Introducción a la bioinformática

Laura Gómez-Romero

Instituto Nacional de Medicina Genómica

lgomez@inmegen.gob.mx

October 15, 2021

Tabla de contenidos

1. Funcionamiento del comando tr
2. Funcionamiento del comando sed
3. Manos a la obra

Tabla de contenidos

1. Funcionamiento del comando tr
2. Funcionamiento del comando sed
3. Manos a la obra

Principal función del comando tr

> "Translate characters"
>
> — Man page tr

Traducir un caracter por otro

La función más básica de tr

```
## traducir un caracter a otro
echo "este es un ejemplo" | tr 'e' 'a'
```

Traducir una lista de caracteres

Explorando las funcionalidades de tr

```
## Traducir una lista de caracteres por otra lista
echo "este es un ejemplo" | tr 'euo' 'EUO'

## Podemos especificar rangos de caracteres
echo "este es un ejemplo" | tr a-z A-Z

## O clases de caracteres POSIX
echo "este es un ejemplo" | tr [:lower:] [:upper:]

## La primer lista puede ser mas larga que la segunda
echo "este es un ejemplo" | tr [:lower:] 9
```

Listas de caracteres POSIX

POSIX class	similar to	meaning
[:upper:]	[A-Z]	uppercase letters
[:lower:]	[a-z]	lowercase letters
[:alpha:]	[A-Za-z]	upper- and lowercase letters
[:digit:]	[0-9]	digits
[:xdigit:]	[0-9A-Fa-f]	hexadecimal digits
[:alnum:]	[A-Za-z0-9]	digits, upper- and lowercase letters
[:punct:]		punctuation (all graphic characters except letters and digits)
[:blank:]	[\t]	space and TAB characters only
[:space:]	[\t\n\r\f\v]	blank (whitespace) characters
[:cntrl:]		control characters
[:graph:]	[^ [:cntrl:]]	graphic characters (all characters which have graphic representation)
[:print:]	[[:graph:]]	graphic characters and space
[:word:]	[[:alnum:]_]	Alphanumeric characters with underscore character _, meaning alnum + _. It is a bash specific character class.

Imagen tomada de:
`https://en.wikibooks.org/wiki/Regular_Expressions/POSIX_Basic_Regular_Expressions`

Algunas opciones del comando tr

Comando	**Función**
-d	Elimina caracteres
-s	Comprime múltiples ocurrencias de los caracteres listados

Eliminando y comprimiendo caracteres

Explorando las -d y -s del comando tr

```
## Eliminar un caracter
echo "este es un ejemplo" | tr -d e

## Eliminando una lista de caracteres
echo "este es un ejemplo" | tr 'euo' 'EUO' | tr -d [:lower:]

## Comprimiendo ocurrencias
echo "este es un ejemplo" | tr 'euo' 'EUO' | \
        tr -d [:lower:] | tr -s 'E'
```

Únicamente se comprimirán las múltiples ocurrencias de un caracter si se encuentran contiguas en la cadena de texto.

Tabla de contenidos

1. Funcionamiento del comando tr
2. Funcionamiento del comando sed
3. Manos a la obra

Descripción del comando sed

> "Sream editor"
>
> — Man page sed

Creando nuestro archivo de prueba

```
## Contenido del archivo
primer linea
segunda linea prueba linea
tercer linea linea
cuarta linea
```

La función más utilizada de sed: sutitución

sed toma una instrucción y la aplica en cada línea del input

```
## Sustitucion, unicamente la primera ocurrencia
sed 's/linea/n/' sed.txt

## Todas las ocurrencias
sed 's/linea/n/g' sed.txt

## Eligiendo el numero de linea
sed '2s/linea/n/g' sed.txt

 ## Eligiendo un rango de lineas
sed '2,4s/linea/n/g' sed.txt
```

Selección de líneas utilizando sed

sed toma una instrucción y la aplica en cada línea del input, imprimiendo una línea de output por cada línea del input

