (16.7)    5  (83.3)   |XXXXXXXX
   600-649      4 (66.7)     4  (66.7)   |XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

mean number of 'trimmed bases' per read:  533.2
Quality Value Distribution
==========================
quality val.      num base  (%)    rcum base (%)
 ------------      -------------    -------------
    0-4             41   (1.1)     3690 (100.0)   |X
    5-9            256   (6.9)     3649  (98.9)   |XXX
   10-14           180   (4.9)     3393  (92.0)   |XX
   15-19            79   (2.1)     3213  (87.1)   |X
   20-24            56   (1.5)     3134  (84.9)   |X
   25-29            69   (1.9)     3078  (83.4)   |X
   30-34            89   (2.4)     3009  (81.5)   |X
   35-39            94   (2.5)     2920  (79.1)   |X
   40-44           434  (11.8)     2826  (76.6)   |XXXXXX
   45-49           227   (6.2)     2392  (64.8)   |XXX
   50-54           510  (13.8)     2165  (58.7)   |XXXXXXX
   55-59          1655  (44.9)     1655  (44.9)   |XXXXXXXXXXXXXXXXXXXXXX

Read Length Distribution
========================
length        num read (%)   rcum read (%)
-----------   ------------   -------------
250-299         1  (16.7)    6 (100.0)   |XXXXXXXX
650-699         5  (83.3)    5 (83.3)    |XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
mean number of bases per read:  615.0
```

实例

峰图文件输入请见光盘：Phred\chromat_dir

```
%ls –l chromat_dir/
total 1228
-rw-r--r--    1 soft    bgi       137332 Dec 20 06:43 23_A8-9.ab1
-rw-r--r--    1 soft    bgi       254559 Dec 20 06:43 22_A8-9.ab1
-rw-r--r--    1 soft    bgi       185602 Dec 20 06:43 21_A8-9.ab1
-rw-r--r--    1 soft    bgi       254615 Dec 20 06:43 15_A8-9.ab1
-rw-r--r--    1 soft    bgi       184235 Dec 20 06:43 11_A8-9_R.ab1
-rw-r--r--    1 soft    bgi       185858 Dec 20 06:43 10_A8-9.ab1
```

建立输出文件目录 phd_dir，命令%mkdir phd_dir

运行命令(在 chromat_dir 上级目录下运行)：%phred -id chromat_dir –pd phd_dir -trim_cutoff 0.01

查看输出结果%ls –l phd_dir/

```
ls -l phd_dir/
total 48
-rw-r--r--    1 soft    bgi        3040 Dec 20 06:58 23_A8-9.ab1.phd.1
-rw-r--r--    1 soft    bgi        6996 Dec 20 06:58 22_A8-9.ab1.phd.1
-rw-r--r--    1 soft    bgi        7013 Dec 20 06:58 21_A8-9.ab1.phd.1
-rw-r--r--    1 soft    bgi        7026 Dec 20 06:58 15_A8-9.ab1.phd.1
-rw-r--r--    1 soft    bgi        6908 Dec 20 06:58 11_A8-9_R.ab1.phd.1
-rw-r--r--    1 soft    bgi        7041 Dec 20 06:58 10_A8-9.ab1.phd.1
```

练习

如实验室有测序项目，请对测序峰图结果进行分析。

参考文献

1. Ewing B, Green P: Basecalling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8:186-194 (1998).
2. Ewing B, Hillier L, Wendl M, Green P: Basecalling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 8:175-185 (1998).

2.3 Phd2Fasta

简介

Phd2fasta 是 phred\phrap 软件包的一部分，phred\phrap 软件包由华盛顿大学分子生物技术学院的 Phil Green 和 Brent Ewing 开发，主要用于学术科研活动。Phd2fasta 将 phred 产生的 phd 文件转换为 fasta 格式的核酸和质量文件，便于 crossmatch 和 phrap 程序应用。

下载

该软件包可以从 phrap 的的网站申请，申请通过后邮件发送，申请链接：http://www.phrap.org/consed/consed.html#howToGet

安装

1、上传 phd2fasta 的压缩包到本地 linux/unix 运算服务器；

2、解压缩：

```
gzip -d phd2fasta-acd-dist.tar.gz
tar -xvf phd2fasta-acd-dist.tar
```

3、编译源程序：

在命令行键入 make，编译完成后，可将执行文件 phd2fasta 拷到统一的可执行程序目录，如：/usr/local/genome/bin 下面，源文件包可删除。

编译成功无提示信息。

使用

程序运行命令行：

```
phd2fasta -id phd_dir -os out.fas -oq out.fas.qual
```

直接键入 phd2fasta 的屏幕提示：

```
no input files specified
no output file specified
```

使用帮助可通过命令 phd2fasta -h 获取：

```
> phd2fasta -h

parameter   argument       default    description
---------   --------       -------    -----------
-if         <filename>     none       read input filenames from file
-id         <dirname>      none       read input files from <dirname>
-ix         <filename>     none       read exclude filenames from file
-is         none           none       read filenames from stdin
-os         <filename>     none       sequence output filename
-oq         <filename>     none       quality output filename
-ob         <filename>     none       base position output file
-oe         <filename>     none       write edit file
-of         <filename>     none       write failure log
-mask       <type>         none       mask vector (types: vector
                                                          sequencing
                                                          cloning
                                                          all)
-halt       none           none       exit on file read/process error

-verbose    none           none       Display some processing
                                      information.
-V          none           none       show version
-help       none           none       help
-h          none           none       help
-doc        none           none       show documentation

no input files specified
no output file specified
```

图 2- 3 phd2fasta 的帮助信息

输入

此程序的输入文件为 Phred 产生的 phd 文件：

```
bash-2.05b$ ls -l phd_dir/
total 44
-rw-r--r--    1 bgi    soft      3040 Dec 20 06:58 23_A8-9.ab1.phd.1
-rw-r--r--    1 bgi    soft      6996 Dec 20 06:58 22_A8-9.ab1.phd.1
-rw-r--r--    1 bgi    soft      7013 Dec 20 06:58 21_A8-9.ab1.phd.1
-rw-r--r--    1 bgi    soft      7026 Dec 20 06:58 15_A8-9.ab1.phd.1
-rw-r--r--    1 bgi    soft      6908 Dec 20 06:58 11_A8-9_R.ab1.phd.1
-rw-r--r--    1 bgi    soft      7041 Dec 20 06:58 10_A8-9.ab1.phd.1
```

输出

此程序运行无屏幕输出，结果为 fasta 格式的序列和质量文件。

序列文件：

```
bash-2.05b$ more out.fas
>10_A8-9.ab1 CHROMAT_FILE: 10_A8-9.ab1 PHD_FILE: 10_A8-9.ab1.phd.1 CHEM: unknown DYE:
unknown TIME: Wed Dec 20 06:58:47 2006
gtgctctggtctctgctcctttcccctaagcaatagtaggcagaatcaac
aaaaacaacccctttctcccctccctacctggggaacagagccaatgagac
aggctcaggaacagggcaccagcacctgcactcaccattcaatctcttta
... ...
acagtctctgaagattcatcctctttccagaaacccaagcccatcttgct
ctcctagaaacctttctataaaaaaaaaaaaan
>11_A8-9_R.ab1  CHROMAT_FILE:  11_A8-9_R.ab1  PHD_FILE:  11_A8-9_R.ab1.phd.1  CHEM:
unknown DYE: unknown TIME: Wed Dec 20 06:58:47 2006
```

```
gggagaggcggagctctggtccttgtcatctaagctgtgtggattgatcg
cctagaacctccctatctaccctccctacctggggaacagagccaatgag
aaaggctcaggaacagggcaccagcacctgcactcaccattcaatctctt
tcaccctcaaacataaaggtgtcagcttctgctcttatgtcctcatcgga
agacagtctctgaagattcatcctctttccagaaacccaagcccatcttg
ctctccagaacccttcttaaa
... ...
```

质量文件：

```
bash-2.05b$ more out.fas.qual
>10_A8-9.ab1 PHD_FILE: 10_A8-9.ab1.phd.1
15 16 15 15 13 20 20 29 40 33 33 32 32 19 13 4 4 4
15 24 32 32 34 34 34 34 40 46 46 46 46 51 51 46 46
... ...
47 42 42 44 44 47 56 56 56 56 56 56 56 40 40 40 40
17 23 18 14 11 8 9 8 12 4 0
>11_A8-9_R.ab1 PHD_FILE: 11_A8-9_R.ab1.phd.1
11 9 13 11 14 13 16 19 13 10 10 11 12 12 14 10 11 10
... ...
56 56 56 56 56 56 56 56 56 56 56 56 51 51 51 51 43 56
56 56 56 56 42 42 42 42 42 56 56 56 56 56 56 56 56 56
56 56 56 56 56 56 51 56 48 48 42 42 44 48 44 56 56
9 10 20 29 32 27 19 6 6 8 9 9
... ...
```

参数

详细的参数说明可以通过键入 phd2fasta -doc 查看：

```
bash-2.05b$ phd2fasta -doc
Input Options
-------------
-id <directory name>    读取目录中的的文件做为输入文件
-if <file name>             读取文件列表中的文件做为输入文件
-is                         读取标准输入做为输入文件
-ix <file name>             读取不需运行的文件列表
Output Options
--------------
-os    <file name>          输出 FASTA 序列到文件.
-oq    <file name>          输出 FASTA 序列的质量到文件
-ob    <file name>          输出序列碱基位置信息到文件
-oe    <file name>          将 phd 文件中的编辑信息提取到文件中
-of    <file name>          phd2fasta 处理失败的文件写入日志文件
Processing Options
------------------
-mask <type>                用 x 屏蔽序列中的载体
-halt                       如有错误则停止继续运行程序
Misc
-verbose                    显示进程信息
-V                          显示 phd2fasta 版本
-h, -help                   显示命令行参数列表
-doc                        显示详细的帮助文档
```

实例

以上一节 phred 中的 phd_dir 里面的文件为输入，光盘 phred\phd_dir

运行命令 phd2fasta –id phd_dir –os out.fas –oq out.fas.qual

输出文件为序列文件 out.fas 和质量文件 out.fas.qual。

％more out.fas

```
>10_A8-9.ab1 CHROMAT_FILE: 10_A8-9.ab1 PHD_FILE: 10_A8-9.ab1.phd.1 CHEM: unknown DYE:
unknown TIME: Wed Dec 20 06:58:47 2006
gtgctctggtctctgctcctttcccctaagcaatagtaggcagaatcaac
```

```
aaaaacaacccctttctcccctccctacctggggaacagagccaatgagac
aggctcaggaacagggcaccagcacctgcactcaccattcaatctcttta
ggctcacggtccttcagaagctcttgtacctcctgccgacagcgctcctg
gtattccgggtgctttgcaaggtggtacaggacccaggagagaccactgg
ctgtggtgtcatggcctgggggggcagcaaggcaggcttgggtctctgggc
tgcttcagcacccgagggtgtacagcaaccttgcattgaggacctcaggg
aggatgggggaaggggatgggaagtgcgaggggtccacccaccctgttc
ctggaatggagatatccaagtccccactctagccccacactggggccctc
accctcaaacataaaggtgtcagcttctgctcttatgtcctcatcggaca
acttcttcccatcttcatcctggagagaaggcaataacccccccaccccca
cccccataaaaagtcacagtacctctgagggctcttgagtctaatctgag
acagtctctgaagattcatcctctttccagaaacccaagcccatcttgct
ctcctagaaacctttctataaaaaaaaaaaaan
……
>23_A8-9.ab1 CHROMAT_FILE: 23_A8-9.ab1 PHD_FILE: 23_A8-9.ab1.phd.1 CHEM: unknown DYE:
unknown TIME: Wed Dec 20 06:58:47 2006
gcatgggattccgatcaggatgatcttcagagactgtctcagattagact
caagagccctcagaggtactgtgacttttatgggggtgggggtgggggg
ttattgccttctctccaggatgaagatgggaagaagttgtcccatccact
ccctctcagcgcacccggacacctttaggtttgcccggcgagacgcgcca
cctggtggctagggtgcgtggctaggggggacaccggatcccaggacagac
cgtgtgctgcgcctgtcatggcctgggggggcagccccgc
```

练习

对实验室测序数据进行峰图转换。

2.4 载体屏蔽 Crossmatch

简介

Phil Green 和 Brent Ewing 开发的 phrap 软件包的一部分，用于比对两套 DNA 序列，如：可以用来找出序列中的载体序列，并产生屏蔽了载体的序列；也可以用于 cDNA 和 cosmid 的比对等。和 blastn 相比速度较慢但敏感度较高（因其允许 gap 存在）

下载

包含在 Phrap 软件包中，Mail to phg@u.washington.edu

安装

1、上传 phrap 的压缩包到本地 linux/unix 运算服务器；

2、解压缩：

```
gzip -d phrap.tar.gz
tar -xvf phrap.tar
```

3、编译源程序：

在命令行键入 make，如果数据集多于 64,000 条序列，或者序列中含有长于 64,000 bp 的序列，则需要使用 cross_match.manyreads 或 cross_match.longreads，这两个程序编译命令为 make manyreads。

使用

命令行：cross_match seq_file1 seq_file2 -minmatch 10 -minscore 20 -screen > screen.out

输入

标准 FASTA 格式的序列文件

参数

```
option name & default value
1. 比对分值控制参数
-penalty -2             Mismatch (substitution) penalty for SWAT comparisons.
-gap_init penalty-2     Gap initiation penalty for SWAT comparisons.
-gap_ext penalty-1      Gap extension penalty for SWAT comparisons.
-ins_gap_ext gap_ext    Insertion gap extension penalty for SWAT comparisons
                        (insertion in subject relative to query).
-del_gap_ext gap_ext    Deletion gap extension penalty for SWAT comparisons (deletion
                        in subject relative to query).
-matrix [None]          Score matrix for SWAT comparisons
-raw *                  Use raw rather than complexity-adjusted Smith-Waterman
                        scores.
2. Banded search
-minmatch 14            Minimum length of matching word to nucleate SWAT comparison.
-maxmatch 30            Maximum length of matching word. For cross_match, the default
                        value is equal to minmatch, instead of 30.
-max_group_size 20      Group size (query file, forward strand words)
-word_raw *             Use raw rather than complexity-adjusted word length, in
                        testing against minmatch
-bandwidth 14           1/2 band width for banded SWAT searches (full width is 2
                        timesbandwidth + 1).
3. 比对筛选
-minscore 30            最小比对分值
-vector_bound 80        序列开头载体的可能碱基数目，默认值为 0 到 80。
-masklevel 80           A match is reported only if at least (100 - masklevel)% of the
                        bases in its "domain" (the part of the query that is aligned)
                        are not contained within the domain of any higher-scoring
                        match.
Special cases:
-masklevel 0            report only the single highest scoring match for each query
-masklevel 100          report any match whose domain is not completely contained
                        within a higher scoring match
-masklevel 101          report all matches
4. 输入相关参数
-default_qual 15        当没有质量文件存在时，设定的每个碱基的质量值，默认为 15
5. 输出相关参数
-tags *                 在标准输出时标记比对的被选行
-screen *               产生".screen" 文件。FASTA 格式，第一输入文件中的序列被第二个
                        文件比上的部分用 x 替代
-alignments *           显示比对情况
-discrep_lists *        显示比对的差异
-discrep_tables *       给出每个比对差异的统计表格
6. 其他
-indexwordsize 10       用于索引的字符数，此参数影响运行时间和内存使用
```

输出

1. *.log files，程序运行日志

2. *.screen 文件,被屏蔽了相应序列后的序列文件，FASTA 格式。(此文件仅当使用-screen 参数时输出)。

3. 标准屏幕输出，可重定向到文件，如>screen.out：

