Práctico 2

Dr Andrés Iriarte

Correr programas desde la línea de comando

PASOS PARA CORRER UN PROGRAMA

1) Establecer donde se encuentra el programa: Los programas, compilados o scripts pueden estar ubicados en:

1. El bin del equipo, /bin/, como ocurre en los casos en los que el programa es instalado por un administrador desde un repositorio utilizando el comando apt-get. También puede estar ubicado en el /bin/ si un usuario administrador, luego de compilarlo o bajarlo lo movió/copio al bin. El bin del equipo está por definición en el path, es decir es accesible desde todos los puntos del equipo por todos los usuarios. Los comandos ls, mkdir, etc, están allí.
2. El bin del usuario (/usr/local/bin/) o el bin sin programas esenciales (/usr/bin). Cada usuario puede instalar programas aquí e instalarlos en este directorio.
3. Cualquier directorio que esté incluido en el path, es decir un programa que está accesible desde cualquier punto del equipo por parte del usuario.
4. Un directorio específico que incluye el ejecutable y que además incluye las librerías y otros scripts asociados necesarios.

2) Identificar cuales son los argumentos y/o opciones mínimos necesarios para correr el programa. En la mayoría de los casos es necesario estudiar cuales son los formatos necesarios en los archivos de entrada y las opciones para obtener los resultados que se buscan.

3) Características del programa y del análisis. Cuanto tiempo correrá el programa? Es necesario usar el comando nohup? Como es el archivo de salida? Hay que dirigir la salida estándar a un archivo o el programa genera un archivo de salida?

4) Correr el programa. ### EJEMPLOS: ### 1. muscle MUSCLE (MUltiple Sequence Comparison by Log- Expectation) es un programa para alinear secuencias aminoacídicas o nucleotídicas. MUSCLE es un programa muy utilizado, combinando de manera muy buena la velocidad y la exactitud de los resultados. Su funcionamiento es muy simple y forma parte de los repositorios de Linux. Manual: www.drive5.com/muscle/manual/

a) Antes de comenzar responda:

b) ¿Donde se ubica el ejecutable?

ii) ¿Cuales son las opciones del programa? y ¿Cuales son sus requerimientos mínimos?

iii) ¿Qué características tienen los archivos de entrada?

b) Como en cada práctico, dentro de su carpeta personal (ubicada dentro de la carpeta general del usuario estudiantes), genere una carpeta correspondiente a este práctico, práctico 2.

```
{bash, echo = T, eval = F} cd & mkdir
```

c) Copie el archivo “para_agregar.fas” y “111816_abi.624.aa” disponibles en la carpeta general MATERIALES/PRACT2 en la carpeta creada en el punto anterior. Luego sume (o concatene) ambos archivos utilizando el comando cat y envíe el resultado a un nuevo archivo.

d) Responda: qué tipo de archivos son ambos?

ii) Utilizando los comandos cp & cat:

```
{bash, echo = T, eval = F} cat para_agregar.fas 111816_abi.624.aa > nuevo_archivo.fas
```

d) Alinea el nuevo archivo utilizando el programa muscle. Para saber sobre el modo de uso utilice la opción –h. “‘{bash, echo = T, eval = F} muscle -in -out

```
> **Nota**:

e) Explore el resultado.

```{bash, echo = T, eval = F}
less output_file
```

f) Vuelva a alinear las secuencias variando uno o varios parámetros (guarde con un nombre distinto la salida).

```
{bash, echo = T, eval = F} muscle -in <input_file> -out <output_file2>
```

g) Compare los resultados del programa y la corrida.

