Nodo Nacional de Bioinformática

Universidad Nacional Autónoma de México – Nodo Mexicano EMBNet

Taller 1. Introducción al biocómputo en Sistemas Linux y su aplicación en filoinformática

Semana 1. Descubriendo el poder del intérprete de comandos (shell)

Profesores:
Romualdo Zayas
Heladia Salgado
George Magklaras

CCG
Centro de Ciencias Genómicas

DIA 3. COMANDOS PARA MANIPULAR EL CONTENIDO DE UN ARCHIVO.

Temario DIA 3

- Comandos para manejar el contenido de un archivo
  - Contando líneas
  - Cortando líneas
  - Ordenando líneas
  - Manejo de líneas repetidos
  - Búsqueda de patrones
- Lenguajes
  - Awk
  - Perl command line
- Práctica 3. Manipulando la información de un archivo

Objetivos

Después de completar esta lección, el alumno será capaz de manipular el contenido de un archivo.

En esta lección usaremos los archivos que se encuentran bajo el directorio de *E_coli*.

- **E_coli_K12_genes.txt** Contiene la lista de genes de *Escherichia coli*, ordenados por posición en el genoma.

- **tf-gene-interactions.txt** Es la red de interacciones de factores de transcripción y sus genes regulados.

- **b????.txt** son las secuencias en formato fasta de ciertos genes de E. coli, el nombre del archivo es el identificador que le han asignado.

- **Gene_sequence.txt**. Es el archivo de secuencias de los genes del genoma de E. coli

- Crear un directorio *data* en tu home

- Copiar todos los directorios y archivos del path *heladia/data/* a tu directorio *data*

Contando datos

Contar líneas, palabras y caracteres usando *wc*

```
$ man wc

$ wc E_coli_K12_genes.txt

$ wc -l E_coli_K12_genes.txt

$ wc -w E_coli_K12_genes.txt

$ wc -m E_coli_K12_genes.txt
```

Cortando columnas

Extraer columnas de datos de un archivo usando *cut*

```
$ man cut

$ cut –f2 E_coli_K12_genes.txt            # campos

$ cut -f2,3,4,5,6 E_coli_K12_genes.txt
$ cut –f4,5,2,3,6 E_coli_K12_genes.txt    # ¿Qué pasa?

$ cut –f2-6 E_coli_K12_genes.txt

$ cut –c1-10 E_coli_K12_genes.txt         # caracteres
```

Ordenando líneas

Ordenar líneas de datos de un archivo usando *sort*

```
$ man sort

$ sort E_coli_K12_genes.txt

$ sort -k2 E_coli_K12_genes.txt       # por nombre del gene
$ sort –k4 E_coli_K12_genes.txt

$ sort –k4 –n E_coli_K12_genes.txt
$ sort –k4 –nr E_coli_K12_genes.txt

$ sort –k4 –n E_coli_K12_genes.txt –o E_coli_K12_genes_sort.txt
$ wc –l E_coli_K12_genes_sort.txt

$ sort –u E_coli_K12_genes.txt –o E_coli_K12_genes_uniq.txt
$ wc –l E_coli_K12_genes_uniq.txt
```

Repeticiones

Reporta u omite líneas repetidas usando *uniq*

```
$ man uniq

$ uniq -c E_coli_K12_genes_sort.txt

$ uniq -d E_coli_K12_genes_sort.txt  # solo las duplicadas

$ uniq -u E_coli_K12_genes_sort.txt  # solo las únicas
```

Búsqueda de patrones en archivos

Buscar líneas que contienen un patrón o cadena usando *grep*

```
$ man grep

$ grep araC E_coli_K12_genes.txt
$ grep ara E_coli_K12_genes.txt

$ grep crp E_coli_K12_genes.txt
$ grep –w crp E_coli_K12_genes.txt

$ grep ‘CTGTATGA’  pc_bn_site_prom__w_ids.txt

$ grep -E 'TACTGTA........CAGT' pc_bn_site_prom__w_ids.txt

$ grep -f lista_genes.txt E_coli_K12_genes.txt

$ grep -v '#' E_coli_K12_genes.txt

$ grep –v -f lista_genes.txt E_coli_K12_genes.txt  # líneas que no
contienen el patrón
```

Escenario 1.