```
$bash more screen.out
cross_match out.fas A8-9 -screen
cross_match version 0.990329

Run date:time  070122:085518
Query file(s):  out.fas
Subject file(s):   A8-9
Presumed sequence type: DNA

Pairwise comparison algorithm: banded Smith-Waterman

Score matrix (set by value of penalty: -2)
    A   C   G   T   N   X
A   1  -2  -2  -2   0  -3
C  -2   1  -2  -2   0  -3
G  -2  -2   1  -2   0  -3
T  -2  -2  -2   1   0  -3
N   0   0   0   0   0   0
X  -3  -3  -3  -3   0  -3

Gap penalties: gap_init: -4, gap_ext: -3, ins_gap_ext: -3, del_gap_ext: -3,
Using complexity-adjusted scores. Assumed background frequencies:
 A: 0.250  C: 0.250  G: 0.250  T: 0.250  N: 0.000  X: 0.000

minmatch: 14, maxmatch: 14, max_group_size: 20, minscore: 30, bandwidth: 14,
indexwordsize: 10
vector_bound: 0
word_raw: 0
masklevel: 80

Sequence file: out.fas    6 entries
Residue counts:
  a     858
  c    1112
  g     914
  n       6
  t     800
Total   3690

Quality file: out.fas.qual

Input quality (quality, n_residues, %,  cum, cum %,  cum expected errs):
 56    1655  44.9    1655  44.9    0.00
 51     362   9.8    2017  54.7    0.01
 50     148   4.0    2165  58.7    0.01
 48      34   0.9    2199  59.6    0.01
 47      63   1.7    2262  61.3    0.01
 46      82   2.2    2344  63.5    0.01
 45      48   1.3    2392  64.8    0.01
 44     148   4.0    2540  68.8    0.02
 43      57   1.5    2597  70.4    0.02
 42     175   4.7    2772  75.1    0.03
… …
  9      90   2.4    3483  94.4   27.31
  8      52   1.4    3535  95.8   35.55
  7      35   0.9    3570  96.7   42.53
  6      79   2.1    3649  98.9   62.37
  4      35   0.9    3684  99.8   76.31
  0       1   0.0    3685  99.9   77.31
 -1       5   0.1    3690 100.0   82.31   (quality -1 = terminal quality 0)

Avg. full length: 615.0, trimmed (qual > -1): 614.2
Avg. quality: 44.0 per base
Maximal single base matches (low complexity regions):

378  0.26 0.00 0.00  10_A8-9.ab1     136   519 (164)  C A8-9   (0)   384    1
378  0.26 0.00 0.00  11_A8-9_R.ab1     138   521 (150)  C A8-9   (0)   384    1
381  0.00 0.00 0.00  15_A8-9.ab1     143   526 (160)  C A8-9   (0)   384    1
```

```
378  0.26 0.00 0.00  21_A8-9.ab1      136   519 (162)  C A8-9   (0)   384     1
377  0.26 0.00 0.00  22_A8-9.ab1      140   523 (157)  C A8-9   (0)   384     1

5 matching entries (first file).
Discrepancy summary:
Qual algn  cum    rcum    (%)    unalgn X    N  sub del ins  total (%)   cum  rcum (%)
56   1315   1315   1920 (100.00)     0  0    0   1   0   0     1 (0.08)    1    4 (0.21)
51    260   1575    605 ( 31.51)     0  0    0   1   0   0     1 (0.38)    2    3 (0.50)
50    103   1678    345 ( 17.97)     0  0    0   0   0   0     0 (0.00)    2    2 (0.58)
48      9   1687    242 ( 12.60)     0  0    0   0   0   0     0 (0.00)    2    2 (0.83)
47      7   1694    233 ( 12.14)     0  0    0   0   0   0     0 (0.00)    2    2 (0.86)
… …
 4      0   1920      0 (  0.00)     0  0    0   0   0   0     0 (0.00)    4    0 (0.00)
 0      0   1920      0 (  0.00)     0  0    0   0   0   0     0 (0.00)    4    0 (0.00)
-1      0   1920      0 (  0.00)     0  0    0   0   0   0     0 (0.00)    4    0 (0.00)

Screened sequences written to  out.fas.screen
```

Query 序列（第一个输入文件）和 subject 序列（第二个输入文件）比对的情况，如果只有一个输入文件，则是这个文件中任意两个序列的比对情况。比对情况通过命令行的-minscore 和 -masklevel 参数控制，另外也受比对分值和 band search 的参数控制。报告按 query 序列顺序输出，例如：

```
440  2.38 1.39 0.79  hh44a1.s1     33   536 (    0)  C 00311   ( 3084)  8277   7771  *
```

对各列阐述如下:

```
440  = smith-waterman 比对分值
2.38 = 比对部分的替换百分比
1.39 = 比对部分的删除百分比
0.79 = 比对部分的插入百分比
hh44a1.s1 = 第一个输入序列的名称
33  = 第一个输入序列比对起始位点
536 = 第一个输入序列比对终止位点
(0) = no. of bases in 1st sequence past the ending position of match
        (so 0 means that the match extended all the way to the end of the 1st sequence)
C 00311: 和输入序列 00311 的互补链比对上
( 3084): 第二个输入序列（互补链）比对开始前共有 3084 个碱基
8277 = 第一个输入序列比对起始位点
7771 =  第一个输入序列比对起始位点
* indicates that there is a higher-scoring match whose domain partly includes the domain
of this match.

Discrepancy summary:
Qual algn  cum    rcum    (%)    unalgn X    N  sub del ins  total (%)   cum  rcum (%)
56   1315   1315   1920 (100.00)     0  0    0   1   0   0     1 (0.08)    1    4 (0.21)
Qual 质量值
Algn 第一个输入序列这个质量值的碱基数
Cum 在 SWAT 比对中比上的碱基数
Rcum 累计比对上的碱基数（包含这个质量即更高的质量）
Unalgn 没有被包含进来的比对部分碱基数
每种类型的不一致的数目（sub 替换、del 删除、ins 插入）
cum(%)差异的总数和百分比
rcum(%)累计差异数和百分比
```

实例

对文件 reads.dat 进行载体屏蔽，所用载体为 puc18。

1. 输入文件 1，需要进行处理的序列：%more reads.dat

```
>gbeod0_000332.z1.scf
taagactaaggatcccgggtacgagctcgaatcaatagcttccttaacct
tctcattaatatttacttttttcaacaatatactcgaaaggtgtatatcgt
```

```
cttttaacttccttttcaacatatacatccaaaggattaattcggctaaa
attacttactacatcatcacttcttaattcaattacaataccacctaatg
cgtatgtgagttccgtgtattctggttcaacatcaagaataatattctta
gttatcccatctaaatatccacgtccaataacggtcttaacaaaaacgcc
tttacctgcacctgaaattccgaatatacccatactgaaactttctaatt
ttgaatcaaaattattgaaatatatatactcgttgttatatacacctatt
ggcattccaccagtatgatttaatgttgaagagttatgaggaaacatagc
tgctaaggcattttcatcaagattcctttcagaattcgtaatcatgtcat
agctgttacctgtgtgaaattgtgatccgctcacaattccacacaacata
cgagccggaagcataaagtgtaaagcctggggtgcctaa
>rgbhoda0_001003.y1.scf
tagtcgacctgcaggcatgcaagcttgcgaagccttcattaaggctaatg
tagcccggggacttactcctaaataaacatagctgttttctctagtttga
gtagcaagctccaccatgtaatttttttactgaatcttcaacgtacactcc
ttgaacaagctcttgtaattcaattaattgagctactgacaatacagcgt
cgatcttttcaattgctttgccattttccgctcgacgtaaaatttctact
tcttgccctctagtagggtagcccatctttattttcaacaaaaaacgatc
aagctgggcttccggcaatggataagtaccttcgtgttctatcggatttt
gcgtggccattacaaagaaaggctgattaatggcaagtgttttaccatca
atagtaacagatgcttcttccatcccctctagtaaagctgattgcgtttt
tggcgaggtacgattaatttcatcagctaaaataacatcgcccattatcg
gtcctggacgaaattcaaactccaaagtctttggattataaatagagatt
cccactacatcggaaggtaataaatctggagtaaattgaattcgtttaaa
ctgtgcatcaaaggatttggctaatgaacgaaccatcattgttttcccaa
caccaggcacatcctctaacaatacgtgccccctcgctaataaagcaaca
aggctcagct
```

2. 输入文件 2，载体序列：%more Vector.seq

```
>PUC18
tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagctcccggagacggtca
cagcttgtctgtaagcggatgccgggagcagacaagcccgtcagggcgcgtcagcgggtg
ttggcgggtgtcggggctggcttaactatgcggcatcagagcagattgtactgagagtgc
accatatgcggtgtgaaataccgcacagatgcgtaaggagaaaataccgcatcaggcgcc
attcgccattcaggctgcgcaactgttgggaagggcgatcggtgcgggcctcttcgctat
tacgccagctggcgaaaggggatgtgctgcaaggcgattaagttgggtaacgccagggt
tttcccagtcacgacgttgtaaaacgacggccagtgccaagcttgcatgcctgcaggtcg
actctagaggatccccgggtaccgagctcgaattcgtaatcatggtcatagctgtttcct
gtgtgaaattgttatccgctcacaattccacacaacatacgagccggaagcataaagtgt
aaagcctggggtgcctaatgagtgagctaactcacattaattgcgttgcgctcactgccc
gctttccagtcgggaaacctgtcgtgccagctgcattaatgaatcggccaacgcgcgggg
agaggcggtttgcgtattgggcgctcttccgcttcctcgctcactgactcgctgcgctcg
gtcgttcggctgcggcgagcggtatcagctcactcaaaggcggtaatacggttatccaca
gaatcaggggataacgcaggaaagaacatgtgagcaaaaggccagcaaaaggccaggaac
cgtaaaaaggccgcgttgctggcgtttttccataggctccgccccctgacgagcatcac
aaaaatcgacgctcaagtcagaggtggcgaaacccgacaggactataaagataccaggcg
tttcccccctggaagctccctcgtgcgctctcctgttccgaccctgccgcttaccggatac
ctgtccgcctttctcccttcgggaagcgtggcgctttctcaaagctcacgctgtaggtat
ctcagttcggtgtaggtcgttcgctccaagctgggctgtgtgcacgaacccccgttcag
cccgaccgctgcgccttatccggtaactatcgtcttgagtccaacccggtaagacacgac
ttatcgccactggcagcagccactggtaacaggattagcagagcgaggtatgtaggcggt
gctacagagttcttgaagtggtggcctaactacggctacactagaagaacagtatttggt
atctgcgctctgctgaagccagttaccttcggaaaaagagttggtagctcttgatccggc
aaacaaaccaccgctggtagcggtggtttttttgtttgcaagcagcagattacgcgcaga
aaaaaaggatctcaagaagatcctttgatcttttctacggggtctgacgctcagtggaac
gaaaactcacgttaagggattttggtcatgagattatcaaaaaggatcttcacctagatc
cttttaaattaaaaatgaagttttaaatcaatctaaagtatatatgagtaaacttggtct
gacagttaccaatgcttaatcagtgaggcacctatctcagcgatctgtctatttcgttca
tccatagttgcctgactccccgtcgtgtagataactacgatacgggagggcttaccatct
ggccccagtgctgcaatgataccgcgagacccacgctcaccggctccagatttatcagca
ataaaccagccagccggaagggccgagcgcagaagtggtcctgcaactttatccgcctcc
atccagtctattaattgttgccgggaagctagagtaagtagttcgccagttaatagtttg
cgcaacgttgttgccattgctacaggcatcgtggtgtcacgctcgtcgtttggtatggct
tcattcagctccggttcccaacgatcaaggcgagttacatgatcccccatgttgtgcaaa
aaagcggttagctccttcggtcctccgatcgttgtcagaagtaagttggccgcagtgtta
tcactcatggttatggcagcactgcataattctcttactgtcatgccatccgtaagatgc
ttttctgtgactggtgagtactcaaccaagtcattctgagaatagtgtatgcggcgaccg
agttgctcttgcccggcgtcaatacgggataataccgcgccacatagcagaactttaaaa
gtgctcatcattggaaaacgttcttcggggcgaaaactctcaaggatcttaccgctgttg
```

```
agatccagttcgatgtaacccactcgtgcacccaactgatcttcagcatcttttactttc
accagcgtttctgggtgagcaaaaacaggaaggcaaaatgccgcaaaaaagggaataagg
gcgacacggaaatgttgaatactcatactcttccttttcaatattattgaagcatttat
cagggttattgtctcatgagcggatacatatttgaatgtatttagaaaaataaacaaata
ggggttccgcgcacatttccccgaaaagtgccacctgacgtctaagaaaccattattatc
atgacattaacctataaaaataggcgtatcacgaggccctttcgtc
```

3. 运行命令：%cross_match reads.dat puc18.fas -minmatch 12 -penalty -2 -minscore 20 -screen > screen.out

4. 输出结果 1，屏蔽载体后的序列文件 reads.dat.screen

%more reads.dat.screen

```
>gbeod0_000332.z1.scf
TAAGACTAAGGATCCCGGGTACGAGCTCGAATCAATAGCTTCCTTAACCT
TCTCATTAATATTTACTTTTTCAACAATATACTCGAAAGGTGTATATCGT
CTTTTAACTTCCTTTTCAACATATACATCCAAAGGATTAATTCGGCTAAA
ATTACTTACTACATCATCACTTCTTAATTCAATTACAATACCACCTAATG
CGTATGTGAGTTCCGTGTATTCTGGTTCAACATCAAGAATAATATTCTTA
GTTATCCCATCTAAATATCCACGTCCAATAACGGTCTTAACAAAAACGCC
TTTACCTGCACCTGAAATTCCGAATATACCCATACTGAAACTTTCTAATT
TTGAATCAAAATTATTGAAATATATATACTCGTTGTTATATACACCTATT
GGCATTCCACCAGTATGATTTAATGTTGAAGAGTTATGAGGAAACATAGC
TGCTAAGGCATTTTCATCAAGATTCCTTTCAXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
>rgbhoda0_001003.y1.scf
TXXXXXXXXXXXXXXXXXXXXXXXXXXCGAAGCCTTCATTAAGGCTAATG
TAGCCCGGGGACTTACTCCTAAATAAACATAGCTGTTTTCTCTAGTTTGA
GTAGCAAGCTCCACCATGTAATTTTTTACTGAATCTTCAACGTACACTCC
TTGAACAAGCTCTTGTAATTCAATTAATTGAGCTACTGACAATACAGCGT
CGATCTTTTCAATTGCTTTGCCATTTTCCGCTCGACGTAAAATTTCTACT
TCTTGCCCTCTAGTAGGGTAGCCCATCTTTATTTTCAACAAAAAACGATC
AAGCTGGGCTTCCGGCAATGGATAAGTACCTTCGTGTTCTATCGGATTTT
GCGTGGCCATTACAAAGAAAGGCTGATTAATGGCAAGTGTTTTACCATCA
ATAGTAACAGATGCTTCTTCCATCCCCTCTAGTAAAGCTGATTGCGTTTT
TGGCGAGGTACGATTAATTTCATCAGCTAAAATAACATCGCCCATTATCG
GTCCTGGACGAAATTCAAACTCCAAAGTCTTTGGATTATAAATAGAGATT
CCCACTACATCGGAAGGTAATAAATCTGGAGTAAATTGAATTCGTTTAAA
CTGTGCATCAAAGGATTTGGCTAATGAACGAACCATCATTGTTTTCCCAA
CACCAGGCACATCCTCTAACAATACGTGCCCCCTCGCTAATAAAGCAACA
AGGCTCAGCT
```

5. 输出结果 2，载体屏蔽信息文件 screen.out

%more screen.out

```
cross_match reads.dat puc18.fas -minmatch 12 -penalty -2 -minscore 20 -screen
cross_match version 0.990329

Run date:time  070309:105149
Query file(s):  reads.dat
Subject file(s):   puc18.fas
Presumed sequence type: DNA

Pairwise comparison algorithm: banded Smith-Waterman

Score matrix (set by value of penalty: -2)
    A   C   G   T   N   X
A   1  -2  -2  -2   0  -3
C  -2   1  -2  -2   0  -3
G  -2  -2   1  -2   0  -3
T  -2  -2  -2   1   0  -3
N   0   0   0   0   0   0
X  -3  -3  -3  -3   0  -3

Gap penalties: gap_init: -4, gap_ext: -3, ins_gap_ext: -3, del_gap_ext: -3,
Using complexity-adjusted scores. Assumed background frequencies:
 A: 0.250  C: 0.250  G: 0.250  T: 0.250  N: 0.000  X: 0.000
```

```
minmatch: 12, maxmatch: 12, max_group_size: 20, minscore: 20, bandwidth: 14,
indexwordsize: 10
vector_bound: 0
word_raw: 0
masklevel: 80

Sequence file: reads.dat    2 entries
Residue counts:
  a     397
  c     288
  g     209
  t     405
Total   1299

NO QUALITY FILE reads.dat.qual WAS FOUND. REMAINING INPUT QUALITIES SET TO 15.
Maximal single base matches (low complexity regions):

  98  1.85 0.93 0.00  gbeod0_000332.z1.scf      482   589 (0)    PUC18      450   558
(2128)

  26  0.00 0.00 0.00  rgbhoda0_001003.y1.scf       2    27 (683)  C PUC18    (2263)
423   398

2 matching entries (first file).

