

Introduction to UNIX environment and command line basics

Stefano Ghignone

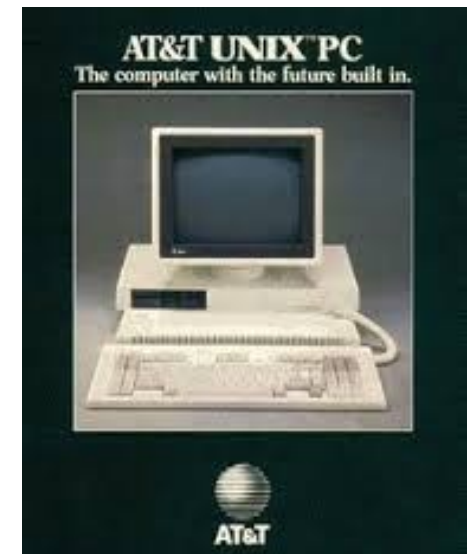
**Institute for Sustainable Plant Protection – Turin Unit
CNR (Italy)**

Why Unix?

The **Unix** operating system (OS) is popular in bioinformatics because of its powerful command-line tools that make scripting and performing automated analyses relatively easy.



- Unics = Uniplexed Information and Computing System
 - Developed in the 1960's by MIT and AT&T Bell Laboratory
 - Unix was widely adopted by the academic community in the 1970s
-
- The first computers all used command line interfaces (no GUIs)
 - Unix is the operating system of choice for servers since it is highly stable and very secure
 - Unix provided one of the first powerful computing system for clients to log into servers and run programs
 - Unix has been developed into Linux, Solaris, Mac OSX and many other flavors



Why Unix?

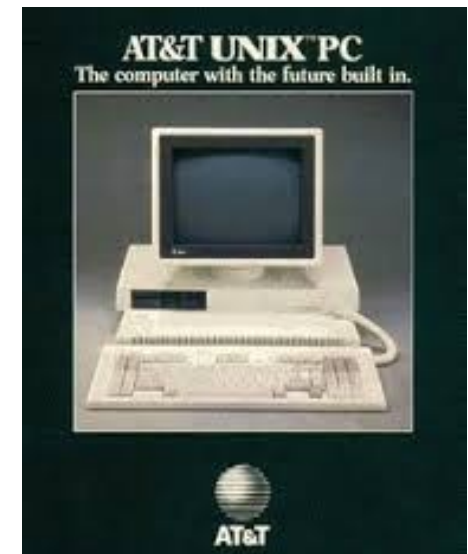
The **Unix** operating system (OS) is popular in bioinformatics because of its powerful command-line tools that make scripting and performing automated analyses relatively easy.



- Unics = Uniplexed Information and Computing System
- Developed in the 1960's by MIT and AT&T Bell Laboratory

Most scientific computing is done in UNIX.

- The first computers all used command line interfaces (no GUIs)
- Unix is the operating system of choice for servers since it is highly stable and very secure
- Unix provided one of the first powerful computing system for clients to log into servers and run programs
- Unix has been developed into Linux, Solaris, Mac OSX and many other flavors



Unix and High-Performance Computing

- Most academic universities have High-Performance Computing Clusters (HPCCs) that allow users to upload data and run programs, allowing **thousands of processors** to be used for a single program
- Parallel Computing = computationally intensive programs that normally take weeks to run can often be broken down into **thousands of small jobs** and run in parallel in a few hours
- HPCCs are often used for the alignment and variant detection of **Next-Generation Sequencing** data

Next-generation sequencing (NGS), also known as high-throughput sequencing, is the catch-all term used to describe a number of different modern sequencing technologies including:

- Illumina (Solexa) sequencing
- Roche 454 sequencing
- Ion torrent: Proton / PGM sequencing
- SOLiD sequencing

These recent technologies allow us to sequence DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing, and as such have revolutionized the study of genomics and molecular biology.

UNIX

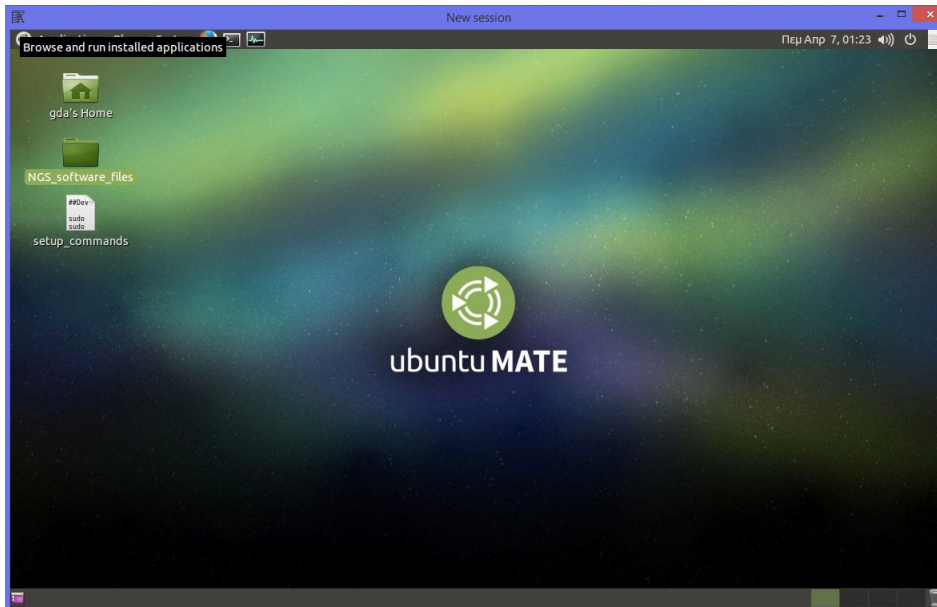
- Linux (Ubuntu, Debian, etc)
- Solaris
- OS X

All have a similar underlying system, and a similar set of command line tools.

