

Ejercicio de parseo de archivos FASTA

Pablo Vinuesa

2019-02-23

Contents

Presentación	1
Licencia y términos de uso	1
Clonación del repositorio	2
Datos de secuencia para la práctica	2
Búsqueda y descarga de secuencias en GenBank usando el sistema ENTREZ	2
Práctica de parseo de archivos FASTA descargados de NCBI mediante ENTREZ	2
Acceso a las secuencias	2
Inspección del formato de las secuencias descargadas	2
Generación de estadísticas básicas a partir de la información disponible en las cabeceras . . .	3
¿Cuántas secuencias hay en el archivo recA_Bradyrhizobium_vinuesa.fna?	3
Veamos las 5 primeras líneas de cabeceras fasta usando grep y head	3
Generación de conteos usando el idioma grep 'REGEX' archivo sort uniq -c .	3
Imprime una lista ordenada de mayor a menor, del número de especies que contiene el archivo FASTA usando sort -nrkN	3
Edición de las cabeceras FASTA mediante herramientas de filtrado de UNIX	4
Exploremos todas las cabeceras FASTA del archivo recA_Bradyrhizobium_vinuesa.fna usando grep	4
Simplificación de las cabeceras FASTA usando sed (el <i>stream editor</i>)	4
Generación automática de archivos FASTA especie-específicos (avanzado)	5
Conversión de archivos FASTA a formato "FASTAB" usando perl 1-liners.	5
Filtrado del archivo en formato fastab con grep	5
Regeneración del formato FASTA a partir de FASTAB con un Perl 1-liner	5
Llamada a un bucle for para repetir operaciones sobre múltiples archivos	5

Presentación

El código presentado aquí corresponde a una adaptación de unas prácticas escritas por Pablo Vinuesa para el libro Sistemática Molecular y Bioinformática - guía práctica editado por la Facultad de Ciencias - UNAM, 2018.

Licencia y términos de uso

Este material didáctico lo distribuyo bajo la **Licencia No Comercial Creative Commons 4.0**

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0

y es parte del curso de **OMICAS_UAEM_UNAM** que imparto cada semestre en el Centro de Ciencias Genómicas - UNAM para alumnos de licenciatura, maestría y doctorado de la Universidad Autónoma del Estado de Morelos (UAEM) y de la Universidad Nacional Autónoma de México (UNAM).

Clonación del repositorio

Si tienes instalado git en tu computadora, puedes clonar el repositorio con todo el material del curso con el comando:

```
git clone https://github.com/vinuesa/OMICAS_UAEM.git
```

En ubuntu es muy fácil instalar git:

```
sudo apt install git
```

Datos de secuencia para la práctica

Para correr los ejercicios, asegúrate de tener el archivo `recA_Bradyrhizobium_vinuesa.fna` en el directorio actual de trabajo.

Búsqueda y descarga de secuencias en GenBank usando el sistema ENTREZ

El archivo `recA_Bradyrhizobium_vinuesa.fna` contiene secuencias del gen *recA* de bacterias del género *Bradyrhizobium* depositadas en GenBank por P. Vinuesa.

El siguiente bloque muestra el comando usado para descargarlas desde NCBI. El comando de filtrado debe pegarse en la ventana de captura de texto de la página que da acceso al sistema ENTREZ.

```
# pega esta sentencia o filtro en la ventana de captura para interrogar la base de datos
# de nucleótidos (nuccore) de NCBI mediante el sistema ENTREZ
Bradyrhizobium[orgn] AND vinuesa[auth] AND recA[gene]
```

Práctica de parseo de archivos FASTA descargados de NCBI mediante ENTREZ

Acceso a las secuencias

```
[ ! -d $HOME/intro2genomics ] && mkdir -p $HOME/intro2genomics
cd $HOME/intro2genomics

cp -r /home/vinuesa/cursos/intro2genomics/sesion1_parseo_fastas .

cd sesion1_parseo_fastas
```

