Ejercicio de parseo de archivos FASTA

Pablo Vinuesa

2018-07-01

Contents

Presentación

Este código corresponde a unas prácticas escritas por Pablo Vinuesa para el manual de Bioinformática y Sistemática Molecular de la Facultad de Ciencias - UNAM, Abril 2015.

Version: 2018-07-01

Adaptada para el Taller de Filoinformática - UNLP, 2-6 Julio 2018.

Preparativos

1. genera el directorio practica2_parseo_fastas

```
### generemos un subdirectorio por debajo del que acabamos de crear y movámonos a él
mkdir -p $HOME/intro2filoinfo/lunes/practica2_parseo_fastas/data && cd $HOME/intro2filoinfo/lunes/practi
```

2. Descarga el archivo recA_Bradyrhizobium_vinuesa.fa en el directorio que acabamos de generar.

Búsqueda y descarga de secuencias en GenBank usando el sistema ENTREZ

El archivo recA_Bradyrhizobium_vinuesa.fa contiene secuencias parciales (amplicones de PCR) del gen *recA* de bacterias del género *Bradyrhizobium* disponibles en GenBank. Este bloque muestra el comando usado para descargarlas. El comando debe pegarse en la ventana superior del sistema ENTREZ.

```
# pega la siguiente sentencia (sin las comillas) en la ventana de captura para interrogar la base de da
# de NCBI mediante el sistema ENTREZ
'Bradyrhizobium[orgn] AND vinuesa[auth] AND recA[gene]'

# Una vez cargada la página, da click en el link 'send to', arriba a la derecha, y guarda en formato FA

# renombra el archivo sequences.fasta a recA_Bradyrhizobium_vinuesa.fa
```

Práctica de parseo de archivos FASTA descargados de NCBI mediante ENTREZ

Inspección y estadísticas básicas de las secuencias descargadas

1. ¿Cuántas secuencias hay en el archivo recA_Bradyrhizobium_vinuesa.fa?

```
grep -c '>' recA_Bradyrhizobium_vinuesa.fa
```

```
## 125
```

2. Veamos las 5 primeras lineas de cabeceras fasta usando **grep** y **head**

```
 grep '>' recA_Bradyrhizobium_vinuesa.fa | 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
```

3. Cuenta el numero de generos y especies que contiene el archivo FASTA

```
grep '>' recA_Bradyrhizobium_vinuesa.fa | cut -d' ' -f3 | sort | uniq -c
```

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

4. Imprime una lista ordenada de mayor a menor, del numero de especies que contiene el archivo FASTA

```
grep '>' recA_Bradyrhizobium_vinuesa.fa | 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

5. Exploremos todas las cabeceras FASTA del archivo recA_Bradyrhizobium_vinuesa.fa usando **grep**

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

```
## >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
```

6. simplifiquemos las cabeceras FASTA usando el comando **sed** (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 numero 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_Bradyrhiz
```

```
## >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
```

8. 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:]]/_
```

Generación automática de archivos FASTA especie-específicos (avanzado)

9. Convertir 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.faed | head
```

```
##
## >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.faed > recA_
```

10. Filtrar el archivo fnaedtab generado en 9 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.faedtab | 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.faedtab > recA_Byuanmingense.fnaedtab
```

11. Estas dos lineas no contienen nada nuevo en cuanto a sintaxis. Simplemente llamamos a perl para sustituir los tabuladores por saltos de linea y asi 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
```

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

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

13. Veamos el resultado

```
ls *fnaedtab
```

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

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

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

14. 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
```

15. 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
```

16. y confirmemos que son fastas regulares

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

```
## >EU574316.1_B_japonicum_NeRa16
## ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCGGGTTCTCTCGGGCTCGACATTGCACTGGGGATCGGCGGTCTGCCCAAGG
## >EU574315.1_B_japonicum_NeRa15
## ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCCGGTTCTCTCGGGCTCGACATTGCACTGGGGATCGGCGGTCTGCCCAAGG
## >EU574314.1_B_japonicum_NeRa14
## ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCCGGTTCTCTCGGGCTCGACATTGCGCTGGGGATCGGCGGTCTGCCCAAGG
```