1 Taller 01: Uso de comandos de shell para filtrar archivos

Para realizar el siguiente taller dos conceptos de la biología molecular deben tenerse en cuenta, el codigo genetico y el proceso de transcribir un mensaje escrito en el genoma a otro llamado RNA mensajero y de este a proteina . El primero está asociado con el desciframiento del código genético pero que es esto? En primer lugar una definición general aceptada es que un código *es una combinación de caracteres que se emplea para crear y entender mensajes secretos.* ¿Cúal es el mensaje secreto en la biología?. Un tema que ha tratado de descifrar Maximo Di Giulio [1], [2]

1.1 Primera idea: El código genético

El código genético en biología, hace parte de los muchos mensajes escritos en el genoma (de modo basico en los genes) y que podriamos entender por analogia como el resultado de permutaciones en matematicas, es decir se construye a partir de un término utilizado en matemáticas y probabilidad que se llama *permutación con repetición.* En el caso de la *permutación con repetición* si se tienen n objetos para elegir (en nuestro caso 4 nucleótidos (nuestro alfabeto estrella)) y r maneras de elegir (en nuestro caso cada una de las 3 posiciones en el codón), entonces la *permutación con repetición* puede ser expresada asi: $n * n * n * \ldots (r \quad veces) = n^r$. Es decir para nuestro caso, el código genético, se puede entender como el resultado de la siguiente permutación con repetición expresada como 4^3. Entonces se tienen 4 posibles caracteres para ser asignados a la primera posición del codon, 4 para la segunda y 4 para la tercera posición. Para una revisión actual sobre el tema consultar en [2]. Ahora esas posibles tripletas obtenidas, en la biologia corresponderan segun unas reglas a uno o varios aminoacidos o unidades constitutivas de las proteinas. Observe la Figura 1 para entender las reglas.

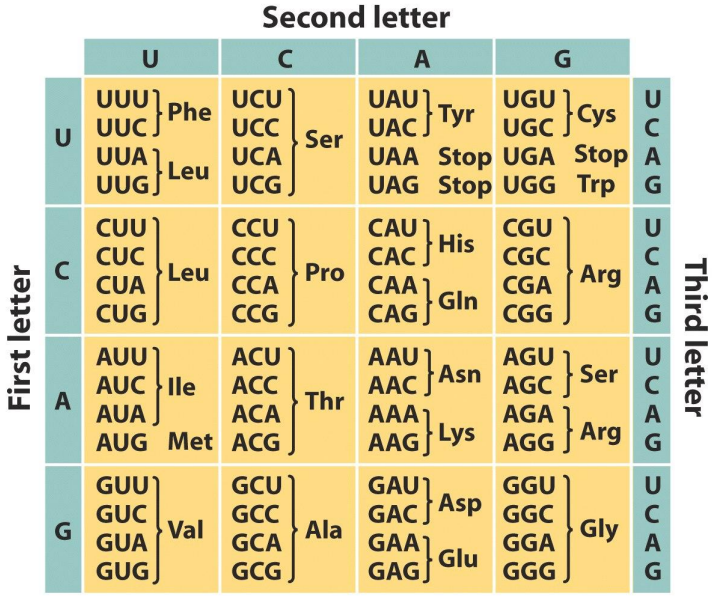

Figure 1: Tabla que respresenta la asignacion de las permutaciones con repeticion posibles del codigo genetico y su correspondiente desciframeinto a cada tipo de aminoacido o unidad constitutiva de las proteinas

1.2 Segunda idea: Molecularmente existen procesos que permitén que es a relación de tripletas (codones) sean correctamente asignadas a los aminoácidos

En biologia el desciframiento del codigo ocurre por medio de proceso moleculares conocidos como la transcripcion paso de la informacion del gen a un intermediario llamado transcrito o RNA mensajero (proceso conocido com transcripcion) y este nuevo mensaje es leido en los ribosomas para trascribir el mensaje ecsrito en el RNA mensajero a proteinas (proceso llamado traduccion). En la figura 2 observe un esquema asociado.

