IMBBC
INSTITUTE OF MARINE BIOLOGY, BIOTECHNOLOGY AND AQUACULTURE

hcmr
ΕΛΚΕΘΕ

Tabular Data File Processing

Evangelos Pafilis

Πανεπιστήμιο Κρήτης
ΙΑΤΡΙΚΗ ΣΧΟΛΗ

FORTH
INSTITUTE OF MOLECULAR BIOLOGY & BIOTECHNOLOGY

IMBBC
INSTITUTE OF MARINE BIOLOGY, BIOTECHNOLOGY AND AQUACULTURE

hcmr
ΕΛΚΕΘΕ

Tabular Data File Processing

Read
Select
Merge (join)
Compute

Πανεπιστήμιο Κρήτης
ΙΑΤΡΙΚΗ ΣΧΟΛΗ

FORTH
INSTITUTE OF MOLECULAR BIOLOGY & BIOTECHNOLOGY

IMBBC
INSTITUTE OF MARINE BIOLOGY, BIOTECHNOLOGY AND AQUACULTURE

hcmr
ΕΛΚΕΘΕ

Evangelos Pafilis

Bioinformatics and Biodiversity Informatics

Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC)

Hellenic Centre for Marine Research (HCMR)

Heraklion Crete, Greece

pafilis@hcmr.gr, http://epafilis.info

IMBBC
INSTITUTE OF MARINE BIOLOGY, BIOTECHNOLOGY AND AQUACULTURE

hcmr
ΕΛΚΕΘΕ

gawk, grep, cut, sort, uniq

Pavlos Pavlidis
Nikolas Papanikolaou


Πανεπιστήμιο Κρήτης
ΙΑΤΡΙΚΗ ΣΧΟΛΗ

This lecture

gawk

- Write gawk script files
- Read multiple files
- How to create associative lists (arrays)
  - and check for containment
- Perform data join operations
- Output selected clauses of previous fragmented data **together** (joined)

Why ?

- Bioinformatics are dominated by data files among others in **tabular** formats
- Data that needs transformation, inspection, selection, reformatting and recombination

Why ?

- Fast and efficient
  - Reads line by line, low on memory if properly coded
  - Works on columns, ideal for tabular data like comma, tab, or other tab separated files
  - Command line environment
  - Fast to code (one-lines replace tens of lines in other programming lange)
  - Awk (or gawk) included in most Unix systems

gawk (GNU awk, ie. “new awk” as opposed to **awk** (the “old awk”))

gawk - version

on the MSc Bioinformatics server

ssh -p 30022 user-your_id@139.91.75.10

Try:

```
$>awk --version

$>gawk --version
```

Question: What is the difference?

gawk - basic example - one input file

```
$>awk 'code' file

e.g.
$>cat /home/pafilis/gawk/colour_hexcode.tsv

$>gawk '{print $2 "\t" $1}' /home/pafilis/gawk/colour_hexcode.tsv
```

Question: what is the output of the above?

Question: what is $0? E.g.

```
$>gawk '{print $0}' /home/pafilis/gawk/colour_hexcode.tsv
```

Question: how can I print only the blue colour details?

gawk - basic example - one input file

```
$>gawk '(condition){ commands }' file

e.g.
$>gawk '($1=="blue"){ print $0; }' colour_hexcode.tsv

Question: what is the output of the above?
```

gawk - one liners

Fast to write, saves a lot of developing time as opposed to programming languages requiring verbose code

Optimized for tabular data files

gawk - basic example - more input files

Tip: a good practice is to always look at the input files and observer the structure (e.g. columns, separator, and corresponding data) e.g.

$>head -3 colour_hexcode.tsv students_favColour.tsv

Now let's try:

$>gawk -F "\t" '(ARGIND==1 && $1=="green"){ print $0; } (ARGIND==2 && $2=="green") {print $0;}' colour_hexcode.tsv students_favColour.tsv

Question: what is the output of the above? Why?

gawk command line parameters

-F field separator

-f awk script file to run

gawk command line parameters

Can the previous one-liner be written in an easier-to-handle way?