UNIX

- Linux (Ubuntu, Debian, etc)
- Solaris
- OS X

Desktop



Terminal

```
gda@snf-703996: ~
stefano@CNR-GHIGNONE:~$ ssh gda@snf-703996.vm.okeanos.grnet.gr
gda@snf-703996.vm.okeanos.grnet.gr's password:
Welcome to Ubuntu 14.04.4 LTS (GNU/Linux 4.2.0-34-generic x86_64)

 * Documentation:  https://help.ubuntu.com/

10 packages can be updated.
10 updates are security updates.

Last login: Thu Apr  7 11:46:01 2016 from cnr-ghignone.bioveg.unito.it
gda@snf-703996:~$
```

For PC download and install '**Putty**' SSH client

<http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html>

Unix environment



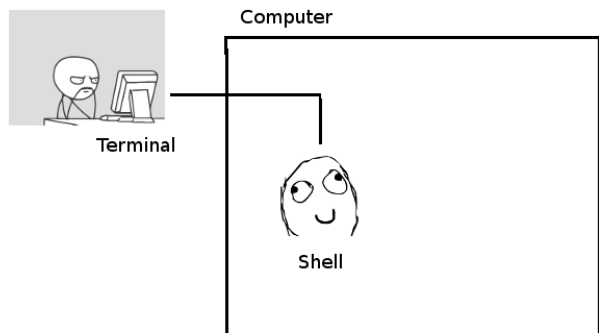
Originally a terminal was a machine like a typewriter.

Many terminals could be plugged into one computer, allowing many people to use it at once.

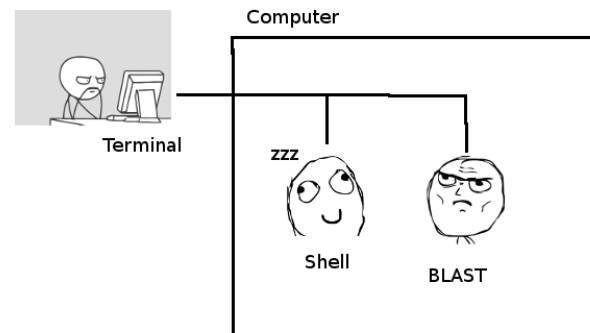
UNIX operating system allows multiple programs to run at once on the computer.

The terminal connects you to a "shell" program which lets you run other programs. Usually the shell program is "bash".

SHELL – is a terminal that allows users to enter commands that instruct the UNIX kernel to run programs

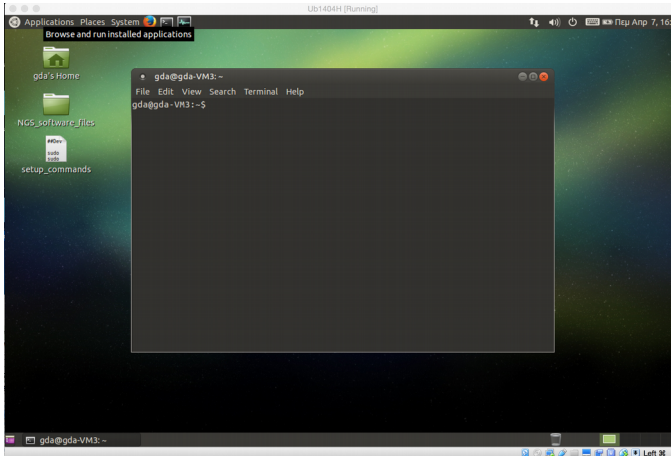


The terminal connects to the computer, and initially talks to a shell program.



When the shell starts up another program it connects it to your terminal as well. The shell goes to sleep until the program finishes.

Using commands



Unix commands are written in the terminal

First example: the **date** command

The current time and date is shown on the **standard output**

Command outputs can be redirected to a not-existing file using **>** (*or overwrites !*) or appended to an existing file with **>>** .

Command syntax (i)

\$ command [options] [files]

Basically, we type the name of the **command**, followed (if needed) by some **parameters/options** that modify the behavior of the command and finally (if needed), the names of the **files/directories** that the command needs to use/change/manipulate. Space is needed between them. Finally, we press ENTER.

One command may take more than one parameters/flags

Usually, command names are shortcuts for English verbs (list = ls, change directory = cd, Print working directory -> pwd, etc...)

Command syntax (ii)

In the same command line, more than one command can be executed, one after the other, using the **;** as separator (e.g.: **date ; pwd**)

The output of one command may be used as input in the next command, with the pipe symbol: **|** (e.g.: **Command1 | command2 > final_results_file**)

Use the **tab** key to *auto-complete* the name of a file/directory. Just type enough words of the file/directory, so that linux can understand what you are referring to.

The history command allows to see which commands has been typed before. (Type: **history**)

Pressing the **up arrow** key, the terminal will show the previous commands executed.

In UNIX there are two types of wildcards that can be used in filenames, to indicate/select multiple files (e.g.: **rm ***):

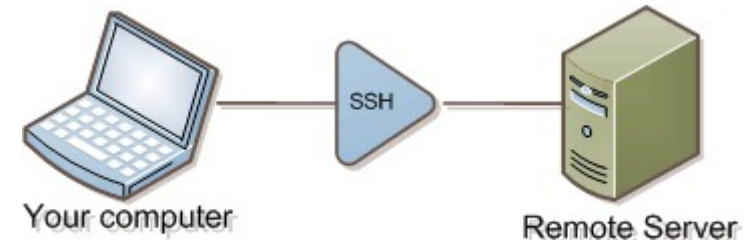
- The ***** wildcard will search for any number of characters before/after the star
- The **?** wildcard will only search for a single character

To find information about a command, use **man** (manual) (e.g. **man ls**)

Login and...what's happening?

