

Deber 2 de bioinformática

1.10.1

1. Change directory to CSB/unix/sandbox.

```
Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub
$ cd CSBbook/

Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook (master)
$ cd unix

Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix (master)
$ cd sandbox/
```

2. What is the size of the file Marra2014_data.fasta?

```
Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/sandbox (master)
$ ls -lh ../data/Marra2014_data.fasta
-rw-r--r-- 1 Inspiron 197121 563K Apr 20 19:02 ../data/Marra2014_data.fasta
```

3. Create a copy of Marra2014_data.fasta in the sandbox and name it my_file.fasta.

```
Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/sandbox (master)
$ cp ../data/Marra2014_data.fasta deber2.fasta

Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/sandbox (master)
$ ls
'Papers and reviews'/' creator.sh  deber2.fasta  g01/  saliz.txt  tc/
```

4. Howmany contigs are classified as isogroup00036?

```
Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/sandbox (master)
$ grep -c isogroup00036 deber2.fasta
16

Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/sandbox (master)
$ grep isogroup00036 deber2.fasta | wc -l
16
```

5. Replace the original “two-spaces” delimiter with a comma.

```
Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/sandbox (master)
$ cat deber2.fasta | tr -s ' ' ',' | head -n 3
>contig00001,length=527,numreads=2,gene=isogroup00001,status=it_thresh
ATCCTAGCTACTCTGGAGACTGAGGATTGAAGTTCAAAGTCAGCTCAAGCAAGAGATTG
TTTACAATTAACCCACAAAAGGCTGTTACTGAAGGTGTGGCTTAAGTGTCAAGCAACAG
```

6. Howmany unique isogroups are in the file?

```
Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/sandbox (master)
$ grep '>' deber2.fasta | cut -d ',' -f 4 | head -n 2
>contig00001 length=527 numreads=2 gene=isogroup00001 status=it_thresh
>contig00002 length=551 numreads=8 gene=isogroup00001 status=it_thresh
```

7. Which contig has the highest number of reads (numreads)? How many

reads does it have?

```
Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/sandbox (master)
$ grep '>' deber2.fasta | cut -d ',' -f 1,3 | head -n 3
>contig00001 length=527 numreads=2 gene=isogroup00001 status=it_thresh
>contig00002 length=551 numreads=8 gene=isogroup00001 status=it_thresh
>contig00003 length=541 numreads=2 gene=isogroup00001 status=it_thresh
```

1.10.2

1. How many times were the levels of individuals 3 and 27 recorded?

3

```
Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/sandbox (master)
$ cut -f 1 deber2.fasta | grep -c -w 3
29
```

27

```
Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/sandbox (master)
$ cut -f 1 deber2.fasta | grep -c -w 27
9
```

2. Write a script taking as input the filename and the ID of the individual, and returning the number of records for that ID.

```
Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/sandbox (master)
$ nano pregunta3.sh

Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/sandbox (master)
$ cat pregunta3.sh
#!/bin/bash

# Solicitar al usuario el nombre del archivo
read -p "Ingrese el nombre del archivo: " archivo

# Solicitar al usuario el número de ID (expresión regular)
read -p "Ingrese el número de ID a buscar: " id

# Realizar la búsqueda en el archivo y obtener el número de registro del individuo de interés
registro=$(grep -n "$id" "$archivo" | cut -d: -f1)

# Verificar si se encontró el registro o no
if [ -n "$registro" ]; then
    echo "El número de registro del individuo con el ID $id es: $registro"
else
    echo "No se encontró ningún registro con el ID $id en el archivo $archivo"
fi

Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/sandbox (master)
$ bash pregunta3.sh
Ingrese el nombre del archivo: deber2.fasta
Ingrese el número de ID a buscar: 3
El número de registro del individuo con el ID 3 es: 22
33
50
131
240
264
346
361
377
381
392
405
409
420
```