Inspección del formato de las secuencias descargadas

- Visualiza la estructura del archivo fasta

```
head recA_Bradyrhizobium_vinuesa.fna
```

```
## >EU574327.1 Bradyrhizobium liaoningense strain ViHaR5 recombination protein A (recA) gene, partial c
## ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCTCCGGCTCGCTCGGGCTCGACA
## TCGCGCTCGGCATCGGCGGCCTGCCAAGGGCGTATCGTCGAGATCTACGGGCCGAATCCTCGGGCAA
## GACCACGCTGGCGCTGCATACGGTGGCGGAAGCGCAGAAGAAGGGCGGCATCTGCGCCTTCATCGACGCC
## GAGCACGCGCTCGACCCGGTCTATGCCCGCAAGCTCGGCGTCAACATCGACGAGCTCCTGATCTCGCAGC
## CCGACACCGCGGAGCAGGCGCTGGAGATCTGCGACACGCTGGTGGCTCGGGCGCTGTCGATGTGCTGGT
## GATCGACTCGGTTGCGGCGCTGGTGCCGAAGGCCGAGCTCGAAGGCGAGATGGGCGATGCGCTGCCAGGC
## TTGCAGGCCCGTCTGATGAGCCAGGCGCTGCGCAAGCTGACGGCCTCCATCAACAAGTCCAACACCATGG
## TGATCTTCATCAACCAGATC
```

- Filtra el archivo con **grep** para desplegar sólo las 10 primeras cabeceras y analizar su estructura

NOTAS MPORTANTES: 1. ¡Fíjate que vamo a usar la expresión **grep** ‘>’ **archivo.fasta** y no **grep > archivo.fasta**! ¿Sabes por qué? 2. Usamos **grep** con la **expresión regular** ‘^>’ en vez de simplemente ‘>’, para que busque específicamente las *líneas que inician con* ‘>’. Este filtrado con **grep** es computacionalmente mucho más eficiente, ya que evita buscar en todas las líneas del archivo, saltándose aquellas que contienen sólo secuencias, ya que no inician con ‘>’.

```
grep '^>' recA_Bradyrhizobium_vinuesa.fna | head
```

```
## >EU574327.1 Bradyrhizobium liaoningense strain ViHaR5 recombination protein A (recA) gene, partial co
## >EU574326.1 Bradyrhizobium liaoningense strain ViHaR4 recombination protein A (recA) gene, partial co
## >EU574325.1 Bradyrhizobium liaoningense strain ViHaR3 recombination protein A (recA) gene, partial co
## >EU574324.1 Bradyrhizobium liaoningense strain ViHaR2 recombination protein A (recA) gene, partial co
## >EU574323.1 Bradyrhizobium liaoningense strain ViHaR1 recombination protein A (recA) gene, partial co
## >EU574322.1 Bradyrhizobium liaoningense strain ViHaG8 recombination protein A (recA) gene, partial co
## >EU574321.1 Bradyrhizobium liaoningense strain ViHaG7 recombination protein A (recA) gene, partial co
## >EU574320.1 Bradyrhizobium liaoningense strain ViHaG6 recombination protein A (recA) gene, partial co
## >EU574319.1 Bradyrhizobium yuanmingense strain ViHaG5 recombination protein A (recA) gene, partial co
## >EU574318.1 Bradyrhizobium yuanmingense strain ViHaG4 recombination protein A (recA) gene, partial co
```

Generación de estadísticas básicas a partir de la información disponible en las cabeceras

¿Cuántas secuencias hay en el archivo `recA_Bradyrhizobium_vinuesa.fna`?

```
grep -c '^>' recA_Bradyrhizobium_vinuesa.fna
```

```
## 125
```