Connection to a remote machine

Managed by the ssh (**S**ecure **S**hell) program



\$ ssh username@hostname.unito.it

"Securely connect me to a shell on a remote machine"

```

stefano@GANDALF: ~
stefano@CNR-GHIGNONE:~$ ssh stefano@130.192.101.115
stefano@130.192.101.115's password:
Welcome to Ubuntu 14.04.4 LTS (GNU/Linux 3.19.0-25-generic x86_64)

 * Documentation:  https://help.ubuntu.com/

System information as of Tue Jul  5 15:17:52 CEST 2016

System load:   16.07           Users logged in:   2
Usage of /home: 55.4% of 7.15TB IP address for em1: 130.192.101.115
Memory usage:   2%            IP address for em2: 192.168.1.2
Swap usage:     0%            IP address for lxcbr0: 10.0.3.1
Processes:      698

Graph this data and manage this system at:
  https://landscape.canonical.com/

7 packages can be updated.
7 updates are security updates.

Last login: Tue Jul  5 15:17:52 2016 from gandalf.unito.it
stefano@GANDALF:~$
  
```

w/who show who's connected

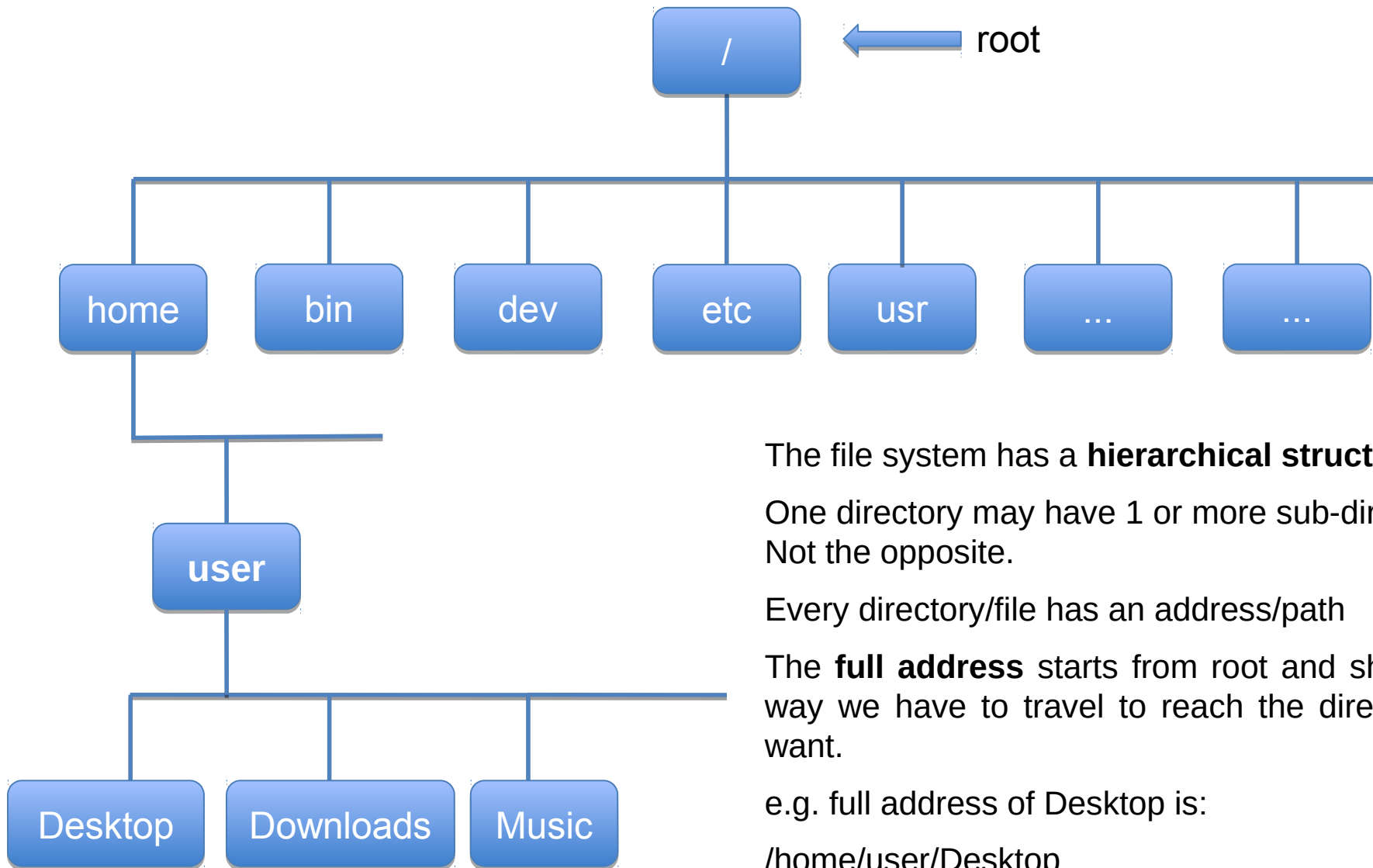
top/htop show processes

df -h report filesystem disk space usage

clear clean the screen

exit terminate the connection

Filesystem basic



The file system has a **hierarchical structure**.

One directory may have 1 or more sub-directories.
Not the opposite.

