

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:

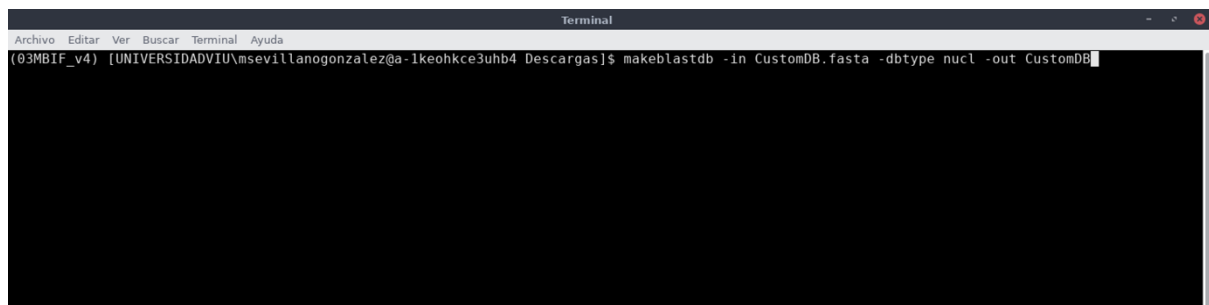
Primero se crea la base de datos en un archivo, con el comando: `makeblastdb -in CustomDB.fasta -dbtype nucl -out CustomDB`

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 -outfmt 1`

Captura:

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

A screenshot of a terminal window titled "Terminal". The window has a menu bar with "Archivo", "Editar", "Ver", "Buscar", "Terminal", and "Ayuda". The terminal shows a command prompt where the user has entered the command: `makeblastdb -in CustomDB.fasta -dbtype nucl -out CustomDB`. The command is partially visible, with the cursor at the end of the line. The terminal background is black, and the text is white.

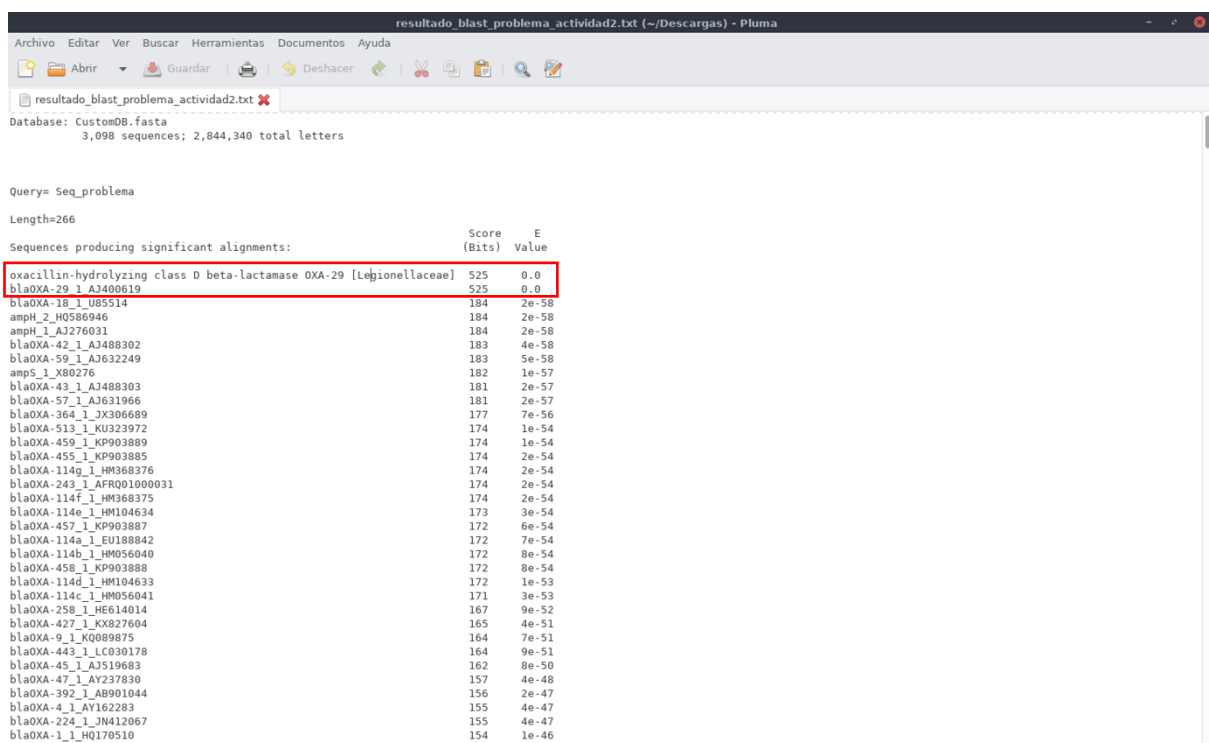
```
Terminal
Archivo Editar Ver Buscar Terminal Ayuda
(03MBIF_v4) [UNIVERSIDADVIU\msevillanogonzalez@a-lkeohkce3uhb4 Descargas]$ tblastn -query problema.fasta -db CustomDB -out resultado_blast_proble
ma_actividad2.txt -outfmt 1
(03MBIF_v4) [UNIVERSIDADVIU\msevillanogonzalez@a-lkeohkce3uhb4 Descargas]$
```