Para inhabilitar este comportamiento default existe la opción -n

```
## Elegir un rango de lineas a imprimir
sed -n '2,4p' sed.txt

## Imprimir las lineas que contienen un patron
sed -n '/tercer/p' sed.txt

## Impirmir las lineas que NO contienen un patron
sed -n '/tercer/!p' sed.txt
```

Expresiones regulares con sed

Encerremos el contenido de un patrón en una subexpresión

```
## Utilizando subexpresiones y sustituciones
sed -E 's/^([a-z]+) ([a-z]+) /\2-\1/' sed.txt
```

Expresiones regulares con sed

Encerremos el contenido de un patrón en una subexpresión

```
## Utilizando subexpresiones y sustituciones
sed -E 's/^([a-z]+) ([a-z]+) /\2-\1/' sed.txt
```

Desmenuzando el contenido:

- s Activa el modo sustitución de sed
- ^ Denota el inicio de la línea
- [a-z]+ Letras, una o más veces
- () Guardan el contenido de una subexpresión
- \1 Imprime el contenido de la segunda subexpresión
- \2 Imprime el contenido de la segunda subexpresión

Tabla de contenidos

1. Funcionamiento del comando tr
2. Funcionamiento del comando sed
3. Manos a la obra

Preparando nuestro directorio de trabajo

- Muévete a tu directorio home en el servidor
- Crea la carpeta practica7
- Entra a la carpeta practica7
- Crea la carpeta data
- Copia los archivos ubicados en /home/lgomez/WelcomeBioinfo/datos/practica7 a tu carpeta data

Familiarizándonos con los datos de trabajo

- Visualiza y explora cada uno de los archivos
- ¿Qué tipo de archivos son? ¿cuál es su contenido?

Verifica la integridad de los datos

- En el mismo directorio existe el archivo md5sum.txt, el contiene contiene la suma md5 para los distintos archivos
- Realiza la verificación de integridad

RECUERDA

El día de hoy tienes una batería de comandos que te pueden ser útiles

Siempre hay más de una forma correcta de llegar al resultado correcto

Ejercicio: buscar un dominio en una familia de proteínas

- ¿Existe el dominio de ferredoxina reductasa en las secuencias del archivo fnr_protein.faa?

Ejercicio: buscar un dominio en una familia de proteínas

¿Existe el dominio de ferredoxina reductasa en las secuencias del archivo fnr_protein.faa?

PROCEDIMIENTO:

- Buscar el patrón

Ejercicio: buscar un dominio en una familia de proteínas

¿Existe el dominio de ferredoxina reductasa en las secuencias del archivo fnr_protein.faa?

```
## Buscar el patrón
sed -n \
        '/SKKNEEGVIVNRYRPKEPYTGKCLLNTKITADDAPGETWHMVFSHQG/p' \
        data/fnr_protein.faa
```

¿A qué proteína pertenece? ¿Seguros que sólo existe en una proteína?

Ejercicio: buscar un dominio en una familia de proteínas

¿Existe el dominio de ferredoxina reductasa en las secuencias del archivo fnr_protein.faa?

- Convertir el header y la secuencia de cada proteína a una única línea
- Buscar el patrón

Ejercicio: buscar un dominio en una familia de proteínas

¿Existe el dominio de ferredoxina reductasa en las secuencias del archivo fnr_protein.faa?

```
## Convertir el header y la secuencia de cada proteína
## a una única línea
cat data/fnr_protein.faa | tr '\n' ' '
```

Recuerda que existen caracteres especiales que pueden confundir nuestros comandos

Ejercicio: buscar un dominio en una familia de proteínas

¿Existe el dominio de ferredoxina reductasa en las secuencias del archivo fnr_protein.faa?

```
## Imprimiendo todos los caracteres de nuestro archivo
cat data/fnr_protein.faa | od -c
```

\n y \r son caracteres especiales para especificar salto de línea.
Sistemas MS-DOS generalmente agregan los caracteres \r al final de las líneas.
Este caracter debe ser eliminado para trabajar en ambientes Linux.