-F field separator

-f awk script file to run

$>cat /home/pafilis/gawk/simple.awk

```
#! /usr/bin/gawk -f
(ARGIND==1){
    if ($1=="green"){
        print $0;
    }
}

(ARGIND==2){
    if ($2=="green"){
        print $0;
    }
}
```

$>chmod +x simple.awk

$> gawk -F "\t" -f simple.awk colour_hexcode.tsv students_favColour.tsv

gawk - basic syntax

```
#! /usr/bin/gawk -f
BEGIN {
    #set up global or environmental variables
    #perfom some actions prior starting reading
    #input files, e.g. instantiate variables
}

( condition_pattern ){
    actions
}

END {
    #perform actions after having read all input files
    #e.g. Statistics like counting or averaging
}
```

gawk built-in parameters

FS the input field separator e.g. “\t” (TAB), “,” (COMMA)
the default value is " " (single space)

RS the input record separator, e.g. “\n” (NEW LINE, default)

FNR This is the current record number in the current file; ~current line if RS == “\n”
Resets to zero each time a new input file start being read

NR the number of input records that have been read since the script started (ie. does not reset to 0 and persists between new file reading

ARGV the array that stores the command-line arguments array

NB: Unlike most awk arrays, **ARGV is indexed from 0** (0-based)

FILENAME the name of the file that is currently being read
FILENAME is changed each time a new file is read (see section Reading Input Files)

Source ftp://ftp.gnu.org/old-gnu/Manuals/gawk-3.1.0/html_node/gawk_115.html#SEC127

ftp://ftp.gnu.org/old-gnu/Manuals/gawk-3.1.0/html_node/gawk_117.html#SEC129

gawk built-in parameters

$0 the complete input record (~line, when RS == "\n")

$1 points to the value of the first field of the current input record (~line if RS == "\n")

$2 points to the value of the second field of the current input record (~line if RS == "\n")

And so forth

NF the number of fields in the current input record (resets at each record)

Source ftp://ftp.gnu.org/old-gnu/Manuals/gawk-3.1.0/html_node/gawk_117.html#SEC129

More **ftp://ftp.gnu.org/old-gnu/Manuals/gawk-3.1.0/html_node/gawk.html**

gawk built-in parameters

ARGIND (by definition) the index in ARGV of the current file being processed. Every time Gawk opens a new data file for processing, it sets ARGIND to the index in ARGV of the file name

in everyday practice words:

gawk -F “\t” ‘(ARGIND==1) { …. process first_input_file.tsv …. } (ARGIND==2) { …. process second_input_file.tsv... } ’ first_input_file.tsv second_input_file.tsv

Source ftp://ftp.gnu.org/old-gnu/Manuals/gawk-3.1.0/html_node/gawk_117.html#SEC129

gawk - variable declaration

Input file: /home/pafilis/gawk/students_favColour.tsv

$>cat /home/pafilis/gawk/simple_count.awk

```
#! /usr/bin/gawk -f
BEGIN {
    FS="\t";
    studentCount=0;
}

#this control statement is checked for every input file line
($2=="green"){
    studentCount++;
     print $0;
}

END {
    print studentCount++;
}
```

Run this as: $> *gawk simple4.awk students_favColour.tsv*

Based on ideas in: http://www.grymoire.com/Unix/Awk.html#uh-1

gawk - conditional clauses

```
if ( condition_pattern ) {
     actions
}
e.g.
if ( $1=="green" ){
     print $0;
}


if ( condition_pattern ) {     actions     }
else { Some other actions }

if ( condition_pattern ) {     actions     }
else if (another_condition_pattern) { Some other actions }
else if (condition_pattern){ Some other actions }
else { Some other actions }

#written in compact form merely to save space on the slide
```

gawk - fop loop

```
for ( initialization_clause ; control_condition_clause ; counter_update_clause) {
     actions
}
```

e.g.

```
for (counter = 1; counter <= 10; counter++) {
    print counter;
}
```

Question: what will happen if one accidentally writes “counter++;” ?