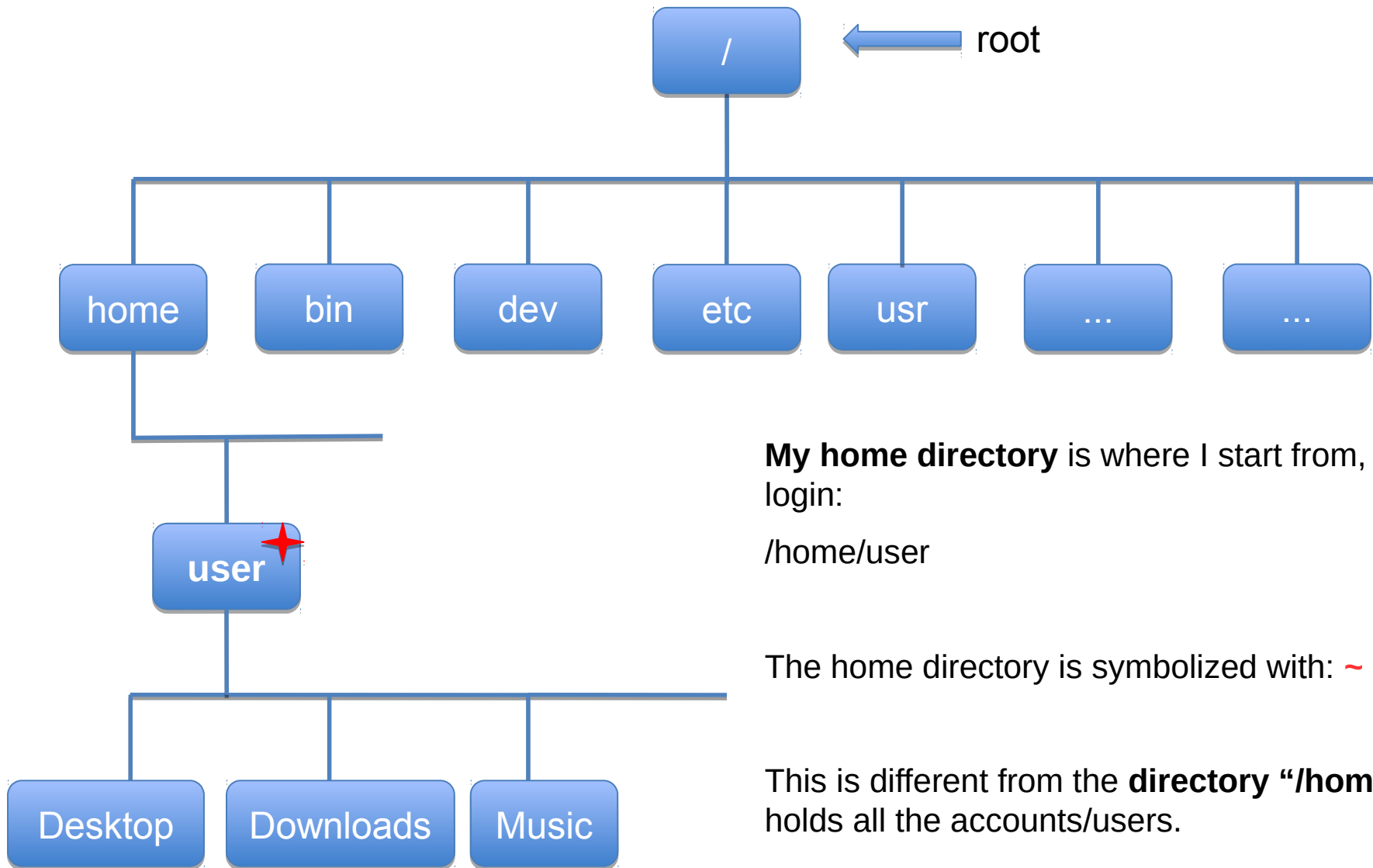
Every directory/file has an address/path

The **full address** starts from root and shows the way we have to travel to reach the directory we want.

e.g. full address of Desktop is:

`/home/user/Desktop`

Filesystem basic



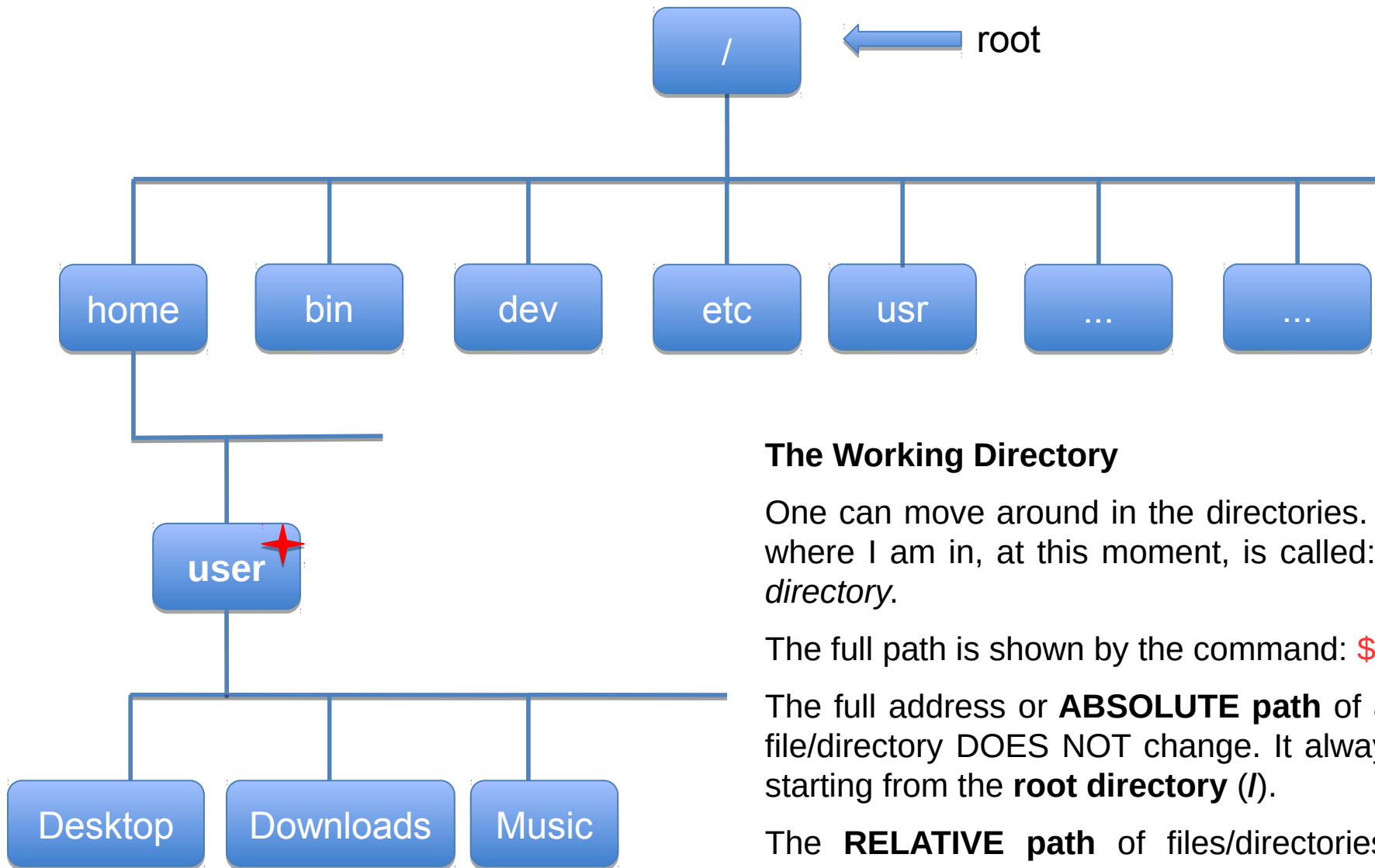
My home directory is where I start from, when I login:

/home/user

The home directory is symbolized with: ~ ("Tilde")

This is different from the **directory "/home"**, that holds all the accounts/users.

Moving around the Filesystem



The Working Directory

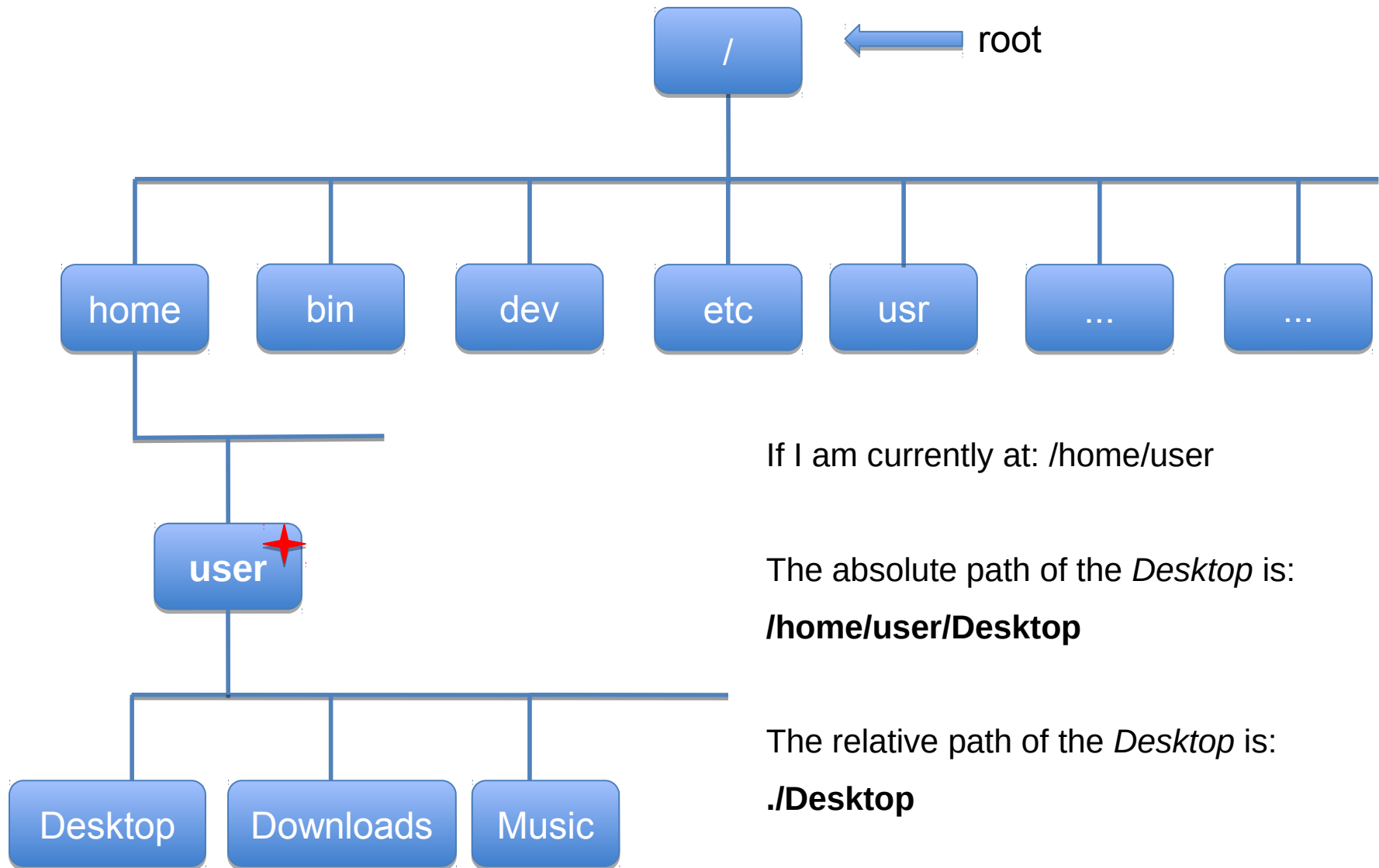
One can move around in the directories. The one where I am in, at this moment, is called: *working directory*.