Discrepancy summary:
Qual algn  cum    rcum    (%)    unalgn X    N  sub del ins  total (%)   cum  rcum (%)


Screened sequences written to  reads.dat.screen
```

练习

对实验室测序数据峰图转换后屏蔽相应的载体。

2.5 序列聚类拼接

2.5.1 Phrap

简介

phrap ("phragment assembly program", or "phil's revised assembly program"), Phrap 是由华盛顿大学分子生物技术学院的 Phil Green 和 Brent Ewing 开发的 phred\phrap 软件包的一部分，主要用于 shotgun 序列的组装。

Key features:

1. 允许使用全长的序列（而不仅仅是高质量部分）
2. 使用质量信息进行组装提高组装的准确度
3. constructs contig sequence as a mosaic of the highest qualityparts of reads (rather than a consensus)
4. 提供广泛的组装信息帮助解决错拼等问题（包括 contig 序列的质量信息）provides extensive information about assembly (including quality values for contig sequence) to assist trouble-shooting;
5. 能够处理比较大的数据集

下载

phrap可通过邮件直接向作者索取：phg@u.washington.edu

安装

1. 上传 phrap 的压缩包到本地 linux/unix 运算服务器；

2. 解压缩：

```
gzip –d phrap.tar.gz
tar –xvf phrap.tar
```

3. 编译源程序：

在命令行键入 make， 屏幕提示如下：

```
bash-2.05b$ make
       cc -O2 -c swat.c
       cc -O2 -c weibull.c
       cc -O2 -c alignments.c
       cc -O2 -c db.c
       cc -O2 -c smith_wat.c
       cc -O2 -c -o full_smith_wat.o  smith_wat.c -DFINDALIGN
       … …
       cc -O2 -c loco.c
       cc  -O2  -o  loco  loco.o   alignments.o  db.o  smith_wat.o  full_smith_wat.o
quick_smith_wat.o utilities.o nw.o full_nw.o profile.o parameters.o  -lm
       chmod o-r loco
```

如果编译器不识别-O2，可将 makefile 文件中 CFLAGS= -O2 行改为 CFLAGS= -O，删除*.o 文件后重新编译。

如果数据集多于 64,000 条序列，或者序列中含有长于 64,000 bp 的序列，则需要使用 phrap.manyreads 或 phrap.longreads, 这两个程序编译命令为:

```
bash-2.05b$ make manyreads
        touch swat.h;
        make CFLAGS="-O2 -DMANYREADS" phrap cross_match;
        cc -O2 -DMANYREADS -c phrap.c
        cc -O2 -DMANYREADS -c call_subs.c
        cc -O2 -DMANYREADS -c contigs.c
        cc -O2 -DMANYREADS -c tig_node.c
        … …
        chmod o-r phrap
        cc -O2 -c cross_match.c
        cc -O2 -o cross_match cross_match.o call_subs.o readin.o words.o segments.o
 recursive_swat.o log_file.o pairs.o cand_pairs.o diffs.o names.o nodes.o anomalies.o
 qual.o  tags.o  alignments.o  db.o  smith_wat.o  full_smith_wat.o  quick_smith_wat.o
 utilities.o nw.o full_nw.o profile.o parameters.o  -lm
        chmod o-r cross_match
```

编译完成后，可用命令 make clean 清除编译过程中的文件，也可用 rm *.c 命令删掉源文件。剩下的有用文件为：

```
bash-2.05b$ ls -l
total 2060
-rwxr-x--x    1 soft  bgi     227380 Jan 11 02:47 cross_match*
-rwxr-x--x    1 soft  bgi     302176 Jan 11 02:47 phrap*
-rw-------    1 soft  bgi      18745 Jan 11 02:46 swat.h
-rwxr-x--x    1 soft  bgi     302176 Jan 11 02:46 phrap.longreads*
-rwxr-x--x    1 soft  bgi     227380 Jan 11 02:46 cross_match.manyreads*
-rwxr-x--x    1 soft  bgi     302176 Jan 11 02:46 phrap.manyreads*
-rwxr-x--x    1 soft  bgi      88348 Jan 11 02:45 loco*
-rwxr-x--x    1 soft  bgi     231476 Jan 11 02:45 cluster*
```

```
-rwxr-x--x    1 soft  bgi     105048 Jan 11 02:45 swat*
-rw-------    1 soft  bgi       6120 Jan 11 02:05 makefile
-rw-------    1 soft  bgi      88444 Mar 25  1999 phrap.doc
-rw-------    1 soft  bgi       6755 Mar  8  1999 general.doc
-rw-------    1 soft  bgi      13729 Nov 17  1997 swat.doc
-rw-r--r--    1 soft  bgi       2083 Jun 20  1997 BLOSUM50
-rwx--x--x    1 soft  bgi      33903 Jul 31  1996 phrapview*
-rw-------    1 soft  bgi       2083 Jun 22  1996 BLOSUM62
-rw-------    1 soft  bgi        192 Jun 22  1996 penalty2
-rw-------    1 soft  bgi        378 Jun 22  1996 mat70
-rw-------    1 soft  bgi        367 Apr 14  1995 mat50
-rw-------    1 soft  bgi     103912 Aug 24  1994 vector.seq
-rw-------    1 soft  bgi       1992 Sep  4  1992 PAM250
```

使用

程序运行命令行：

```
phrap [sequence file] -new_ace > phrap.out
```

输入

Fasta 格式的核酸序列，如：pp.seq.screen：

```
>10_A8-9.ab1
gtgctctggtctctgctcctttcccctaagcaatagtaggcagaatcaac
aaaaacaaccccttctcccctccctacctggggaacagagccaatgagac
aggctcaggaacagggcaccagcacctgcactcaccattcaatctcttta
ggctcacggtccttcagaagctcttgtacctcctgccgacagcgctcctg
gtattccgggtgctttgcaaggtggtacaggacccaggagagaccactgg
ccccccataaaaagtcacagtacctctgagggctcttgagtctaatctgag
acagtctctgaagattcatcctctttccagaaacccaagcccatcttgct
ctcctagaaacctttctataaaaaaaaaaaaan
>11_A8-9_R.ab1
gggagaggcggagctctggtccttgtcatctaagctgtgtggattgatcg
cctagaacctccctatctaccctccctacctggggaacagagccaatgag
aaaggctcaggaacagggcaccagcacctgcactcaccattcaatctctt
taggctcacggtccttcagaagctcttgtacctcctgccgacagcgctcc
caacttcttcccatcttcatcctggagagaaggcaataacccccccacccc
caccccccataaaaagtcacagtacctctgagggctcttgagtctaatctg
agacagtctctgaagattcatcctctttccagaaacccaagcccatcttg
ctctccagaacccttcttaaa
>15_A8-9.ab1
aagactggcagnggatctctgcatctagtcacctaagctatagctggtag
actcgaccaaaacaacccttttctaccctccctacctggggaacagagcca
atgagacaggctcaggaacag
… …
```

如有质量文件，则质量文件需和序列文件放在同一目录下，且名字为[序列文件名.qual]，如，序列文件名为 pp.seq.screen，质量文件名必须为 pp.seq.screen.qual，质量文件不需要在命令行中。并且质量文件中的序列和序列文件中的序列必须一一对应，包括顺序和碱基个数。

输出

在程序运行目录，除屏幕输出外，会产生一系列相关文件，分别为：

1. *.contigs 文件。组装好的 contig 序列，格式为 FASTA 格式。其中包括单个 read 的 contig（这类 reads 和其他 contig 有比对上的部分，但达不到连上的标准）(without pads; bases in this file are upper case if and only if the quality is >= qual_show). These include singleton contigs consisting of single reads with a match to some other contig, but that couldn't be merged consistently with it.
2. *.contigs.qual 文件。Contig 组装的质量文件，FASTA 格式。此文件记录每个 contig 的碱基质

量信息。

3. *.singlets 文件。和任何其他 reads 没有 overlap 的序列，FASTA 格式。
4. *.log 文件和*.problems 文件。对使用者基本没用。
5. *.ace 文件。当使用参数-new_ace 或-old_ace 时才会产生的文件，用 consed 查看组装结果时需要，It's format is described in the consed documentation.
6. `*.view` 文件。当使用`-view`参数时产生的文件，用 `phrapview` 查看组装结果时需要。
7. 除以上文件外，`phrap` 还有屏幕输出，可重定向到文件，如 `phrap > phrap.out`

```
/usr/local/genome/bin/phrap pp.fasta.screen -new_ace -view
phrap version 0.990329

Run date:time  061230:074329
Query file(s):  pp.fasta.screen
Presumed sequence type: DNA

Pairwise comparison algorithm: banded Smith-Waterman

Score matrix (set by value of penalty: -2)
   A   C   G   T   N   X
A   1  -2  -2  -2   0  -3
C  -2   1  -2  -2   0  -3
G  -2  -2   1  -2   0  -3
T  -2  -2  -2   1   0  -3
N   0   0   0   0   0   0
X  -3  -3  -3  -3   0  -3

Gap penalties: gap_init: -4, gap_ext: -3, ins_gap_ext: -3, del_gap_ext: -3,
Using complexity-adjusted scores. Assumed background frequencies:
 A: 0.250  C: 0.250  G: 0.250  T: 0.250  N: 0.000  X: 0.000

minmatch: 14, maxmatch: 30, max_group_size: 20, minscore: 30, bandwidth: 14,
indexwordsize: 10
vector_bound: 80
word_raw: 0
trim_penalty: -2, trim_score: 20, trim_qual: 13, maxgap: 30
repeat_stringency: 0.950000
qual_show: 20
confirm_length: 8, confirm_trim: 1, confirm_penalty: -5, confirm_score: 30
node_seg: 8, node_space: 4
forcelevel: 0, bypasslevel: 1
max_subclone_size: 5000

Sequence file: pp.fasta.screen    8 entries
Residue counts:
  A    1032
  C    1332
  G    1138
  N       6
  T     950
Total   4458

Read name analysis:
 # Reads     # templates
   1           6
   2           1

 Suffix counts:
  a           6
  r           2

Templates inferred from description field:     0
Templates inferred from name field:           8
```

```
Read-template multiplicity analysis:
 # Reads     # templates
   1           6
   2           1

Chemistries inferred from description field:
    0  dye-primer
    0  old-dye-terminator
    0  big-dye-terminator
    0  other

Chemistries inferred from name:
    2  dye-primer
    0  old-dye-terminator
    0  big-dye-terminator
    6  other

Directions inferred from description field:
    0  fwd
    0  rev
    0  unknown (set to fwd)

Directions inferred from name:
    0  fwd
    2  rev
    6  unknown (set to fwd)

Quality file: pp.fasta.screen.qual

Input quality (quality, n_residues, %,  cum, cum %,  cum expected errs):
 59    384   8.6    384   8.6    0.00
 56   2019  45.3   2403  53.9    0.01
 51    362   8.1   2765  62.0    0.01
 50    148   3.3   2913  65.3    0.01
 48     52   1.2   2965  66.5    0.01
 47     63   1.4   3028  67.9    0.01
 46     82   1.8   3110  69.8    0.01
 45     48   1.1   3158  70.8    0.02
… …
  7     35   0.8   4338  97.3   42.53
  6     79   1.8   4417  99.1   62.38
  4     35   0.8   4452  99.9   76.31
  0      1   0.0   4453  99.9   77.31
 -1      5   0.1   4458 100.0   82.31   (quality -1 = terminal quality 0)

Avg. full length: 557.2, trimmed (qual > -1): 556.6
Avg. quality: 46.3 per base

Following regions converted to N's

Exact duplicate reads:
A8-9.ref.scf        A8-9.ref       (perfect)

2d read in each pair excluded from assembly.

Probable unremoved sequencing vector (matches excluded from assembly, quality reduced
to 0):  None.

Near duplicate reads:
10_A8-9.ab1         22_A8-9.ab1    (imperfect: 2-652 (30)   6-656 (23) )
10_A8-9.ab1         21_A8-9.ab1    (imperfect: 6-679 (3)   7-679 (1) )
10_A8-9.ab1         15_A8-9.ab1    (imperfect: 24-655 (27)   31-662 (23) )
15_A8-9.ab1         22_A8-9.ab1    (imperfect: 15-675 (10)   13-671 (8) )
15_A8-9.ab1         21_A8-9.ab1    (imperfect: 15-659 (26)   10-652 (28) )
21_A8-9.ab1         22_A8-9.ab1    (imperfect: 7-663 (17)   10-667 (12) )

Internal read matches (same orientation) :  None.
```

```
No. of node-rejected pairs: None.

Multi-segment reads (initially rejected segments in parentheses) -- XXX means segments
flank X'd region:

0 reads with multiple segments.

Probable deletion reads (excluded from assembly): None.

