Introduction to Bioinformatics

The unix/linux system and basic commands

What is a shell?

- Logging in Linux -$\rightarrow$ an internal program starts (The shell)
- The shell interprets commands the user types
- The shell is the bridge between you and the computer!!
- The shell make you to be connected with the Kernell!!
- There are many types of "shells" tcsh, korn, bash etc.
- We will use a shell called Bash!

Connecting to the Linux system

- Open the terminal.
- Now you could use the shell. My path

jose@biolinuxJoxcar[jose] → My path

The user   The host   The current directory
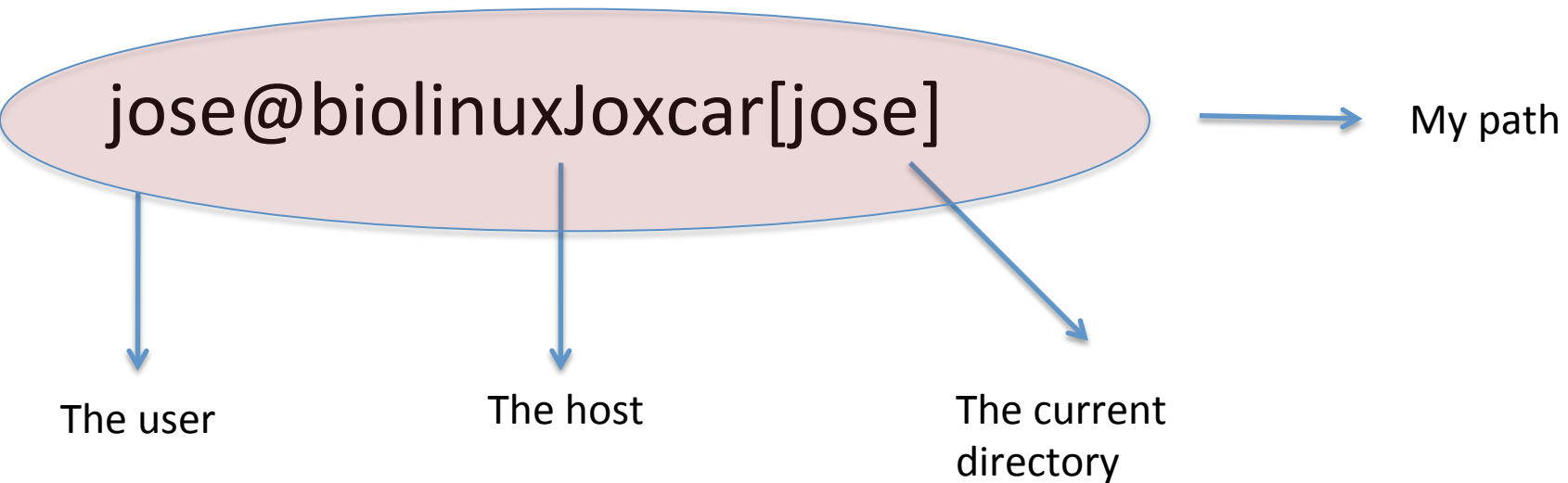
Moving in linux …

- ls
- List the directory contents
- ls –a -> list all
- ls –l -> long listing format

File permissions

- Each file in Unix/Linux has an associated permission level

- This allows the user to prevent others from reading/writing/executing their files or directories

- Use "ls -l *filename*" to find the permission level of that file

Permission levels

- "r" means "read only" permission
- "w" means "write" permission
- "x" means "execute" permission

Moving

- pwd -> shows the current directory
- cd -> moves from one directory to a new one
- cd mydirectory -> moves to the next directory
- cd .. -> moves to the previous directory, one step back
- cd ~/ -> moves to the home directory

Making and removing directories

mkdir

- mkdir mydir -> Create a directory mydir

- rmdir mydir -> removes the directory (the directory has to be empty!!)

- rm –R mydir -> Removes everything!!! CAUTION!!!!

Copy and move

- cp   copy
- cp file1 file2 -> copies file1 to file2
- cp –R mydir1 mydir2 -> copies recursively, everything will be copied in mydir2

- mv   move
- mv file1 file2 -> you will rename the file

Deleting files

- rm   remove

- rm file1 file2 -> it will delete file1 and file1

- Protect yourself rm –I -> It will ask for confirmation

File content handling

- cat   displays a file or concatenates files
- cat file1 -> It displays the content of the file
- cat file1 file2 -> it concatenates file1 and file2
- cat file1 file2 > file3 -> > it concatenates file1 and file2 in file3

- IMPORTANT : when you use ">", this will direct your out put to a new file
- Example :
- ls –l > list1 -> This lists the files in the file list1

More or less

- more file1 -> Displays the file in segments
- less file1 -> Displays the file in segments and you could move back and forward

