**AP2 Herramientas de alineamiento, calidad y pre-procesamiento de secuencias**

*Instrucciones: Contesta en el espacio a la pregunta y proporciona una captura de pantalla cuando se solicite. Ajusta el espacio para incluir la captura de pantalla y que sea legible.*

**AP2.1 Caracterización de una proteína mediante la herramienta blast (5 puntos)**

Queremos caracterizar una proteína problema que se encuentra en el archivo problema.fasta y vamos a llevarlo a cabo mediante Blast, usando dos bases de datos diferentes. Para ello vamos a utilizar una base de datos personalizada que podemos obtener mediante el archivo CustomDB.fasta y la base de datos del GenBank del NCBI.

Empezaremos por un análisis en local, realiza un blast en local:

**2.1.1** Proporciona los comandos utilizados para realizar el blast en local en texto junto a una captura de pantalla en la que parezca el terminal con: el prompt, el comando utilizado y el resultado obtenido.

Comando:

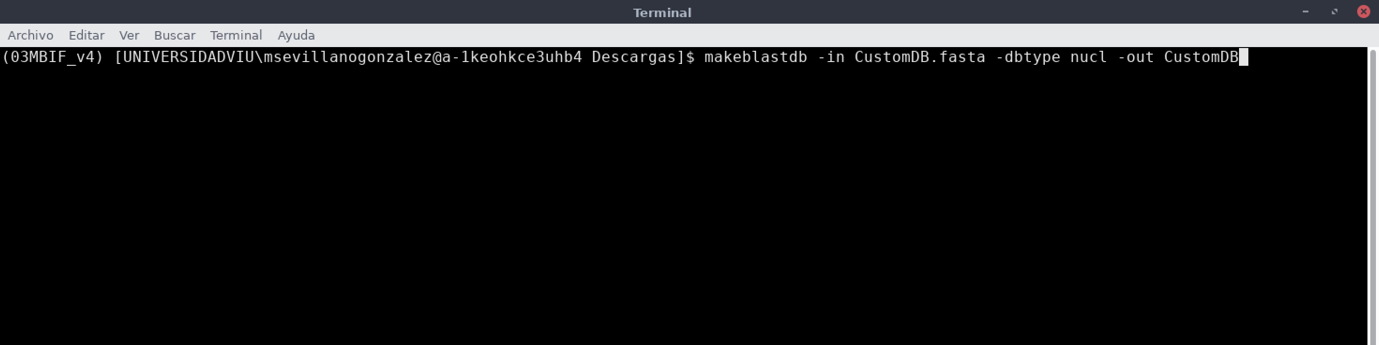
Primero se crea la base de datos en un archivo, con el comando: makeblastdb -in CustomDB.fasta -dbtype nucl -out CustomDB

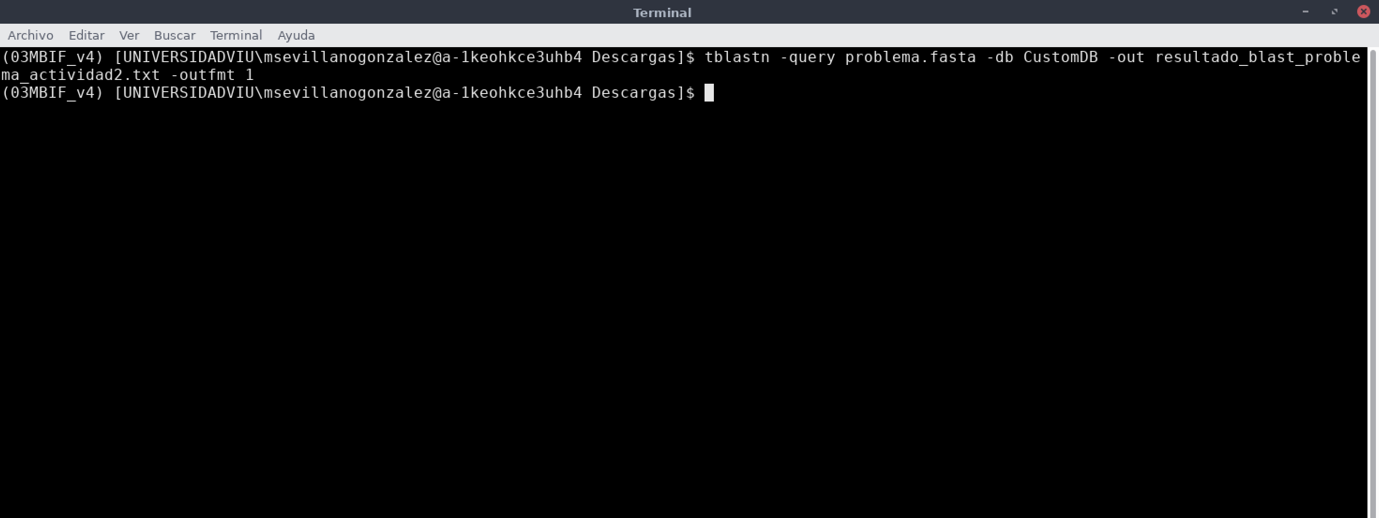
Después se realiza el blast, que en este caso al enfrentar una secuencia problema de aminoácidos de una proteína frente a una base de datos de nucleótidos la herramienta a utilizar es tblastn. Comando completo:

tblastn -query problema.fasta -db CustomDB -out resultado\_blast\_problema\_actividad2.txt -outfm 1

Captura:

En la siguiente captura no aparece el prompt porque me acordé después y ya había hecho clear.