gawk - while loop

```
control_variable_initialization
while ( control_condition_clause ) {
     actions
     control_variable_update
}

e.g.

w_counter = 20;
while ( w_counter  > 11) {
   print w_counter;
   w_counter--;
}
```

gawk - continue, break

Question: what will be printed in the following two cases?

```
for (counter = 1; counter <= 10; counter++) {
    if (counter == 4) { continue; }
    if (counter == 8) { continue; }
    print counter;
}
```

```
for (counter = 1; counter <= 10; counter++) {
    if (counter == 5) { break; }
    print counter;
}
```

gawk - arrays

gawk associative arrays

NB: in GAWK all arrays are **associative arrays** (equivalent to hashtables, dictionaries, etc in other programming languages) e.g.

[0] = “user-1120016”

[1] = “user-1120017”

is not created automatically, you need to: associate “0” with “user-1120016”, “1” with “user-1120017” and so on: e.g.

```
student_list [“0”] = “user-1120016”;
student_list [“1”] = “user-1120017”;
```

Syntax **arrayname[key_string] = value_string ;**

e.g. `colour_hexcode_map [ “blue” ] = “#0000ff” ;`

Very useful in associating fields read from different column of an input file:

e.g. `colour_hexcode_map [ $1 ] = $2 ;`

Source http://www.thegeekstuff.com/2010/03/awk-arrays-explained-with-5-practical-examples/

gawk accessing all elements of an array

```
for (array_element in arrayname) {
     print array_element" contains the value "arrayname[array_element ];
}

for (colour in colour_hexcode_map) {
     print colour", "colour_hexcode[colour];
}
```

Source http://www.thegeekstuff.com/2010/03/awk-arrays-explained-with-5-practical-examples/

Containment control: gawk check if element in array and access its value

```
if ( "blue" in colour_hexcode_map) {  print $colour_hexcode_map["blue"]; }
```

```
if ($1 in colour_hexcode_map)
{
     print $1"\t"colour_hexcode_map[$1];
}
```

Source http://www.thegeekstuff.com/2010/03/awk-arrays-explained-with-5-practical-examples/

Example

on the MSc Bioinformatics server

ssh -p 30022 user-your_id@139.91.75.10

under:

/home/pafilis/gawk

selected_colours.tsv

colour_hexcode.tsv

print_selected_colour_hex.awk

print_selected_colour_hex_reverse_order.awk

Source https://www.rapidtables.com/web/color/RGB_Color.html

print_selected_colour_hex.awk

```
#! /usr/bin/gawk -f
BEGIN {
  FS="\t"
    ### global variables are defined here
    IGNORECASE=1; #built-in variable to make string matches case insensitive or sensitive (1 or 0
}

## usage: ./print_selected_colour_hex.awk selected_colours.tsv colour_hexcode.tsv

(ARGIND==1){  #selected_colours.tsv
    selected_colours[ $1 ] = "1";
}


(ARGIND==2){  #colour_hexcode.tsv
    if ( $1 in selected_colours) { print "the selected colour is: "$1" and its HEX code is:
  "$2; }
}
```

print_selected_colour_hex_reverse_order.awk

```
#! /usr/bin/gawk -f
BEGIN {
  FS="\t";
    ### global variables are defined here
    IGNORECASE=1; #built-in variable to make string matches case insensitive or sensitive (1 or 0
}
#usage: ./print_selected_colour_hex_reverse_order.awk colour_hexcode.tsv selected_colours.tsv
(ARGIND==1){ #colour_hexcode.tsv
    colour_hexcode_map[ $1 ] = $2;
}

(ARGIND==2){  #selected_colours.tsv
        print "the selected colour is: "$1" and its HEX code is: "colour_hexcode_map[$1];
}

END {
    #output all colour info as reference, for loop example
    for (colour in colour_hexcode_map)
        {
            print colour", "colour_hexcode_map[colour];
        }
}
```

Discussion

What is the difference between ?