1.3 Introducción al lenguaje de la shell: shell programming o shell scripting

Para comunicarnos con el sistema operativo de linux y construir scripts y tuberias de procesos (pipelines) es necesario aprender el lenguaje de la shell. La shell es referida como *un intermediario entre el sistema operativo y el usuario gracias a líneas de comando que el usuario introduce. Su función es la de leer la línea de comandos, interpretar su significado y mediante instrucciones de la shell el usuario puede comunicarse con el nucleo del sistema operativo. Existen diferentes versiones como Bourne shell (sh), Almquist shell (ash), Bourne-Again*

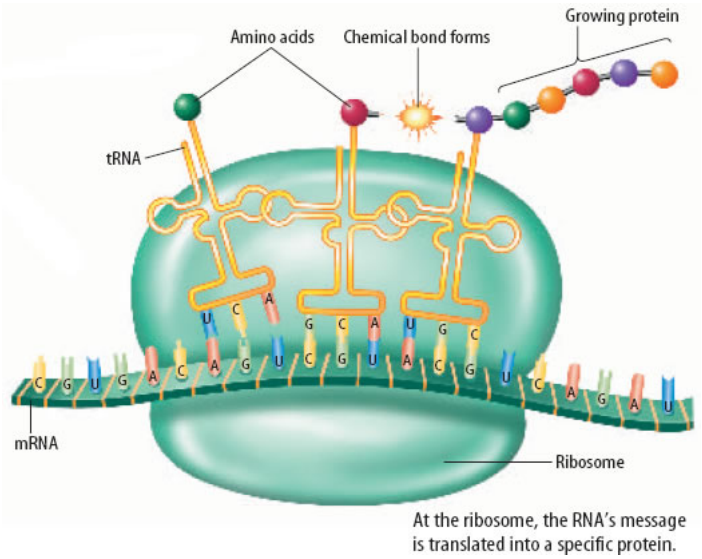

Figure 2: Proceso de traduccion

shell (bash) entre otras. La shell se inicia cuando se leen las configuraciones de inicio del sistema. Posteriormente aparece el siguiente indicador (prompt en inglés): equipo:/directorio/actual$ en donde "$" siginifica un usuario normal que no es root .

1.4 Sintaxis general

Para poder escribir lineas de comandos, es necesario familiarizarse con estos. Una línea de comandos es una cadena de caracteres formada por un comando que corresponde a un archivo ejecutable del sistema y tiene una sintaxis:

- Sintaxis: comando < *option* > file

1.5 Comandos wc

Imprime en pantalla: 1 columna número de lineas, 2 columna numero de palabras 3 número de caracteres

1.5.1 Ejemplo

wc ../Data/genetic_code.txt

1.6 tail

Imprime las últimas 10 lineas de un archivo. Cuando hay mas de un archivo, en la salida hay un encambezamiento dando el nombre del archivo.

1.6.1 Ejemplo

- tail -n 4 ../Data/genetic_code.txt
- tail -n 4 ../Data/genetic_code.txt ../Data/genetic_code.txt
- tail -n 4 -v ../Data/genetic_code.txt ../Data/genetic_code.txt

1.7 cat

Concatenación. La salida estandar "The standard output" se ve en pantalla, pero si usa el operador (>) se redirecciona a una nuevo archivo.

1.7.1 Ejemplo

- cat ../Data/genetic_code.txt
- cat ../Data/genetic_code.txt ../Data/genetic_code.txt

1.8 grep

Utilidad de la linea de comandos que busca un patrón e imprime la linea que concuerda.

1.8.1 Ejemplo

- grep "Serine" ../Data/genetic_code.txt
- grep "AA" ../Data/genetic_code.txt

1.8.2 Expresiones regulares

1.8.3 Ejemplo

Para buscar concordancias mas complejas. Por ejemplo , solo imprima las lineas que comienzan con la letra A, seguida por cualquier otra palabra y Lys.