```
{bash, echo = T, eval = F} less output_file2
```

2. fasttree

FastTree es un programa para inferir árboles filogenéticos por máxima verosimilitud en base a secuencias aminoacídicas o nucleotídicas alineadas. FastTree fue especialmente desarrollado para manejar un gran número de secuencias, en el orden de miles o decenas de miles. Su funcionamiento es muy simple y su guía esta en: www.microbesonline.org/fasttree/. Al igual que muscle el programa fasttree está instalado utilizando apt-get por un administrador y por lo tanto está disponible en todo el equipo para todos los usuarios incluyendo estudiantes.

a) Genere un árbol filogenético utilizando el programa fasttree en base a alguno de los alineamientos generados en la actividad anterior.

b) Puede visualizar la salida del programa, árbol filogenético en formato newick, en alguna herramienta online como etetoolkit.org/treeview/, iubio.bio.indiana.edu/treeapp/treeprint-form.html o www.trex.uqam.ca/index.php?action=newick

3. abyss-fac.pl

abyss-fac.pl es un script escrito en perl que es parte del paquete ABySS para ensamblado y análisis de genomas. Este script permite obtener estadísticos sobre la continuidad del ensamblaje (medida de calidad del mismo) y que corre en la carpeta donde se ubica (aunque también puede estar ubicado en un directorio que está en el path o en el bin, haciéndose accesible a todos los usuarios desde cualquier lugar del equipo). El paquete completo también puede instalarse utilizando apt-get. Ver guía en computing.bio.cam.ac.uk/local/doc/abyss.html. Uso: abyss-fac.pl assembly.fas

a) Primero bajaremos un borrador de genoma (draft genome) del Genbank. Desde la web https://www.ncbi.nlm.nih.gov busque “Salmonella sp. HMSC13B08”.

b) Seleccione la sección “Assembly”.
```

c) Cliquee en “Download the GenBank assembly”. Eso lo llevará a una página web en protocolo ftp. Ftp es un protocolo al igual que http y está optimizado para la transferencia de archivos.

d) Cliquee con Botón Derecho en “GCA_001808015.1_ASM180801v1_genomic.fna.gz” y copie la dirección del enlace.

e) Vuelva a la terminal y utilizando el comando wget con la dirección del enlace opiada en el punto anterior baje el archivo del genoma.

```
{bash, echo = T, eval = F} wget ftp://ftp.ncbi.nlm.nih.gov/.../.../.../GCA_001808015.1_ASM180801v1_genor
```

f) Utilizando el programa gunzip descomprima el archivo. Y explore el archivo.

```
{bash, echo = T, eval = F} gunzip GCA_001808015.1_ASM180801v1_genomic.fna.gz less GCA_001808015.1_ASM18
```

g) Utilice el script abyss-fac.pl para sacar los estadísticos de continuidad de este genoma.

h) ¿Porque piensa usted que a diferencia de los casos anteriores debe utilizarse un “./” antes del nombre del script.

```
{bash, echo = T, eval = F} ./abyss-fac.pl GCA_001808015.1_ASM180801v1_genomic.fna
```

Guía para interpretar el resultado: (1) n N total de contigs en el ensamblaje (2) n:100 N contigs mayores al límite (o threshold, se puede establecer con la opción “-t”) Nota: todos los estadísticos luego de este valor se calculan solamente para los contigs mayores al límite establecido en el punto 2. (3) n:N50 N de contigs contenido en el set N50 (4) min Tamaño del contig más pequeño (5) median Mediana del tamaño de los contigs (6) mean Promedio del tamaño de los contigs (7) N50 Es definido como la longitud de los contigs tal que usando contigs de igual o mayor tamaño produce la mitad de las bases del genoma. El tamaño N50 se calcula ordenando todos los contigs de mayor a menor, y determinando el conjunto mínimo de contigs cuyo tamaño total sea el 50% de todo el genoma. (8) max Contig mayor (9) sum Suma total de los contigs en el ensamblado

Las únicas opciones de la línea de comando son –t para establecer el límite y –j para obtener el resultado en formato JIRA.

iii) Vuelva a calcular los estadísticos de continuidad variando el largo mínimo de los contigs a considerar. Compare los resultados.

PARA HACER EN EL DOMICILIO

4. Analice los programas SPAdes y BWA.

a) ¿Para qué sirve cada uno de ellos?
b) ¿Qué archivos son necesarios para utilizarlos?
c) ¿Cómo correría estos programas?