Revised quality (quality, n_residues, %,  cum, cum %,  cum expected errs):
 90     373   8.4     373   8.4    0.00
 88       3   0.1     376   8.4    0.00
 73       4   0.1     380   8.5    0.00
 59     304   6.8     684  15.3    0.00
 56    1655  37.1    2339  52.5    0.00
 51     362   8.1    2701  60.6    0.01
… …
 36       6   0.1    3590  80.5    0.05
 35      16   0.4    3606  80.9    0.06
 34      24   0.5    3630  81.4    0.07
-1     358   8.0    4458 100.0  395.33   (quality -1 = terminal quality 0)

Avg. full length: 557.2, trimmed (qual > -1): 512.5
Avg. quality: 47.6 per base

LLR score histogram:
Score    #   cum #
  0.0     7     7
  5.0     3    10
 10.0    10    20

LLR score histogram:
Score    #   cum #
  0.0     7     7
  5.0     3    10
 10.0    10    20

2d revised quality (quality, n_residues, %,  cum, cum %,  cum expected errs):
 90     373   8.4     373   8.4    0.00
 88       3   0.1     376   8.4    0.00
 73       4   0.1     380   8.5    0.00
 59     304   6.8     684  15.3    0.00
 56    1655  37.1    2339  52.5    0.00
 51     362   8.1    2701  60.6    0.01
 50     148   3.3    2849  63.9    0.01
 48      36   0.8    2885  64.7    0.01
… …
  8      26   0.6    4019  90.2   15.85
  7      15   0.3    4034  90.5   18.84
  6      53   1.2    4087  91.7   32.16
  4      13   0.3    4100  92.0   37.33
 -1     358   8.0    4458 100.0  395.33   (quality -1 = terminal quality 0)

Avg. full length: 557.2, trimmed (qual > -1): 512.5
Avg. quality: 47.6 per base

No. confirmed reads: 7
Avg. length: 582.0, confirmed: 523.6, str. confirmed: 419.7, trimmed: 542.3
Preliminary clone size estimate: 655 bp, depth of coverage: 5.6

Depth histogram (max_depth, #reads, cum #reads):

 6     5       5
 5     2       7

Forward confirmed bases: 0
```

```
Substitutions by nucleotide:
      A     C     G     T     N     X     Z    Total
A     0     0     0     0     0     0     0       0
C     0     0     0     0     0     0     0       0
G     0     0     0     0     0     0     0       0
T     0     0     0     0     0     0     0       0
N     0     0     0     0     0     0     0       0
X     0     0     0     0     0     0     0       0
Z     0     0     0     0     0     0     0       0

Substitutions by quality:
      Total
Histogram of spacings between adjacent indel pairs:
Reverse confirmed bases: 0
Substitutions by nucleotide:
      A     C     G     T     N     X     Z    Total
A     0     0     0     0     0     0     0       0
C     0     0     0     0     0     0     0       0
G     0     0     0     0     0     0     0       0
T     0     0     0     0     0     0     0       0
N     0     0     0     0     0     0     0       0
X     0     0     0     0     0     0     0       0
Z     0     0     0     0     0     0     0       0

Substitutions by quality:
      Total
Histogram of spacings between adjacent indel pairs:

Blocked reads:
10_A8-9.ab1 39 655   right
15_A8-9.ab1 59 662   right
21_A8-9.ab1 10 651  left

3 blocked reads: 1 left only, 2 right only, 0 both.
0 reads (not shown) lack a high-quality segment.

1 perfect duplicates:
  Read        Length
 A8-9.ref      384

0 isolated singlets (having no non-vector match to any other read)

Contig 1.  7 reads; 685 bp (untrimmed), 653 (trimmed).  Isolated contig.
     -1   682 15_A8-9.ab1   604 (  0)  1.55 0.31 0.00   15 ( 58)   23 ( 23)
      1   679 22_A8-9.ab1   635 (  0)  0.15 0.30 0.15    0 (  6)   23 ( 19)
      2   673 11_A8-9_R.ab1  580 (  0)  0.67 0.00 0.17   65 ( 65)    6 ( 15)
      5   686 10_A8-9.ab1   662 (  0)  0.44 0.15 0.00    2 (  2)    1 ( 27)
      4   684 21_A8-9.ab1   648 (  0)  0.59 0.15 0.15    7 (  7)    1 ( 24)
C   139   522 A8-9.ref.scf  381 (  0)  0.00 0.00 0.00    0 (  0)    0 (  0)
C   352   641 23_A8-9.ab1   120 (  0)  0.00 0.00 0.79  147 (147)   16 ( 16)

Overall discrep rates (%):           0.58 0.16 0.11

Contig quality (quality, n_residues, %,  cum, cum %,  cum expected errs):
 90    373  54.5    373  54.5    0.00
 88      3   0.4    376  54.9    0.00
 73      4   0.6    380  55.5    0.00
 56    148  21.6    528  77.1    0.00
 51     34   5.0    562  82.0    0.00
 50      1   0.1    563  82.2    0.00
 48      8   1.2    571  83.4    0.00
… …
  9      1   0.1    647  94.5    0.28
  8      1   0.1    648  94.6    0.44
  7      3   0.4    651  95.0    1.04
  4      2   0.3    653  95.3    1.84
 -1     32   4.7    685 100.0   33.84   (quality -1 = terminal quality 0)
```

```
Avg. full length: 685.0, trimmed (qual > -1): 653.0
Avg. quality: 69.3 per base

Initial, terminal qual 0 segments:  1-6, 660-685

Regions of LLR- adjusted quality < 2.0:
1-14, 28-30, 659-685,

3 regions, avg size 14.7, avg spacing 228.3

First_start: 7, last_end: 660

Slack, # used pairs (max_score), unused
 0    14  (14.0)     0 ( 0.0)       20
 1     6  (12.1)     0 ( 0.0)        0

LLR histograms (used, unused pairs):

   DS Gap        Size     Closest read (Start)   Covers   Read length required
                                                  now?       to cover
Top strand:
 left -     0       0+
  686 - right       0+     10_A8-9.ab1  (   5)    No           680+

Bottom strand:
 left -   138     138+     A8-9.ref.scf ( 522)    No           522+
  626 - right      60+

Read/contig alignment summary, by read base; trace qualities
Qual algn  cum    rcum    (%)    unalgn X    N  sub del ins  total (%)   cum  rcum (%)
56   2019   2019   3760 (100.00)     0  0    0   1   0   0     1 (0.05)    1   32 (0.85)
51    362   2381   1741 ( 46.30)     0  0    0   1   0   0     1 (0.28)    2   31 (1.78)
50    148   2529   1379 ( 36.68)     0  0    0   0   0   0     0 (0.00)    2   30 (2.18)
48     52   2581   1231 ( 32.74)     0  0    0   0   0   0     0 (0.00)    2   30 (2.44)
47     63   2644   1179 ( 31.36)     0  0    0   0   0   0     0 (0.00)    2   30 (2.54)
46     82   2726   1116 ( 29.68)     0  0    0   0   0   0     0 (0.00)    2   30 (2.69)
45     48   2774   1034 ( 27.50)     0  0    0   0   0   0     0 (0.00)    2   30 (2.90)
… …
 8     21   3702     79 (  2.10)     5  0    0   0   0   1     1 (4.76)   22   11 (13.92)
 4      9   3760      9 (  0.24)     4  0    0   0   1   0     1 (11.11)  32    1 (11.11)
-1      8   3768      0 (  0.00)   267  0    0   0   0   0     0 (0.00)   32    0 (0.00)


Read/contig alignment summary, by read base; adjusted qualities
Qual algn  cum    rcum    (%)    unalgn X    N  sub del ins  total (%)   cum  rcum (%)
90    373    373   3661 (100.00)     0  0    0   0   0   0     0 (0.00)    0   18 (0.49)
88      3    376   3288 ( 89.81)     0  0    0   0   0   0     0 (0.00)    0   18 (0.55)
73      4    380   3285 ( 89.73)     0  0    0   0   0   0     0 (0.00)    0   18 (0.55)
… …
22     19   3491    189 (  5.16)     0  0    0   0   0   0     0 (0.00)    4   14 (7.41)
21      8   3499    170 (  4.64)     0  0    0   1   0   0     1 (12.50)   5   14 (8.24)
20      9   3508    162 (  4.43)     0  0    0   0   0   0     0 (0.00)    5   13 (8.02)
19     18   3526    153 (  4.18)     0  0    0   0   0   0     0 (0.00)    5   13 (8.50)
4       8   3661     8 (  0.22)     1  0    0   0   1   0     1 (12.50)  18    1 (12.50)

Depth 0 regions:

Block histogram:
Qual   bases    cum    blocks
  0      32      32       2
  4       2      34       3
  7       3      37       4
  8       1      38       3
… …
 47       7     114      15
 48       8     122      12
 50       1     123      13
 51      34     157      14
```

```
 56    148     305       2
 73      4     309       2
 88      3     312       3
 90    373     685       1

SS region: 198 (28.91%), flagged: 1 (0.15%)

Sites with total LLR scores < -3.0  [max pos LLR read, max neg LLR read]  (#discrep
top reads, #discrep bottom reads):
  180   -15.1  [-5.6,  0.0]  (3, 0)

Read/contig discrepancies (* = higher-quality):
*   23  D    22_A8-9.ab1    (111)/(87)  21 TCCT / TCT
   663  S    11_A8-9_R.ab1   (0)/(0)  662 AAC / ACC
   668  I    21_A8-9.ab1    (0)/(0)  668 TC / TTC
   677  S    21_A8-9.ab1    (0)/(0)  676 AAA / ATA

1 HQ discrepancies in 1 reads.
3 lower quality discrepant sites.

Reads with neg LLR score, or confirmed or high-qual unaligned seg > 20 bases, or other
problem: None.

Gaps in unique-read coverage:  None.

Subclone/read contig links and consistency checks (* = inconsistency; Contig 0 =
singlets)
Max subclone size: 5000

Contig 1 same sense  LEFT LINK:   complement Contig 0
   C A8-9.ref.scf      A8-9.ref          519    521     -2
Size histogram for consistent forward-reverse pairs (*** = inconsistent pairs)
  ***    0
Consistent opp sense links (* = not used in chain, ** = multiple non-zero):
```

对于 phrap.out 的结果，我们可以用 phraplist 程序提取有用的信息。命令如下：

```
phraplist phrap.out > phrap.lis
```

phraplist 代码如下：

```
#!/usr/bin/perl
die "Usage:$0 phrap.out\n" if (@ARGV!=1);
open(PhrapOut, "$ARGV[0]") ||die "could not open $ARGV[0]";
@line=<PhrapOut>;
$real=0;
foreach $hang (@line) {
       if($hang =~/^Contig\s\d+.\s+\d+\s\w+;\s\d+\sbp/ ) {
              $real=1;
       }
       $real=0  if($hang  =~/Contig  quality  (.*):$/  ||  $hang  =~/^Overall  discrep
rates/);
    $real=0 if($hang=~"Overall");
    print $hang if($real);
}
close(PhrapOut);
```

提取到的信息 phrap.lis 包含了每个 contig 的组成、长度等信息，格式如下：

```
Contig 1.  7 reads; 685 bp (untrimmed), 653 (trimmed).  Isolated contig.
    -1   682 15_A8-9.ab1   604 (  0)  1.55 0.31 0.00   15 ( 58)   23 ( 23)
     1   679 22_A8-9.ab1   635 (  0)  0.15 0.30 0.15    0 (  6)   23 ( 19)
     2   673 11_A8-9_R.ab1  580 (  0)  0.67 0.00 0.17   65 ( 65)    6 ( 15)
     5   686 10_A8-9.ab1   662 (  0)  0.44 0.15 0.00    2 (  2)    1 ( 27)
     4   684 21_A8-9.ab1   648 (  0)  0.59 0.15 0.15    7 (  7)    1 ( 24)
```

```
C   139   522 A8-9.ref.scf  381 (  0)  0.00 0.00 0.00    0 (  0)    0 (  0)
C   352   641 23_A8-9.ab1   120 (  0)  0.00 0.00 0.79  147 (147)   16 ( 16)
```

参数

详细的参数列表可以查看 phrap 文档：

```
option name & default value
1. Scoring of pairwise alignments
-penalty -2             Mismatch (substitution) penalty for SWAT comparisons.
-gap_init penalty-2     Gap initiation penalty for SWAT comparisons.
-gap_ext  penalty-1     Gap extension penalty for SWAT comparisons.
-ins_gap_ext gap_ext    Insertion gap extension penalty for SWAT comparisons
                        (insertion in subject relative to query).
-del_gap_ext gap_ext    Deletion gap extension penalty for SWAT comparisons (deletion
                        in subject relative to query).
-matrix [None]          Score matrix for SWAT comparisons (if present, supersedes
                        -penalty) Matrix format: (TO BE ADDED)
-raw *                  Use raw rather than complexity-adjusted Smith-Waterman
                        scores.
2. Banded search
-minmatch 14            Minimum length of matching word to nucleate SWAT comparison.
-maxmatch 30            Maximum length of matching word. For cross_match, the default
                        value is equal to minmatch, instead of 30.
-max_group_size 20      Group size (query file, forward strand words)
-word_raw *             Use raw rather than complexity-adjusted word length, in
                        testing against minmatch (N.B. maxmatch always refer to raw
                        lengths).
-bandwidth 14           1/2 band width for banded SWAT searches (full width is 2 times
                        bandwidth + 1).
3. Filtering of matches
-minscore 30            Minimum alignment score.
-vector_bound 80        Number of potential vector bases at beginning of each read.
 Special cases:
-masklevel 0            report only the single highest scoring match for each query
-masklevel 100          report any match whose domain is not completely contained
                        within a higher scoring match
-masklevel 101          report all matches
4. Input data interpretation
-default_qual 15        Quality value to be used for each base, when no input .qual
                        file is provided.
-subclone_delim .       Subclone name delimiter
-n_delim 1              Indicates which occurrence of the subclone delimiter character
                        denotes the end of the subclone name
-group_delim _          Group name delimiter: Character used to indicate end of that
                        part of the read name that corresponds to the group name
                        (relevant only if option -preassemble is used);
-trim_start 0           No. of bases to be removed at beginning of each read.
5. Assembly
-forcelevel 0           Relaxes stringency to varying degree during final contig merge
                        pass.
-bypasslevel 1          Controls treatment of inconsistent reads in merge.
-maxgap 30              Maximum permitted size of an unmatched region in merging
                        contigs, during first (most stringent) merging pass.
-repeat_stringency .95  Controls stringency of match required for joins.
-revise_greedy *        Splits initial greedy assembly into pieces at "weak joins",
                        and then tries to reattach them to give higher overall score.
-shatter_greedy *       Breaks assembly at weak joins (as with -revise_greedy) but does
                        not try to reattach pieces.
-preassemble *          Preassemble reads within groups, prior to merging with other
                        groups.
-force_high *           Causes edited high-quality discrepancies to be ignored during
                        final contig merge pass.
6. Consensus sequence construction
-node_seg 8             Minimum segment size (for purposes of traversing weighted
                        directed graph).
-node_space 4           Spacing between nodes (in weighted directed graph).
7. Output
```

```
-tags *                Tag selected lines in the standard output, to facilitate
                       parsing.
-screen *               when the -old_ace or -new_ace option is specified (see below),
                       this option causes parts of the read sequences that consist
                       of phrap-inferred sequencing vector and chimeric segments to
                       be replaced by X's in the .ace file.
-old_ace *             Create ".ace" file in old style format.
-new_ace *             Create ".ace" file in a new style format
-ace *                 Same as -new_ace.
-view *                Create ".view" file suitable for input to phrapview.
-qual_show 20          Cutoff for flagging "low_quality" regions in contig sequence
                       and "high quality" discrepancies between read and contig.
-print_extraneous_matches *Print information about non-local matches between
                       contigs.
-exp [None]            (gcphrap only). Name of a directory in which output experiment
                       files are to be placed.
8. Miscellaneous
-retain_duplicates *   Retain exact duplicate reads, rather than eliminating them.
-max_subclone_size 5000Maximum subclone size.
-trim_penalty -2       Penalty used for identifying degenerate sequence at beginning
                       & end of read.
-trim_score 20         Minimum score for identifying degenerate sequence at beginning
                       & end of read.
-trim_qual 13          Quality value used in to define the "high-quality" part of a
                       read.
-confirm_length 8      Minimum size of confirming segment.
-confirm_trim 1        Amount by which confirming segments are trimmed at edges.
-confirm_penalty -5    Penalty used in aligning against "confirming" reads.
-confirm_score 30      Minimum alignment score for a read to be allowed to "confirm"
                       part of another read.
-indexwordsize 10      Size of indexing (hashing) words, used in finding word matches
                       between sequences.
```

运行问题

1. 内存不足：

如果程序运行提前终止，并给出以下错误信息提示：