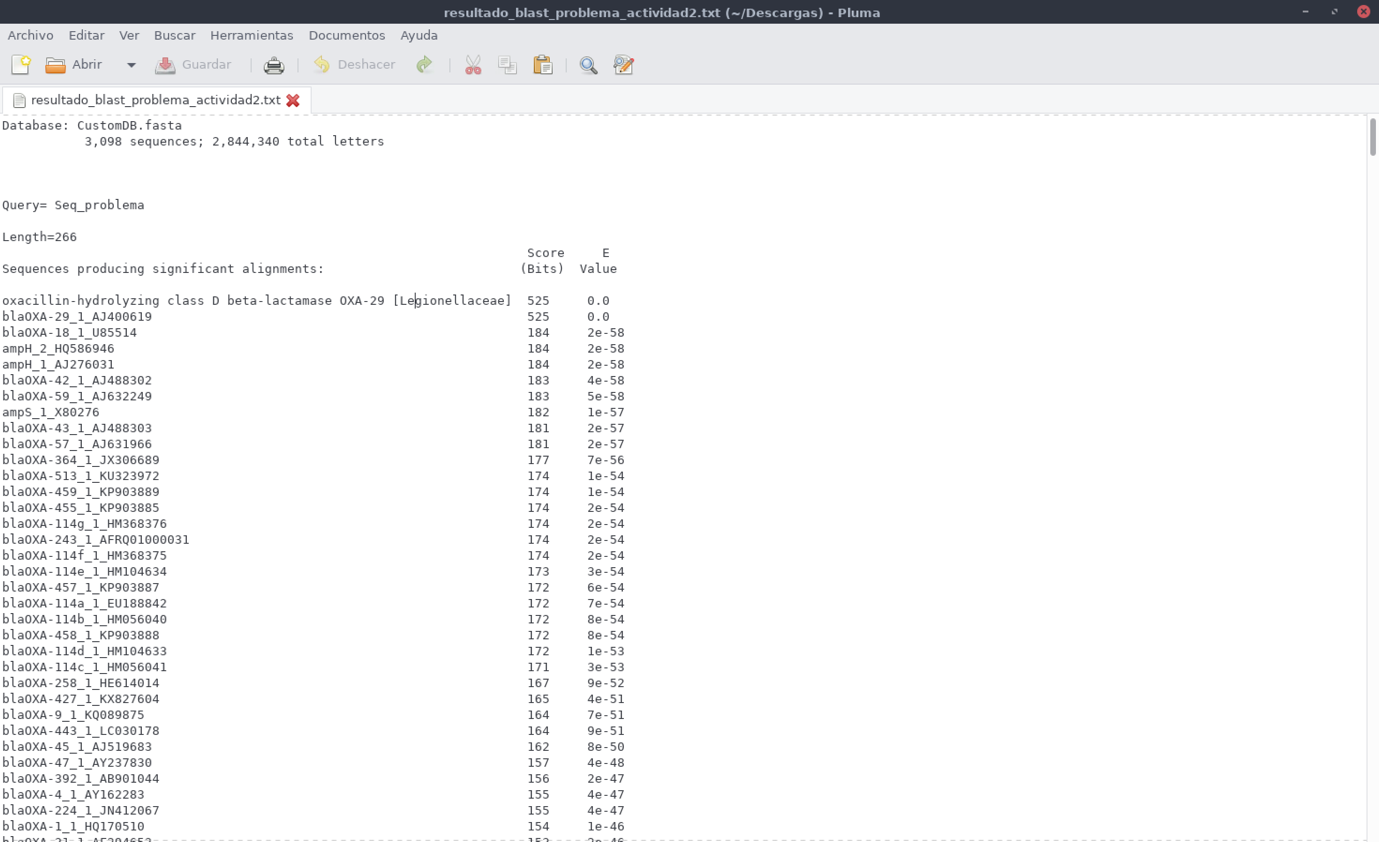


**2.1.2** Indica y justifica cual es la proteína más relacionada y una captura de pantalla del resultado.

Respuesta:

La proteína más relacionada es la beta-lactamasa OXA-29 (oxacilina), o blaOXA-29, una betalactamasa de la clase D, ya que es la que aparece con un e-value de 0.0, que calcula la probabilidad de obtener un falso positivo en el alineamiento, siendo en este caso de 0. Por otro lado, también es la que tiene un valor de Bitscore más alto, que valora la identidad y la similitud de las dos secuencias, lo que la hace también la más relacionada de todas las que aparecen. En la siguiente imagen se marca con un cuadrado en rojo.

Captura:



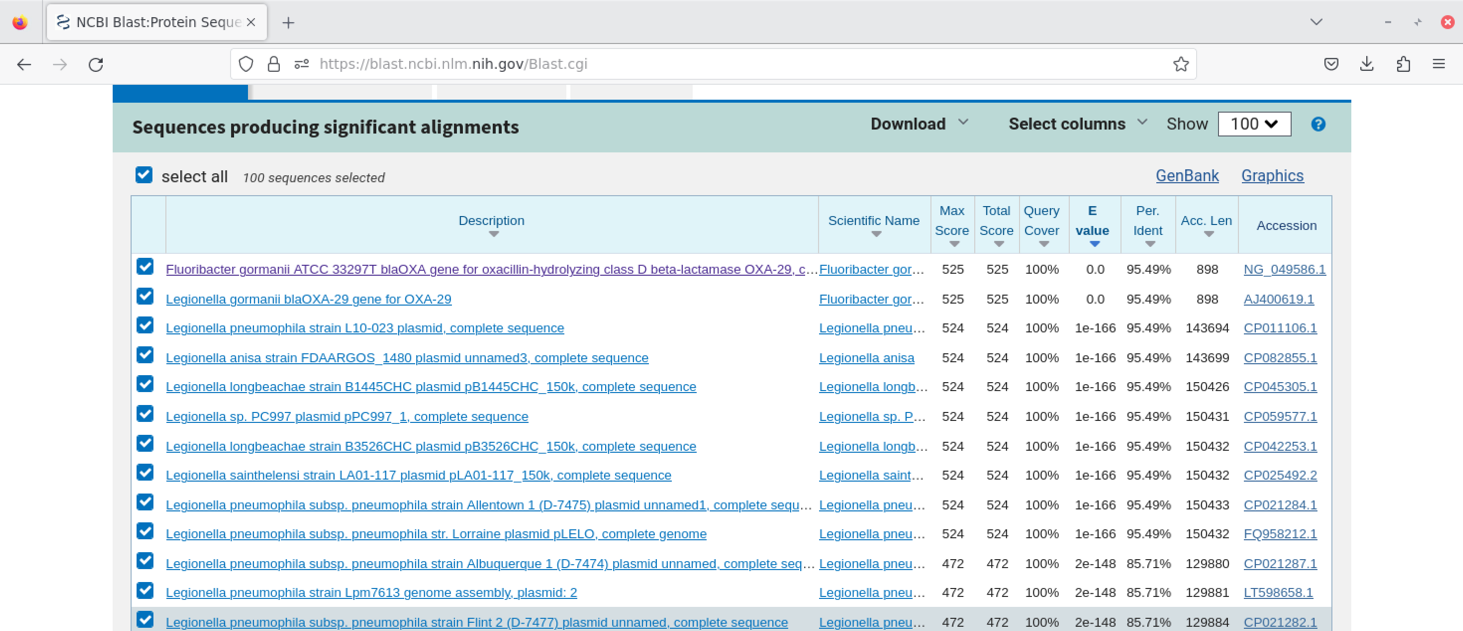
Seguiremos con un análisis en la web, blast en la web mediante la herramienta online del NCBI https://blast.ncbi.nlm.nih.gov/Blast.cgi