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:



| Sequences producing significant alignments: | Score (Bits) | E Value |
|---|--------------|---------|
| oxacillin-hydrolyzing class D beta-lactamase OXA-29 [Leptionellaceae] | 525 | 0.0 |
| blaOXA-29_1_AJ400619 | 525 | 0.0 |
| blaOXA-18_1_U85514 | 184 | 2e-58 |
| ampH_2_HQ586946 | 184 | 2e-58 |
| ampH_1_AJ276031 | 184 | 2e-58 |
| blaOXA-42_1_AJ488302 | 183 | 4e-58 |
| blaOXA-59_1_AJ632249 | 183 | 5e-58 |
| ampS_1_X80276 | 182 | 1e-57 |
| blaOXA-43_1_AJ488303 | 181 | 2e-57 |
| blaOXA-57_1_AJ631966 | 181 | 2e-57 |
| blaOXA-364_1_XJ306689 | 177 | 7e-56 |
| blaOXA-513_1_KU323972 | 174 | 1e-54 |
| blaOXA-459_1_KP903889 | 174 | 1e-54 |
| blaOXA-455_1_KP903885 | 174 | 2e-54 |
| blaOXA-114g_1_HM368376 | 174 | 2e-54 |
| blaOXA-243_1_AFRQ01000031 | 174 | 2e-54 |
| blaOXA-114f_1_HM368375 | 174 | 2e-54 |
| blaOXA-114e_1_HM104634 | 173 | 3e-54 |
| blaOXA-457_1_KP903887 | 172 | 6e-54 |
| blaOXA-114a_1_EU188842 | 172 | 7e-54 |
| blaOXA-114b_1_HM056040 | 172 | 8e-54 |
| blaOXA-458_1_KP903888 | 172 | 8e-54 |
| blaOXA-114d_1_HM104633 | 172 | 1e-53 |
| blaOXA-114c_1_HM056041 | 171 | 3e-53 |
| blaOXA-258_1_HE614014 | 167 | 9e-52 |
| blaOXA-427_1_KX827604 | 165 | 4e-51 |
| blaOXA-9_1_KQ889875 | 164 | 7e-51 |
| blaOXA-443_1_LC030178 | 164 | 9e-51 |
| blaOXA-45_1_AJ519683 | 162 | 8e-50 |
| blaOXA-47_1_AY237830 | 157 | 4e-48 |
| blaOXA-392_1_AB901044 | 156 | 2e-47 |
| blaOXA-4_1_AY162283 | 155 | 4e-47 |
| blaOXA-224_1_JN412067 | 155 | 4e-47 |
| blaOXA-1_1_HQ170510 | 154 | 1e-46 |

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.