```
FATAL ERROR: REQUESTED MEMORY UNAVAILABLE
```

2. 程序长时间运行：

可以试着提高参数-minmatch 的值

实例

练习

参考文献

2.5.2 Cap3

简介

Huang, X. 和 Madan, A 开发的一套用于核酸序列拼接的软件，它有如下特征。

1. 应用正反向信息更正拼接错误、连接 contigs。

2. 在序列拼接中应用 reads 的质量信息。

3. 自动截去 reads5`端、3`端的低质量区。

4. 产生 Consed 程序可读的 ace 格式拼接结果文件。

5. CAP3 能用于 Staden 软件包的中的 GAP4 软件。

下载

通过 email 联系作者 Xiaoqiu Huang at huang@mtu.edu

CAP3 详细参考文档可见：http://genome.cs.mtu.edu/sas.html

安装

1. 上传 cap3 的压缩包到本地 linux/unix 运算服务器；

2. 解压缩：

```
bash-2.05b$ uudecode cap3.tar.uuencode-sgi
uudecode: cap3.tar.uuencode-sgi: No `end' line
bash-2.05b$ tar xvf cap3.tar
CAP3/
CAP3/README
CAP3/cap3
CAP3/doc
CAP3/aceform
CAP3/formcon
```

3. 查看解压缩后的文件：

```
bash-2.05b$ ls -l
total 240
-rwxr-xr-x    1 soft  bgi     25844 Sep  2  2002 formcon*
-rwxr-xr-x    1 soft  bgi    169836 Sep  2  2002 cap3*
-rw-r-----    1 soft  bgi       513 Aug 22  2002 README
-rw-------    1 soft  bgi     18448 Aug 22  2002 acefo
```

使用

程序运行命令行：

```
cap3  <dna-file in fasta format> <options> >cap3.out
```

输入

输入序列是普通的 FASTA 格式，如果序列文件名为“xyz”，则质量文件应命名为“xyz.qual”，约束文件应命名为“xyz.con”。

“xyz”格式如下：

```
>Sequence1
ACGTGCGCGATCGCCTGCTAGGCGTACGTCGCAGGCGATCGATGTGCTAGATCAGATGACA
>Sequence2
GGGCTAGATTAGCACCACATACATCGCTCA
```

“xyz.qual”格式如下：

```
>R1
  6  8  8  8 15 17 17 17 12 12 20 20 29 31 34 34 38 38 40 40 49 49 37 33 33
 33 33 30 31 24 24 34 45 45 45 45 38 38 38 45 40 40 40 40 40 40 40 40 40 40
33 33 33 33 33 33 40 37 40 40 45 45 45 40 40 40 45 45 45 45 49 49 49 49 45
 49 45 45 45 45 40 40 43 43 43 40 40 40 37 40 49 49 40 40 37 37 37 42 45 40
 36 36 36 36 33 33 27 27 21 19 19 27 33 33 34 36 36 36 36 38 36 36 40 33 35
>R2
 98 98 98 98 98 98 98 98 98 98 98 98 98 98 98 98 98 98 98 98 98 98 98 98 98
 37 37 37 37 37 37 37 37 37 37 37 37 37 34 34 34 34 37 37 37 37 34 34 37 38
34 37 34 37 37 37 37 37 45 37 37 37 37 37 37 37 40 37 37 32 45 41 45 45 41
```

约束文件“xyz.con”中每一行都以如下格式指定了正反向的约束：

```
ReadA ReadB MinDistance MaxDistance
```

其中“ReadA”和“ReadB”是两个 reads 的名称；“MinDistance”、“MaxDistance”是最小、最大距离（bp）。

输出

输出文件格式：

1. xyz.cap.ace： ace 格式文件，注意：reads 的 5`、3`的低质量区没有被显示在 ace 格式中。
2. xyz.cap.contigs：生成的 contigs 序列文件
3. xyz.cap.contigs.qual：生成的 contigs 质量文件
4. xyz.cap.singlets：没有用于拼接的 reads 文件
5. xyz.cap.info：关于拼接的额外信息文件

```
    No file of constraints (.con) is found.
R1      5   849   860
R2     55   888  1022
R3     55   870   918
R4      4   790   799
R5     17   599   920
R6     70   789   850

R1-       2    427   R4+     374     799
R1-       2    239   R5+     390     628
R1-       2    311   R6+     518     827
R1+      12    829   R2+      55     873
R1+      12    829   R3+      55     874
R2-      27    545   R4+     282     799
R2-      34    357   R5+     304     628
R2-      27    429   R6+     426     827
R2+      55    838   R3+      55     838
R3-       4    441   R4+     364     799
R3-      90    253   R5+     464     628
R3-       4    325   R6+     508     827
R4+      46    611   R5+      52     628
R4+      18    683   R6+     152     827
R5+      52    628   R6+     181     755

R1     12   859   860
R2     55   996  1022
R3     55   915   918
R4     18   799   799
R5     52   628   920
R6    152   827   850
Number of overlaps saved: 15
ComputeOverlap done
IdentifyChimeras done
Number of overlaps removed: 0
RemovePoorOverlaps done
R1+       1     848     848   R3+       1     850     861     Containment   53378
R3+       1     861     861   R2+       1     861     942     Containment   51554
R1+       1     848     848   R2+       1     849     942     Containment   50668
R6+       1     676     676   R4+       1     666     782     Containment   36206
R4+     357     782     782   R1-       1     426     848         Overlap   19848
R4+     265     782     782   R2-       1     519     942         Overlap   19133
R4+     347     782     782   R3-       1     438     861         Overlap   18011
R5+       1     577     577   R4+      29     594     782     Containment   10311
R6+     275     676     676   R2-       1     403     942         Overlap   10308
R5+       1     577     577   R6+      30     604     676     Containment   10211
R6+     367     676     676   R1-       1     310     848         Overlap   10047
R6+     357     676     676   R3-       1     322     861         Overlap    8516
R5+     246     577     577   R2-       1     331     942         Overlap    3021
R5+     339     577     577   R1-       1     238     848         Overlap    2473
R5+     327     577     577   R3-       1     250     861         Overlap    1871
ASSEM done
ComputDistForContigs done
```

```
PresentLayout done
Clip R2 left clip: 55, right clip: 996,  length: 1022,  right size 26
Clip R3 left clip: 55, right clip: 915,  length: 918,  right size 3
Clip R1 left clip: 12, right clip: 859,  length: 860,  right size 1
Clip R4 left clip: 18, right clip: 799,  length: 799,  right size 0
Clip R5 left clip: 52, right clip: 628,  length: 920,  right size 292
Clip R6 left clip: 152, right clip: 827,  length: 850,  right size 23
```

6. cap3.out：拼接的结果文件

```
Number of segment pairs = 30; number of pairwise comparisons = 15
'+' means given segment; '-' means reverse complement

Overlaps          Containments  No. of Constraints Supporting Overlap

******************** Contig 1 ********************
R2+
                 R3+ is in R2+
                 R1+ is in R3+
R4-
                 R5- is in R4-
                 R6- is in R4-

DETAILED DISPLAY OF CONTIGS
******************** Contig 1 ********************
                   .    :    .    :    .    :    .    :    .    :    .    :
R2+             AGTTTTAGTTTTCCTCTGAAGCAAGCACACCTTCCCTTTCCCGTCTGTCTATCCATCCCT
R3+             AGTTTTAGTTTTCCTCTGAAGCAAGCACACCTTCCCTTTCCCGTCTGTCTATCCATCCCT
R1+             AGTTTTAGTTTTCCTCTGAAGCAAGCACACCTTCCCTTTCCCGTCTGTCTATCCATCCCT
                ____________________________________________________________
consensus        AGTTTTAGTTTTCCTCTGAAGCAAGCACACCTTCCCTTTCCCGTCTGTCTATCCATCCCT

… …
                     .    :    .    :    .    :    .    :    .    :    .    :
R4-             ATATTATAT-ACATATCACATT
R6-             ATATTATATTACATATCACATT
                ____________________________________________________________
consensus        ATATTATATTACATATCACATT
```

参数

```
Options (default values):
  -a  N  specify band expansion size N > 10 (20)
  -b  N  specify base quality cutoff for differences N > 15 (20)
  -c  N  specify base quality cutoff for clipping N > 5 (12)
  -d  N  specify max qscore sum at differences N > 20 (200)
  -e  N  specify clearance between no. of diff N > 10 (30)
  -f  N  specify max gap length in any overlap N > 1 (20)
  -g  N  specify gap penalty factor N > 0 (6)
  -h  N  specify max overhang percent length N > 2 (20)
  -m  N  specify match score factor N > 0 (2)
  -n  N  specify mismatch score factor N < 0 (-5)
  -o  N  specify overlap length cutoff > 20 (40)
  -p  N  specify overlap percent identity cutoff N > 65 (80)
  -r  N  specify reverse orientation value N >= 0 (1)
  -s  N  specify overlap similarity score cutoff N > 400 (900)
  -t  N  specify max number of word matches N > 30 (300)
  -u  N  specify min number of constraints for correction N > 0 (3)
  -v  N  specify min number of constraints for linking N > 0 (2)
  -w  N  specify file name for clipping information (none)
  -x  N  specify prefix string for output file names (cap)
  -y  N  specify clipping range N > 5 (250)
  -z  N  specify min no. of good reads at clip pos N > 0 (3)
```

实例

练习

参考文献

Huang, X. and Madan, A. (1999) CAP3: A DNA Sequence Assembly Program. Genome Research, 9: 868-877.

2.6 Consed

简介

Consed 是一款非常强大的图形化 finish 软件，由 David Gordon 等人于 1998 年发布，目前已更新至 15.0 版本。

现在 consed 已经成为基因组 finish 的标准工具，它为组装正确性的验证提供了一个直观的界面，能够方便的进行组装的各项统计并绘图，对结果进行比对分析，并能实现对组装结果进行拆分、重组等等功能。同时还可以通过峰图的比较来查找或者验证 SNP。

该软件需要在支持图形界面的 X-win32 环境下操作，软件的使用需要获得作者的授权。

下载

Consed 软件需要到 phrap 网站申请，申请成功后下载相应操作系统的版本，如 consed_linux.tar.z。

申请地址：

http://bozeman.mbt.washington.edu/consed/consed.html#howToGet

安装

1．将软件包上传到大型机上

2．解压缩 zcat consed_linux.tar.Z | tar -xvf -

3．环境变量配置

1）默认 CONSE_HOME 为/usr/local/genome，如果不使用这个目录，请建立相关链接，并修改环境变量设置（.cshrc 或其他 shell 的配置文件）：setenv CONSED_HOME xxx，xxx 为 consed 安装的目录。

2）建立 $CONSED_HOME/bin 和 $CONSED_HOME/lib 目录，可执行文件全部放到 $CONSED_HOME/bin 目录下

Consed 需要使用其他的一些软件，如：phred，phrap，crossmatch，这些文件需放到 /usr/local/genome/bin 目录下，或$CONSED_HOME/bin。

对于软件 phred，联系：bge@u.washington.edu (Brent Ewing)

对于软件 phrap 和 crossmatch， 联系：phg@u.washington.edu (Phil Green)

3） 编译 phd2fasta:

到 misc/phd2fasta 目录，键入命令'make'编译 phd2fasta，然后将 phd2fasta 可执行文件移到目录 /usr/local/genome/bin 或 $CONSED_HOME/bin)

4) 编译 mktrace:

到 misc/mktrace 目录，键入命令'make'编译 mktrace，然后将 mktrace 可执行文件移到目录 /usr/local/genome/bin 或 $CONSED_HOME/bin)

5) 将所有的 perl 程序（scripts 目录和 contributions 目录下）移到目录 /usr/local/genome/bin 或 $CONSED_HOME/bin)，并修改权限为可执行(chmod a+x *)

6）如果系统 perl 不是安装在/usr/bin/下，需将每个 perl 程序的开头位置 #!/usr/bin/perl -w 改成相应的路径。

7) 建立子目录 /usr/local/genome/lib/screenLibs 或 $CONSED_HOME/lib/screenLibs 将目录 misc 下的文件 primerCloneScreen.seq 和 primerSubcloneScreen.seq 拷到此目录下。

8)建立载体序列文件（FASTA 格式）：

/usr/local/genome/lib/screenLibs/vector.seq

(或$CONSED_HOME/lib/screenLibs/vector.seq。此文件包含所有载体序列。

9)建立重复序列文件（FASTA 格式）：

/usr/local/genome/lib/screenLibs/repeats.fasta，

(或 $CONSED_HOME/lib/screenLibs/repeats.fasta)。如果不想标注任何重复序列，将 phredPhrap 相关的行屏蔽掉即可（行前加＃号），即：

```
!system( "$tagRepeats $szAceFileToBeProduced" )
  || die "some problem running $tagRepeats";
```

改为：

```
#!system( "$tagRepeats $szAceFileToBeProduced" )
#  || die "some problem running $tagRepeats";
```

输入

Consed 的输入文件是 phrap 组装生成的*.ace 文件和组装用到的 reads 的 phd、峰图文件。这些文件必须以如下方式存放：

一个存放峰图文件的目录，目录名必须是 chromat_dir；

一个存放 phred 读取峰图输出的文件——phd 文件的目录，目录名必须是 phd_dir；

一个供 consed 编辑的工作目录，目录名任意（通常命名为 edit_dir），里面存放 ace 文件。

三个目录必须同级放置。如：

```
[liudy@119 bash /disk2/team06/liudy/test/test_consed]$ls -lFt
total 72
drwxr-xr-x    2 liudy    prj0327     4096 Sep 22 02:01 edit_dir/
drwxr-xr-x    2 liudy    prj0327    20480 Sep 18 03:21 phd_dir/
```