Other useful tools

- head file1 -> first of the file

- tail file1 -> last lines of the file

- wc –l file1 -> Counts lines (-l), words (-w) and characters (-m) of file1
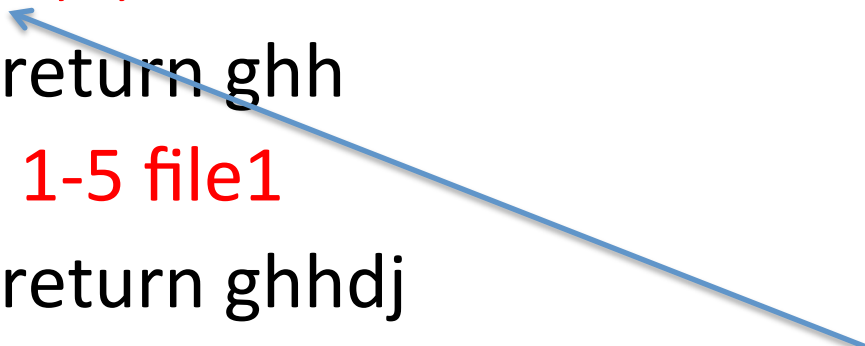
More commands

- **echo** -> repeats or shows the input
- **clear** -> clear the screen
- **whoami** -> returns your username
- **ps** -> shows the jobs that are running in your system
- **kill PID** -> ends the process with the ID you give
- **passwd** -> lets you change your password
- **man X** -> shows the usage of the X command
- **whatis** -> shows a quick reference of X command
- **which X** -> shows where the file X is located
- **man ifconfig** -> linux manual

- Many of these commands can be used with
- THE WILDCARD *
- THE OUTPUT to new file >

How to create a file

- Text editors:
- Nano, emacs, pico, vi

Text manipulation in UNIX

- cut -> Extracts various fields of data from a data file.
- Example:
- In the file1:
- ghhdjskhd
- cut –c 1,2,3 file1
- It will return ghh
- cut -c 1-5 file1
- It will return ghhdj

Delimiter –c character

Using cut to extract columns

- File2 is:
- One two three
- Four five six
- Seven eight nine

- cut –d ' ' –f 1 file2
- one
- Four
- seven

- cut –d ' ' –f1,3 file
- One three
- Four six
- Seven nine

Fields –d delimiter

cat to concatenate files

- cat file1 file2 file3 > new_file -> This concatenates all the files in a new one
- File1+file2+file3 in file new_file

Sort files

- sort -> sorts the characters of a file
- sort file1 -> sorts the characters in alphabetical order
- sort –r file1 -> sorts the characters in alphabetical order (reverse order)
- sort –n file2 -> sorts the characters in order (this is for numbers)

Replacing with sed

- sed 's/pattern/replace/g' file1

- sed 's/plant/animal/g' list

- Replace all plant for animal

grep to search and extract

- grep 'your word' file1
- This show your word in the document
- grep –n 'your word' file1
- This show your word in the document with the line number
- grep –n '^your word' file1
- Your word at the beginning of the line
- grep –n 'your word$; file1
- Your word at the end of the line

More search with grep

- grep 'your word .' file1 -> this will show your word plus any other character in file1
- grep '[0-9]' file1 -> this shows lines containing numeric data in file1

Piping in UNIX

- Most of the commands could be linked in only one using the connector "|"

Re-directing in Unix

- The output of most of the commands could be re-directed using ">"

Search and extract with grep

- grep 'your word' > newfile
- This searches for your word and print it in newfile
- grep –w your word > newfile
- This searches for your word (whole word) and print it in newfile

Create your folders

- In your home, move to your desktop or documents

- Create the folder lab1

- Move the files for the first lab to this folder

Linux for bioinformatics
Exercise 1: *S. scabies* gene expression

- *S. scabies* is a plant pathogen that causes necrosis in potatoes. Most of the virulence factors are located in a region of the chromosome that is called the genomic island PAISsc1. Gene expression analysis indicate that some genes are expressed at high levels during the infection. Column 1 the gene_id, column2 log2 fold change expression, column, position in the island (1, in the island, 2 not in the island)

Questions

- Using unix commands list the scabies expression file.

- Sort the scabies expression file by the level of expression (second column)

- Create a file that contains the top 10 highly expressed genes

- Using the new highly expressed gene file, list genes that are not locate in the island (column 3 equals 0)

- Create a new file that contains only the gene names

Linux for bioinformatics
Exercise 2 : *P. syringae* proteome

- *P. syringae* pv tomato DC3000 is a model pathogen for probing disease susceptibility

- The genome is composed by One chromosome and two plasmids

- Files chro_PS.faa, Plasmid1_PS.faa and Plasmid2_PS.faa contain the whole predicted set of proteins in PS (the proteome).

Questions

- Using Linux commands list the files containing the proteome
- What do you see?, describe the files?
- Using Linux commands create a single file that contains the whole proteome of Ps.
- Change the name *Pseudomonas syringae* to *P. syringae*
- Type III effectors are important virulent factors in Ps. List all the genes associated with this name
- How many predicted proteins are in the files?
- Pectin lyases are important in virulence. How many Pectin lyases are in the proteome?