3. [Advanced]17 Write a script that returns the number of times each

individual was sampled.

```
Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/sandbox (master)
$ nano pregunta23.sh

Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/sandbox (master)
$ cat pregunta23.sh
#!/bin/bash

# Pedir al usuario el nombre del archivo y el número de ID
echo "Ingrese el nombre del archivo:"
read archivo
echo "Ingrese el número de ID:"
read id

# Obtener los registros de interés
registros=$(tail -n +2 "$archivo" | cut -d ',' -f 1 | sort -n | uniq)

# Iterar sobre cada registro de interés
for registro in $registros
do
    # Contar el número de registros para el ID actual
    num_registros=$(grep -c "^$registro," "$archivo")

    # Mostrar el número de ID y los registros correspondientes
    echo "ID: $registro - Registros: $num_registros"
done

Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/sandbox (master)
$ bash pregunta23.sh
Ingrese el nombre del archivo:
deber2.fasta
Ingrese el número de ID:
27
ID: >contig00002 - Registros: 0
ID: length=551 - Registros: 0
ID: numreads=8 - Registros: 0
ID: gene=isogroup00001 - Registros: 0
- Registros: 0hresh
ID: >contig00003 - Registros: 0
ID: length=541 - Registros: 0
ID: numreads=2 - Registros: 0
ID: gene=isogroup00001 - Registros: 0
- Registros: 0hresh
ID: >contig00004 - Registros: 0
ID: length=291 - Registros: 0
ID: numreads=3 - Registros: 0
ID: gene=isogroup00001 - Registros: 0
```

```
- Registros: 0GGCGGCGGGTGTCTCTCGGCGGCGGCGGAGGCAGTAGCCAGTATGTCG
- Registros: OCCAaTAGATCAGCAGTTTTGGAGACTGGCCAGGcTTCTGCAAGTACCAa
- Registros: 0Ag
- Registros: 0ACGTACAGTGAGTGTTTTGGACCCCTACTACAGGCCACGAGATGGTGCT
- Registros: 0CACAACCTTCTTCGACCCCGCAGGAGGTGGAGACCCAGTACTTTACCAa
- Registros: 0TCACCGTCGTTGCCACAGTCACCGTCACATTGAAGGCTCGGGGGTCAC
- Registros: 0CCAGGCCTAGGAATTGAACCAACTGCTGTCAACCACTGAGGATGAGAGGG
- Registros: 0GGGACACCTCTGATCCACTCACACTATCATCACAAGTCATCAAAACAGC
- Registros: 0AGAGTTTTCTTCCTCATCTGCTTGCCCC
```

1.10.3 Plant–Pollinator Networks

Saavedra and Stouffer (2013) studied several plant–pollinator networks. These can be represented as rectangular matrices where the rows are pollinators, the columns plants, a 0 indicates the absence and 1 the presence of an interaction between the plant and the pollinator. The data of Saavedra and Stouffer (2013) can be found in the directory CSB/unix/data/Saavedra2013.

1. Write a script that takes one of these files and determines the number of rows (pollinators) and columns (plants). Note that columns are sep-arated by spaces and that there is a space at the end of each line. Your script should return

```
Inspiron@ElizaSalazar20 MINGW64 ~/Documents/GitHub/CSBbook/unix/data/Saavedra2013 (master)
$ nano pregunta31.sh

Inspiron@ElizaSalazar20 MINGW64 ~/Documents/GitHub/CSBbook/unix/data/Saavedra2013 (master)
$ cat pregunta31.sh
# Pedir al usuario el nombre del archivo
echo "Ingrese el nombre del archivo:"
read archivo

# Contar el número de filas en el archivo
num_filas=$(wc -l < "$archivo")

# Mostrar el número de filas
echo "El número de filas en el archivo es: $num_filas"

# Pedir al usuario el nombre del archivo
echo "Ingrese el nombre del archivo:"
read archivo

# Obtener la primera fila del archivo
primera_fila=$(head -n 1 "$archivo")

# Eliminar espacios y saltos de línea
sin_espacios=$(echo "$primera_fila" | tr -d '[:space:]')

# Contar el número de caracteres
num_columnas=$(echo -n "$sin_espacios" | wc -c)

# Mostrar el número de columnas
echo "El número de columnas en la primera fila del archivo es: $num_columnas"

Inspiron@ElizaSalazar20 MINGW64 ~/Documents/GitHub/CSBbook/unix/data/Saavedra2013 (master)
$ bash pregunta31.sh
Ingrese el nombre del archivo:
n1.txt
El número de filas en el archivo es: 97
Ingrese el nombre del archivo:
n1.txt
El número de columnas en la primera fila del archivo es: 80
```

2. [Advanced]18 Write a script that prints the numbers of rows and columns for each network:

```

Inspiron@ElizaSalazar20 MINGW64 ~/Documents/GitHub/CSBbook/unix/data/Saavedra2013 (master)
$ nano pregunta32.sh

Inspiron@ElizaSalazar20 MINGW64 ~/Documents/GitHub/CSBbook/unix/data/Saavedra2013 (master)
$ cat pregunta32.sh
# Pedir al usuario el nombre del archivo
echo "Ingrese el nombre del archivo:"
read archivo

# Contar el número de filas en el archivo
num_filas=$(wc -l < "$archivo")

# Mostrar el número de filas
echo "El número de filas en el archivo es: $num_filas"

# Realizar el proceso para cada fila
for ((i=1; i<=num_filas; i++))
do
    # Obtener la fila actual
    fila=$(head -n $i "$archivo" | tail -n 1)

    # Eliminar espacios y saltos de línea
    sin_espacios=$(echo "$fila" | tr -d '[:space:]')

    # Contar el número de columnas
    num_columnas=$(echo -n "$sin_espacios" | wc -c)

    # Mostrar el número de columnas para la fila actual
    echo "Número de columnas en la fila $i: $num_columnas"
done