```
drwxr-xr-x    2 liudy    prj0327    16384 Sep 18 03:21 chromat_dir/
```

使用

满足上述输入条件以后，在目录“edit_dir”下直接键入"consed"即可运行程序，程序打开以后会弹出一个选择输入的 ace 文件的窗口：

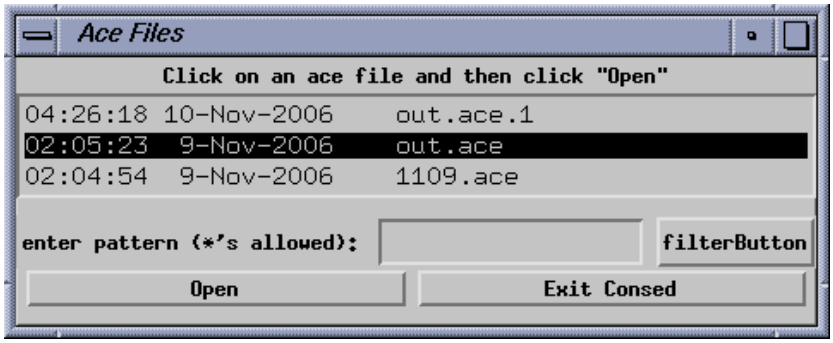

图 2-5 consed 的输入选择界面

如果 phd_dir 目录缺失却需要强行打开 consed，必须加"-nophd"参数运行才能打开 consed 界面，否则会报错退出。而在"-nophd"参数下，consed 的很多功能都无法实现，包括查看每个 read 的质量、调整组装结果等等。而如果 chromat_dir 缺失，则不能查看 reads 的原始峰图。通常运行 consed 的时候都要求至少绝大多数 reads 的 phd 文件都存在。

以下的所有功能的实现都是在 consed 目录结构完整，reads 路径对应正确，并且参数配备无误的情况进行的。

1．主界面布局：

主界面"Consed Main Window" 从上到下依次排列了菜单区、功能键、contig 列表和 read 列表。

Contig 列表中的所有 contigs 按照包含 reads 从少到多的顺序排列。窗口中显示了 contig 名称、拼成 contig 的 reads 数和 contig 的总长度等信息。

Read 列表中显示了每一个 read 在拼接结果中属于哪一个 contig、read 长度和在 contig 上的拼接位置。

Contig 列表和 read 列表的下方分别有一个搜索区，可以输入 contig 或者 read 的名称进行模糊，搜索区支持模糊搜索的功能。

图 2-6 就是 consed 的主界面：

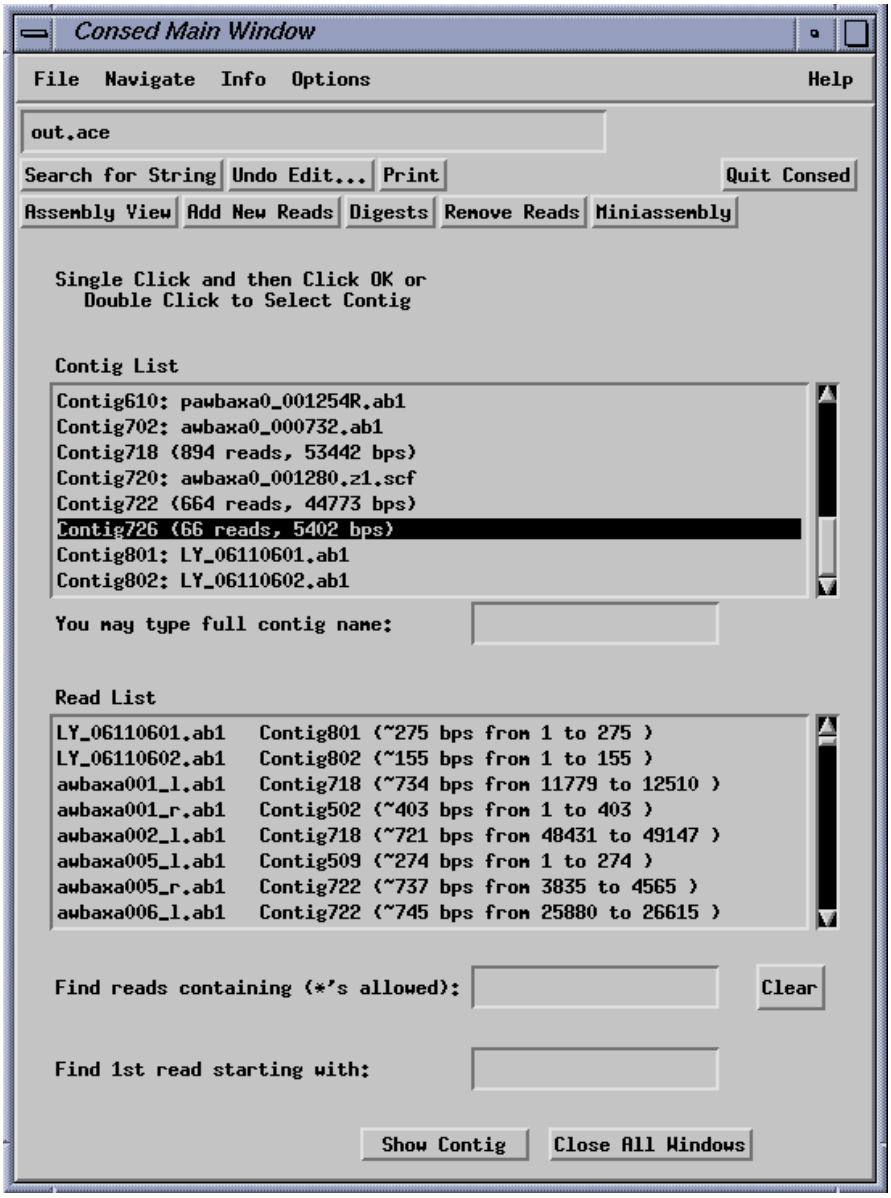

图 2-6 consed 的主界面

2．检查 contig 的组装质量：

在 contig 列表中双击一个 contig 的名字，会弹出这个 contig 的窗口。窗口中以图形的方式显示了此 contig 的组装情况。最上面一行的碱基表示组装完成的 contig 序列（consensus），下面的每一行表示组成 contig 的每一条 read，在窗口的左端显示了每一条 read 的名字，名字后面的箭头代表 read 的测序方向。拼接质量是由碱基的背景色表示的，背景色浅表示质量好，反之表示质量差。通过拖动滚动条，可以查看到整个 contig 的拼接情况。如果需要查看某一个 read 的峰图，只需选中这个 read 上的碱基点击鼠标中键，就会弹出峰图（双键鼠标可以通过同时点击左右键来实现中键功能）。如图 2-7：

图 2-7 contig 窗口和 reads 峰图

对于比较大的 contig，手动检查的效率是很低的，所以 consed 提供了一系列统计以辅助检查 contig 的拼接：

第一是提供了 contig 的平均单碱基错误率统计，以衡量 contig 的整体质量。这个信息显示在 contig 窗口按键区"Err/10kb"的右边。如上图显示就是万分之 3.38 的错误率。

第二是提供了查找 contig 上组装有问题区域的功能。点击"navigate"按钮，下拉菜单中有很多查找选项，其中第一个选项"Low Cons/High Qual Descrep/Single Stranded/Single Subclone/Unaligned High"选项，即查找全部有问题的组装区域。相比于这种一网打尽的找法，分类寻找往往更有针对性，所以最常用的是如下选项："Low consensus quality"、"Region covered by only 1 subclone"和"High quality discrepancies/>5bp from unaligned region"，即低质量、单覆盖和高质量错配。

以查找低质量区为例，依次点击"navigate"->"Low consensus quality"，会弹出一个窗口显示所有低于指定质量值（默认为 25）的区域，双击其中的任意一个结果，contig 窗口就会显示这个位置附近的组装情况。点击"save"按键，弹出窗口显示的统计结果可以保存。如图 2-8：

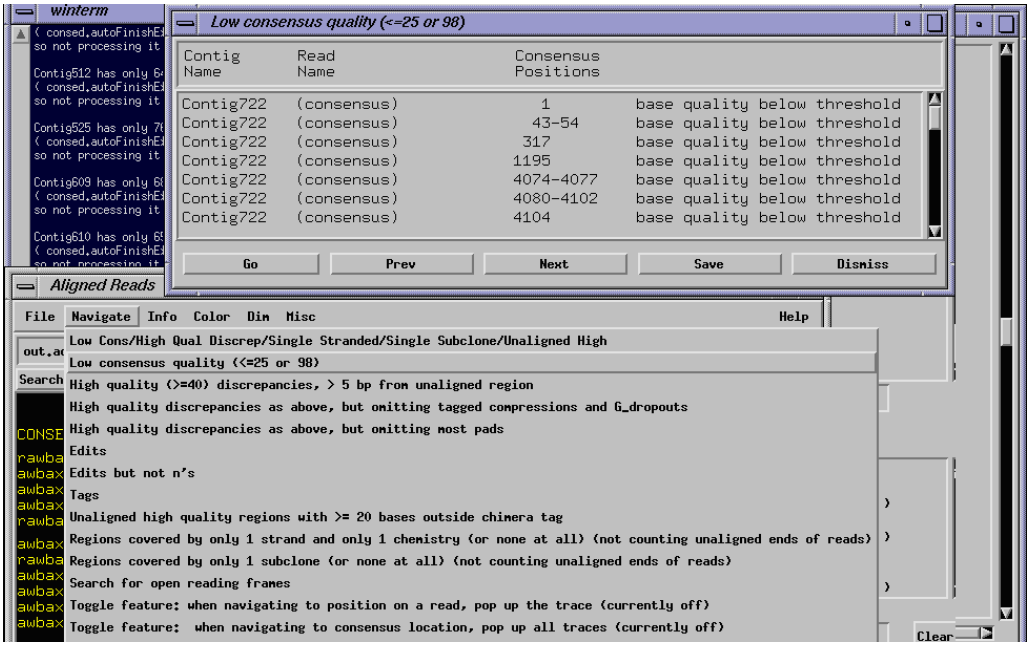

图 2-8 寻找 contig 的低质量区

3．提取组成 contig 的所有 reads 的位置信息：

在 contig 窗口上点击"Info"按钮，选择"Show Contig Information"，就会弹出"Contig Information"窗口，显示所有 reads 在这个 contig 上的位置和方向。可以点击"Save"输出这些信息。如图 2-9

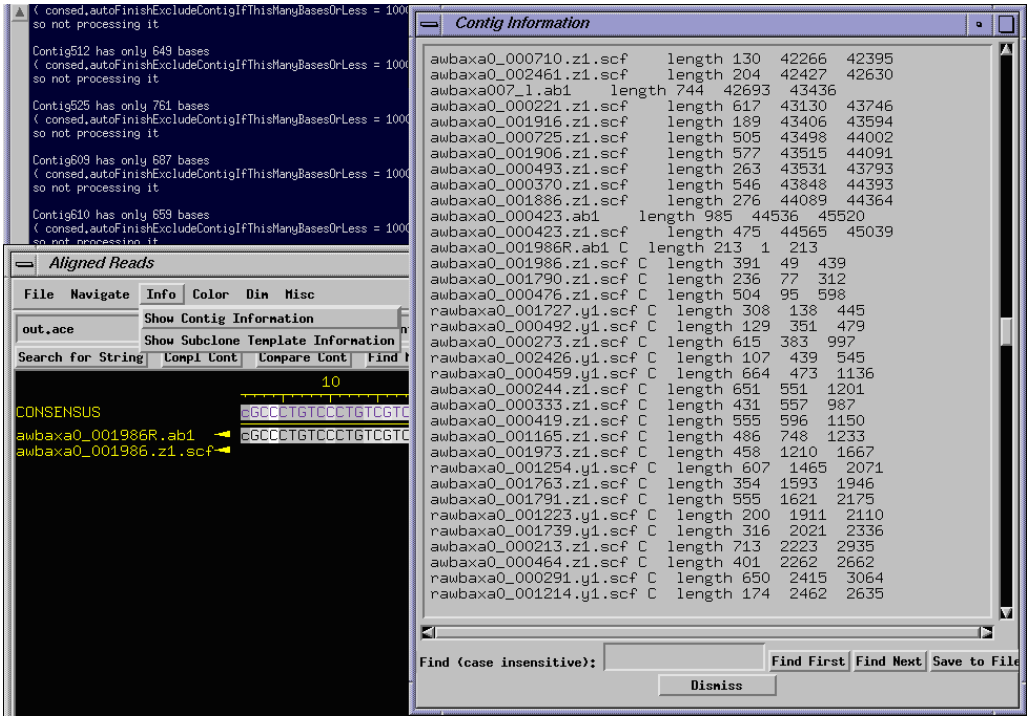

图 2-9 查看 contig 上 reads 的位置

4．查看 contig 之间的关系和正反向 reads 的覆盖情况：

在主窗口上点击按钮"Assembly View"会弹出一个窗口显示 contig 之间的正反向 reads 关系，并将关系足够多的正反向连成 scaffold。在 contig 的上方会出现两条起伏的线，较高的一条是浅绿色，表示亚克隆的覆盖度曲线；较低的一条是深绿色，表示组装的 reads 覆盖度

曲线。这两条曲线突然降低的位置往往是组装结果中连接较弱的位置，甚至是错拼。因此这两条曲线能够用来粗略的检验序列组装的可靠性。如图 2-10：

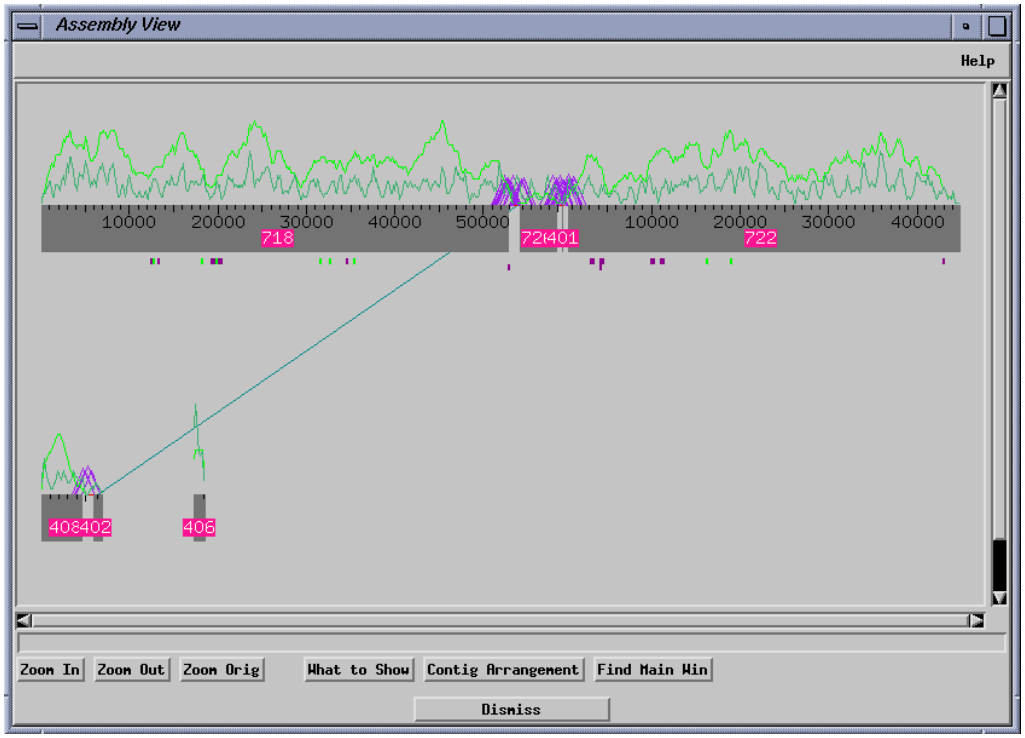

图 2-10 Assembly View