**2.1.3** Indica cual es la proteína más relacionada y una captura de pantalla del resultado.

Respuesta:

La proteína más relacionada es la beta-lactamasa OXA 29, oxalicina, de clase D.

Captura: Realizando un tblastn en web.

****

**2.1.4** Indica evalue e identidad en el mejor hit.

Respuesta:

e-value: 0.0

Identidad (Bitscore): 525

**2.1.5** ¿Qué método presenta un mejor resultado, web o local? Justifica la respuesta.

Respuesta: Presenta un mejor resultado en web, pues el alineamiento de las secuencias que se obtiene con la web (Imagen 1) es mucho mejor que el alineamiento de las secuencias que se obtiene mediante blast local (Imagen 2), ya que en el alineamiento local los puntos señalan las coincidencias y las letras los cambios o diferencias, por lo que hay bastantes cambios. Mientras que en el alineamiento web se puede observar que el alineamiento coincide en toda la secuencia.

Por otro lado, si realizamos un blastp en web, en lugar de un tblastn, se obtiene un resultado mejor con un Bitscore de 552, un e-value de 0.0 y un porcentaje de identidad del 100% (Imagen 3), frente al porcentaje de identidad obtenido al realizar tblasn que es de un 95. 49%.

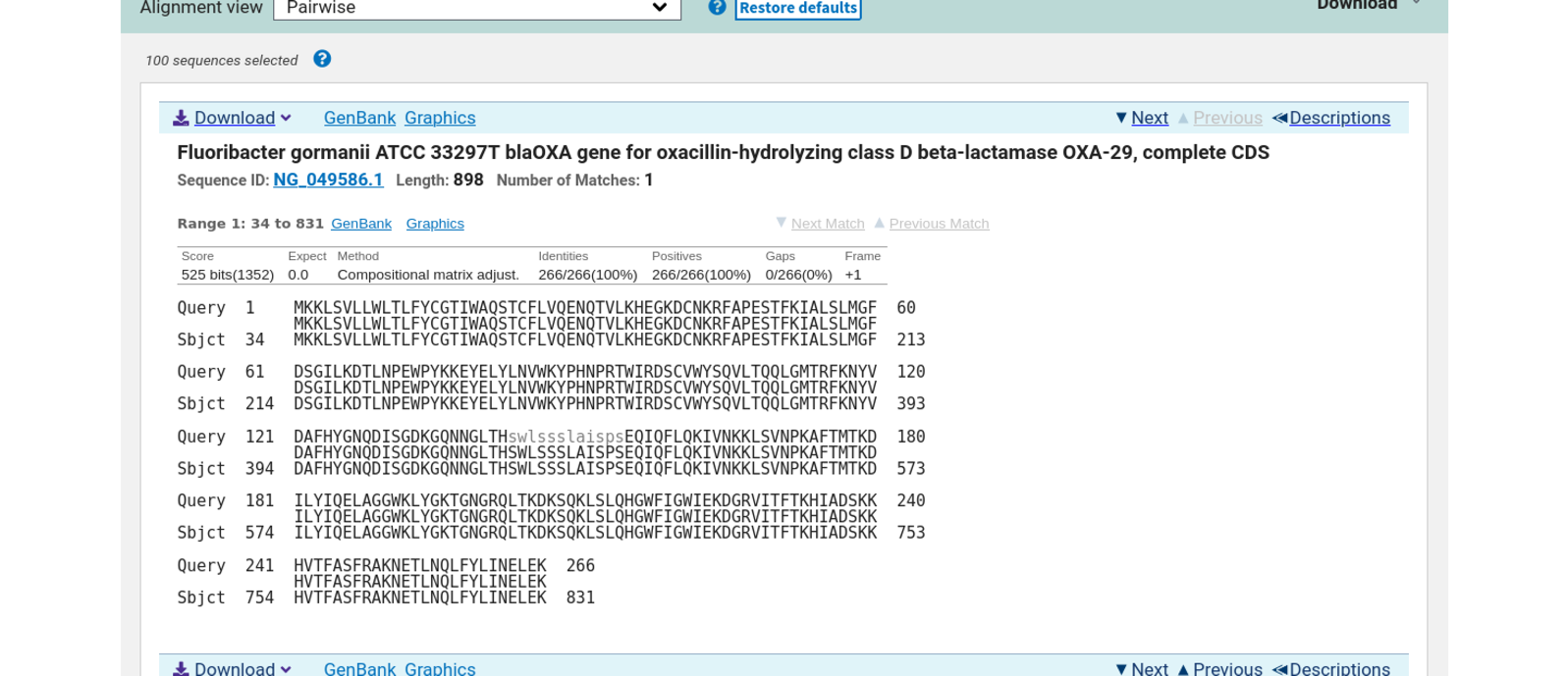


Imagen 1. Alineamiento con la herramienta tblastn obtenido en web.