`print_selected_colour_hex.awk`

`Print_selected_colour_hex_reverse_order.awk`

Case study

Processing the complete tabular output of the WOrld Registry of Marine Species (WORMS, VLIZ, Belgium)

IMBBC
INSTITUTE OF MARINE BIOLOGY, BIOTECHNOLOGY AND AQUACULTURE

hcmr
ΕΛΚΕΘΕ

Tabular Data File Processing

Read
Select
Merge (join)
Compute


Πανεπιστήμιο Κρήτης
ΙΑΤΡΙΚΗ ΣΧΟΛΗ

Final Objective

species	genus	family	order	class	phylum	GR-CR01-IT	GR-CR01-ST	
Autonoe spir	Autonoe	Aoridae	Amphipoda	Malacostraca	Arthropoda	75.1879699	375.93985	
Capitella mir	Capitella	Capitellidae		Polychaeta	Annelida		75.1879699	
Autonoe spir	Autonoe	Aoridae	Amphipoda	Malacostraca	Arthropoda	75.1879699	375.93985	
Capitella mir	Capitella	Capitellidae		Polychaeta	Annelida		75.1879699	
Capitellethus	Capitellethus	Capitellidae		Polychaeta	Annelida		75.1879699	
Caulleriella v	Caulleriella	Cirratulidae	Terebellida	Polychaeta	Annelida	112.781955		
Protodorville	Protodorville	Dorvilleidae	Eunicida	Polychaeta	Annelida	150.37594	75.1879699	TCTACTTCGTTCTACGACTGAGCCGGACTTCTAGGGA
Schroederell	Schroederell	Orbiniidae		Polychaeta	Annelida		150.37594	
Sphaerosyllis	Sphaerosyllis	Syllidae	Phyllodocida	Polychaeta	Annelida	150.37594		ACTATATATATAATGATCGGAATATGAAGAGGCCTCA
Ctena decuss	Ctena	Lucinidae	Lucinida	Bivalvia	Mollusca	75.1879699		
Salvatoria yr	Salvatoria	Syllidae	Phyllodocida	Polychaeta	Annelida	75.1879699		

classification

Abundance (individuals/m^2)

COI sequence (if available)

NB: unpublished material, please do not distribute further

IMBBC
INSTITUTE OF MARINE BIOLOGY, BIOTECHNOLOGY AND AQUACULTURE

hcmr
ΕΛΚΕΘΕ

Tabular Data File Processing

Read
Select
Merge (join)
Compute


Πανεπιστήμιο Κρήτης
ΙΑΤΡΙΚΗ ΣΧΟΛΗ

Reverse Start: today's practical: Species X Stations (ack: EMBOS team)

WoRMS_scientificName	CY-SFDA-SUB	CY-SFDB-SUB	EE-Kanissaare-SUB	EE-Kõiguste-SUB	ES-ES1-LWL
Abra alba					
Abra tenuis					
Acrocnida brachiata					
Alkmaria romijni					
Ampelisca brevicornis					
Amphipholis squamata					
Antalis vulgaris					75.18796992
Aonides oxycephala					125.3132832
Apherusa bispinosa					
Apseudopsis latreillii					75.18796992
Arcuatula senhousia					
Arenicola marina					
Aricidea (Acmira) cerrutii					
Armandia cirrhosa					
Athanas nitescens					
Autonoe spiniventris					
Bathyporeia guilliamsoniana	225.5639098				
Bathyporeia pilosa					
Bathyporeia sunnivae	75.18796992	75.18796992			
Bittium reticulatum					
Boccardiella hamata					
Calyptraea chinensis					75.18796992
Capitella capitata					112.7819549