Ejercicio: buscar un dominio en una familia de proteínas

¿Existe el dominio de ferredoxina reductasa en las secuencias del archivo fnr_protein.faa?

```
## Eliminando el caracter \r
## Utilizando la sustitucion in-place de sed
sed -i 's/\r//g' data/fnr_protein.faa

## Verificando nuestro resultado
cat data/fnr_protein.faa | od -c
```

Ahora si regresemos a nuestra pregunta

Ejercicio: buscar un dominio en una familia de proteínas

¿Existe el dominio de ferredoxina reductasa en las secuencias del archivo fnr_protein.faa?

```
## Convertir el header y la secuencia de cada proteina
## a una unica linea

## Eliminamos los saltos de linea, tenemos una sola linea enorme
cat data/fnr_protein.faa | tr -d '\n'

## Agregando saltos de linea en posiciones convenientes
cat data/fnr_protein.faa | tr -d '\n' | tr '>' '\n'

## Buscando nuestro dominio
cat data/fnr_protein.faa | tr -d '\n' | tr '>' '\n' | \
 sed -n '/SKKNEEGVIVNRYRPKEPYTGKCLLNTKITADDAPGETWHMVFSHQG/p'

## Buscando parte de nuestro dominio
cat data/fnr_protein.faa | tr -d '\n' | tr '>' '\n' | \
 sed -n '/CLLNTKIT/p'
```

Este dominio no está totalmente conservado

```
CLUSTAL O(1.2.4) multiple sequence alignment

gi|15239282|ref|NP_201420.1|        MAAAISAAVSLPS----SKSSSLLTKISSVSPQRIFLKKS--------TVCYRRVVSVKA      48
gi|145323954|ref|NP_001077566.1|    MATTMNAAVSLTS----SNSSSFPATSCAIAPERIRFTKGAFYYKSNNVVTGKRVFSIKA      56
gi|115465942|ref|NP_001056570.1|    MAAVTAAAVSTSAAAAVTKASPSPAHCFLPCPPRTRAA----------HQRGLLLRA      47
                                    **:.  ****  :    :::*   :     .* *                 :* . ::*

gi|15239282|ref|NP_201420.1|        QVTTD---TTEAPPVKVVKESKKQEEGIVVNKFKPKNPYTGRCLLNTKITGDDAPGETWH     105
gi|145323954|ref|NP_001077566.1|    QITTE--TDTPTPAKKVEKVSKKNEEGVIVNRYRPKEPYTGKCLLNTKITADDAPGETWH     114
gi|115465942|ref|NP_001056570.1|    QVSTTDAAAVAAAPAKKEKISKKHDEGVVTNKYRPKEPYVGKCLLNTKITADDAPGETWH     107
                                    *::*     . :   *  * ***::**::.*:::**:**.*:********.*********

gi|15239282|ref|NP_201420.1|        IVFTTEGEVPYREGQSIGVIPEGIDKNGKPHKLRLYSIASSAIGDFGDSKTVSLCVKRLV     165
gi|145323954|ref|NP_001077566.1|    MVFSHQGKIPYREGQSVGVIADGIDKNGKPHKVRLYSIASSALGDLGNSETVSLCVKRLV     174
gi|115465942|ref|NP_001056570.1|    MVFSTEGEIPYREGQSIGVIADGVDKNGKPHKLRLYSIASSALGDFGDSKTVSLCVKRLV     167
                                    :**: :*::*******:*** :*:********:*********:**:*:*:**********

gi|15239282|ref|NP_201420.1|        YTNDGGEIVKGVCSNFLCDLKPGDEAKITGPVGKEMLMPKDPNATIIMLGTGTGIAPFRS     225
gi|145323954|ref|NP_001077566.1|    YTNDQGETVKGVCSNFLCDLAPGSDVKLTGPVGKEMLMPKDPNATVIMLATGTGIAPFRS     234
gi|115465942|ref|NP_001056570.1|    YTNDQGEIVKGVCSNFLCDLKPGSDVKITGPVGKEMLMPKDPNANIIMLATGTGIAPFRS     227
                                    **** ** ************ **.:.*:***************.:***.**********

gi|15239282|ref|NP_201420.1|        FLWKMFFEEHEDYKFNGLAWLFLGVPTSSSLLYKEEFEKMKEKNPDNFRLDFAVSREQTN     285
gi|145323954|ref|NP_001077566.1|    FLWKMFFEKHDDYKFNGLAWLFLGVPTTSSLLYQEEFDKMKAKAPENFRVDYAISREQAN     294
gi|115465942|ref|NP_001056570.1|    FLWKMFFEKYDDYKFNGLAWLFLGVPTSSSLLYKEEFDKMKAKAPENFRVDYAVSREQTN     287
                                    ********:::***************:*****:***:*** * *:***:*:*:*****:*

gi|15239282|ref|NP_201420.1|        EKGEKMYIQTRMAEYAEELWELLKKDNTFVYMCGLKGMEKGIDDIMVSLAAKDGIDWLEY     345
gi|145323954|ref|NP_001077566.1|    DKGEKMYIQTRMAQYAAELWELLKKDNTFVYMCGLKGMEKGIDDIMVSLAANDGIDWFDY     354
gi|115465942|ref|NP_001056570.1|    AQGEKMYIQTRMAEYKEELWELLKKDHTYVYMCGLKGMEKGIDDIMVSLAAKDGIDWADY     347
                                     :***********:*  *********:*:**********************:***** :*

gi|15239282|ref|NP_201420.1|        KKQLKRSEQWNVEVY   360
gi|145323954|ref|NP_001077566.1|    KKQLKKAEQWNVEVY   369
gi|115465942|ref|NP_001056570.1|    KKQLKKGEQWNVEVY   362
                                    *****:.********
```