```

```

Inspiron@ElizaSalazar20 MINGW64 ~/Documents/GitHub/CSBbook/unix/data/Saavedra2013 (master)
$ bash pregunta32.sh
Ingrese el nombre del archivo:
n1.txt
El número de filas en el archivo es: 97
Número de columnas en la fila 1: 80
Número de columnas en la fila 2: 80
Número de columnas en la fila 3: 80
Número de columnas en la fila 4: 80
Número de columnas en la fila 5: 80
Número de columnas en la fila 6: 80
Número de columnas en la fila 7: 80
Número de columnas en la fila 8: 80
Número de columnas en la fila 9: 80
Número de columnas en la fila 10: 80
Número de columnas en la fila 11: 80
Número de columnas en la fila 12: 80
Número de columnas en la fila 13: 80
Número de columnas en la fila 14: 80
Número de columnas en la fila 15: 80
Número de columnas en la fila 16: 80
Número de columnas en la fila 17: 80
Número de columnas en la fila 18: 80
Número de columnas en la fila 19: 80
Número de columnas en la fila 20: 80
Número de columnas en la fila 21: 80
Número de columnas en la fila 22: 80
Número de columnas en la fila 23: 80
Número de columnas en la fila 24: 80

```

3. Which file has the largest number of rows? Which has the largest number of columns?

```
MINGW64:/c:/Users/Inspiron/Documents/GitHub/CSBbook/unix/data/Saavedra2013
Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/data/Saavedra2013 (master)
$ nano pregunta33.sh

Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/data/Saavedra2013 (master)
$ cat pregunta33.sh
# Pedir al usuario el nombre del archivo
echo "Ingrese el nombre del archivo:"
read archivo

# Contar el número de filas en el archivo
num_filas=$(wc -l < "$archivo")

# Mostrar el número de filas
echo "El número de filas en el archivo es: $num_filas"

# Inicializar variables para el número máximo de filas y columnas
max_filas=0
max_columnas=0

# Realizar el proceso para cada fila
for ((i=1; i<=num_filas; i++))
do
    # Obtener la fila actual
    fila=$(head -n $i "$archivo" | tail -n 1)

    # Eliminar espacios y saltos de línea
    sin_espacios=$(echo "$fila" | tr -d '[:space:]')

    # Contar el número de columnas
    num_columnas=$(echo -n "$sin_espacios" | wc -c)

    # Actualizar el número máximo de columnas si es necesario
    if ((num_columnas > max_columnas))
    then
        max_columnas=$num_columnas
    fi

    # Actualizar el número máximo de filas si es necesario
    if ((i > max_filas))
    then
        max_filas=$i
    fi
done

# Mostrar el número máximo de filas y columnas
echo "El número máximo de filas es: $max_filas"
echo "El número máximo de columnas es: $max_columnas"
```

```
Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/data/Saavedra2013 (master)
$ bash pregunta33.sh
Ingrese el nombre del archivo:
n10.txt
El número de filas en el archivo es: 14
El número máximo de filas es: 14
El número máximo de columnas es: 20

Inspiron@ElizaSalazarm20 MINGW64 ~/Documents/GitHub/CSBbook/unix/data/Saavedra2013 (master)
$ bash pregunta33.sh
Ingrese el nombre del archivo:
n11.txt
El número de filas en el archivo es: 270
El número máximo de filas es: 270
El número máximo de columnas es: 91
```

1.10.4 Data Explorer

Buzzard et al. (2016) collected data on the growth of a forest in Costa Rica. In

the file Buzzard2015_data.csv you will find a subset of their data, including taxonomic information, abundance, and biomass of trees.

1. Write a script that, for a given CSV file and column number, prints

- the corresponding column name;
- the number of distinct values in the column;
- the minimum value;
- the maximum value.

```
Inspiron@ElizaSalazar20 MINGW64 ~/Documents/GitHub/CSBbook/unix/data (master)
$ nano pregunta4.sh

Inspiron@ElizaSalazar20 MINGW64 ~/Documents/GitHub/CSBbook/unix/data (master)
$ cat pregunta4.sh
#Extraccion del nombre de la columna
cut -d ',' -f 7 Buzzard2015_data.csv | head -n 1
#Obteccion de valores distintos
cut -d ',' -f 7 Buzzard2015_data.csv | tail -n +2 | sort | uniq | wc -l
# Obtener el Mínimo
cut -d ',' -f 7 Buzzard2015_data.csv | tail -n +2 | sort -n | head -n 1
# Obtener el Máximo
cut -d ',' -f 7 ../data/Buzzard2015_data.csv | tail -n +2 | sort -n | tail -n 1

Inspiron@ElizaSalazar20 MINGW64 ~/Documents/GitHub/CSBbook/unix/data (master)
$ bash pregunta4.sh
biomass
285
1.048466198
14897.29471
```