NB: unpublished material, please do not distribute further

Second Input: Taxonomy Classification

species	genus	family	order	class	phylum
Acanthocephaloides distinc	Acanthocephaloides	Arhythmacanthidae	Echinorhynchida	Palaeacanthocephala	Acanthocephala
Acanthocephaloides geneti	Acanthocephaloides	Arhythmacanthidae	Echinorhynchida	Palaeacanthocephala	Acanthocephala
Acanthocephaloides incrass	Acanthocephaloides	Arhythmacanthidae	Echinorhynchida	Palaeacanthocephala	Acanthocephala
Acanthocephaloides propin	Acanthocephaloides	Arhythmacanthidae	Echinorhynchida	Palaeacanthocephala	Acanthocephala
Acanthocephalus anguillae	Acanthocephalus	Echinorhynchidae	Echinorhynchida	Palaeacanthocephala	Acanthocephala
Acanthocephalus clavula	Acanthocephalus	Echinorhynchidae	Echinorhynchida	Palaeacanthocephala	Acanthocephala
Acanthocephalus lucii	Acanthocephalus	Echinorhynchidae	Echinorhynchida	Palaeacanthocephala	Acanthocephala
Andracantha phalacrocorac	Andracantha	Polymorphidae	Polymorphida	Palaeacanthocephala	Acanthocephala
Andracantha tunitae	Andracantha	Polymorphidae	Polymorphida	Palaeacanthocephala	Acanthocephala
Arhythmorhynchus longicol	Arhythmorhynchus	Polymorphidae	Polymorphida	Palaeacanthocephala	Acanthocephala
Echinorhynchus armoricanu	Echinorhynchus	Echinorhynchidae	Echinorhynchida	Palaeacanthocephala	Acanthocephala
Echinorhynchus bothniensis	Echinorhynchus	Echinorhynchidae	Echinorhynchida	Palaeacanthocephala	Acanthocephala
Echinorhynchus brayi	Echinorhynchus	Echinorhynchidae	Echinorhynchida	Palaeacanthocephala	Acanthocephala
Echinorhynchus calloti	Echinorhynchus	Echinorhynchidae	Echinorhynchida	Palaeacanthocephala	Acanthocephala
Echinorhynchus cinctulus	Echinorhynchus	Echinorhynchidae	Echinorhynchida	Palaeacanthocephala	Acanthocephala
Echinorhynchus gadi	Echinorhynchus	Echinorhynchidae	Echinorhynchida	Palaeacanthocephala	Acanthocephala
Echinorhynchus trachyrinci	Echinorhynchus	Echinorhynchidae	Echinorhynchida	Palaeacanthocephala	Acanthocephala
Aspersentis johni	Aspersentis	Heteracanthocephalidae	Echinorhynchida	Palaeacanthocephala	Acanthocephala
Bolbosoma capitatum	Bolbosoma	Polymorphidae	Polymorphida	Palaeacanthocephala	Acanthocephala
Bolbosoma physeteris	Bolbosoma	Polymorphidae	Polymorphida	Palaeacanthocephala	Acanthocephala
Bolbosoma vasculosum	Bolbosoma	Polymorphidae	Polymorphida	Palaeacanthocephala	Acanthocephala

NB: unpublished material, please do not distribute further

Species - Gene Information

A	B	C	D	E
Protodorvillea kefersteini	>KF808171.1 Protodorvillea kefersteini cytochrome oxidase subunit 1 (COI) gene,	CTCTACTTCGTTCTACGACTGAGCCGGAC		
Protodorvillea kefersteini	>DQ779759.1 Protodorvillea kefersteini histone H3 gene, partial cds	CGTAAATCTACCGGAGGCAAGGCCCCCA		
Ctena decussata	>FR686826.1 Ctena decussata partial 28S rRNA gene, specimen voucher NHM BMI	TATTAATAAGCGGAGGAAAAGAAACTAA		
Ctena decussata	>FR686825.1 Ctena decussata partial 28S rRNA gene, specimen voucher NHM BMI	TATTAATAAGCGGAGGAAAAGAAACTAA		
Ctena decussata	>FR686706.1 Ctena decussata partial 18S rRNA gene, specimen voucher NHM BMI	AGTCATATGCTTGTCTCAAAGATTAAGCC		
Ctena decussata	>FR686705.1 Ctena decussata partial 18S rRNA gene, specimen voucher NHM BMI	AGTCATATGCTTGTCTCAAAGATTAAGCC		
Protodorvillea kefersteini	>DQ779670.1 Protodorvillea kefersteini 18S ribosomal RNA gene, partial sequence	CGGCATGCGCCCGATCCGGCAACGGAGA		
Protodorvillea kefersteini	>DQ779634.1 Protodorvillea kefersteini 16S ribosomal RNA gene, partial sequence	GCATAATCATTTGCCCTTTAATTGGGGGC		
Sphaerosyllis glandulata	>EF123765.1 Sphaerosyllis glandulata cytochrome oxidase subunit I (COI) gene, pa	CACTATATATATAATGATCGGAATATGAA		
Sphaerosyllis glandulata	>EF123840.1 Sphaerosyllis glandulata 18S ribosomal RNA gene, partial sequence	GTTGTGCGTTCCTTAGATCGTTTTACAGTT		
Ctena decussata	>FR686621.1 Ctena decussata mitochondrial partial cytB gene for cytochrome B pı	TATGGTAACAGTTATTCCATATGTGGGGA		
Ctena decussata	>FR686620.1 Ctena decussata mitochondrial partial cytB gene for cytochrome B pı	TATGGTAACAGTTATTCCATATGTGGGGA		
Protodorvillea kefersteini	>DQ779708.1 Protodorvillea kefersteini 28S ribosomal RNA gene, partial sequence	ATCACTAAGCGGAGGAAAAGAAACTAAC		
Protodorvillea kefersteini	>AY598738.1 Protodorvillea kefersteini cytochrome c oxidase subunit I gene, parti	GTCTGAGCACACCATATGTTTACCGTCGG		
Protodorvillea kefersteini	>AF412799.1 Protodorvillea kefersteini 18S ribosomal RNA gene, partial sequence	TCTCAAAGATTAAGCCATGCATGTCTAAG		
Protodorvillea kefersteini	>AY732230.1 Protodorvillea kefersteini 28S ribosomal RNA gene, partial sequence	CAAGTACCGTGAGGGAAAGTTGAAAGCA		