| Description | Scientific Name | Max Score | Total Score | Query Cover | E-value | Per. Ident. | Acc. Len | Accession |
|--|------------------------|-----------|-------------|-------------|---------|-------------|----------|-------------|
| <input checked="" type="checkbox"/> Fluoribacter gormanii ATCC 33297T blaOXA gene for oxacillin-hydrolyzing class D beta-lactamase OXA-29... | Fluoribacter gormanii | 525 | 525 | 100% | 0.0 | 95.49% | 898 | NG_049586.1 |
| <input checked="" type="checkbox"/> Legionella pneumophila strain L10-023 plasmid, complete sequence | Legionella pneumophila | 524 | 524 | 100% | 1e-166 | 95.49% | 143694 | CP011106.1 |
| <input checked="" type="checkbox"/> Legionella pneumophila strain B1445CHC plasmid pB1445CHC_150k, complete sequence | Legionella pneumophila | 524 | 524 | 100% | 1e-166 | 95.49% | 150426 | CP045305.1 |
| <input checked="" type="checkbox"/> Legionella pneumophila strain B3526CHC plasmid pB3526CHC_150k, complete sequence | Legionella pneumophila | 524 | 524 | 100% | 1e-166 | 95.49% | 150432 | CP042253.1 |
| <input checked="" type="checkbox"/> Legionella pneumophila strain LA01-117 plasmid pLA01-117_150k, complete sequence | Legionella pneumophila | 524 | 524 | 100% | 1e-166 | 95.49% | 150432 | CP025492.2 |
| <input checked="" type="checkbox"/> Legionella pneumophila subsp. pneumophila strain Allentown 1 (D-7475) plasmid unnamed1, complete sequence | Legionella pneumophila | 524 | 524 | 100% | 1e-166 | 95.49% | 150433 | CP021284.1 |
| <input checked="" type="checkbox"/> Legionella pneumophila subsp. pneumophila strain Lorraine plasmid pLELO, complete genome | Legionella pneumophila | 524 | 524 | 100% | 1e-166 | 95.49% | 150432 | FQ958212.1 |
| <input checked="" type="checkbox"/> Legionella pneumophila subsp. pneumophila strain Albuquerque 1 (D-7474) plasmid unnamed, complete sequence | Legionella pneumophila | 472 | 472 | 100% | 2e-148 | 85.71% | 129880 | CP021287.1 |
| <input checked="" type="checkbox"/> Legionella pneumophila strain Lpm7613 genome assembly, plasmid 2 | Legionella pneumophila | 472 | 472 | 100% | 2e-148 | 85.71% | 129881 | LT598658.1 |
| <input checked="" type="checkbox"/> Legionella pneumophila subsp. pneumophila strain Flint 2 (D-7477) plasmid unnamed, complete sequence | Legionella pneumophila | 472 | 472 | 100% | 2e-148 | 85.71% | 129884 | CP021282.1 |

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 tblastn que es de un 95.49%.