Alineamiento realizado con Clustal Omega:
`https://www.ebi.ac.uk/Tools/msa/clustalo/`

Ejercicio: identificar cuáles genes podrían ser blanco de un miRNA

- ¿Existe la semilla del miRNA mirBio en las secuencias del archivo fnr_gene.fa?
- ¿Existe la secuencia complementaria reversa de la semilla del miRNA mirBio en las secuencias del archivo fnr_gene.fa?

Ejercicio: identificar cuáles genes podrían ser blanco de un miRNA

¿Existe la semilla del miRNA mirBio en las secuencias del archivo fnr_gene.fa?

El procedimiento, implementando lo que hemos aprendido es:

- Explorar el archivo por caracteres especiales
- Si existen caracteres especiales indeseados, eliminarlos
- Convertir el header y la secuencia de cada gene a una línea
- Buscar nuestra secuencia

Ejercicio: identificar cuáles genes podrían ser blanco de un miRNA

¿Existe la semilla del miRNA mirBio en las secuencias del archivo fnr_gene.fa?

```
## Explorar el archivo por caracteres especiales
cat data/fnr_gene.fa | od -c

## Si existen caracteres especiales indeseados, eliminarlos
sed -i 's/\r//g' data/fnr_gene.fa
cat data/fnr_gene.fa | od -c

## Convertir el header y la secuencia de cada gene a una línea
cat data/fnr_gene.fa | tr -d '\n' | tr '>' '\n'

## Buscar nuestra secuencia
cat data/fnr_gene.fa | tr -d '\n' | tr '>' '\n' | \
 sed -n '/GACCATAATGTCATC/p'
```