The full path is shown by the command: **\$ pwd**

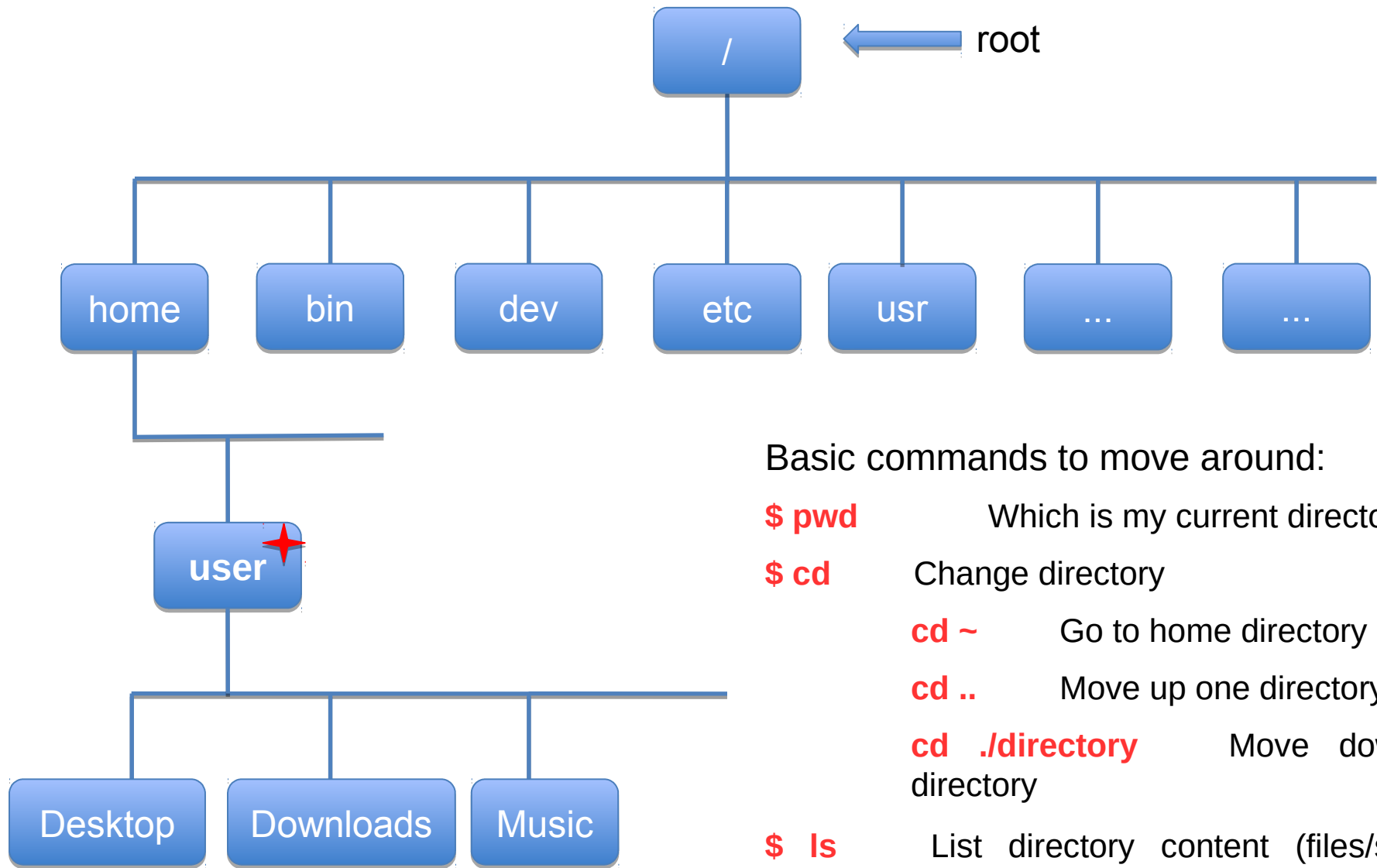
The full address or **ABSOLUTE path** of a certain file/directory DOES NOT change. It always refers starting from the **root directory (/)**.

The **RELATIVE path** of files/directories DOES change, depending on the current position in the filesystem.

Moving around the Filesystem



Moving around the Filesystem



\$ ls -lh

drwxrwr-x	2	root	root	4096	apr 1	16:51	dir
-rwxrwr-x	1	root	root	1092	apr 1	16:53	file
1 2	3	4	5	6		7	8

1. Type: directory (d), file (-) or link (l)
2. Permissions: read (r), write (w) or execute (x); defined in blocks of 3 characters: first block – owner's rights; second block – group's rights; last block – rights of any user. The dash character (-) means that the respective permission is not set.
3. Number of links: directory (2), file (1). Dir links increase in case of presence of subdirs.
4. Owner
5. Group
6. Size in Byte. If the option -h (human) is used, then file size is rounded and K (Kilobyte), M (Megabyte) or G (Gigabyte) are used.
7. Modification Date and Time
8. Name

Special files: . (*dot*) and .. (*dotdot*); they point to the current directory and to the parent directory, respectively.

- (1) Open the terminal and print your current 'position' in the filesystem**
- (2) Move to the Desktop and print the Directory content**
- (3) Print the content of the Download directory**
- (4) From Desktop, move to the Download directory**
- (5) Move back to your home directory**
- (6) List the content of the `/usr/lib/perl` directory
using absolute and relative paths**

(1) Open the terminal and print your current 'position' in the filesystem

(1) **\$ pwd**

(2) Move to the Desktop and print the Directory content

(1) **\$ cd ./Desktop**

(2) **\$ ls -lh**

(3) Or **\$ cd ./Desktop ; ls -lh**

(3) Print the content of the Download directory

(1) **\$ ls -lh ../Download** (relative)

(2) **\$ ls -lh ~/Download**

(4) From Desktop, move to the Download directory

(1) **\$ cd ../Download**

(5) Move back to your home directory

(1) **\$ cd**

(6) List the content of the /usr/lib/perl directory
using absolute and relative paths

(1) **\$ ls -lh /usr/lib/perl**

(2) **\$ ls -lh ../../usr/lib/perl**

Viewing and Editing Files

<i>more/less</i>	<p>browse or page through a text file. Usage: press the space bar to go forward a page, type b to go back a page and type q to quit. \$ less file1</p>
<i>cat</i>	<p>display and concatenate arguments to standard output \$ cat file1 file2 > file3</p>
<i>head</i>	<p>Show the first 10 lines of a file Basic Options: -# show first # lines of the specified file \$ head -20 file</p>
<i>tail</i>	<p>show last 10 lines of a file Basic Options: -f keep the file open and whenever lines are appended to the file, print them. \$ tail -f file</p>
<i>wc</i>	<p>display the number of lines, words and characters in a file \$ wc file</p>
<i>grep</i>	<p>search for a character string in a file Basic Options: -c print only a count of matching lines \$ grep -c ">" file.fasta or grep -c "@HWI-ST365" file.fastq</p>

Viewing and Editing Files

nano/pico
vi/vim
emacs

text editors. *vi* is the standard UNIX editor. Try to use them, and choose your preferred one.

awk

pattern-matching language. Useful to process text files, and extract data, also by means of regular-expression.

#Print the second and the fourth word on each line of a file

\$ awk '{print \$2, \$4}' file

sed

pattern-matching engine, that can perform manipulation on lines of text.

#print on terminal the file with all occurrences of the string 'red' changed to 'hat'

\$ sed 's/red/hat/g' file

#change all occurrences of the string 'red' to 'hat' in the file

\$ sed -i 's/red/hat/g' file

Manipulating Files and Directories

<i>cp</i>	copy files \$ cp fromfile tofile
<i>mv</i>	move or rename files \$ mv fromfile tofile
<i>ln</i>	create a link between source and destination files Basic Options: -s creates symbolic link \$ ln -s source dest
<i>mkdir</i>	create a new subdirectory in the current directory \$ mkdir subdir
<i>touch</i>	update the time stamp on existing files, create new empty files \$ touch file
<i>wget</i>	hits a URL and download the information to a file \$ wget remotefile
<i>rm</i>	remove (delete) files (<i>rmdir</i> removes empty directories) Basic Options: -r recursive, descend into subdirs removing files \$ rm file \$ rm -r NotEmptyDirectory

File Packaging and Compression

gzip/gunzip

compress and uncompress files in GNU Zip format. Compressed files have suffix `.gz`.

#Compress *file* to create *file.gz*. Original *file* is deleted

\$ gzip file

#uncompress *file.gz* to create *file*. Original *file.gz* is deleted

\$ gunzip file.gz

compress/ uncompress

compress and uncompress files in standard Unix compression format. Compressed files have suffix `.Z`

#Compress *file* to create *file.Z*. Original *file* is deleted

\$ compress file

#uncompress *file.Z* to create *file*. Original *file.Z* is deleted

\$ uncompress file.gz

tar

compact a directory and all its content (files and subdirs) into a single file with name of the packed directory, and with `.tar` extension.

\$ tar cvf myarchive.tar mydir

#create

\$ tar xvf myarchive.tar

#extract

#gzipped tar files

\$ tar czvf myarchive.tar.gz mydir

#create

\$ tar xzvf myarchive.tar.gz

#extract

File Packaging and Compression

zip/unzip

compress and uncompress files in Windows Zip format.

Compressed files have suffix `.zip`. Does not delete original file(s)

\$ zip myfiles.zip file1 file2 file3.... #pack

\$ zip -r myfiles.zip directory #pack recursively

\$ unzip myfile.zip #unpack

(1) Retrieve a fasta (**example.fasta**) and a fastq (**Cont-2-9_R1.subset.fastq**) files from github

```
wget -c https://github.com/mchiapello/2017\_PhD\_Toolbox\_course/raw/master/Resources/Day1/'file'
```

(2) Inspect the files (cat/less)

(3) View the beginning (head) and the end of the files (tail)

(4) Calculate how many sequences they contain

```
$ grep -c ">" or grep "..."
```

(5) Create a subset of sequences. Which strategy would you use?

(6) Count how many occurrences of the pattern "ATTCAAAT" appear in the .fastq file

(7) Retrieve the sequences (with all the fastq data) containing "ATTCAAAT"

- (1) Retrieve a fasta (**example.fasta**) and a fastq (**Cont-2-9_R1.subset.fastq**) files from github

```
wget -c https://github.com/mchiapello/2017_PhD_Toolbox_course/raw/master/Resources/Day1/'file'
```

- (2) Inspect the files (cat/less)

- (3) View the beginning (head) and the end of the files (tail)

- (4) Calculate how many sequences they contain

```
$ grep -c ">" or grep "..."
```

- (5) Create a subset of sequences. Which strategy would you use?

```
awk 'BEGIN { RS = "\n>"; FS = "\n"; OFS = "" };
{
  if (NR == 1) {
    print $1
  } else
  if (NR > 1) {
    print ">"$1
  }
  $1=""
  print
}' infile.fasta > outfile.fasta
```

(1) Retrieve a fasta (**example.fasta**) and a fastq (**Cont-2-9_R1.subset.fastq**) files from github

```
wget -c https://github.com/mchiapello/2017\_PhD\_Toolbox\_course/raw/master/Resources/Day1/'file'
```

(2) Inspect the files (cat/less)

(3) View the beginning (head) and the end of the files (tail)

(4) Calculate how many sequences they contain

```
$ grep -c ">" or grep "..."
```

(5) Create a subset of sequences. Which strategy would you use?

(6) Count how many occurrences of the pattern "ATTCAAAT" appear in the .fastq file

(7) Retrieve the sequences (with all the fastq data) containing "ATTCAAAT"

Basic Bioinformatic Examples

Blastn

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

[Learn more](#)

NEWS

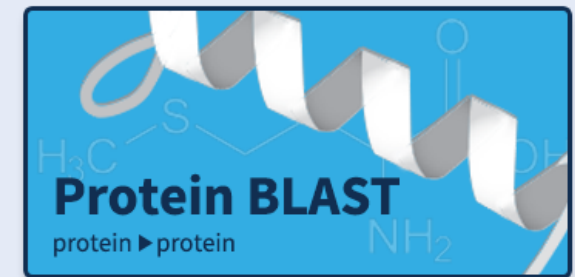
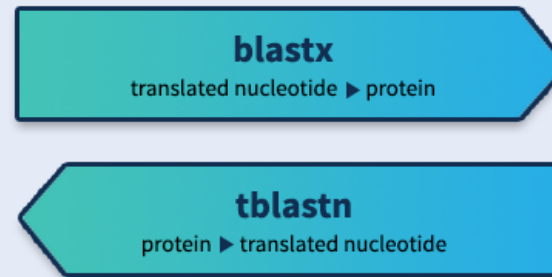
Magic-BLAST 1.2.0 released

A new version of the BLAST RNA-seq mapping tool is now available.

Mon, 27 Feb 2017 14:00:00 EST

[More BLAST news...](#)

Web BLAST



Database: any collection of sequences in fasta can be formatted as blast usable database
Common db are: nr (proteins), nt (nucleotides), swissprot (proteins), ref_seq serie...

```
stefano@RxR:~$ blastn --help
USAGE
  blastn [-h] [-help] [-import_search_strategy filename]
    [-export_search_strategy filename] [-task task_name] [-db database_name]
    [-dbsize num_letters] [-gilist filename] [-seqidlist filename]
    [-negative_gilist filename] [-entrez_query entrez_query]
    [-db_soft_mask filtering_algorithm] [-db_hard_mask filtering_algorithm]
    [-subject subject_input_file] [-subject_loc range] [-query input_file]
    [-out output_file] [-evalvalue evalvalue] [-word_size int_value]
    [-gapopen open_penalty] [-gapextend extend_penalty]
    [-perc_identity float_value] [-qcov_hsp_perc float_value]
    [-max_hsps int_value] [-xdrop_ungap float_value] [-xdrop_gap float_value]
    [-xdrop_gap_final float_value] [-searchsp int_value]
    [-sum_stats bool_value] [-penalty penalty] [-reward reward] [-no_greedy]
    [-min_raw_gapped_score int_value] [-template_type type]
    [-template_length int_value] [-dust DUST_options]
    [-filtering_db filtering_database]
    [-window_masker_taxid window_masker_taxid]
    [-window_masker_db window_masker_db] [-soft_masking soft_masking]
    [-ungapped] [-culling_limit int_value] [-best_hit_overhang float_value]
    [-best_hit_score_edge float_value] [-window_size int_value]
    [-off_diagonal_range int_value] [-use_index boolean] [-index_name string]
    [-lcase_masking] [-query_loc range] [-strand strand] [-parse_deflines]
    [-outfmt format] [-show_gis] [-num_descriptions int_value]
    [-num_alignments int_value] [-line_length line_length] [-html]
    [-max_target_seqs num_sequences] [-num_threads int_value] [-remote]
    [-version]

DESCRIPTION
  Nucleotide-Nucleotide BLAST 2.2.31+

Use '-help' to print detailed descriptions of command line arguments
=====
```

Remote blast (only with blast 2.6.0)

```
$ mkdir Programs ; cd Programs
$ wget ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ncbi-blast-2.6.0+-x64-linux.tar.gz
$ tar xzf ncbi-blast-2.6.0+-x64-linux.tar.gz
$ cd ncbi-blast-2.6.0+/bin/
$ pwd
$ export PATH=/home/user/Programs/ncbi-blast-2.6.0+/bin:$PATH
  or (to make blast suite available permanently)
$ echo "export PATH=/home/user/Programs/ncbi-blast-2.6.0+/bin:$PATH" >> ~/.bashrc
$ cd

$ blastn -query example.2.fasta -db nt -remote
```

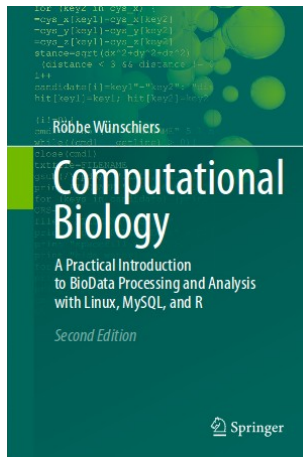
Local blast

Create your sequence collection and select a query

```
$ makeblastdb -in input_file -dbtype molecule_type -title database_title -out
database_name
```

```
$ blastn -query myquery.fasta -db mydb -evaluate 0.0005 -dust no -num_descriptions 1
-num_alignments 1 -num_threads 4 -outfmt 6 -out myquery.basltm.out.txt
```

...try awk to reorganize output columns

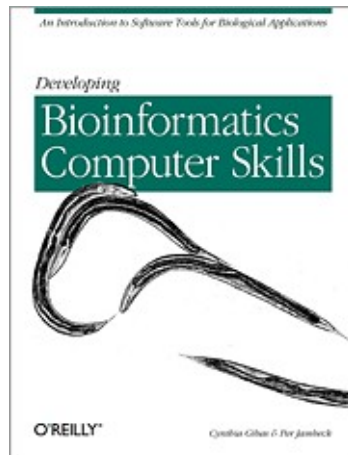


Computational Biology

A Practical Introduction to BioData Processing and Analysis with Linux, MySQL, and R

Authors: R  bbe W  nschiers

ISBN: 978-3-642-43097-8 (Springer)



Developing Bioinformatics Computer Skills

An Introduction to software Tools for Biological Applications

Authors: Cynthia Gibas, Per Jambeck

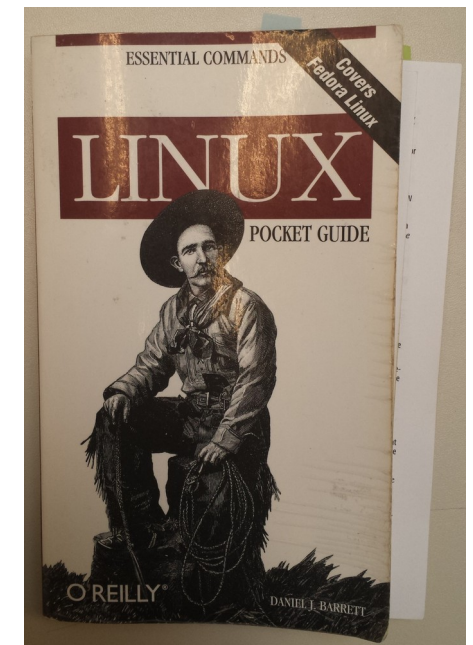
ISBN: 978-1-56592-664-6 (O'Reilly Media)

Linux Pocket Guide

Essential Commands

Authors: Daniel J. Barrett

ISBN: 978-0-596-00628-0 (O'Reilly Media)



Overused real sample