Final Objective

species	genus	family	order	class	phylum	GR-CR01-IT	GR-CR01-ST	
Autonoe spir	Autonoe	Aoridae	Amphipoda	Malacostraca	Arthropoda	75.1879699	375.93985	
Capitella mir	Capitella	Capitellidae		Polychaeta	Annelida		75.1879699	
Autonoe spir	Autonoe	Aoridae	Amphipoda	Malacostraca	Arthropoda	75.1879699	375.93985	
Capitella mir	Capitella	Capitellidae		Polychaeta	Annelida		75.1879699	
Capitellethus	Capitellethus	Capitellidae		Polychaeta	Annelida		75.1879699	
Caulleriella v	Caulleriella	Cirratulidae	Terebellida	Polychaeta	Annelida	112.781955		
Protodorville	Protodorville	Dorvilleidae	Eunicida	Polychaeta	Annelida	150.37594	75.1879699	CTCTACTTCGTTCTACGACTGAGCCGGACTTCTAGGGAC
Schroederell	Schroederell	Orbiniidae		Polychaeta	Annelida		150.37594	
Sphaerosyllis	Sphaerosyllis	Syllidae	Phyllodocida	Polychaeta	Annelida	150.37594		CACTATATATATAATGATCGGAATATGAAGAGGCCTCA
Ctena decuss	Ctena	Lucinidae	Lucinida	Bivalvia	Mollusca	75.1879699		
Salvatoria yr	Salvatoria	Syllidae	Phyllodocida	Polychaeta	Annelida	75.1879699		

NB: unpublished material, please do not distribute further

Final Objective

species	genus	family	order	class	phylum	GR-CR01-IT	GR-CR01-ST	
Autonoe spir	Autonoe	Aoridae	Amphipoda	Malacostraca	Arthropoda	75.1879699	375.93985	
Capitella mir	Capitella	Capitellidae		Polychaeta	Annelida		75.1879699	
Autonoe spir	Autonoe	Aoridae	Amphipoda	Malacostraca	Arthropoda	75.1879699	375.93985	
Capitella mir	Capitella	Capitellidae		Polychaeta	Annelida		75.1879699	
Capitellethus	Capitellethus	Capitellidae		Polychaeta	Annelida		75.1879699	
Caulleriella v	Caulleriella	Cirratulidae	Terebellida	Polychaeta	Annelida	112.781955		
Protodorville	Protodorville	Dorvilleidae	Eunicida	Polychaeta	Annelida	150.37594	75.1879699	TCTACTTCGTTCTACGACTGAGCCGGACTTCTAGGGA
Schroederell	Schroederell	Orbiniidae		Polychaeta	Annelida		150.37594	
Sphaerosyllis	Sphaerosyllis	Syllidae	Phyllodocida	Polychaeta	Annelida	150.37594		ACTATATATATAATGATCGGAATATGAAGAGGCCTCA
Ctena decuss	Ctena	Lucinidae	Lucinida	Bivalvia	Mollusca	75.1879699		
Salvatoria yr	Salvatoria	Syllidae	Phyllodocida	Polychaeta	Annelida	75.1879699		