Veamos las 5 primeras líneas de cabeceras fasta usando **grep** y **head**

```
grep '^>' recA_Bradyrhizobium_vinuesa.fna | head -5
```

```
## >EU574327.1 Bradyrhizobium liaoningense strain ViHaR5 recombination protein A (recA) gene, partial co
## >EU574326.1 Bradyrhizobium liaoningense strain ViHaR4 recombination protein A (recA) gene, partial co
## >EU574325.1 Bradyrhizobium liaoningense strain ViHaR3 recombination protein A (recA) gene, partial co
## >EU574324.1 Bradyrhizobium liaoningense strain ViHaR2 recombination protein A (recA) gene, partial co
## >EU574323.1 Bradyrhizobium liaoningense strain ViHaR1 recombination protein A (recA) gene, partial co
```

Generación de conteos usando el idioma **grep** ‘REGEX’ **archivo** | **sort** | **uniq -c**

- Cuenta el número de taxones que contiene el archivo FASTA

```
grep '^>' recA_Bradyrhizobium_vinuesa.fna | cut -d' ' -f2,3 | sort | uniq -c
```

```
##      18 Bradyrhizobium canariense
##      18 Bradyrhizobium elkanii
##       6 Bradyrhizobium genosp.
##      28 Bradyrhizobium japonicum
##      15 Bradyrhizobium liaoningense
##       8 Bradyrhizobium sp.
##      32 Bradyrhizobium yuanmingense
```

Imprime una lista ordenada de mayor a menor, del número de especies que contiene el archivo FASTA usando **sort -nrkN**

```
grep '^>' recA_Bradyrhizobium_vinuesa.fna | cut -d' ' -f2,3 | sort | uniq -c | sort -nrk1
```

```
##      32 Bradyrhizobium yuanmingense
##      28 Bradyrhizobium japonicum
##      18 Bradyrhizobium elkanii
##      18 Bradyrhizobium canariense
##      15 Bradyrhizobium liaoningense
##       8 Bradyrhizobium sp.
##       6 Bradyrhizobium genosp.
```

Edición de las cabeceras FASTA mediante herramientas de filtrado de UNIX

Exploremos todas las cabeceras FASTA del archivo `recA_Bradyrhizobium_vinuesa.fna` usando `grep`

```
# grep '^>' recA_Bradyrhizobium_vinuesa.fna | less # para verlas por página
grep '^>' recA_Bradyrhizobium_vinuesa.fna | head # para no hacer muy extensa la salida
```

```
## >EU574327.1 Bradyrhizobium liaoningense strain ViHaR5 recombination protein A (recA) gene, partial cds
## >EU574326.1 Bradyrhizobium liaoningense strain ViHaR4 recombination protein A (recA) gene, partial cds
## >EU574325.1 Bradyrhizobium liaoningense strain ViHaR3 recombination protein A (recA) gene, partial cds
## >EU574324.1 Bradyrhizobium liaoningense strain ViHaR2 recombination protein A (recA) gene, partial cds
## >EU574323.1 Bradyrhizobium liaoningense strain ViHaR1 recombination protein A (recA) gene, partial cds
## >EU574322.1 Bradyrhizobium liaoningense strain ViHaG8 recombination protein A (recA) gene, partial cds
## >EU574321.1 Bradyrhizobium liaoningense strain ViHaG7 recombination protein A (recA) gene, partial cds
## >EU574320.1 Bradyrhizobium liaoningense strain ViHaG6 recombination protein A (recA) gene, partial cds
## >EU574319.1 Bradyrhizobium yuanmingense strain ViHaG5 recombination protein A (recA) gene, partial cds
## >EU574318.1 Bradyrhizobium yuanmingense strain ViHaG4 recombination protein A (recA) gene, partial cds
```

Simplificación de las cabeceras FASTA usando `sed` (el *stream editor*)

El objetivo es eliminar redundancia y los campos `gb|no.de.acceso`, así como todos los caracteres `(, ; :)` que impedirían el despliegue de un árbol filogenético, al tratarse de caracteres reservados del formato NEWICK. Dejar solo el número GI, así como el género, especie y cepa indicados entre corchetes.