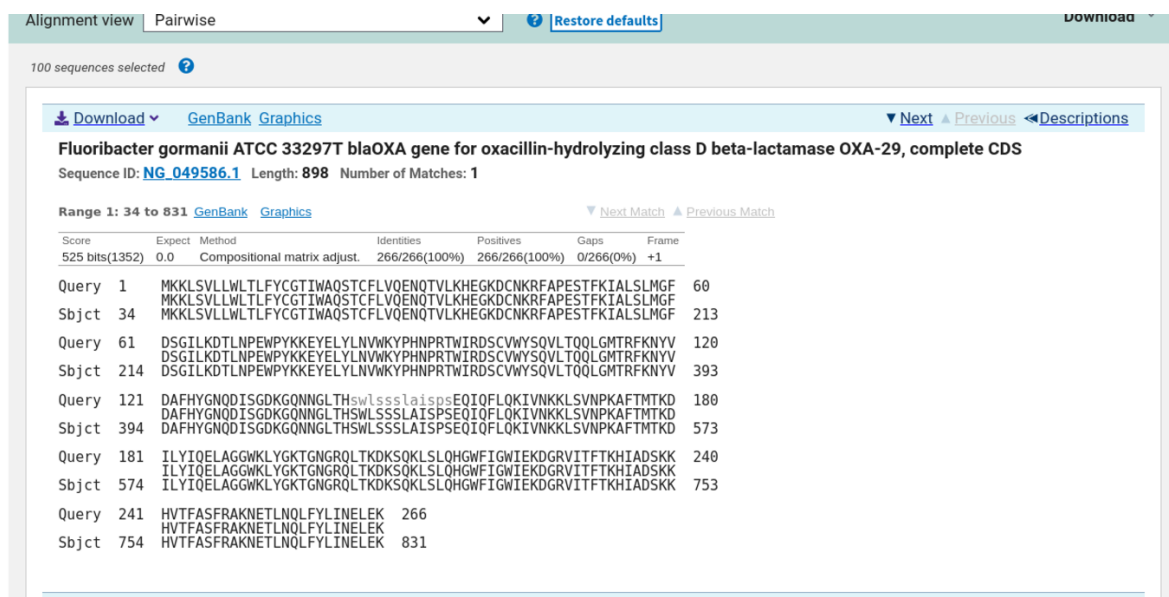


Imagen 1. Alineamiento con la herramienta tblastn obtenido en web.

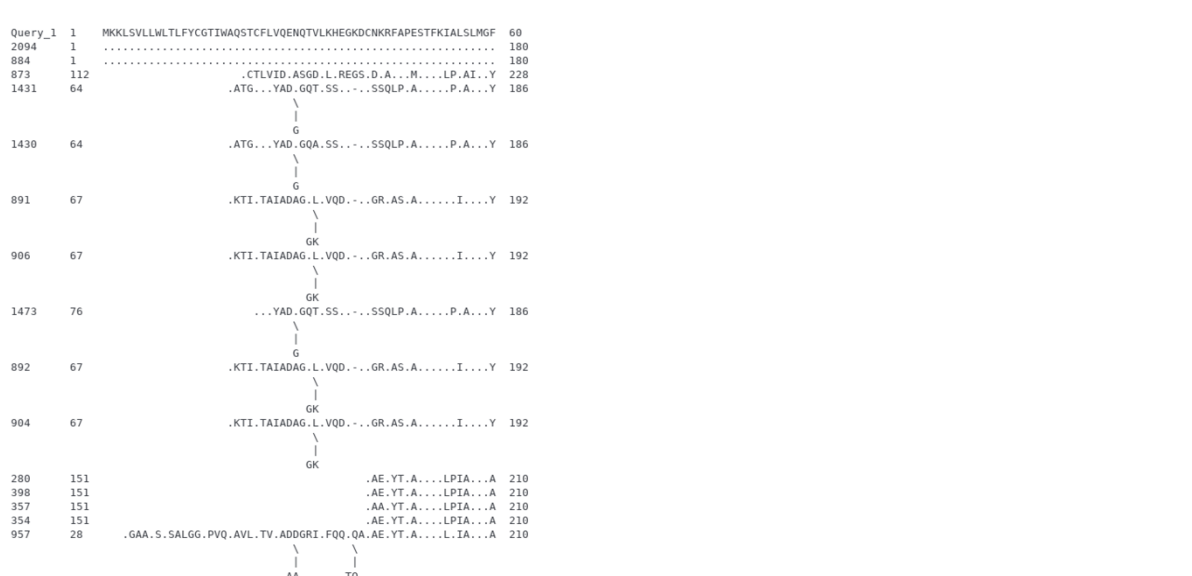


Imagen 2. Alineamiento en local.