- grep ^A.*Lys ../Data/genetic_code.txt ../Data/genetic_code.txt

Para buscar concordancias mas complejas asi como ^ representa el inicio de la linea $ representa el final de la linea

- grep ^T.*Ser.*S$ ../Data/genetic_code.txt
- grep -i ine ../Data/genetic_code.txt. Imprime todas las lineas que contienen el patrón sin importar la letra mayuscula o minuscula. -i como argumento indica "ignore case".
- grep -w Lysine ../Data/genetic_code.txt. Con el comando -w se buscan coincidencias exactas de la palabra Por ejemplo "Lysine".

- grep -v Lysine ../Data/genetic_code.txt. El comando -v imprime todas las lineas que no coinciden con el patrón.
- grep -E “pattern1|pattern2” ../Data/genetic_code.txt. En este caso | funciona como el operador OR pero usando -E es evaluado dentro de la expresión
- grep -E “pattern1.*pattern2” filename. Es una alternativa para usar AND. No se tiene un operador AND in grep. Con esta idea se imprimen las lineas que contienen pattern1 y pattern2
- grep -E “TCT.*Ser” genetic_code.txt
- grep -E “pattern1” filename | grep -E “pattern2”. Es una alternativa para simular el escenario de AND tambien, usando el pipe |
- grep -E “Ser” genetic_code.txt — grep -E “AGT”
- grep -v “pattern1” filename. Puede usarse para negar la coincidencia del patron. Funcionaría como un operador NOT.

1.9 echo

Comando para la impresión de un texto, actua como la función print de otros lenguajes.

1.9.1 Ejemplo

- echo “aaaaObbbbbbOcccccOdd”
- echo “aaaaObbbbbbOcccccOdd” | cut -dO -f2

1.10 cut

Remueve campos de cada linea.

1.10.1 Ejemplo

- cut -d -f3 ../Data/genetic_code.txt

2 Combinations

2.1 | pipe symbol

Operador que envia la salida de una linea de comando previa a una nueva linea de comandos

2.1.1 Ejemplo

- grep “Serine” ../Data/genetic_code.txt | cut -d -f3
- echo “aaaaObbbbbbOcccccOdd” | cut -dO -f2
- cat ../Data/genetic_code.txt ../Data/genetic_code.txt | wc

2.2 >

Operator para enviar la salida de cat a otro archivo.

2.2.1 Ejemplo

- cat ../Data/genetic_code.txt ../Data/genetic_code.txt > ../Results/2vecescodigo.txt

2.3 *

jocker, wild cart

3 Ejercicio

Utilizando los comandos anteriores: 1. Cree un nuevo archivo con información de interés para usted. Archívelo solo en el directorio Data/. Indique que combinaciones y operador usted ha utilizado 2. Utilize al menos tres combinaciones de comandos para generar nueva información? 3. Ideas de uso.

4 Guia rápida de compilación del código fuente LATEX

1. Escritura del texto científico en lenguaje LATEX (extensión .tex)
2. Escritura de la bibliografia en lenguaje LATEX (extensión .bib). *Puede consultar sus articulos para converir la cita a formato bibtex consultando en http://www.bioinformatics.org/texmed/*
3. Una vex revizada la sintaxix válida para el lenguaje se procede a compilar el código asi
   - pdflatex file.tex
   - bibtex file **Nota: es el mismo nombre que le puso a file.tex pero sin esa extensión**
   - bibtex file
   - pdflatex file.tex (Dos o tres veces hasta que actualice la bibliografia en el arte final)
4. abrir el archivo con un visualizador de formato pdf como acroread, evince, okular, xpdf etc. Es decir evince file.pdf

References

[1] M. Di Giulio. On the origin of the genetic code. *Trends Ecol. Evol. (Amst.)*, 7(6):176–178, Jun 1992.

[2] M. Di Giulio. An Autotrophic Origin for the Coded Amino Acids is Concordant with the Coevolution Theory of the Genetic Code. *J. Mol. Evol.*, 83(3-4):93–96, Oct 2016.