BIOINFORMATICS

Lecture 2

Linux Basics & Data Management

Today's Lecture

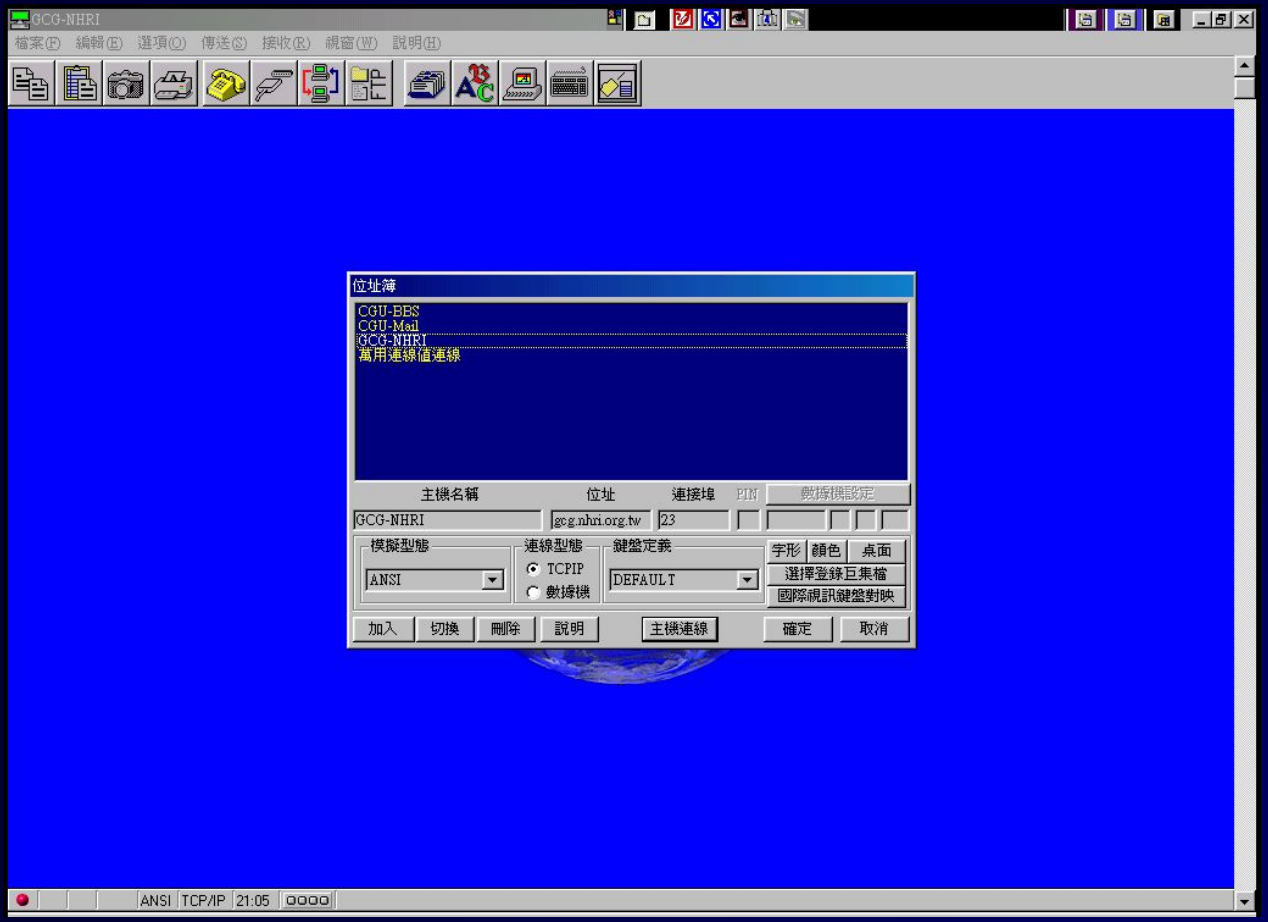
Unix
Unix commands

遠端載入 (Telnet) to GCG
PCMAN

檔案傳輸 (File Transfer Protocol－FTP)
FileZilla

UNIX SYSTEM

UNIX 是個多人多工作業系統。另外，UNIX 多種不同之版本如 SunOS、Linux 等等。但各系統之基本操作指令大同小異。



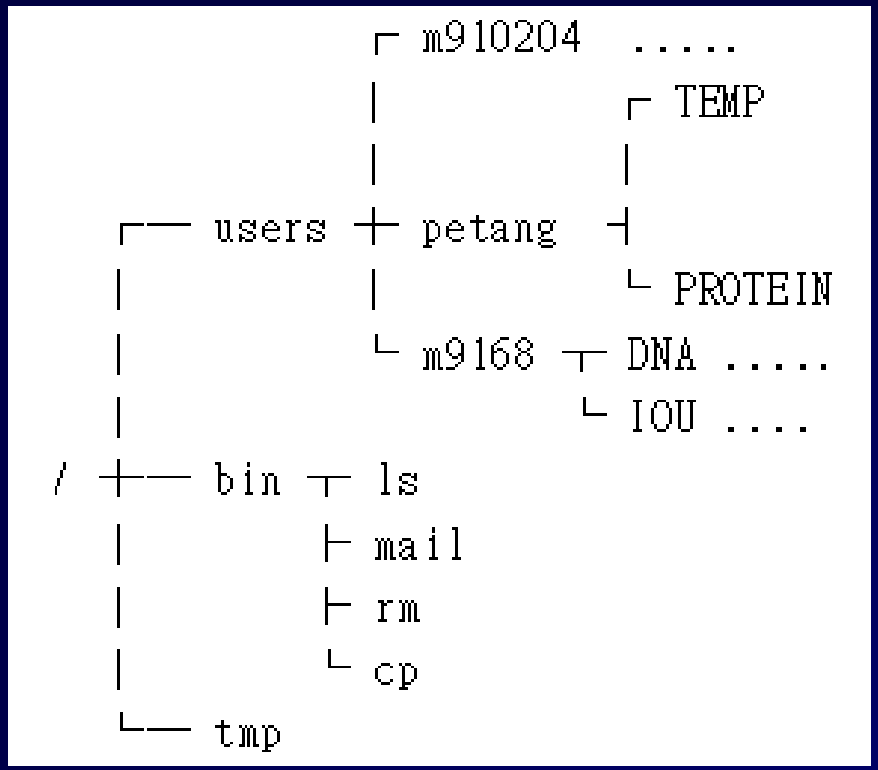
GCG UNIX 作業系統操作簡介-1

檔案、目錄、與檔案系統

- 每一個使用者在 UNIX 下面均有一個 HOME directory ，做為這個使用者個人資料的存放處。使用者 login 時，系統管理者設定進入工作目錄的位置。

- UNIX 使用一個和 DOS 相似的 single root 樹狀結構檔案系統，由根目錄(/)開始一層層將子目錄建下去，各階層目錄以 / 隔開。

- 最頂層稱為根目錄(root)，包含許多目錄，每一個目錄底下均存放許多的檔案與子目錄。



GCG UNIX 作業系統操作簡介-2

指令格式

指令 [參數] [處理對象]
gcg2% [command] [options] [file]
例：ls –l aminoacid.dat
　　除了列出檔名外，並列出檔案屬性及擁有者、檔案大小及建立時間等資訊

- 在 UNIX 之下英文字母的大小寫是有差別的，命令一般是小寫字串。
  本課程建議所有"根目錄"以大寫命名。其它檔案全部用小寫。
- 選項通常以減號(-)再加上一個或數個字元表示，
  用來選擇一個命令的不同操作
- 同一行可下數個命令，命令間應以分號隔開。
- 線上求助指令 -- man　可在線上用來查詢各種命令用法(manual page)的指令
  例：man ls　　查詢 ls 這個指令的用法

GCG UNIX 作業系統操作簡介-3

Download

基本指令

ls	list files
more	display contents of file on screen page by page
cat	display contents the whole file
cp	make a new copy of a file
rm	remove (delete) a file
mv	rename (move) a file
mkdir	make a new directory
rmdir	remove a directory