如果想仔细观察正反向的覆盖情况，可以点击"Assembly view"窗口的"What to Show"，在菜单中选择"Fwd/Rev Pairs"，选中正反像选项中的"Show each consistent fwd/rev pair within contigs"和"Show legs on squares for consistent fwd/rev pairs"并点击"Apply"，就会在显示 contigs 之间的关系的同时也显示 contigs 内部的正反向关系，能够比较方便的找到正反向覆盖异常的区域。

5．寻找组装结果中的重复区：

在"Assembly View"窗口点击"Sequence Matches"，会弹出 cross_match 比对的参数选项窗口。点击"run crossmatch"，程序会在所有的 contigs 之间进行比对，并把比对结果绘制在"Assembly View"窗口里面的 contig 上，其中橙色线条代表正向比对的结果，黑色代表反向比对。如图 2-11：

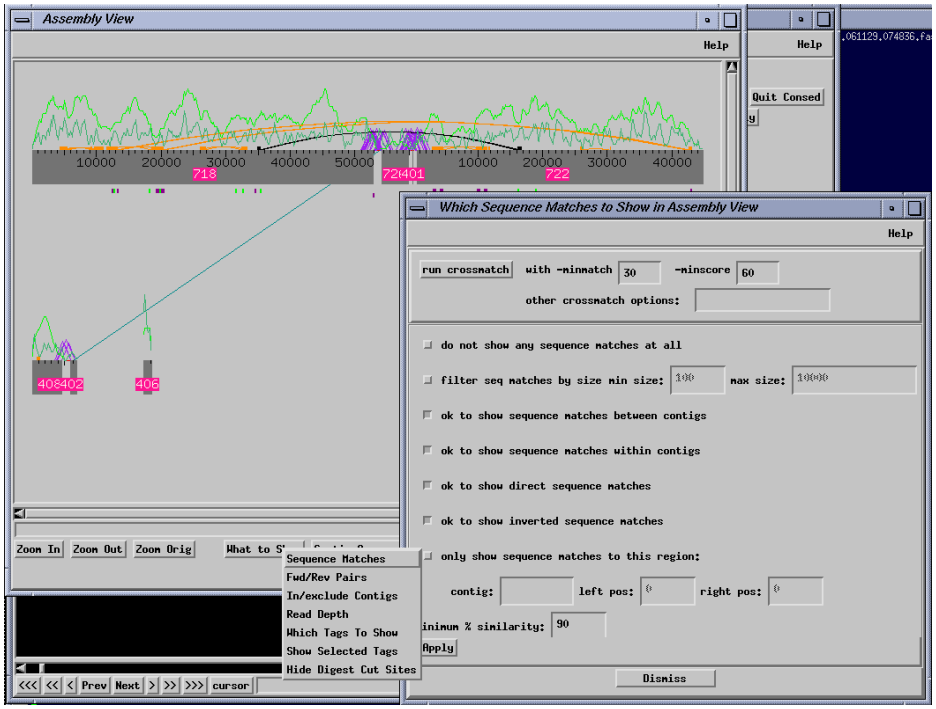

图 2-11 Assembly View 的比对功能

6. 在 consed 中搜索序列：

打开"Search for String"窗口，从一个 contig 中选中一段序列（consed 设置为选中复制），用鼠标中键粘贴在"Query String"内（也可以键盘输入），然后点击"OK"，程序就会找出这一段序列在所有结果中出现的位置。如图 2-12：

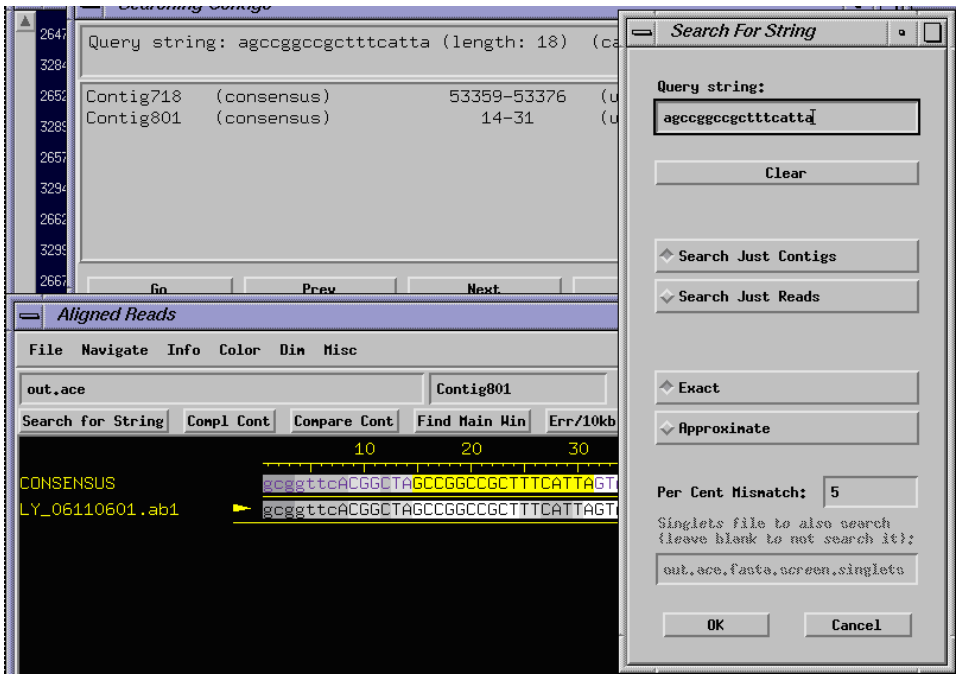

图 2-12 搜索序列

7. 连接 contigs：

对于有重复区域的两个 contigs，我们可以把鼠标的焦点定在两个 contig 重复区域的同一个碱基上，在两个 contig 窗口里分别点击"Compare Cont"弹出比对窗口。点击窗口中间

的"Align"比对。查验比对结果没有问题可以接受以后，点击比对窗口右下角的"join Contigs"，两个 contigs 就连起来了，如图 2-13 和 2-14。需要注意的是，如果两个 contigs 是反向比对，则必须用按钮“Compl Cont”把其中一个 contig 变成互补序列，才能进行连接。

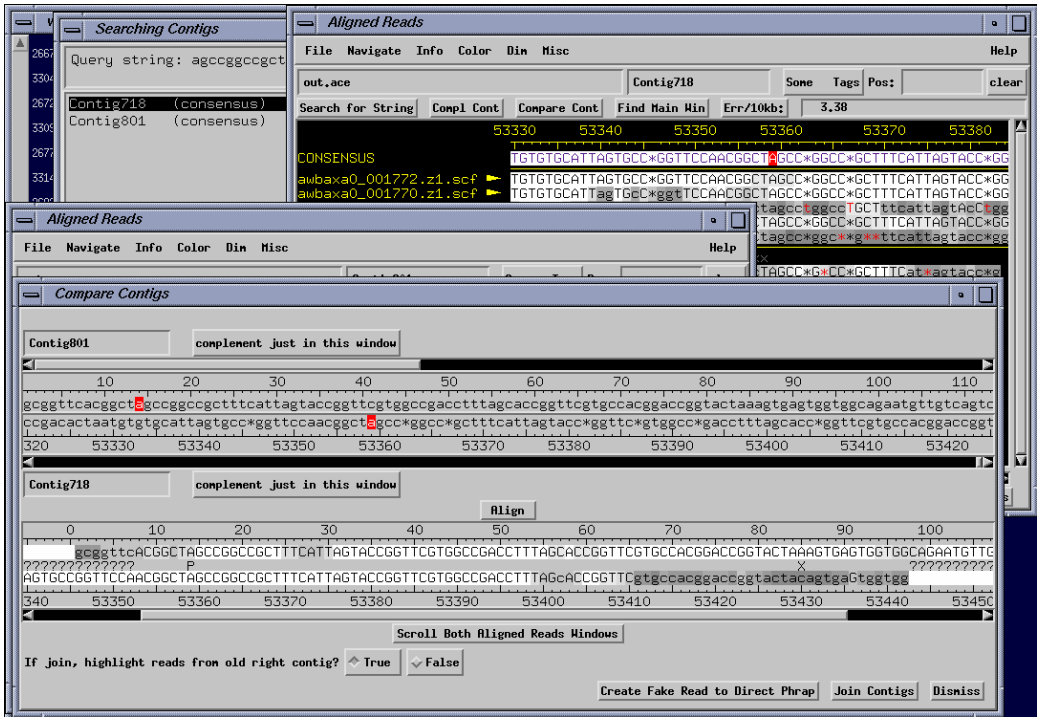

图 2-13 连接 contigs

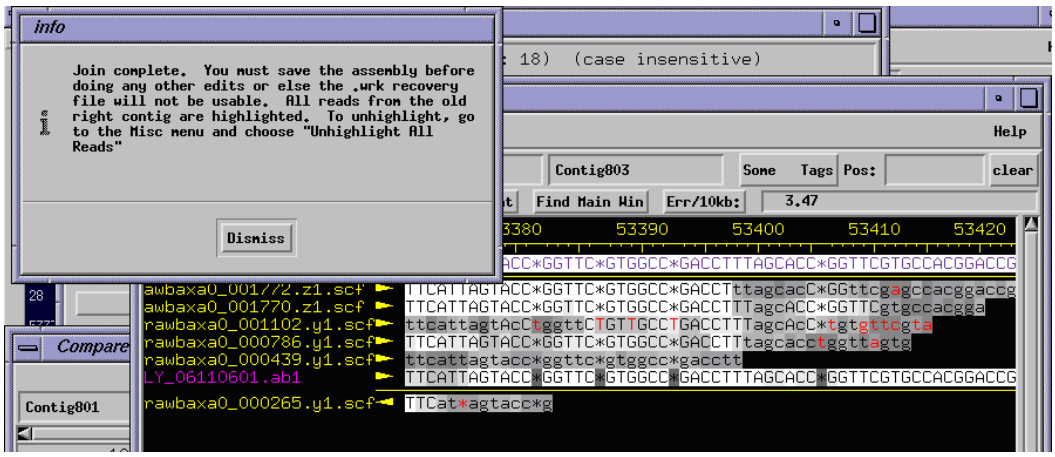

图 2-14 连接以后的 contig

8. 拆分 contig:

在 contig 窗口里选中选一个位置按右键，选择"Tear contig at this consensus position"，就会弹出一个窗口以供选择跨过这一碱基的每一个 reads 应该划分到上游还是下游。选定之后点击"Do Tear"，原来的 contig 就拆成了 2 个。如果 2-15 和 2-16

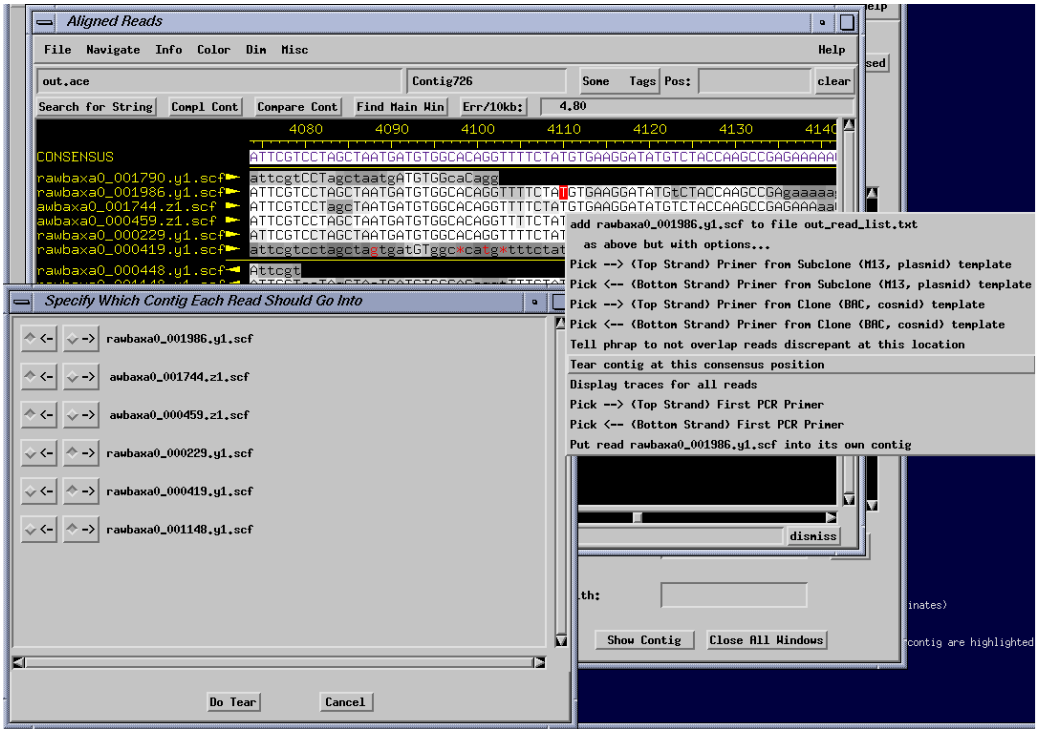

图 2-15 拆分 contig

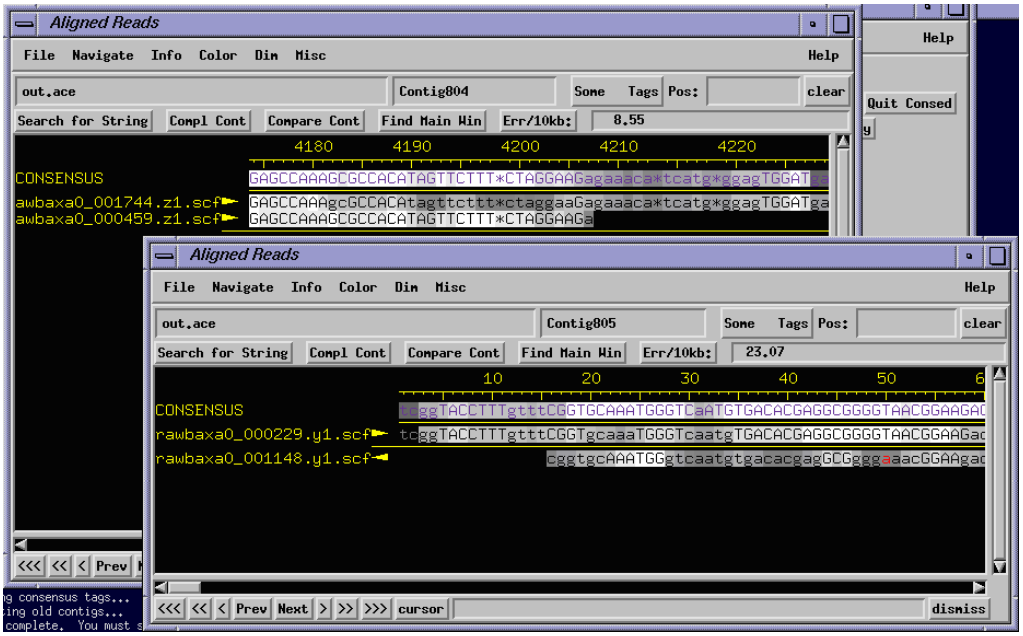

图 2-16 拆分后的 contigs

9．把一个 read 从 contig 中分离出来：

在 contig 窗口中选中需要分出来的 read，点鼠标右键，选择"Put read *** into its own contig"，即可把这条 read 从中分离出来。如图 2-17 和 2-18：

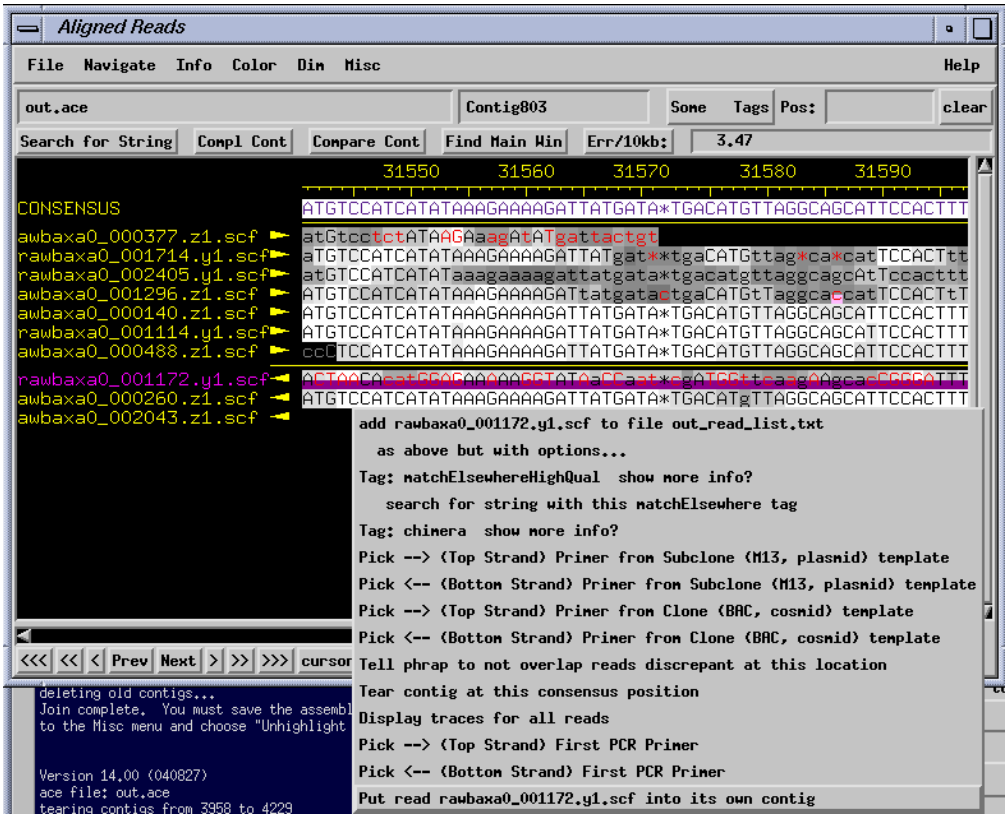

图 2-17 从 contig 中分离 reads

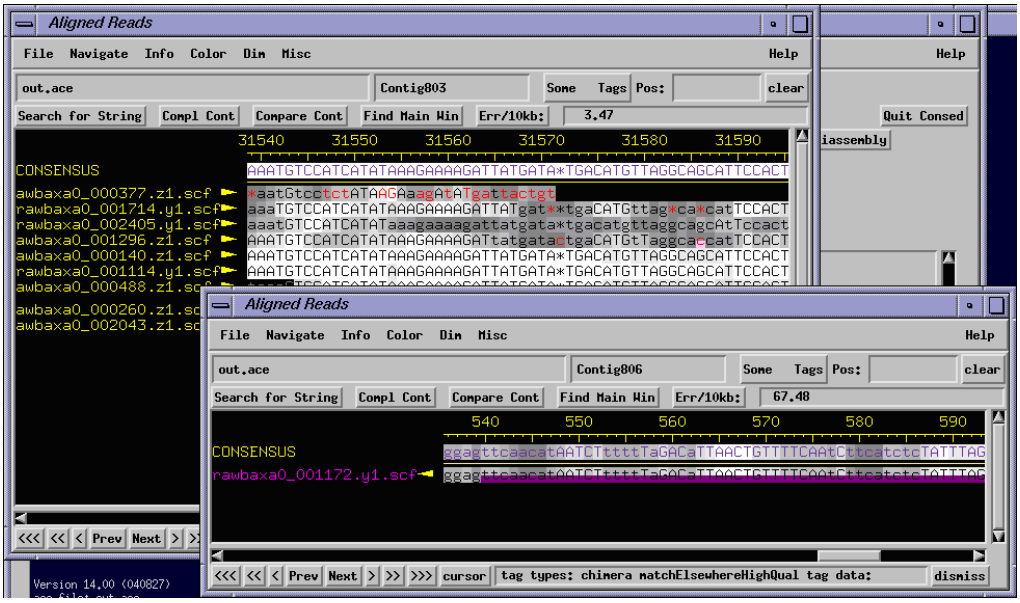

图 2-18 分离出来的 read 单独成为一个 contig

以上是一些常用的基本功能，其他的扩展功能读者可以慢慢摸索。需要注意的是，以上的功能都是在参数配备完整的情况下实现的。如果 consed 实现某一功能的调用程序路径不对，会弹出类似于这样的错误窗口：

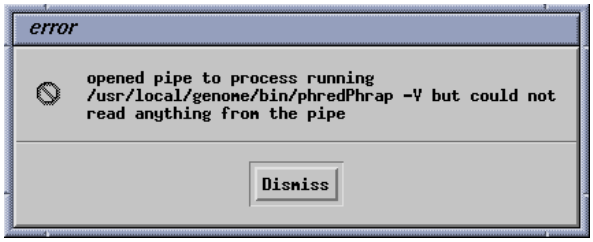

图 2-19 错误 1

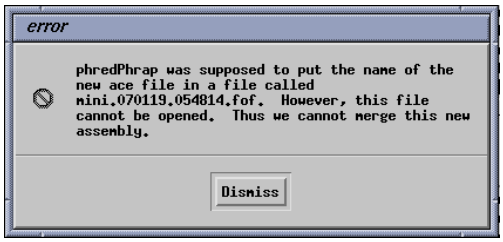

图 2-20 错误 2

遇到这种情况的需要重新配置 consed 的参数调用列表，方法如图 2-21，在主界面上点击"Options"，选择"Edit Consed/Autofinish Parameters"，把报错的调用程序路径修改为当前系统内的有效路径即可。使用 consed 时多数配置问题可以通过这种方法解决。

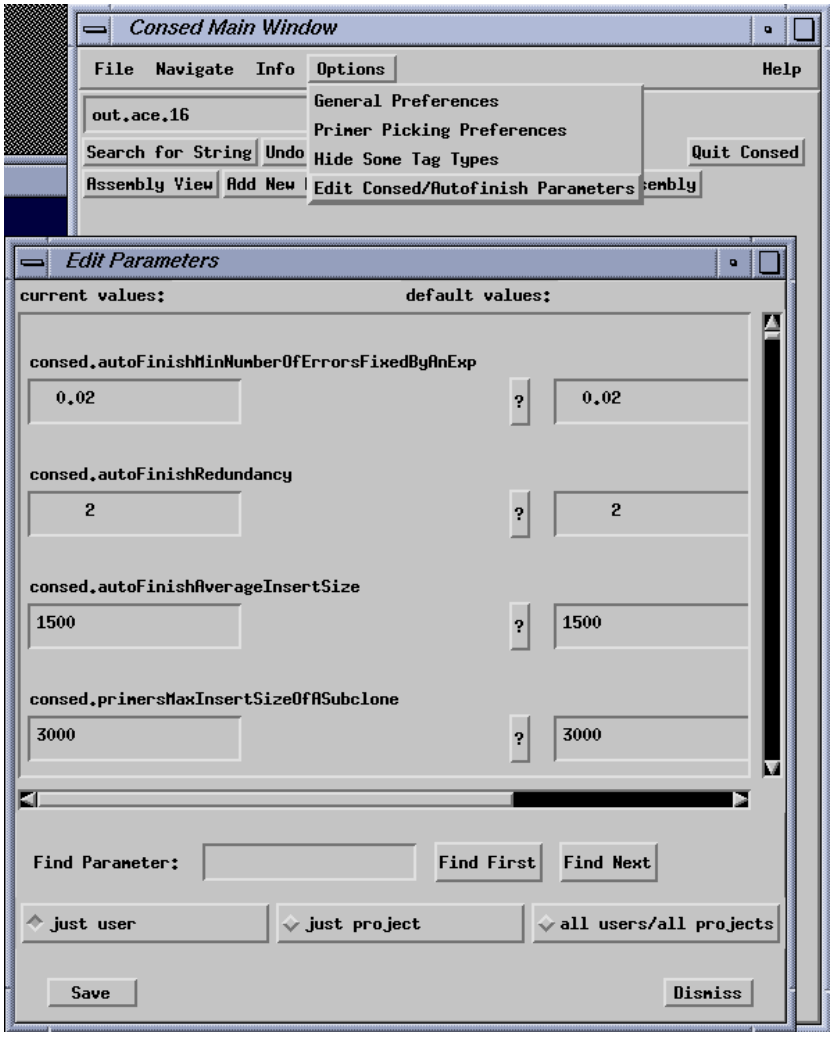

图 2-21 调整 consed 参数

输出

1．保存 ace 文件：

点击主窗口的“File”按钮，在菜单中选择"Save assembly"选项，可以用来保存修改后的 ace 文件。见图 2-22

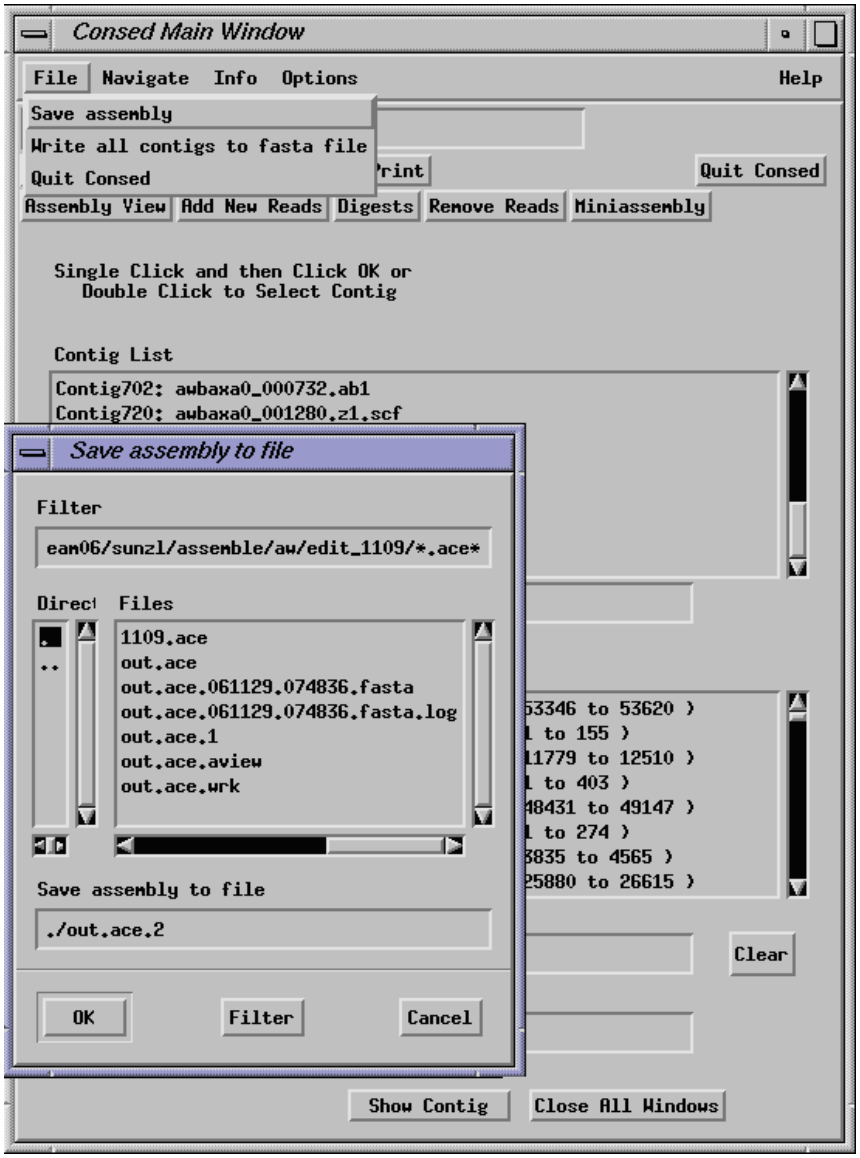

图 2-22 保存 ace 文件

2．输出 contigs 序列：

点击主界面的“File”，选择"Write all contigs to fasta file"可以输出所有 contigs。如果需要单独输出某一个 contig，可以在相应的 contig 窗口内点击"File"，选择 "Export consensus sequence" 或者 "Export consensus sequence (with options)"来指定输出完整 contig 还是部分序列、输出起止位点、是否输出质量、输出格式是 fasta 还是 phd 等等。如图 2-23：

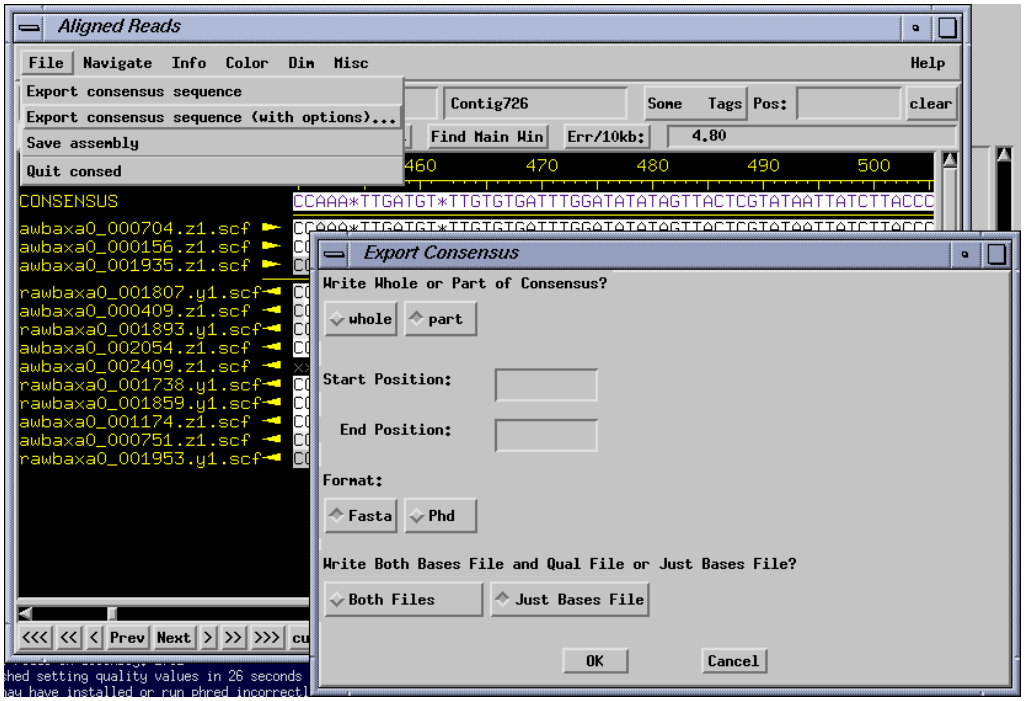

图 2-23 输出 contig 序列

常见问题

1．运行 consed 时报下列错误：