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:

```
Terminal
Archivo  Editar  Ver  Buscar  Terminal  Ayuda
(03MBIF_v4) [UNIVERSIDADVIU\msevillanogonzalez@a-lkeohkce3uhb4 Descargas]$ muscle -in frataxin_mamiferos.fasta -out muscle_frataxina_results -html
MUSCLE v3.8.1551 by Robert C. Edgar

http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

frataxin mamiferos 11 seqs, lengths min 172, max 288, avg 215
00:00:00 24 MB(-4%) Iter 1 100.00% K-mer dist pass 1
00:00:00 24 MB(-4%) Iter 1 100.00% K-mer dist pass 2
00:00:00 27 MB(-5%) Iter 1 100.00% Align node
00:00:00 27 MB(-5%) Iter 1 100.00% Root alignment
00:00:00 27 MB(-5%) Iter 2 100.00% Refine tree
00:00:00 27 MB(-5%) Iter 2 100.00% Root alignment
00:00:00 27 MB(-5%) Iter 2 100.00% Root alignment
00:00:00 27 MB(-5%) Iter 3 100.00% Refine biparts
00:00:00 27 MB(-5%) Iter 4 100.00% Refine biparts
00:00:00 27 MB(-5%) Iter 5 100.00% Refine biparts
00:00:00 27 MB(-5%) Iter 5 100.00% Refine biparts
00:00:00 27 MB(-5%) Iter 6 100.00% Refine biparts
00:00:00 27 MB(-5%) Iter 7 100.00% Refine biparts
00:00:00 27 MB(-5%) Iter 8 100.00% Refine biparts
00:00:00 27 MB(-5%) Iter 9 100.00% Refine biparts
(03MBIF_v4) [UNIVERSIDADVIU\msevillanogonzalez@a-lkeohkce3uhb4 Descargas]$
```

Imagen del resultado obtenido:

file:///home/msevillanogonzalez/Descargas/muscle_frataxina_results

| | |
|-------------|---|
| Raton | -----Mwa-----fGgRAavGL-----lPrt--asRAsawvgnP-- |
| Rata | -----MWT-----fGRRaAaAGL-----lPrt--asRAsawvRnP-- |
| Caballo | -----Miy-----rspaATsGLgErvdvwrErAaPc--lARgRa--iP-- |
| Vaca | -----MWT-----LGRRsVAsf-----lPrSA1PgFApTragaP-- |
| Perro | -----MWT-----LGRRaAaAGL-----lPrSApPqsAaagagtr-- |
| Rhesus | -----MWT-----fGRRAVAGL-----LASPS--PAqAQTlTRaP-- |
| Macaco | -----MWT-----fGRRAVAGL-----LASPS--PAqAQTlTRaP-- |
| Humano | -----MWT-----LGRRAVAGL-----LASPS--PAqAQTlTRvP-- |
| Chimpance | -----MWT-----LGRRAVAGL-----LASPS--PAqAQTlTRvP-- |
| Ornitoringo | mavgaacalgswLeadWakpailLphteskqLtKltsggsQvAarg--qqRrRpggalPal |
| Zarigueyas | -----M-----ssRcQTqgsvh-- |
| Raton | -----RwrEpivtCGRR-----GL-- |
| Rata | -----RgrErigtCGRR-----GL-- |
| Caballo | -----svSrlpspGKR-----Givsr |
| Vaca | -----RpAkdLsLsGLp-----GL-- |
| Perro | -----gptraApLhGgR-----GL-- |
| Rhesus | -----RLAE1AqLCsRR-----GL-- |
| Macaco | -----RLAE1AqLCsRR-----GL-- |
| Humano | -----RpAE1ApLCGR-----GL-- |
| Chimpance | -----RpAE1ApLCGR-----sL-- |
| Ornitoringo | rdqrvvgreeeggasatpgrlrLggaaSaaCgpgwaagwLasfspgerapvgsGLppp |
| Zarigueyas | -----katknSsqfQGr-----G----- |
| Raton | --hvtvNagaTrHah--lNLhyL-QlLNiKKQSVcvvhLRnlGTLdnPsSLDETaYERLA |
| Rata | --hvtvNAdairHsh--lNLhyLQlLNiKKQSVcvvhLRnSGTLGnPsSLDETaYERLA |
| Caballo | sasaGktgkggreggqSssLhflsQlLNiKKQSVcVhLrttGTLGDPGSLEDTTYERLA |
| Vaca | --Rigtakaparsqs-SlsLRcLNQtlldVKKQSVcwiLRtaGTLGDaGtLddTTYERLA |
| Perro | --RvGtgAargPshA-nlsLnhLNQlvNVKKQSVCLMnRtvGtvssPGSLEDTTYERLA |
| Rhesus | --RTGiNATcTtHht-SsNLrgLNQlRNVKqQSVyLMNLRKSGTLGhPGSLddTTYERLA |
| Macaco | --RTGiNATcTtHht-SsNLrgLNQlRNVKqQSVyLMNLRKSGTLGhPGSLddTTYERLA |
| Humano | --RTdiDATcTPrra-SsNqRgLNQlWNVKKQSVyLMNLRKSGTLGhPGSLDETTTYERLA |
| Chimpance | --RTdiDATcTPrra-SsNqRgLNQlWNVKKQSVyLMNLRKSGTLGhPGLLgsnpYERLA |
| Ornitoringo | grRdGrSAasePerq-gptcayyvQlkiKtQSiqfihLRKaGTLsDksSLDETTTYekLA |
| Zarigueyas | -----QlPkVKKQSVLlLNLRKSGTLGDknSLDETTTYekLA |
| Raton | EETLDSLAEFFEDLADKPYTLEDYDVSFGdGVLTiKLGGDLGTYVINKOTPNKQIWLSSP |
| Rata | EETLDaLAEFFEDLADKPYTLkDYDVSFGdGVLTiKLGGDLGTYVINKOTPNKQIWLSSP |
| Caballo | EETLDSLAEFFEDLADKPYTFEDYDVSFGSGVLTiKmGGDLGTYVINKOTPNKQIWLSSP |
| Vaca | EETLDSLAEFFEDLADKPYTFEDYDVSFGSGVLTVKLGGDLGTYVINKOTPNKQIWLSSP |
| Perro | EtLDSLAEFFEDLADKPYTLEDYDVSFGSGVLTVKLGGDLGTYVINKOTPNKQIWLSSP |
| Rhesus | EETLDSLAEFFEDLADKPYTFEDYDVSFGSGVLTVKLGGDLGTYVINKOTPNKQIWLSSP |
| Macaco | EETLDSLAEFFEDLADKPYTFEDYDVSFGSGVLTVKLGGDLGTYVINKOTPNKQIWLSSP |
| Humano | EETLDSLAEFFEDLADKPYTFEDYDVSFGSGVLTVKLGGDLGTYVINKOTPNKQIWLSSP |
| Chimpance | EETLDfLAEFFEDLADKPYTFEDYDVSFGSGVLTVKLGGDLGTYVINKOTPNKQIWLSSP |
| Ornitoringo | EETLDSLsdFFEDLADKPYTsEdfDVSFGSGVLTVKLGGDLGTYVINKOTPNKQIWLSSP |
| Zarigueyas | EETLDSLadFFEDLgDKPftSkDYDVSIGSGVLTiKLGGDLGTYVINKOTPNKQIWLSSP |
| Raton | SSGPKRYDWTGKNwVYSHDGVSLHELLAReLTkALnTKLDLSSLAYSGKgT----- |
| Rata | SSGPKRYDWTGKNwVYSHDGVSLHELLAReLTaALnTKLDLSSLAYSGKgT----- |
| Caballo | SSGPKRYDWTGKNwVYSHDGVSLHELLAAELTKALKTKLDLSSLAYSGKgT----- |
| Vaca | SSGPKRYDWTGrnwVYSHDGVSLHELLAtELtqALKTKLDLSaLAYSGKDTccpaqc |
| Perro | SSGPKRYDWTGKNwVYSHDGVSLHELLAtELTKAfKlKLdLSSLAYSGKgT----- |
| Rhesus | SSGPKRYDWTGKNwVYSHDGVSLHELLgAELTKALKTKLDLSSLAYSGKDa----- |
| Macaco | SSGPKRYDTrGKNwVYSHDGVSLHELLgAELTKALKTKLDLSSLAYSGKDa----- |
| Humano | SSGPKRYDWTGKNwVYSHDGVSLHELLAAELTKALKTKLDLSSLAYSGKDa----- |
| Chimpance | SSGPKRYDWTGKNwVYSHDGVSLHELLAAELTKALKTKLDLSSLAYSGKDa----- |
| Ornitoringo | SSGPKRYDWTGKNwVYSHDgMSHELLA1ELsKALKtTLdLSSLYSGKDT----- |
| Zarigueyas | tSGPKRYDWTGKNwVYSHDGVSLHELLemEfseqtLKTqLDLSSLYSGKDT----- |