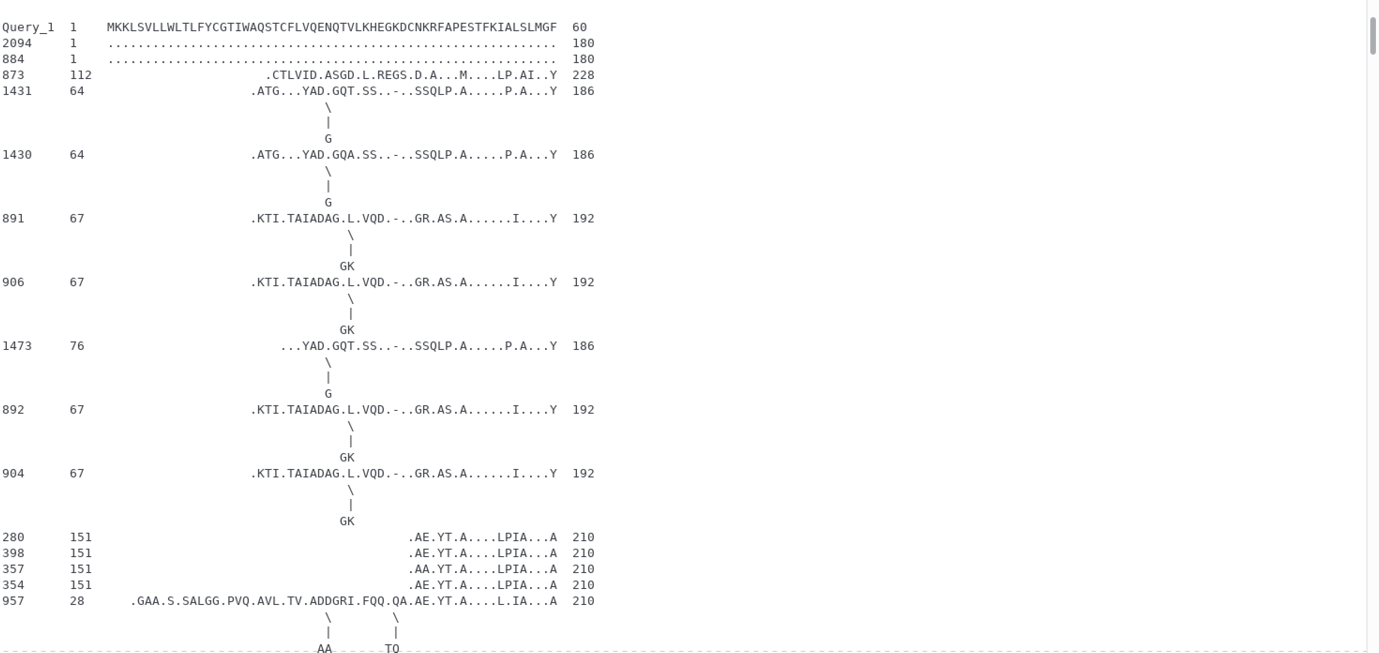


Imagen 2. Alineamiento en local.

**AP2.2 Caracterización de regiones conservadas en proteínas (3 puntos)**

Vamos a analizar varias proteínas homólogas a la frataxina humana, proteína implicada en la enfermedad ataxia de Friedreich. *Archivo: frataxin\_mamiferos.fasta*

Empezaremos por un alineamiento en local, mediante la herramienta muscle:

**2.2.1** Proporciona los comandos utilizados para realizar un alineamiento con salida en html en texto junto a una captura de pantalla en la que parezca el terminal con: el prompt, el comando utilizado y el resultado obtenido.

Comando:

muscle -in frataxin\_mamiferos.fasta -out muscle\_frataxina\_results -html

Captura:

Imagen del comando utilizado y el prompt:

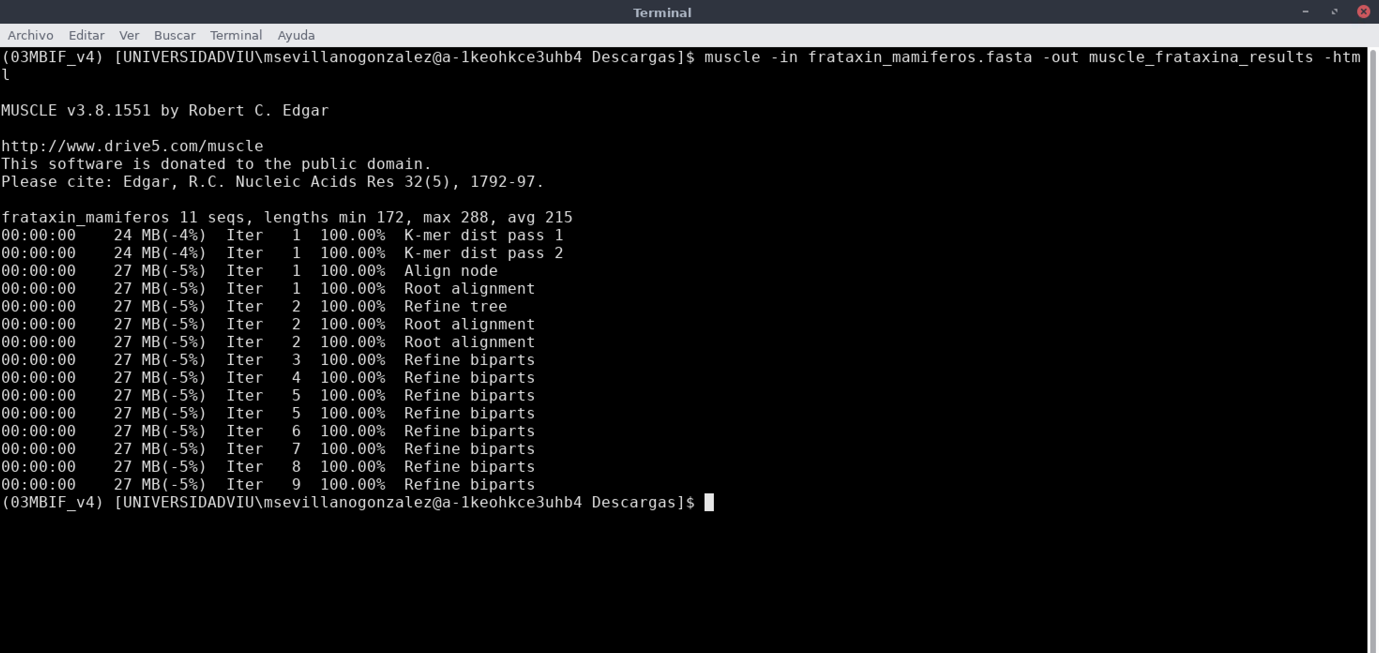
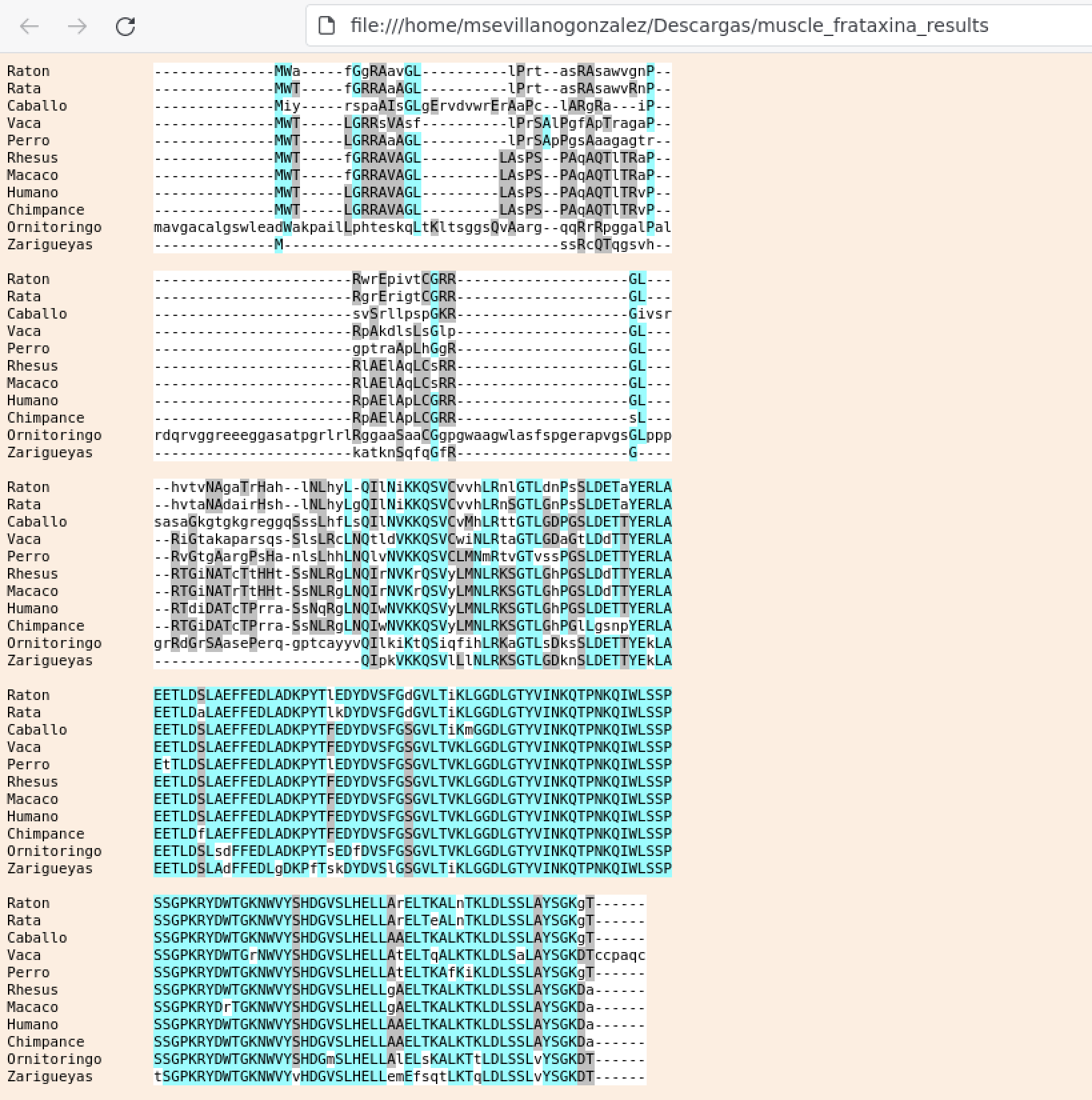