GCG UNIX 基本指令 - ls

ls 的意義為 "list"，也就是列出目錄內檔案名稱（如 DOS 的 dir/w）

- 如果你在下 ls 指令後頭沒有跟著任何的檔名，
  它將會顯示出目前目錄中所有檔案。
- 也可以在 ls 後面加上所要察看的目錄名稱或檔案的名稱，如

```
gcg2% ls filename
gcg3% ls ROOT
```

ls -[參數]

ls -l 除了列出檔名外，並列出檔案屬性及擁有者、檔案大小及建立時間等資訊
ls -a 列出所有檔案，包括隱藏檔
ls -R 地列出所有子目錄內所有檔案

ls -[參數] [filename 萬用字元(wildcard)]

萬用字元有三種：

*	代表任意長度的字串
?	代表任何的單一字元
[...]	代表區間內定義的任一字元

c* 代表cat ，cp ，cal ...任意長度且開頭為c 。
ca? 代表cat，cal ...三個字且前二字為 ca。
ca[a-z] 代表caa，cab，caz ... 三個字且頭二字為 ca 第三字為 a-z。

GCG UNIX 基本指令 – cat/file

以上二個指令均為察看檔案內容的指令。

- cat [filename]

是 concatenate ，在字典上的意思是 “連結,將…串成鎖狀” 。就是把檔案的內容顯示出來的意思。
問題在於 cat 它是不會停下來的，因此並不好用(試想如果一個螢幕二十四行，而一個檔案四百行，cat 一出來將會劈哩啪啦不斷的捲上去，使用者很難據此得到他們所需的資訊。
cat 可配合萬用字元使用。

- more [filename]

可以將所觀察的檔案跟據終端機的形態一頁頁的顯示出來，
再根據使用者的要求換頁或捲行。
如果你在使用中覺得已經看到了所要看的部份，
可以按 'q‘ 離開 more 的使用。

GCG UNIX 基本指令 - cp

這個指令的意義是複製 ("COPY") ,
也就是將一個或多個檔案複製成另一個檔案或者是將其複製到另一個目錄。

- cp 的用法如下:

  cp f1 f2 : 將檔名為 f1 的檔案複製一份為檔名為 f2 的檔案。
  cp f1 f2 f3 ... dir : 將檔案 f1 f2 f3 ... 都以相同的檔名複製一份放到目錄 dir 裡面。
  cp -r dir1 dir2 : 將 dir1 的全部內容全部複製到 dir2 裡面。

- cp -[參數]:

  -i : 此參數是當已有檔名為 f2 的檔案時，若逕自使用 cp 將會將原來 f2 的內容掩蓋過去，因此在要蓋過之前必須先詢問使用者一下。如使用者的回答是y(yes)才執行複製的動作。

  -r : 此參數是用來做遞迴複製用，可將一整顆子樹都複製到另一個目錄中。

GCG UNIX 基本指令 - mv

mv 的意義為 move , 主要是將一檔案改名或換至另一個目錄。
如同 cp ，它也有三種格式:

mv f1 f2 : 將檔名為 f1 的檔案變更成檔名為 f2 的檔案。
mv dir1 dir2 : 將檔名為 dir1 的目錄變更成檔名為 dir2 的目錄。
mv f1 f2 f3 ... dir : 將檔案 f1 f2 f3 ... 都移至目錄 dir 裡面。

- mv-[參數]

-i：此參數的意義與 cp 中的相同，均是 interactive 詢問之意。
-f：此參數為強迫(force) , 就是不管有沒有同名的檔案，反正我就是要搬過去，所有其他的參數遇到 -f 均會失效。

GCG UNIX 基本指令 - rm

rm 的意義是 remove ，也就是用來殺掉一個檔案的指令。

在 UNIX 中一個被殺掉的檔案除非是系統恰好有做備份，否則是無法像 DOS 裡面一樣還能夠救回來的。所以在做 rm 動作的時要特別小心。

rm 的格式如下:
rm f1 f2 f3 .....

- rm-[參數]
  - -f : 將會使得系統在刪除時，不提出任何警告訊息。
  - -i : 在除去檔案之前均會詢問是否真要除去。
  - -r : 遞迴式的刪除。

GCG UNIX 基本指令 – mkdir/rmdir

- mkdir 是一個讓使用者建立一個目錄的指令。

你可以在一個目錄底下使用 midir 建立一個子目錄，使用的方法如下:

mkdir dirname1 [dirname2 ...]