classification

Abundance (individuals/m^2)

COI sequence (if available)

NB: unpublished material, please do not distribute further

IMBBC
INSTITUTE OF MARINE BIOLOGY, BIOTECHNOLOGY AND AQUACULTURE

hcmr
ΕΛΚΕΘΕ

Practical

Πανεπιστήμιο Κρήτης
ΙΑΤΡΙΚΗ ΣΧΟΛΗ

Practical I

NB: unpublished material
Please do not distribute further

Work on the MSc Bioinformatics server

ssh -p 30022 your_user_name@139.91.75.10

Practical I

NB: unpublished material
Please do not distribute further

Copy from */home/pafilis/gawk* to your home folder

the following files and create the ones indicated in green

- */home/pafilis/gawk/species_stations.tsv* (tab delimited version of species_stations.xlsx)
- ("cut" down the species name and GReek station columns => **species_stations.GR.tsv** ; semi-filled in commands given below:)

```
>> cut -f1,14,15
>> gawk -F '\t' '($2!="" || $3!="")
```

- */home/pafilis/gawk/classification.csv* (needs conversion to => **classification.tsv**)
  - How many lines does this file contain?

```
>>  perl -pe 's/,/\t/g' classification.csv > classification.tsv
```

- */home/pafilis/gawk/genes.GR.fa.tsv* (select only CO1 or COI gene entries => **co1.GR.fa.tsv**)

```
>> grep -i coi genes.GR.fa.tsv  > co1.GR.fa.tsv
```

Practical II

- Generate the “Final objective” slide tabular data file

If the gawk command were an one-liner it would look something like:

```
gawk -F "\t" '(ARGIND==1){  ......  }(ARGIND==2){  ......
} (ARGIND==3){ ......}  }' species_stations.GR.tsv
co1.GR.fa.tsv classification.tsv  >
classification_stations_co1.tsv
```

Practical III

```
#! /usr/bin/gawk -f
BEGIN {
  FS="\t";
    ### global variables are defined here
    IGNORECASE=1; #built-in variable to make string matches case insensitive
or sensitive (1 or 0)
}
#usage: ./script_name.awk species_stations.GR.tsv co1.GR.fa.tsv classification.tsv >
classification_stations_co1.tsv

(ARGIND==1){ #species_stations.GR.tsv
   #load stuff in memory as an associative list
}

(ARGIND==2){  #co1.GR.fa.tsv
        #load stuff in memory as an associative list
}

(ARGIND==3){  #classification.tsv
   #check some condition and print the output as wish
}
```

Practical III - solution

```
#! /usr/bin/gawk -f
  BEGIN {
      FS="\t";
   ### global variables are defined here
   IGNORECASE=1; #built-in variable to make string matches case insensitive or sensitive (1 or 0)
  }

  #usage: ./solution.awk species_stations.GR.tsv co1.GR.fa.tsv classification.tsv  > classification_stations_co1.tsv

  (ARGIND==1){ #species_stations.GR.tsv
   species_station1_station2[$1]=$2"\t"$3;

  }

  (ARGIND==2){  #co1.GR.fa.tsv
   sp_gene[$1]=$3;
  }

  (ARGIND==3){  #classification.tsv
   if ($1 in species_station1_station2 ) {
       print $0"\t"species_station1_station2[$1]"\t"sp_gene[$1];
   }
  }
```