El archivo **E_coli_K12_genes.txt** contiene información sobre los genes de E. coli.

1.- Queremos obtener el total de los genes.

Tips:

- Las líneas de los genes empiezan con ECK12
- Puede haber líneas repetidas
- Las líneas que inician con # son comentarios

2.- Queremos saber el total de genes que van en orientación 5’-3´(forward) y 3´-5´(reverse).

3.- Cómo obtenemos aquellos genes a los que no les anotaron la orientación (forward, reverse).

4.- Con lo que hasta ahora hemos visto, ¿Podríamos obtener el tamaño de los genes?

```
# 1.1

$ grep ECK12 E_coli_K12_genes.txt |sort –u | wc

# 1.2
$ grep ECK12 E_coli_K12_genes.txt | grep forward |wc

$ grep ECK12 E_coli_K12_genes.txt | grep reverse |wc

# 1.3
$ grep ECK12 E_coli_K12_genes.txt | grep -v -E "forward|reverse”

# 1.4
$
```

Escenario 2

El archivo tf-gene-interactions.txt contiene la lista de interacciones entre las proteínas y los genes a los que regulan, donde

- \+ indica que lo activa,
- el – que lo reprime,
- +- que lo activa y reprime en distintas condiciones y
- ? que se desconoce como la proteína afecta al gene

1.- Obtener el total de proteínas únicas que son las que regulan a los genes.

2.- Obtener el total de genes únicos regulados por las proteínas.

3.- Cuantas interacciones +,-,+- y ? Existen ?

4.- Obtener el total de genes regulados por cada proteína.

5.- Obtener el total de proteínas que regulan al gene.

6.- Obtener la distribución de genes regulados por 1,2,3,etc proteínas.

```
# 2.1

$ grep -v "#" tf-gene-interactions.txt  | cut -f1 | sort -u |wc

# 2.2
$ grep -v "#" tf-gene-interactions.txt  | cut -f2 | sort -u |wc

# 2.3
$ grep -v "#" tf-gene-interactions.txt  |cut -f1-3 |sort -u |cut -f3 |
sort | uniq -c | sort -nr

# 2.4
$ grep -v "#" tf-gene-interactions.txt  |cut -f1,2 |sort -u |cut -f1 |
uniq -c | sort -nr |more

# 2.5
$ grep -v "#" tf-gene-interactions.txt  |cut -f1,2 |sort | uniq | cut -f2
| sort |uniq -c | sort -nr |more

# 2.6
$grep -v "#" tf-gene-interactions.txt  |cut -f1,2 |sort | uniq | cut -f2 |
sort |uniq -c | sort -nr |cut -c1-4 |uniq -c |sort -nr
```

Escenario 3

En RegulonDB se han curado/revisado las referencias(papers) asociados al proceso de regulación de la bacteria E. coli disponibles en la base de datos pubmed.

Usando el identificador de las referencias en pubmed, se extrajo información importante tal como el año de la publicación, los autores, el abstract de la referencia, etc… en formato medline. Esta información esta en el archivo PMIDs_RegulonDB_8.5_medline.txt

1.- Queremos obtener la distribución de papers revisados asociados al proceso, por año (se puede usar el campo DP)

2.- Queremos obtener la distribución de papers por journal. Esto nos ayudará a determinar cuales son las revistas de donde más hay información del proceso.

```
# 3.1

$ grep "DP  -" PMIDs_RegulonDB_8.5_medline.txt |cut -c7-10 | sort | uniq –c


# 3.2
$ grep "JT  -" PMIDs_RegulonDB_8.5_medline.txt | sort | uniq -c |sort -nr |
more
```