如此你就可以建立一個或多個目錄

- 相對於 mkdir ，rmdir 是用來將一個"空的"目錄殺掉的。如果一個目錄下面沒有任何檔案，你就可以用 rmdir 指令將其除去。rmdir 的使用法如下:

rmdir dirname1 [dirname2 ....]

如果一個目錄底下有其他的檔案， rmdir 將無法將這個目錄殺掉，除非使用 rm 指令的 -r 選項。

GCG UNIX 終端機常用控制鍵

Ctrl-C　中斷程式的執行。

Ctrl-Z　暫停程式的執行，稍後可下 fg 或 bg 指令繼續，若未下 fg 或 bg 指令繼續執行，該 process 仍會留在系統內。

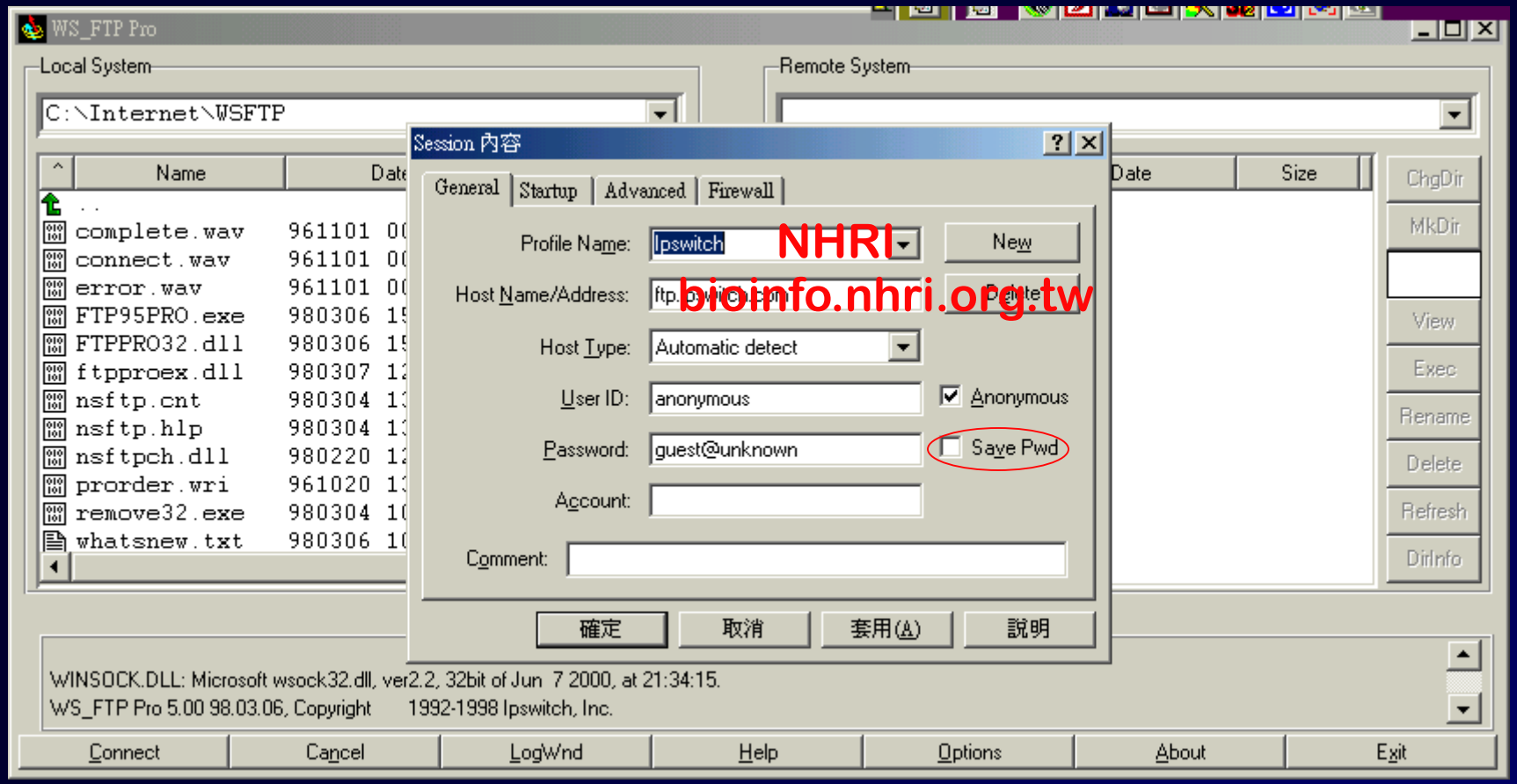
啟動 **PCMAN**，選擇「地址簿」，輸入下列設定，按加入

主机名稱：GCG-NHRI　位址：bioinfo.nhri.org.tw　連接埠：23
模擬型態：ANSI　連線型態：TCPIP

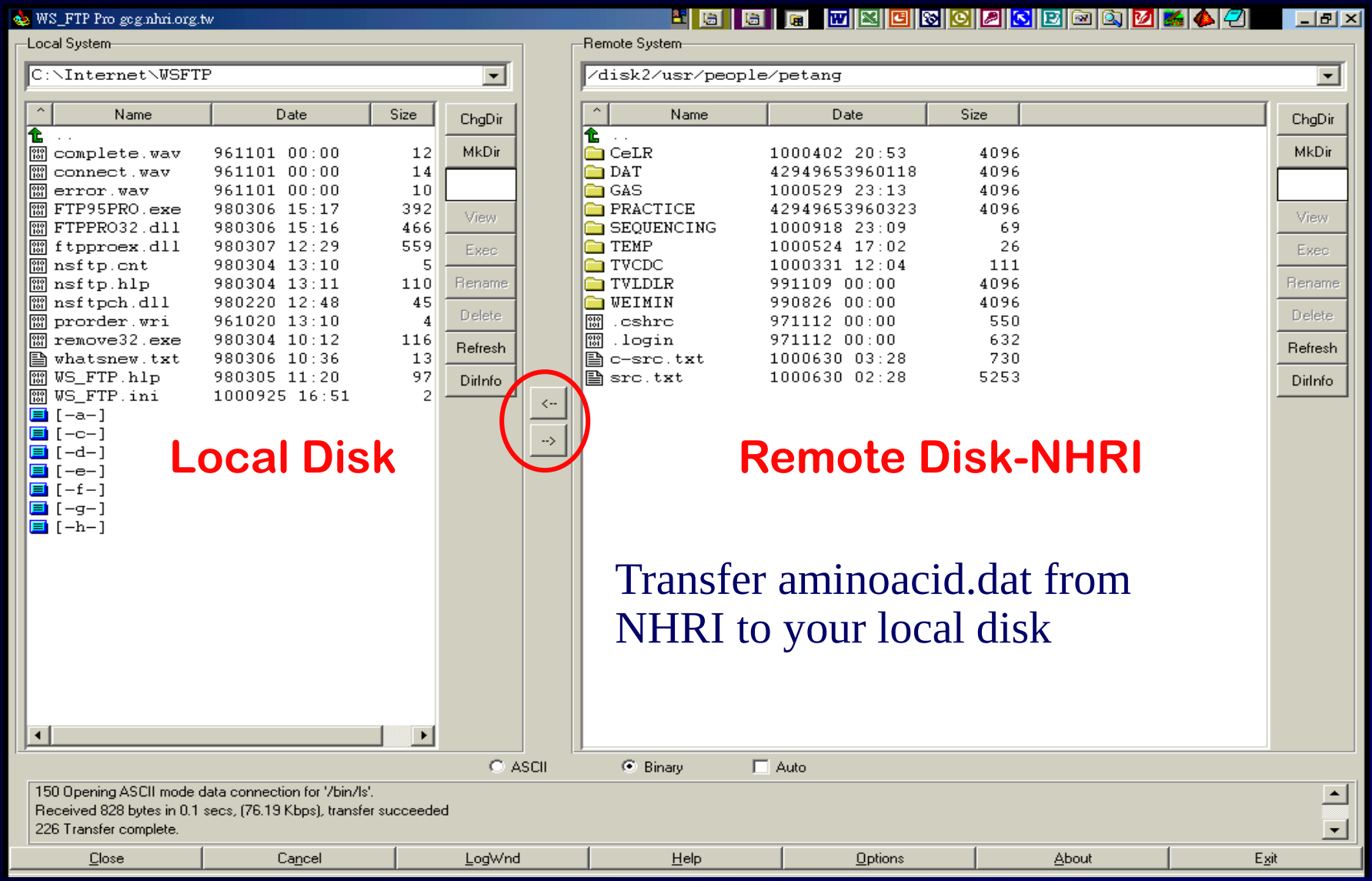


Exercise02

File Transfer Protocol (FTP)