```
no ~/.consedrc file so no user resources will be used--that's ok
no ./.consedrc file so no project-specific resources--that's ok
couldn't open readOrder.txt--that's ok
Error: Can't open display:
```

这种情况通常是使用的远程登陆工具不支持图形界面。使用 X-win32 登陆即可解决。

2．运行 consed 时报下列错误：

```
no ~/.consedrc file so no user resources will be used--that's ok
no ./.consedrc file so no project-specific resources--that's ok
couldn't open readOrder.txt--that's ok
Fatal: The parent directory must contain phd_dir and chromat_dir, but it doesn't.
A typical directory structure is a directory named after the project, with
subdirectories named edit_dir (containing the ace files), phd_dir (containg the phd
files), and chromat_dir (containing the chromatogram files). Consed would then be
run from within edit_dir.
Version 14.00 (040827)
```

这是由于上级目录没有“phd_dir”。

练习

1．对一个组装结果进行调整，连接原来未连接起来的 contigs，并统计每个 contig 的平均单碱基错误率，对已知排列顺序的 contig 设计引物，输出 contig 序列。

数据存放在：

光盘:\consed\example\test.ace

参考文献

Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. Genome Res, 1998, 8(3):195-202

2.7 Primer3

简介

Primer3 是一款以命令行形式运行的引物设计软件。它源于引物设计程序 primer0.5，由 Steve Lincoln、Mark Daly 和 Eric Lander 开发。

Primer3 的功能是做 PCR 引物设计，它能够比较严格的控制引物发夹结构和二聚体。命令行的操作方式也使 primer3 很容易嵌入流程，适合做大规模的 PCR 引物设计。

下载

下载地址：

http://sourceforge.net/project/showfiles.php?group_id=112461，下载包：primer3-1.1.0-beta.tar.gz，以往版本下载地址：

http://fokker.wi.mit.edu/primer3/old_releases.html。

安装

```
$ unzip primer3_1.0.1.tar.gz  ——解压gz包
   $ tar xvf primer3_1.0.1.tar    ——解压 tar 包
   $ cd primer3_1.0.1/src         ——转到安装操作目录
   $ make all                     ——生成文件
```

输入参数

Primer3 的参数全部写在输入文件里，具体如下：

PRIMER_SEQUENCE_ID=引物的名字

SEQUENCE=要设计引物的序列

TARGET= 指定一个位置及长度作为标靶，引物对必须在它的两侧

PRIMER_MIN_SIZE=引物最小长度

PRIMER_OPT_SIZE=引物最适长度

PRIMER_MAX_SIZE=引物最大长度

PRIMER_MIN_TM=引物最小退火温度

PRIMER_OPT_TM=引物最适退火温度

PRIMER_MAX_TM=引物最大退火温度

PRIMER_MAX_GC=引物最大 GC 含量

PRIMER_MIN_GC=引物最小 GC 含量

PRIMER_PRODUCT_SIZE_RANGE=产物长度范围（格式：min-max）

PRIMER_PRODUCT_OPT_SIZE=最适产物长度

PRIMER_NUM_RETURN=返回的引物数量,默认为 5

PRIMER_FILE_FLAG=是否输出过程文件(推荐值 0,不输出)