

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:

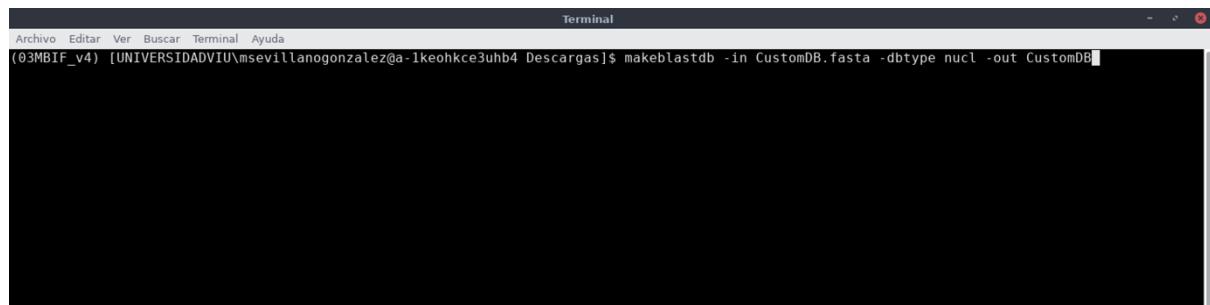
Primero se crea la base de datos en un archivo, con el comando: makeblastdb -in CustomDB.fasta -dbtype nucl -out CustomDB