Practical III - solution

```
#! /usr/bin/gawk -f
  BEGIN {
      FS="\t";
   ### global variables are defined here
   IGNORECASE=1; #built-in variable to make string matches case insensitive or sensitive (1 or 0)
  }

  #usage: ./solution.awk species_stations.GR.tsv co1.GR.fa.tsv classification.tsv  > classification_stations_co1.tsv

  (ARGIND==1){ #species_stations.GR.tsv
   species_station1_station2[$1]=$2"\t"$3;

  }

  (ARGIND==2){  #co1.GR.fa.tsv
   sp_gene[$1]=$3;
  }

  (ARGIND==3){  #classification.tsv
   if ($1 in species_station1_station2 ) {
       print $0"\t"species_station1_station2[$1]"\t"sp_gene[$1];
   }
  }
```

join

Homework

A. Report two sites where GAWK manual-like information is available

B. Some species have gaps in their classification. In */home/pafilis/gawk/classification.csv*:

Can you spot these with a **grep** search?
Can you spot these with a **gawk** script?

Regular expressions in gawk have not been presented today. Can you describe how a regular expression applied in gawk can answer the above? (in your reply include both a gawk-one-liner and an alternative version in a gawk script. In the gawk script SET the FIELD SEPARATOR (FS) in the BEGIN clause)

C. Based on the */home/pafilis/gawk/homework/classification_10_entries.tsv*, create a: "species X family" 2D matrix (see next two slides) (*classification_10_entries.tsv* contains only the top 10 entries of *classification.csv* -- for development and educational purposes). Please report the final output and the script used

Tip: You may have to read the classification file twice and Store family names in an array initiated in BEGIN so could print the in the header in the output

Tip: consider composite keys

Tip: Gawk built-in parameters can help you handle the header lines (first line in each file)

NB: unpublished material, please do not distribute further

Homework

Input file: */home/pafilis/gawk/**homework/classification_10_entries.tsv*** *(NB: not /home/pafilis/gawk/classification*)*

```
row_names,species,genus,family,order,class,phylum
Acanthocephaloides distinctus,Acanthocephaloides distinctus,Acanthocephaloides,Arhythmacanthidae,Echinorhynchida,Palaeacanthocephala,Acanthoceph
Acanthocephaloides geneticus,Acanthocephaloides geneticus,Acanthocephaloides,Arhythmacanthidae,Echinorhynchida,Palaeacanthocephala,Acanthocephal
Acanthocephaloides incrassatus,Acanthocephaloides incrassatus,Acanthocephaloides,Arhythmacanthidae,Echinorhynchida,Palaeacanthocephala,Acanthoce
Acanthocephaloides propinquus,Acanthocephaloides propinquus,Acanthocephaloides,Arhythmacanthidae,Echinorhynchida,Palaeacanthocephala,Acanthoceph
Acanthocephalus anguillae,Acanthocephalus anguillae,Acanthocephalus,Echinorhynchidae,Echinorhynchida,Palaeacanthocephala,Acanthocephala
Acanthocephalus clavula,Acanthocephalus clavula,Acanthocephalus,Echinorhynchidae,Echinorhynchida,Palaeacanthocephala,Acanthocephala
Acanthocephalus lucii,Acanthocephalus lucii,Acanthocephalus,Echinorhynchidae,Echinorhynchida,Palaeacanthocephala,Acanthocephala
Andracantha phalacrocoracis,Andracantha phalacrocoracis,Andracantha,Polymorphidae,Polymorphida,Palaeacanthocephala,Acanthocephala
Andracantha tunitae,Andracantha tunitae,Andracantha,Polymorphidae,Polymorphida,Palaeacanthocephala,Acanthocephala
```

Homework

Requested file: species X family 2D matrix example

```
species_name,Echinorhynchidae,Arhythmacanthidae,Polymorphidae
Acanthocephaloides distinctus,0,1,0
Acanthocephaloides geneticus,0,1,0
Acanthocephaloides incrassatus,0,1,0
Acanthocephaloides propinquus,0,1,0
Acanthocephalus anguillae,1,0,0
Acanthocephalus clavula,1,0,0
Acanthocephalus lucii,1,0,0
Andracantha phalacrocoracis,0,0,1
Andracantha tunitae,0,0,1
```