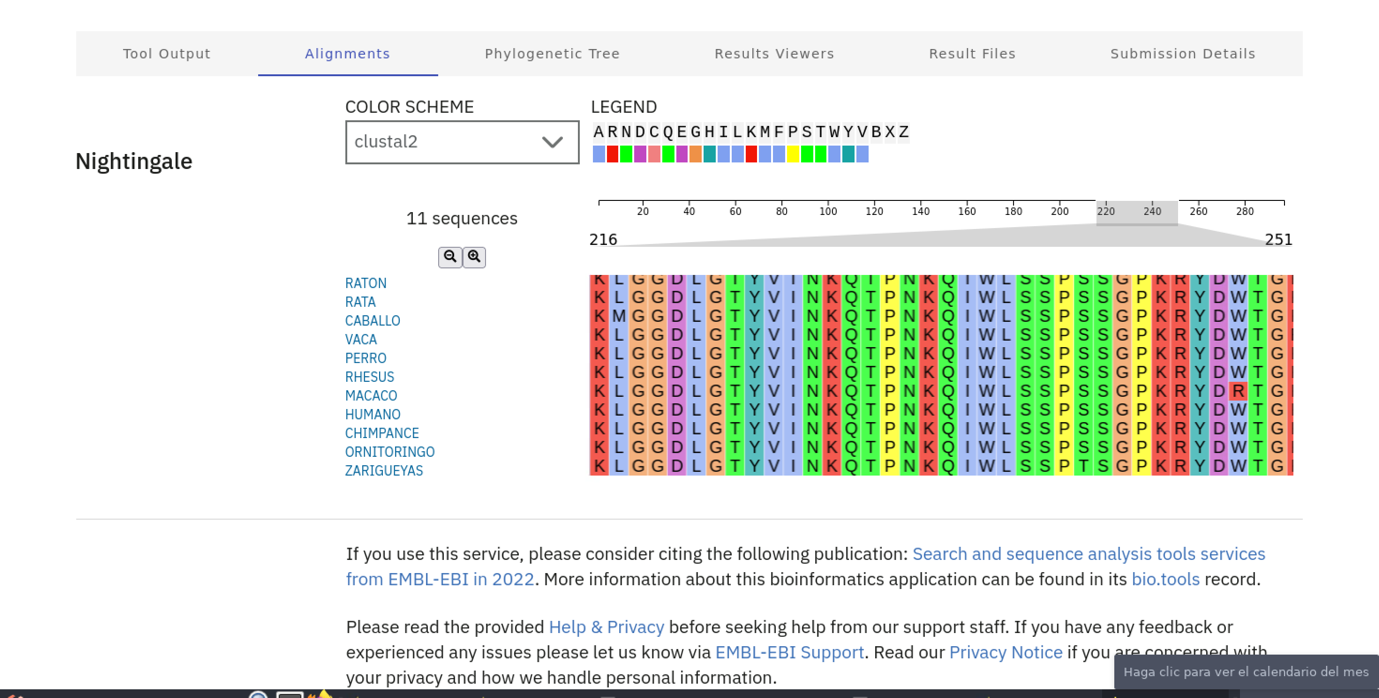
Imagen del resultado obtenido:

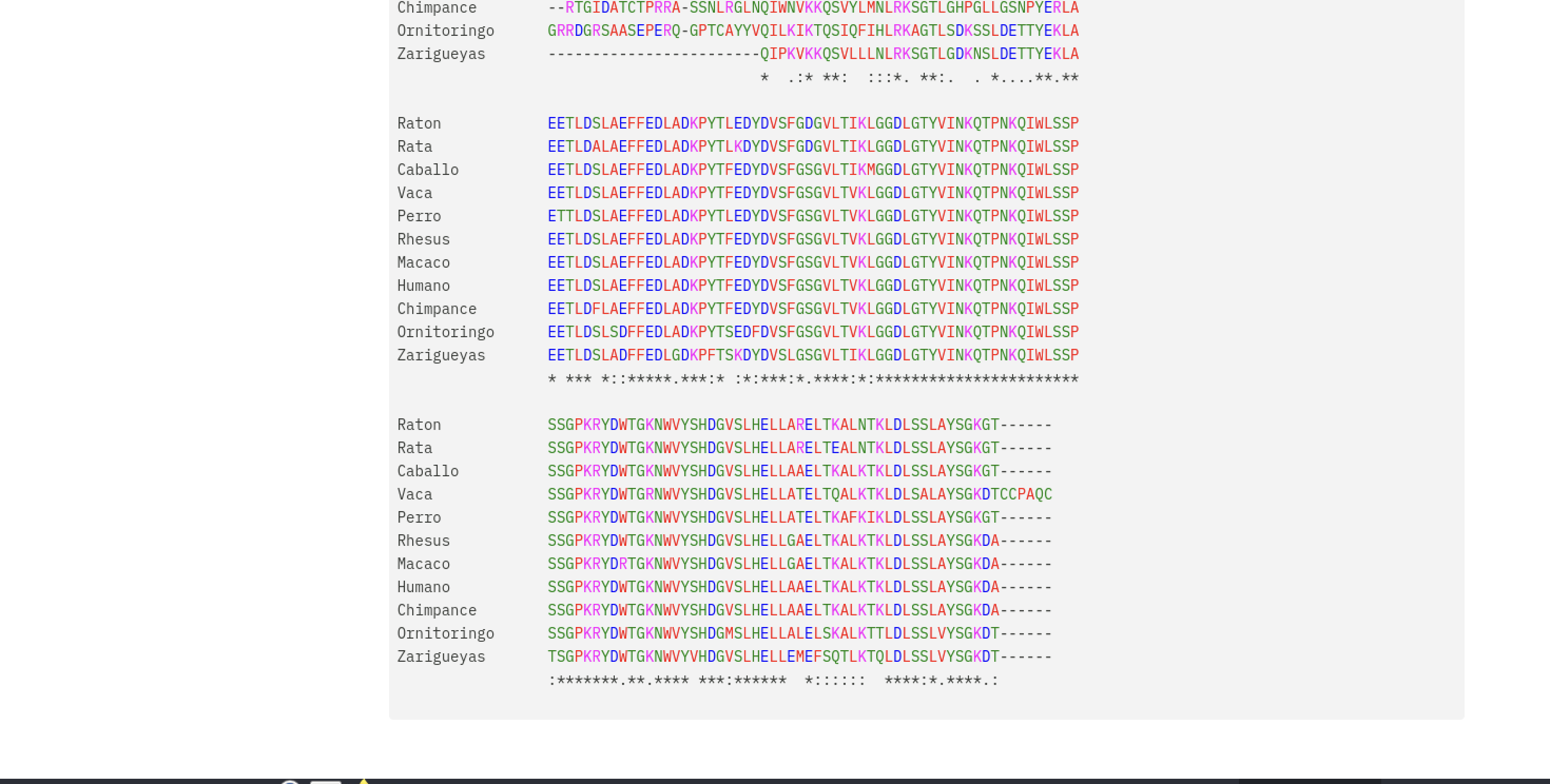


Seguiremos con un análisis en la web, alineamiento en la web mediante la herramienta online del EBI: https://www.ebi.ac.uk/Tools/msa/muscle/

**2.2.2** Proporciona una captura de pantalla con el resultado e indica y justifica en que mitad o región el alineamiento es más óptimo.