Ejercicio: identificar cuáles genes podrían ser blanco de un miRNA

¿Existe la semilla del miRNA mirBio en las secuencias del archivo fnr_gene.fa?

```
## Explorar el archivo por caracteres especiales
cat data/fnr_gene.fa | od -c

## Si existen caracteres especiales indeseados, eliminarlos
sed -i 's/\r//g' data/fnr_gene.fa
cat data/fnr_gene.fa | od -c

## Convertir el header y la secuencia de cada gene a una línea
cat data/fnr_gene.fa | tr -d '\n' | tr '>' '\n'

## Buscar nuestra secuencia
cat data/fnr_gene.fa | tr -d '\n' | tr '>' '\n' | \
 sed -n '/GACCATAATGTCATC/p'
```

Esta secuencia no existe en nuestros genes

Ejercicio: identificar cuáles genes podrían ser blanco de un miRNA

¿Existe la secuencia complementaria reversa de la semilla del miRNA mirBio en las secuencias del archivo fnr_gene.fa?

El procedimiento, implementando lo que hemos aprendido es:

- Explorar el archivo por caracteres especiales, HECHO
- Si existen caracteres especiales indeseados, eliminarlos, HECHO
- Convertir el header y la secuencia de cada gene a una línea, HECHO
- Obtener la secuencia complementaria reversa de nuestro miRNA [HINT, comando tr y rev]
- Buscar nuestra secuencia

Ejercicio: identificar cuáles genes podrían ser blanco de un miRNA

¿Existe la secuencia complementaria reversa de la semilla del miRNA mirBio en las secuencias del archivo fnr_gene.fa?

```
## Obtener la secuencia complementaria reversa

## Obtener la secuencia complementaria
cat data/mirBio.txt | tr 'GACT' 'ctga' | tr [:lower:] [:upper:]

## Obtener la secuencia complementaria reversa
cat data/mirBio.txt | tr 'GACT' 'ctga' | tr [:lower:] [:upper:] | rev

## Convertir el header y la secuencia de cada gene a una línea
cat data/fnr_gene.fa | tr -d '\n' | tr '>' '\n'

## Buscar nuestra secuencia, en el archivo de genes modificado
cat data/fnr_gene.fa | tr -d '\n' | tr '>' '\n' | \
 sed -n '/GATGACATTATGGTC/p'
```

Ejercicio: identificar cuáles genes podrían ser blanco de un miRNA

¿Existe la secuencia complementaria reversa de la semilla del miRNA mirBio en las secuencias del archivo fnr_gene.fa?

```
## Obtener la secuencia complementaria reversa

## Obtener la secuencia complementaria
cat data/mirBio.txt | tr 'GACT' 'ctga' | tr [:lower:] [:upper:]

## Obtener la secuencia complementaria reversa
cat data/mirBio.txt | tr 'GACT' 'ctga' | tr [:lower:] [:upper:] | rev

## Convertir el header y la secuencia de cada gene a una línea
cat data/fnr_gene.fa | tr -d '\n' | tr '>' '\n'

## Buscar nuestra secuencia, en el archivo de genes modificado
cat data/fnr_gene.fa | tr -d '\n' | tr '>' '\n' | \
 sed -n '/GATGACATTATGGTC/p'
```

Ejercicio: identificar cuáles genes podrían ser blanco de un miRNA

¿Existe la secuencia complementaria reversa de la semilla del miRNA mirBio en las secuencias del archivo fnr_gene.fa?

Los dos homólogos en A. thaliana podrían ser regulados por nuestro miRNA putativo, sin embargo el homólogo en O. sativa no podría ser un blanco

Es importante mencionar que esta conclusión es una simplificación del problema real, la complementaridad de los miRNAs con sus genes blanco puede no ser perfecta.