Es decir vamos a: - reducir *Bradyrhizobium* a `'B.'` - eliminar `'RNA poly ...'` y reemplazarlo por `']'` - eliminar `'genosp.'` - sustituir espacios por guiones bajos

Noten el uso de **expresiones regulares** como `'*y'[:space:]'`

```
sed 's/|gb.*| /|/; s/Bradyrhizobium /B./; s/genosp\./ /; s/ RNA.*\]/]; s/[[:space:]]/_/g;' recA_Bradyrhizobium_vinuesa.fna
```

```
## >EU574327.1_B.liaoningense_strain_ViHaR5_recombination_protein_A_(recA)_gene,_partial_cds
## >EU574326.1_B.liaoningense_strain_ViHaR4_recombination_protein_A_(recA)_gene,_partial_cds
## >EU574325.1_B.liaoningense_strain_ViHaR3_recombination_protein_A_(recA)_gene,_partial_cds
## >EU574324.1_B.liaoningense_strain_ViHaR2_recombination_protein_A_(recA)_gene,_partial_cds
## >EU574323.1_B.liaoningense_strain_ViHaR1_recombination_protein_A_(recA)_gene,_partial_cds
```

Cuando estamos satisfechos con el resultado, guardamos la salida del comando en un archivo usando `'^>'` para redirigir el flujo de STDOUT a un archivo de texto

```
sed 's/ recom.*cds//; s/|gb.*| /|/; s/Bradyrhizobium /B /; s/genosp\./ /; s/ RNA.*\]/]; s/[[:space:]]/_/g;' recA_Bradyrhizobium_vinuesa.fna >recA_Bradyrhizobium_vinuesa.fna_simplified
```

Generación automática de archivos FASTA especie-específicos (avanzado)

Conversión de archivos FASTA a formato “FASTAB” usando perl 1-liners.

Vamos a transformar los FASTAS de tal manera que las secuencias queden en la misma línea que su cabecera, separada de ésta por un tabulador. Esto puede ser muy útil para filtrar el archivo resultante con grep. Veamos un ejemplo:

```
perl -pe 'unless(/^>){s/\n//g}; if(/^>){s/\n\t/g}; s/>/\n>/' recA_Bradyrhizobium_vinuesa.fnaed | head -5

##
## >EU574327.1_B_liaoningense_ViHaR5      ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGG
## >EU574326.1_B_liaoningense_ViHaR4      ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGG
## >EU574325.1_B_liaoningense_ViHaR3      ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGG
## >EU574324.1_B_liaoningense_ViHaR2      ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGG
## >EU574323.1_B_liaoningense_ViHaR1      ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGG
## >EU574322.1_B_liaoningense_ViHaG8      ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGG
## >EU574321.1_B_liaoningense_ViHaG7      ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGG
## >EU574320.1_B_liaoningense_ViHaG6      ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGG
## >EU574319.1_B_yuanmingense_ViHaG5      ATGAAGCTCGGCAAGAACGACCGCTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGG
perl -pe 'unless(/^>){s/\n//g}; if(/^>){s/\n\t/g}; s/>/\n>/' recA_Bradyrhizobium_vinuesa.fnaed > recA_Byuanmingense.fnaed
```

Filtrado del archivo en formato fastab con grep

Filtremos el archivo *fnaedtab (en formato “fastab”) generado en el punto anterior para obtener solo las secuencias de B._yuanmingense del mismo, guardarlo en un archivo y convertirlo de nuevo a formato FASTA.

```
grep yuanmingense recA_Bradyrhizobium_vinuesa.fnaedtab | head -5

## >EU574319.1_B_yuanmingense_ViHaG5      ATGAAGCTCGGCAAGAACGACCGCTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGG
## >EU574318.1_B_yuanmingense_ViHaG4      ATGAAGCTCGGCAAGAACGACCGCTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGG
## >EU574297.1_B_yuanmingense_InRo02      ATGAAGCTCGGCAAGAACGACCGCTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGG
## >EU574296.1_B_yuanmingense_InKo02      ATGAAGCTCGGCAAGAACGATCGCTCCATGGACATCGAGGCGGTCTCCTCCGGCTCGCTCGGG
## >EU574295.1_B_yuanmingense_InKo01      ATGAAGCTCGGCAAGAACGATCGCTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGG
grep yuanmingense recA_Bradyrhizobium_vinuesa.fnaedtab > recA_Byuanmingense.fnaedtab
```