File Transfer Protocol (FTP)

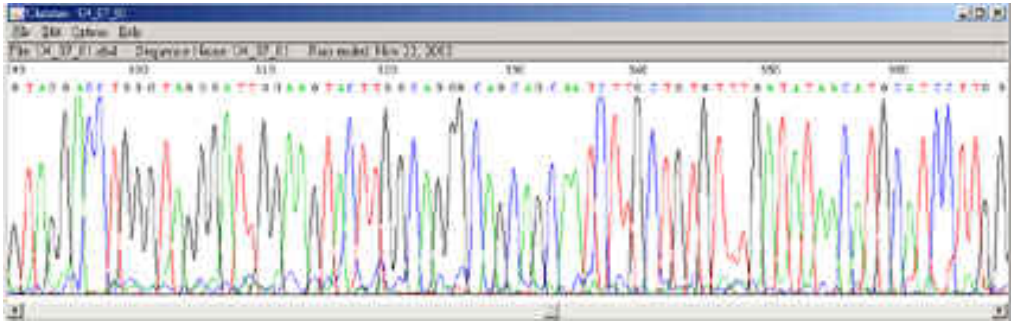


EXERCISE

Exercise 02-1 Linux Commands
Exercise 02-2 File Transfer via FTP
Exercise 02-3 Sequence Manipulation in GCG UNIX
Exercise 02-4 Sequence Manipulation in GCG UNIX : LEGUMAIN

ASSIGNMENT 1

Softwares & Sequence Formats



Program	Formats Default	Accept	Multiple sequence
WWW	text file	paste & Copy	
SeqWEB	text file	paste & copy	
GCG	● GCG file	FASTA ● GenBANK ● ● EMBL ● ● Staden SwissProt	Multiple sequence file (msf) Rich sequence file (rsf) List files (lst)
VectorNTI			
CLC Genomics			

Importing and Exporting

You need a FTP program to transfer files between your PC and GCG.
The sequence file must be in “plain text” format.

chopup: converts a non-GCG format sequence file containing lines longer than 511 characters and as long as 32,000 characterters into a new file containing no longer than 50 characters.

breakup: reads a non-GCG format sequence file containing more than 350,000 sequence characterters and writes it as a set of separate, shorter, overlapping sequence files than can be analyzed by GCG.

reformat: rewrites sequence files, scoring matrix files, or enzyme data files so than they can be read by GCG programs.

fromfasta: reformats one or more sequences from FastA format into single sequence files in GCG format.

Exercise 02-2

(A) Transfer sequence files from your PC to GCG
(B) Chopup the sequence
(C) Reformat the sequence
(D) Edit the sequence

Create a folder “BIO” in your hard disk
Start WsFTP (ftp://bioinfo.nhri.org.tw)
Upload “naq.txt” & “psq.txt” to GCG
Start Netterm
Start GCG
Chopup “naq.txt” & “psq.txt”
Reformat “naq.dat” or “psq.dat”
Cat “naq.txt” or “psq.txt”

Exercise 02-3

Sequence Manipulation in GCG UNIX

Use the database searching techniques you learned today to retrieve the reference sequence

Homo sapiens LEGUMAIN

and the amino acid sequence of

ALL LEGUMAIN

From NCBI
And then transfer the sequence(s) to
GCG Unix (in GCG format)

There are many different ways to DO it.
You can have your lunch now if you can make it.

ASSIGNMENT 1.

Use the Entrez searching techniques you learned today to retrieve the
Reference sequence and
the corresponding amino acid sequences of

All the subclasses of *Homo sapiens* cyclophilin

Transfer the sequences to GCG Unix,
Transform the sequences to GCG format

E-mail
1. The steps (including URL of WWW sites) you used and
2. The sequences in GCG format as attached file to
petang@mail.cgu.edu.tw before next Thursday 1200
****郵件主旨： ASS1 bioinfo – (學號)