Verifiquemos la vecindad de nuestra secuencia blanco

```
gi|115465941|ref|NM_001063105.1|      GCTCAAGGAGAGAAGATGTACATTCAGACCAGGATGGCAGAGTACAAGGAAGAGCTGTGG      989
gi|145323953|ref|NM_001084097.1|      GATAAAGGAGAGAAAATGTATATCCAGACTCGGATGGCGCAATACGCAGCTGAGTTATGG      1076
gi|145359720|ref|NM_126017.5|         GAGAAGGGAGAGAAAATGTACATTCAGACAAGAATGGCAGAGTATGCAGAAGAGCTGTGG      1102
                                      *   * ******** ***** ** *****  * *****  * **    *  *** * ***

gi|115465941|ref|NM_001063105.1|      GAGCTCCTGAAGAAGGACCACACCTATGTGTACATGTGTGGACTGAAAGGCATGGAGAAG      1049
gi|145323953|ref|NM_001084097.1|      GAGTTGTTGAAGAAAGACAACACTTTTGTTTACATGTGTGGACTCAAGGGAATGGAGAAA      1136
gi|145359720|ref|NM_126017.5|         GAGTTGCTGAAGAAAGACAACACCTTTGTTTACATGTGTGGTCTTAAGGGTATGGAGAAG      1162
                                      *** *  ******* *** **** * *** ********** ** ** ** ********

gi|115465941|ref|NM_001063105.1|      GGTATTGATGACATTATGGTGTCATTGGCTGCAAAAGATGGAATCGACTGGGCTGATTAC      1109
gi|145323953|ref|NM_001084097.1|      GGAATTGATGACATTATGGTCTCATTGGCTGCAAATGACGGTATTGACTGGTTTGATTAC      1196
gi|145359720|ref|NM_126017.5|         GGTATCGATGACATTATGGTCTCGCTTGCTGCTAAAGATGGGATCGATTGGTTGGAGTAC      1222
                                      ** ** ************** **  * ***** ** ** ** ** ** ***   ** ***

gi|115465941|ref|NM_001063105.1|      AAGAAGCAACTGAAGAAGGGCGAGCAATGGAACGTGGAAGTCTACTAATTCTTCCAATTT      1169
gi|145323953|ref|NM_001084097.1|      AAGAAGCAGTTGAAGAAGGCAGAGCAATGGAACGTTGAAGTCTACTGATCAAAAAGCCTT      1256
gi|145359720|ref|NM_126017.5|         AAGAAGCAATTGAAGAGGAGTGAACAGTGGAATGTTGAAGTCTACTAAGGAAGCTTCTGA      1282
                                      ********  ****** *   ** ** ***** ** ********** *

gi|115465941|ref|NM_001063105.1|      TCCTCACATCTGTTTCTTTTTTTTCTTCCATTTGTATCTGTGTGCACATCTGTGCCTGTG      1229
gi|145323953|ref|NM_001084097.1|      TGACATTTCTGTAGCAAAGTATAGC--------TGAACAAAACTGTAATTTTCGCTTCTG      1308
gi|145359720|ref|NM_126017.5|         GGGAGTAATTATATAATGTAGATAAAAAGCTTCTGATGCATTGTGAAATCTTCATATCTG      1342
                                                                        *          ** *     * **

gi|115465941|ref|NM_001063105.1|      ATCACTCTATAATGTTAGATAGGCGTATATATATACTGTTTGTCATGTTGGTTAAATTCA      1289
gi|145323953|ref|NM_001084097.1|      AATTTCTGTATTTTGAAGATAAGTTTTTTAGATATGTTATACTAAAAAAAGAGTTCTTTA      1368
gi|145359720|ref|NM_126017.5|         CTTCTTTTTTCTTTCTC--AAGGATTTTCAATCAAAACATTCAAA----AGAGAGCATCA      1396
                                            *       * *  * *           *    *     *      * *

gi|115465941|ref|NM_001063105.1|      AGCTT---CATATAAGAATTACTACTTAT---GTCTGATCCAAATACTACTATGGTCAAG      1343
gi|145323953|ref|NM_001084097.1|      ATCATTATCATTTACTTCTTCCCA----------------------------------      1392
gi|145359720|ref|NM_126017.5|         ATACCACCAAACTACTCTTTCTCACTTGTTGTTCATGTTATATTTCTTGCAACAATATAG      1456
                                      *        *  **    **   *
```