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:

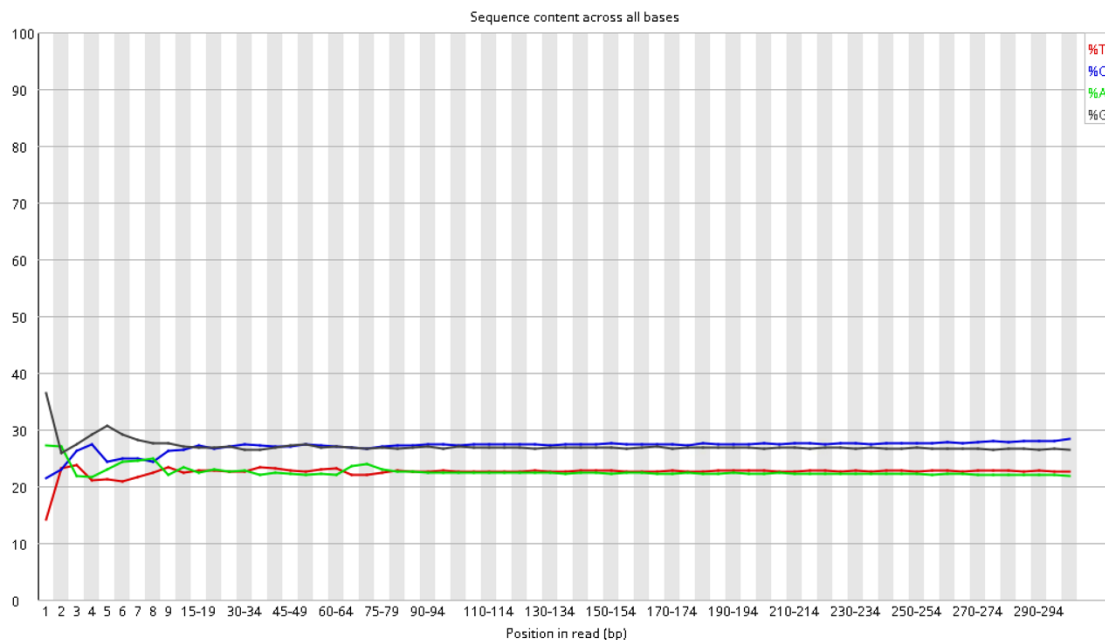
2.3.1 Realiza un primer análisis de calidad con la herramienta fastqc. Proporciona el comando utilizado y una captura de pantalla de la sección del informe que muestra “Per base sequence content”.

Comando: fastqc Illumina_MySeq.fastq

Captura:

```
Terminal
Archivo Editar Ver Buscar Terminal Ayuda
(03MBIF_v4) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Descargas]$ less Illumina_MySeq.fastq
(03MBIF_v4) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Descargas]$ fastqc Illumina_MySeq.fastq
Started analysis of Illumina_MySeq.fastq
Approx 5% complete for Illumina_MySeq.fastq
Approx 10% complete for Illumina_MySeq.fastq
Approx 15% complete for Illumina_MySeq.fastq
Approx 20% complete for Illumina_MySeq.fastq
Approx 25% complete for Illumina_MySeq.fastq
Approx 30% complete for Illumina_MySeq.fastq
Approx 35% complete for Illumina_MySeq.fastq
Approx 40% complete for Illumina_MySeq.fastq
Approx 45% complete for Illumina_MySeq.fastq
Approx 50% complete for Illumina_MySeq.fastq
Approx 55% complete for Illumina_MySeq.fastq
Approx 60% complete for Illumina_MySeq.fastq
Approx 65% complete for Illumina_MySeq.fastq
Approx 70% complete for Illumina_MySeq.fastq
Approx 75% complete for Illumina_MySeq.fastq
Approx 80% complete for Illumina_MySeq.fastq
Approx 85% complete for Illumina_MySeq.fastq
Approx 90% complete for Illumina_MySeq.fastq
Approx 95% complete for Illumina_MySeq.fastq
Approx 100% complete for Illumina_MySeq.fastq
Analysis complete for Illumina_MySeq.fastq
(03MBIF_v4) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Descargas]$
```

! Per base sequence content



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

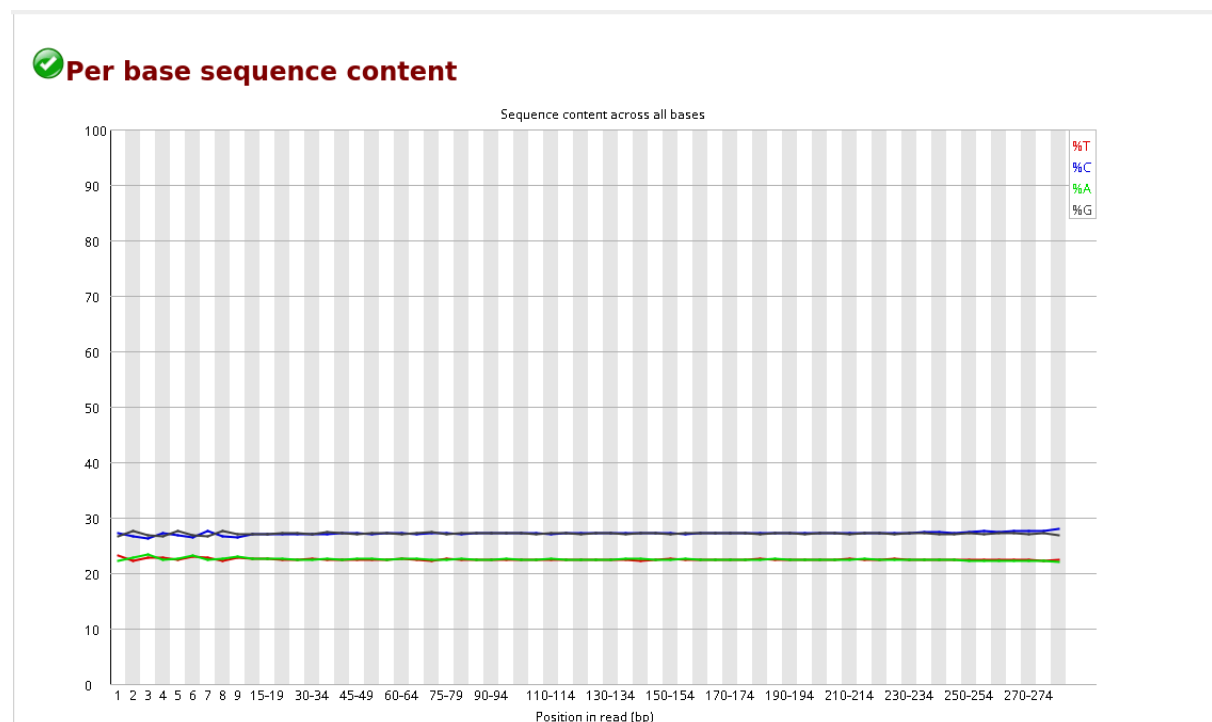
Y fastqc del nuevo archivo: fastqc Illumina_MySeq_prinseq_good_DHXy_prinseq_good_Kw5Q.fastq

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:

```
(03MBIF_v4) [UNIVERSIDADVIU\msevillanogonzalez@a-lkeohkce3uhb4 Descargas]$ prinseq-lite.pl -fastq Illumina_MySeq.fastq -min_qual_mean 25 -trim_le
ft 20
Input and filter stats:
  Input sequences: 250,000
  Input bases: 75,000,000
  Input mean length: 300.00
  Good sequences: 213,546 (85.42%)
  Good bases: 59,792,880
  Good mean length: 280.00
  Bad sequences: 36,454 (14.58%)
  Bad bases: 10,936,200
  Bad mean length: 300.00
  Sequences filtered by specified parameters:
  min_qual_mean: 36454
(03MBIF_v4) [UNIVERSIDADVIU\msevillanogonzalez@a-lkeohkce3uhb4 Descargas]$
```