Regeneración del formato FASTA a partir de FASTAB con un Perl 1-liner

- Estas dos líneas no contienen nada nuevo en cuanto a sintaxis. Simplemente llamamos a perl para sustituir los tabuladores por saltos de línea y así reconstituir el FASTA.

```
perl -pe 'if(/^>){s/\t/\n/}' recA_Byuanmingense.fnaedtab | head -5

## >EU574319.1_B_yuanmingense_ViHaG5
## ATGAAGCTCGGCAAGAACGACCGCTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGGCTCGATATCGCGCTCGGCATCGGCGGCTTGCCCAAGG
## >EU574318.1_B_yuanmingense_ViHaG4
## ATGAAGCTCGGCAAGAACGACCGCTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGGCTCGATATCGCGCTCGGCATCGGCGGCTTGCCCAAGG
## >EU574297.1_B_yuanmingense_InRo02
perl -pe 'if(/^>){s/\t/\n/}' recA_Byuanmingense.fnaedtab > recA_Byuanmingense.fna
```

Llamada a un bucle for para repetir operaciones sobre múltiples archivos

- Llamar a un bucle for de shell para generar archivos fastab para todas las especies

```
for sp in $(grep '^>' recA_Bradyrhizobium_vinuesa.fnaed | cut -d_ -f3); do
  grep "$sp" recA_Bradyrhizobium_vinuesa.fnaedtab > recA_B${sp}.fnaedtab
done
```

- Veamos el resultado

```
ls *fnaedtab
```

```
## recA_Balpha.fnaedtab
## recA_Bbeta.fnaedtab
## recA_Bcanariense.fnaedtab
## recA_Belkanii.fnaedtab
## recA_Bjaponicum.fnaedtab
## recA_Bliaoningense.fnaedtab
## recA_Bradyrhizobium_vinuesa.fnaedtab
## recA_Bsp..fnaedtab
## recA_Byuanmingense.fnaedtab
```

```
head -5 recA_Bjaponicum.fnaedtab
```

```
## >EU574316.1_B_japonicum_NeRa16 ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCGGGTTCTCTCGGGCTCG
## >EU574315.1_B_japonicum_NeRa15 ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCCGGTTCTCTCGGGCTCG
## >EU574314.1_B_japonicum_NeRa14 ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCCGGTTCTCTCGGGCTCG
## >EU574313.1_B_japonicum_NeRa12 ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCCGGTTCTCTCGGGCTCG
## >EU574312.1_B_japonicum_NeRa11 ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCCGGTTCTCTCGGGCTCG
```

- Finalmente convertimos todos los archivos fnatabed a FASTA con el siguiente bucle for:

```
for file in *fnaedtab; do perl -pe 'if(</>){s/\t/\n/}' $file > ${file%.*}.fna; done
```

- Visualizemos las cabeceras de dos archivos FASTA especie-específicos

```
grep '^>' recA_Bjaponicum.fna | head -5
```

```
## >EU574316.1_B_japonicum_NeRa16
## >EU574315.1_B_japonicum_NeRa15
## >EU574314.1_B_japonicum_NeRa14
## >EU574313.1_B_japonicum_NeRa12
## >EU574312.1_B_japonicum_NeRa11
```

- Confirmemos que son fastas regulares

```
head -6 recA_Bjaponicum.fna
```

```
## >EU574316.1_B_japonicum_NeRa16
## ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCGGGTTCTCTCGGGCTCGACATTGCACTGGGGATCGGCGGTCTGCCAAGG
## >EU574315.1_B_japonicum_NeRa15
## ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCCGGTTCTCTCGGGCTCGACATTGCACTGGGGATCGGCGGTCTGCCAAGG
## >EU574314.1_B_japonicum_NeRa14
## ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCCGGTTCTCTCGGGCTCGACATTGCGCTGGGGATCGGCGGTCTGCCAAGG
```