Alineamiento realizado con Clustal Omega:
`https://www.ebi.ac.uk/Tools/msa/clustalo/`

Ejercicio: calcular el contenido de GC

Los genes con los que hemos trabajado son homólogos entre ellos, lo cual quiere decir que provienen de un **ancestro común**.

Averigüemos que tanto han divergido, ¿cuál es su contenido de GC?

El contenido de GC se calcula:

- sumando la cantidad de G's + la cantidad de C's
- dividiendo este número entre la longitud de la secuencia

Ejercicio: calcular el contenido de GC

En equipo:

- Describan el procedimiento que seguirían, incluyendo el comando que aplicarían para cada paso, para resolver el problema.
- Pueden proponer generar tantos archivos intermediarios como sea necesario
- En su propuesta puede no estar incluida la operación aritmética final: G + C / longitud, sin embargo, si deben obtener cada elemento de la fórmula
- Preparen una diapositiva con el procedimiento propuesto para presentarla al resto de la clase.

Pueden utilizar todos los comandos vistos durante el curso.

Ejercicio: calcular el contenido de GC

¿Cuál es el contenido de GC para cada gene del archivo fnr_gene.fa?

El procedimiento propuesto por mi:

- Generar archivos independientes para cada gene
- Obtener el contenido de GC
  - Traducir G's y C's a saltos de línea
  - Contar los saltos de linea
- Obtener la longitud de las secuencias
  - Eliminar saltos de línea
  - Contar caracteres

Ejercicio: calcular el contenido de GC

¿Cuál es el contenido de GC para cada gene del archivo fnr_gene.fa?

```
## Generar archivos independientes para cada gene
grep -n ">" data/fnr_gene.fa
wc -l  data/fnr_gene.fa

sed -n '2,21p' data/fnr_gene.fa > thaliana_fnr2.txt
sed -n '23,44p' data/fnr_gene.fa > thaliana_fnr1.txt
sed -n '46,65p' data/fnr_gene.fa > sativa_fnr.txt
```

Ejercicio: calcular el contenido de GC

¿Cuál es el contenido de GC para cada gene del archivo fnr_gene.fa?

```
### Obtener el contenido de GC
### Traducir G's y C's a saltos de linea
### Contar los saltos de linea

cat sativa_fnr.txt | tr -d '\n' | tr [CG] '\n' | wc -l
cat thaliana_fnr1.txt | tr -d '\n' | tr [CG] '\n' | wc -l
cat thaliana_fnr2.txt | tr -d '\n' | tr [CG] '\n' | wc -l
```

Ejercicio: calcular el contenido de GC

¿Cuál es el contenido de GC para cada gene del archivo fnr_gene.fa?

```
### Obtener la longitud de las secuencias
### Eliminas saltos de linea
### Contar caracteres

cat sativa_fnr.txt | tr -d '\n' | wc
cat thaliana_fnr1.txt | tr -d '\n' | wc
cat thaliana_fnr2.txt | tr -d '\n' | wc
```

Ejercicio: calcular el contenido de GC

¿Cuál es el contenido de GC para cada gene del archivo fnr_gene.fa?

Contenido de GC = cantidad de G + C/longitud de la secuencia

- A. thaliana FNR1: 648/1509 = 42.9%
- A. thaliana FNR2: 563/1392 = 40.44%
- O. sativa: 719/1373 = 52.36%

Los genes homólogos en A. thaliana son muy parecidos entre ellos (en términos de contenido de GC) y ligeramente diferentes al homólogo en O. sativa

References I