Respuesta: De la región 220 a la 249 aproximadamente es la que se observa más coincidencia entre todas las secuencias de todos los individuos, como se puede apreciar en las siguientes imágenes señalado por un recuadro. Los asteriscos debajo de las secuencias indican las coincidencias y parece haber una serie de aminoácidos que se encuentran conservados entre las especies.

Captura:



**2.2.3** Elimina las dos secuencias más alejadas, ornitorrinco y zarigüeya, repite con ambas técnicas el alineamiento proporcionando los comandos utilizados y capturas de pantalla y contesta justificando la respuesta: ¿ha mejorado o empeorado el alineamiento? ¿Se puede identificar regiones más conservadas y menos conservadas?

Respuesta: Sí ha mejorado el alineamiento, quitando por ejemplo las dos primeras partes, que eran en las que menos coincidencias había. Y han aparecido más regiones que parecen conservadas, o las que ya parecían conservadas ahora son regiones algo más largas. Se señalan en las siguientes imágenes con un recuadro azul:





------------------------------------------------------------------------------------------------------------------------------

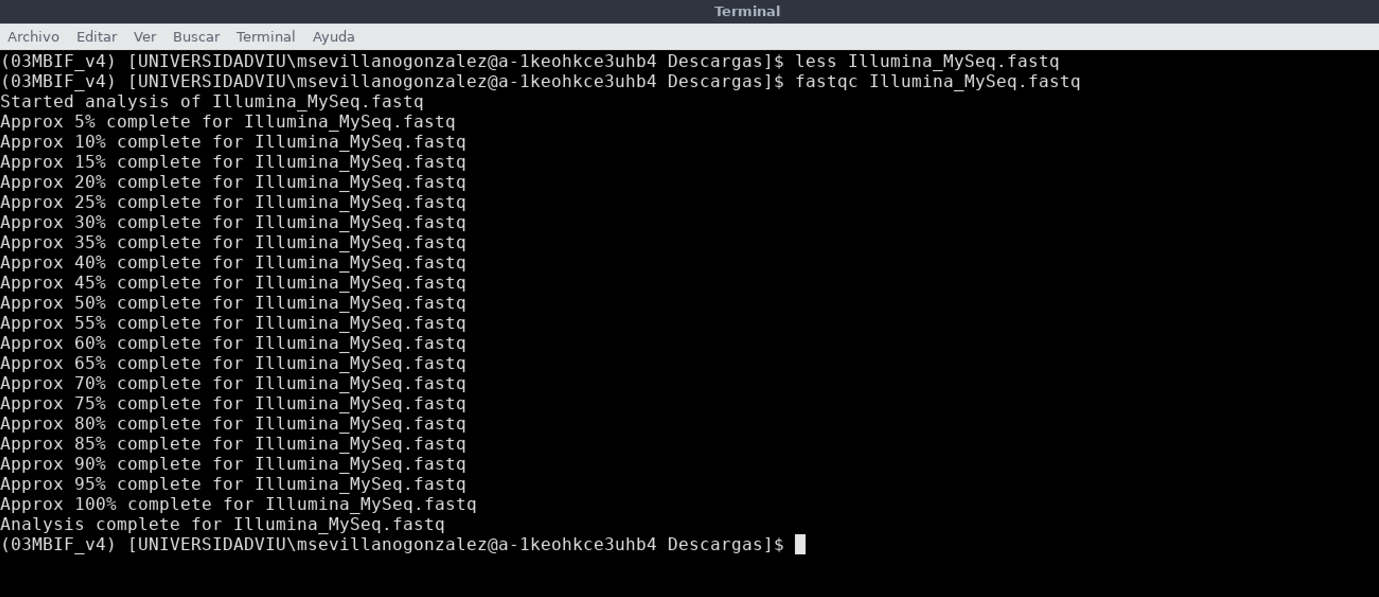
**AP2.3 Calidad y pre-procesamiento de secuencias (2 puntos)**

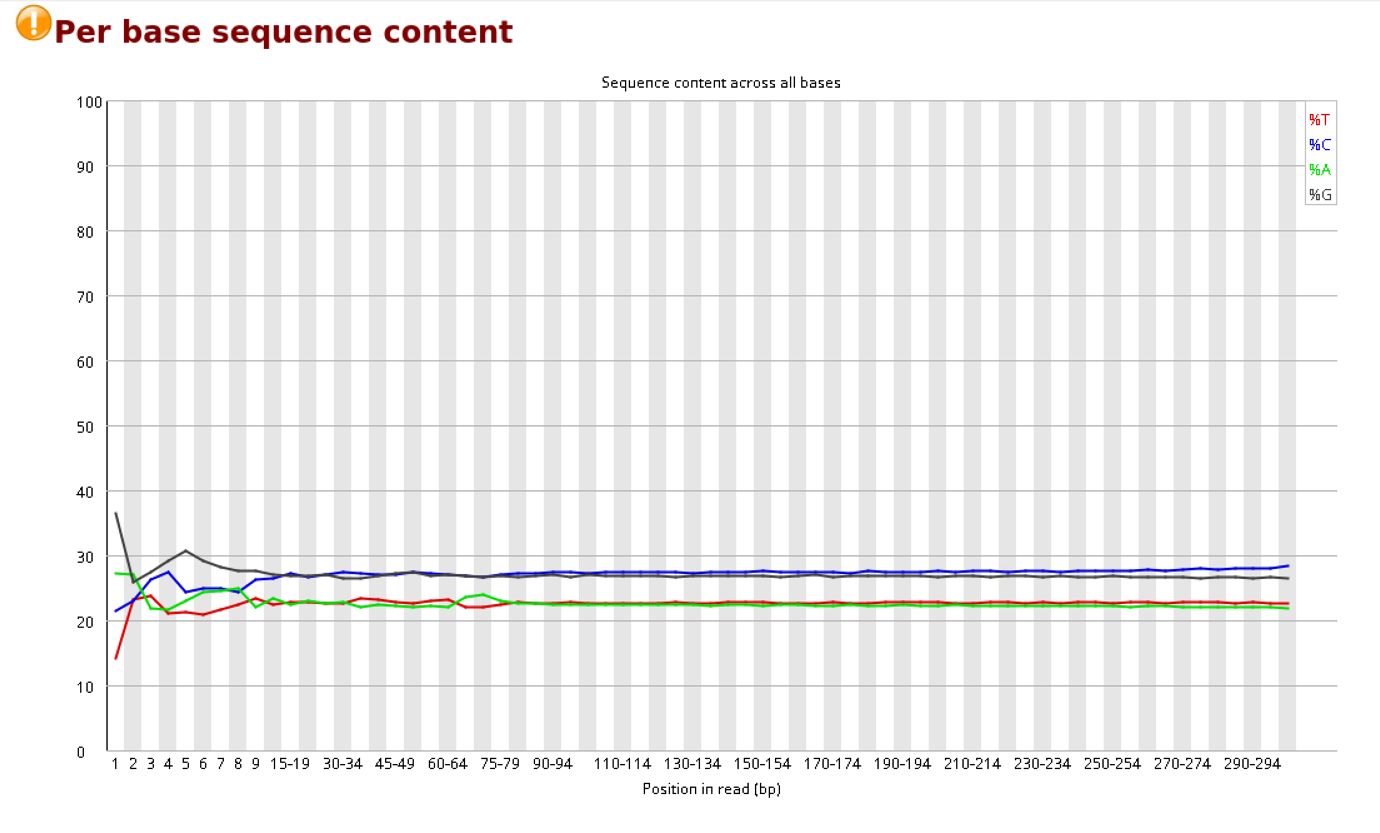
Utilizando el archivo de Illumina\_MySeq.fastq:

**2.3.1** Realiza un primer análisis de calidad con la herramienta fastqc. Proporciona el comando utilizado y una captura de pantalla de la sección del informe que muestra “*Per base sequence content*”.

Comando: fastqc Illumina\_MySeq.fastq

Captura:





**2.3.2** Mediante la herramienta prinseq-lite.pl realiza un filtrado por calidad mínima media Q25 y recorta los primeros 20 nucleótidos de las secuencias en 5’.  Proporciona el comando utilizado y una captura de pantalla de la sección del informe que muestra “*Per base sequence content*”.

Comando: Los comandos a seguir serían:

Primero un fastqc Illumina\_MySeq.fastq

Después se haría el filtrado por calidad: prinseq-lite.pl -fastq Illumina\_MySeq.fastq -min\_qual\_mean 25

fastq Illumina\_MySeq\_prinseq\_good\_DHXy.fastq

Y, por último, para recortar los 20 nucleótidos de las secuencias en 5’: prinseq-lite.pl -fastq Illumina\_MySeq\_prinseq\_good\_DHXy.fastq -trim\_left 20

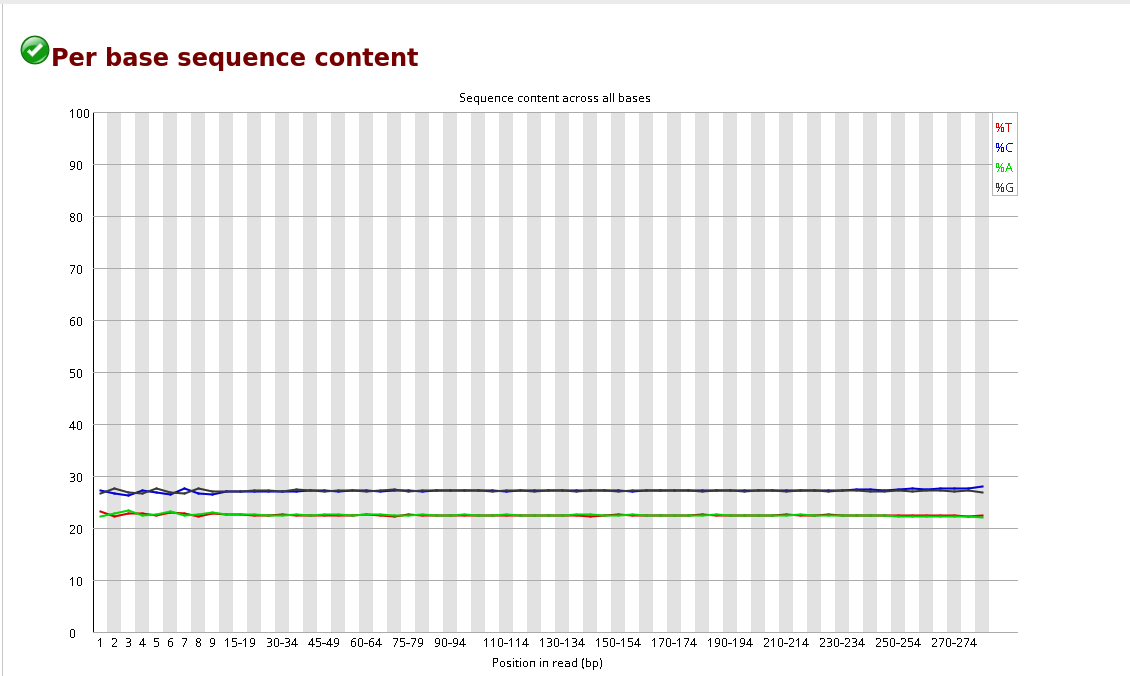
Y fastqc del nuevo archivo: fastqc Illumina\_MySeq\_prinseq\_good\_DHXy\_prinseq\_good\_Kw5Q.fastq

Otra forma de hacerlo podría ser combinando ambos comandos a la vez, para filtrar por calidad y recortando los 20 primeros nucleótidos con el comando:

prinseq-lite.pl -fastq Illumina\_MySeq.fastq -min\_qual\_mean 25 -trim\_left 20

Captura:

Captura de pantalla del resultado de *“Per base sequence content”* de la primera forma de hacerlo.



Capturas de pantalla de la segunda forma de hacerlo, combinando el filtrado por calidad y el recorte de los 20 primeros nucleótidos en un mismo comando:

