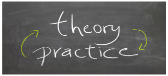
Unix/Linux Tutorial for Beginners
Session V

Mihaela Martis

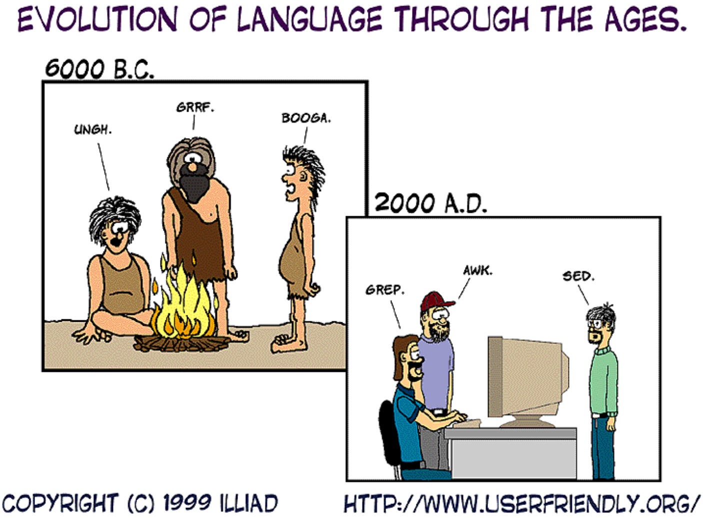
NBIS & Faculty of Medicine and Health Sciences
Division Cell Biology, IKE

Today's schedule



$9^{00} - 10^{30}$	*Practical examples & pair exercises*
$10^{30} - 10^{40}$	**Coffee break**
$10^{40} - 12^{10}$	*Finding & compressing files*
$12^{10} - 13^{10}$	**Lunch**
$13^{10} - 15^{00}$	*Other util commands*
$15^{00} - 15^{20}$	**Coffee break**
$15^{20} - 17^{00}$	*Shell scripting*

Evolution


EVOLUTION OF LANGUAGE THROUGH THE AGES.
6000 B.C.
UNGH.
GRRF.
BOOGA.
2000 A.D.
GREP.
AWK.
SED.
COPYRIGHT (C) 1999 ILLIAD
HTTP://WWW.USERFRIENDLY.ORG/

Shell commands

Filesystem	Text processing	Filters	Documentation	Editors	I/O redirection
cat	basename	grep	help	nano	>
cd	cut		man	vi	>>
cp	dirname				<
file	head				<<
ls	less				|
mv	sed				
mkdir	sort				
pwd	tail				
rm	tr				
rmdir	uniq				
touch	wc				
tree					

Pair exercises - guidance

1. say hello to your next neighbor ☺→ he/she will be your partner for the next exercises
2. the remote participants will work alone → use the chat/video conference to communicate with each other
3. you need only the terminal integrated in the e-learning platform
4. solve the tasks by combining the commands taught yesterday
5. the exercises can be found on the e-learning platform under session 5

Pair exercises - guidance

1. say hello to your next neighbor ☺→ he/she will be your partner for the next exercises
2. the remote participants will work alone → use the chat/video conference to communicate with each other
3. you need only the terminal integrated in the e-learning platform
4. solve the tasks by combining the commands taught yesterday
5. the exercises can be found on the e-learning platform under session 5

   There's more than one way to skin a cat!

Practical examples

1. reverse complement a fasta file
2. transform lower case nucleotides to upper case nucleotides in a fasta file
3. extract sequences from a FASTA file based on supplied identifier
4. count the number of reads in a FASTQ file
5. convert a FASTQ file to a FASTA file
6. determine all feature types annotated in a GFF file
7. determine the number of genes annotated in a GFF3 file

Reverse complement a fasta file

```
$ paste -d "\n" <(grep ">" myCDS.fa) <(grep -v ">" myCDS.fa|rev|tr "acgtACGT" "
    tgcaTGCA") >myCDS_revCompl.fa
```

1. extract the sequence identifier:

```
$ grep ">" myCDS.fa
```

2. reverse complement the sequences:

```
$ grep -v ">" myCDS.fa | rev | tr "acgtACGT" "tgcaTGCA"
```

3. combine the outcome from step 1) and 2) using the command paste

```
$ paste -d "\n" <(grep ">" myCDS.fa) <(grep -v ">" myCDS.fa|rev|tr "
    acgtACGT" "tgcaTGCA")
```

4. redirect the output to a file

```
$ paste -d "\n" <(grep ">" myCDS.fa) <(grep -v ">" myCDS.fa|rev|tr "
    acgtACGT" "tgcaTGCA") >myCDS_revCompl.fa
```

Transform all lower case nucleotides to upper case nucleotides

```
$ paste -d "\n" <(grep ">" myCDS_revCompl.fa) <(grep -v ">" myCDS_revCompl.fa |
    tr "a-z" "A-Z")  >myCDS_revCompl_lc.fa
```

1. extract the sequence identifier

```
$ grep ">" myCDS_revCompl.fa
```

2. transform lower case nucleotides to upper case nucleotides

```
$ grep -v ">" myCDS_revCompl.fa | tr "a-z" "A-Z"
```

3. combine the outcome from step 1) and 2)

```
$ paste -d "\n" <(grep ">" myCDS_revCompl.fa) <(grep -v ">" myCDS_revCompl
    .fa | tr "a-z" "A-Z")
```

4. redirect the output to a new file

```
$ paste -d "\n" <(grep ">" myCDS_revCompl.fa) <(grep -v ">" myCDS_revCompl
    .fa | tr "a-z" "A-Z")  >myCDS_revCompl_lc.fa
```

Extracting sequences from a Fasta file based on supplied IDs

1. extract 7 sequence identifiers from the file *barley_CDS.fa*

```
$ grep ">" barley_CDS.fa | tr -d '>' | head -7 > myIds.txt
```

2. extract for these identifiers the sequences from the file *barley_CDS.fa*

```
$ grep -A1 -F -f myIds.txt barley_CDS.fa | tr -d '-' > my_seq_ids.fa

OR

$ grep -A1 -F -f myIds.txt barley_CDS.fa | grep -v '^--$' > my_seq_ids.fa
```

grep -F – fixed strings

grep -A1 – print matching line plus the next line

How many reads are in a fastq file?

```
$ cat DRR001013.fastq | echo $((‘wc -l‘/4))
```

→ works only with 'back-ticks'

Convert FASTQ to FASTA

```
$ sed -n '1~4s/^@/>/p; 2~4p' DRR001013.fastq > DRR001013.fasta
```

- M∼N with p – prints every N'th line starting with line first
- -n suppress automatic printing of pattern – will not print anything, unless an explicit request to print is found

Determine all feature types annotated in a GFF file

```
grep -v "^#" Pygoscelis_adeliae.gff | cut -s -f 3 | sort | uniq
```

Determine the number of genes annotated in a GFF3 file

```
grep -c $’\tmRNA\t’ Pygoscelis_adeliae.gff
```