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 dark background and light-colored text. At the top, there is a menu bar with options: Archivo, Editar, Ver, Buscar, Terminal, Ayuda. Below the menu, the window title is "Terminal". The main area of the terminal shows a command-line session. The user is in a directory "(03MBIF_v4) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Descargas]". The user has run the command "makeblastdb -in CustomDB.fasta -dbtype nucl -out CustomDB". The output of the command is visible, showing the progress of the database creation process.

```

Archivo Editar Ver Buscar Terminal Ayuda
(03MBIF_v4) [UNIVERSIDADADIVU\msevillanogonzalez@a-1keohkce3uhb4 Descargas]$ blastn -query problema.fasta -db CustomDB -out resultado_blast_problema_actividad2.txt -outfmt 1
(03MBIF_v4) [UNIVERSIDADADIVU\msevillanogonzalez@a-1keohkce3uhb4 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:

Query= Seq_problema	Length=266	Score	E	Value
		(Bits)		
Sequences producing significant alignments:				
oxacillin-hydrolyzing class D beta-lactamase OXA-29 [Legionellaceae]	525	0.0		
blaOXA-29_1_AJ400619	525	0.0		
blaOXA-18_1_U085514	184	2e-58		
ampH_2_H056946	184	2e-58		
ampH_1_A3276031	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_JX306689	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_AFRQ000000031	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_HM04633	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_K0089875	164	7e-51		
blaOXA-443_1_LC030178	164	9e-51		
blaOXA-45_1_AU519683	162	8e-50		
blaOXA-47_1_AY237838	157	4e-48		
blaOXA-392_1_AB9901044	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.

Sequences producing significant alignments									Download	Select columns	Show 100	?
<input checked="" type="checkbox"/> select all 100 sequences selected		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, c...	Fluoribacter gor...	525	525	100%	0.0	95.49%	898	NG_049586.1			
<input checked="" type="checkbox"/>	Legionella gormanii blaOXA-29 gene for OXA-29	Fluoribacter gor...	525	525	100%	0.0	95.49%	898	AJ400619.1			
<input checked="" type="checkbox"/>	Legionella pneumophila strain L10-023 plasmid _complete sequence	Legionella pneu...	524	524	100%	1e-166	95.49%	143694	CP011106.1			
<input checked="" type="checkbox"/>	Legionella anisa strain FDAARGOS_1480 plasmid unnamed3 _complete sequence	Legionella anisa	524	524	100%	1e-166	95.49%	143699	CP082855.1			
<input checked="" type="checkbox"/>	Legionella longbeachae strain B1445CHC plasmid pB1445CHC_150k _complete sequence	Legionella longb...	524	524	100%	1e-166	95.49%	150426	CP045305.1			
<input checked="" type="checkbox"/>	Legionella sp. PC997 plasmid pPC997_1 _complete sequence	Legionella sp. P...	524	524	100%	1e-166	95.49%	150431	CP059577.1			
<input checked="" type="checkbox"/>	Legionella longbeachae strain B3526CHC plasmid pB3526CHC_150k _complete sequence	Legionella longb...	524	524	100%	1e-166	95.49%	150432	CP042253.1			
<input checked="" type="checkbox"/>	Legionella saintheliensi strain LA01-117 plasmid plA01-117_150k _complete sequence	Legionella saint...	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 sequ...	Legionella pneu...	524	524	100%	1e-166	95.49%	150433	CP021284.1			
<input checked="" type="checkbox"/>	Legionella pneumophila subsp. pneumophila str. Lorraine plasmid pLELO _complete genome	Legionella pneu...	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 seq...	Legionella pneu...	472	472	100%	2e-148	85.71%	129880	CP021287.1			
<input checked="" type="checkbox"/>	Legionella pneumophila strain Lpm7613 genome assembly _plasmid: 2	Legionella pneu...	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 pneu...	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%.

Alignment view Pairwise Download

100 sequences selected ?

Download GenBank Graphics Next Previous Descriptions

Fluoribacter gormanii ATCC 33297T blaOXA gene for oxacillin-hydrolyzing class D beta-lactamase OXA-29, complete CDS
Sequence ID: [NG_049586.1](#) Length: 898 Number of Matches: 1

Range 1: 34 to 831 [GenBank](#) [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
525 bits(1352)	0.0	Compositional matrix adjust	266/266(100%)	266/266(100%)	0/266(0%)	+1
Query 1 MKKLSVLLWLTLYCGTIWAQSTCFLVQENOTVLKHEGKDCNKRFAPESTFKIALSLMGF 60						
MKKLSVLLWLTLYCGTIWAQSTCFLVQENOTVLKHEGKDCNKRFAPESTFKIALSLMGF 213						
Query 61 DSGILKDTLNPEWPYKKEYEYLNVWKYPHNPRTHIRDSCWVSYQVLTOQLGMTRFKNYV 120						
DSGILKDTLNPEWPYKKEYEYLNVWKYPHNPRTHIRDSCWVSYQVLTOQLGMTRFKNYV 393						
Sbjct 394 DAFHYGNQDISGDKGQNNGLTHswlssslaispsEQIOFLQKIVNKKLSVNPKAFTMTKD 573						
Query 121 DAFHYGNQDISGDKGQNNGLTHswlssslaispsEQIOFLQKIVNKKLSVNPKAFTMTKD 180						
Sbjct 394 DAFHYGNQDISGDKGQNNGLTHswlssslaispsEQIOFLQKIVNKKLSVNPKAFTMTKD 573						
Query 181 ILYIQELAGGWLYGKGTGNGROLTKDKSQKLSLQHGWFIGWIEKDRVITFTKHIADSKK 240						
Sbjct 574 ILYIQELAGGWLYGKGTGNGROLTKDKSQKLSLQHGWFIGWIEKDRVITFTKHIADSKK 753						
Query 241 HVTFASFRAKNETLNQLFYLINELEK 266						
Sbjct 754 HVTFASFRAKNETLNQLFYLINELEK 831						

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

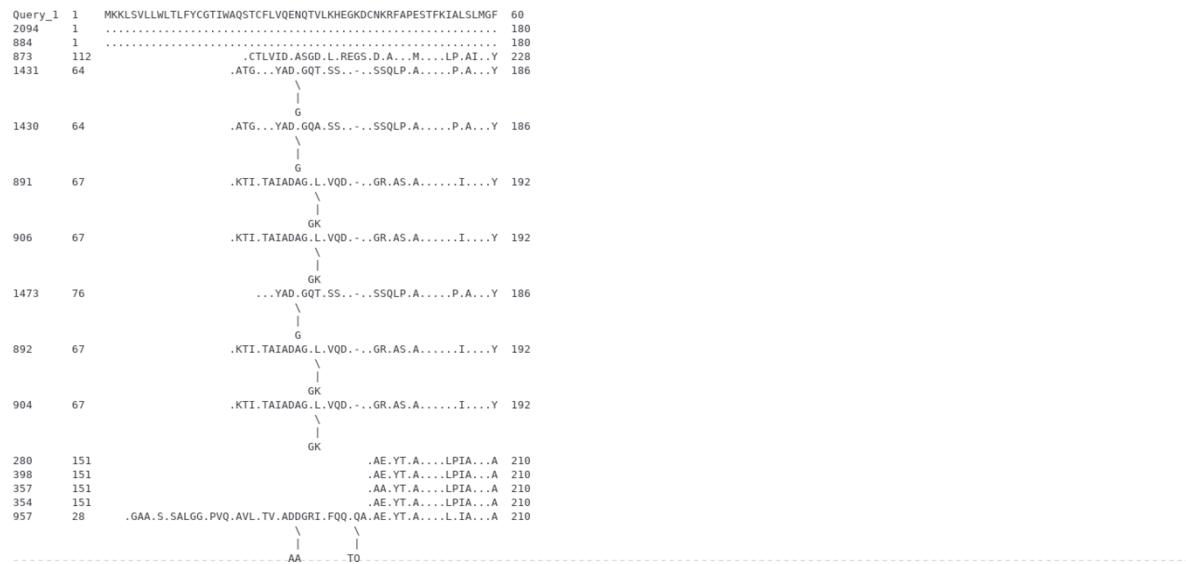


Imagen 2. Alineamiento en local.

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-1keohkce3uhb4 Descargas]$ muscle -in frataxin_mamiferos.fasta -out muscle_frataxina_results -htm
l

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 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-1keohkce3uhb4 Descargas]$
```

Imagen del resultado obtenido:

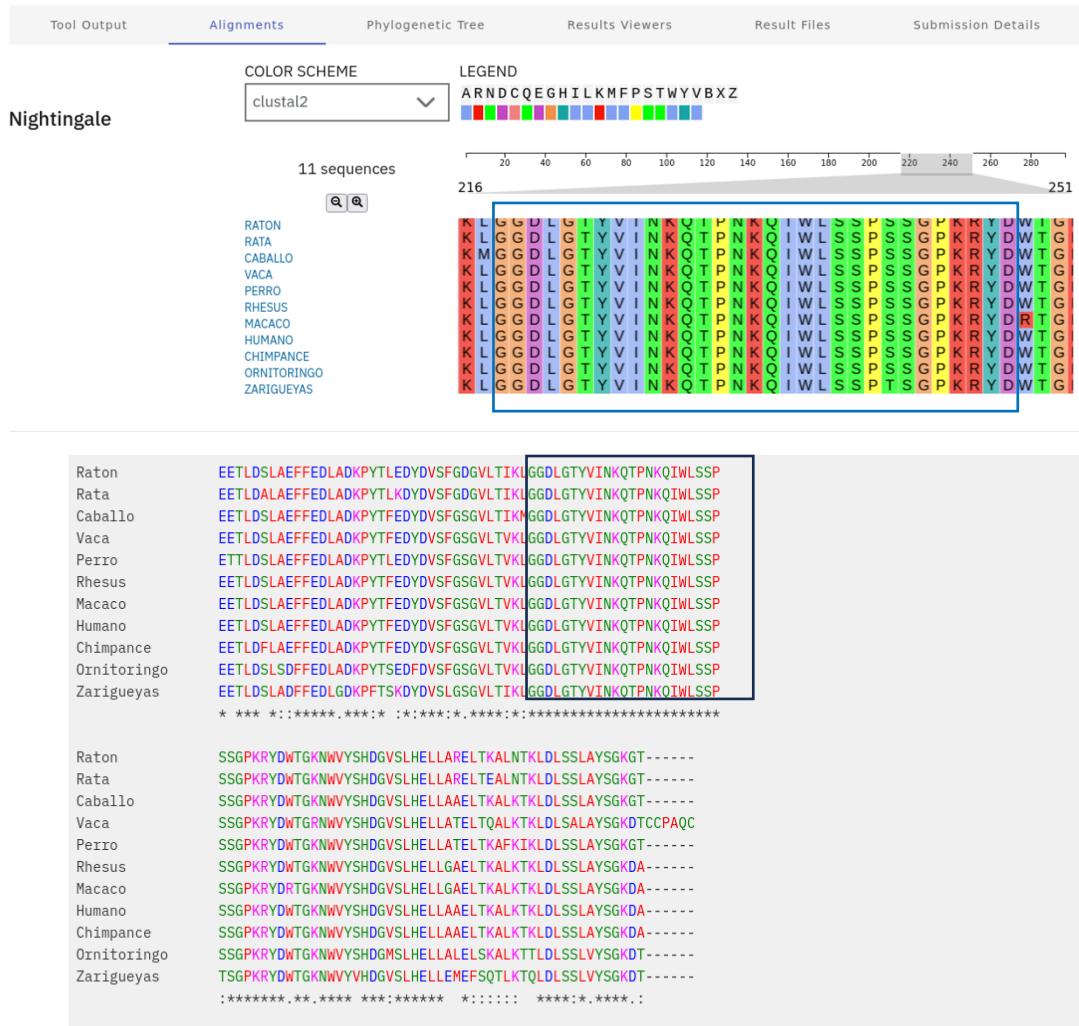
The screenshot displays a multiple sequence alignment of the frataxin gene across 11 mammalian species. The sequences are aligned vertically, and conserved regions are highlighted in green, while variable regions are highlighted in blue. The species listed are: Raton, Rata, Caballo, Vaca, Perro, Rhesus, Macaco, Humano, Chimpance, Ornitoringo, and Zarigueyas. The alignment shows a highly conserved start site (ATG) at the top, followed by a large variable region (blue) before the first conserved region (green). Subsequent conserved regions (green) are found throughout the sequence, with some additional variable regions (blue) interspersed.

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:

The screenshot shows a sequence alignment interface with the following data:

	Raton	Rata	Caballo	Vaca	Perro	Rhesus	Macaco	Humano	Chimpance
1	MWaFGgRAavGL-----LPRTASrAsAvgn----PRwrEpivtCGRRGL-----hvtV	MWTFGRRAaAGL-----LPRTASrAsAvwRn---PRgrErigtCGRRGL-----hvtA	MiyrspaAIsGLgerdvwrerAaPclArgRai---PsvsrlppspGKRGivrsasaGk	MWTlGRRsVAsf-----LPRSALPgfAptRaga--PRpAkdlLsGlpGL-----R1gt	MWTlGRRAaAGL-----LPRSAPPgaaAgagTRgP--traApLhGgRGL-----Rvgt	MWTFGRRAVAGL-----L---ASPSpAqAQTLTRAPRlAELaQlCsRRGL-----RTGI	MWTFGRRAVAGL-----L---ASPSpAqAQTLTRAPRlAELaQlCsRRGL-----RTGI	MWTFGRRAVAGL-----L---ASPSpAqAQTLTrvPrpAElaPlCGRRL-----RTdI	MWTFGRRAVAGL-----L---ASPSpAqAQTLTrvPrpAElaPlCGRRL-----RTGI
2	NAgATRHa--hlNLhyL-QilNiKKQSVcvhLRnlGldnPsSLDETaYERLAEETLDS	NAdairhs--hlNLhyLgQilNiKKQSVcvhLRnSGTLGnPSSLDETtaYERLAEETLDA	gtkgReggaSSlshflsQilNVKKQSVcvMhLrtTGTlgdPGSLDETTYERLAEETLDS	akapaRsqs-SlsLrcLnQtlDVKKQSVciNLRatGTLGdaGtLdTTYERLAEETLDS	gAargphsA-nlsLhhlnQlvNVKKQSVCLMNmRtvGtvsPGSLDETTYERLAEETLDS	NATCtHht-SSNLRglNQIrNVkrQSVyLMNLRKSGTlGHPGSLddTTYERLAEETLDS	NATRtHht-SSNLRglNQIrNVkrQSVyLMNLRKSGTlGHPGSLddTTYERLAEETLDS	dATCtprrA-SSNqRglNQiwNVKKQSVyLMNLRKSGTlGHPGSLDETTYERLAEETLDS	dATCtprrA-SSNLRglNQiwNVKKQSVyLMNLRKSGTlGHPGllgsnpyERLAEETLdf
3	LAEFFEDLADKPYTLEDYDVSFGGGVLT1KLGGDLGTYVINKQTPNKQIWLSPPSSGPKR	LAEFFEDLADKPYTLDKYDVSFGGGVLT1KLGGDLGTYVINKQTPNKQIWLSPPSSGPKR	LAEFFEDLADKPYTLEDYDVSFGSGVLT1KmGGDLGTYVINKQTPNKQIWLSPPSSGPKR	LAEFFEDLADKPYTLEDYDVSFGSGVLTVKLGGDLGTYVINKQTPNKQIWLSPPSSGPKR	LAEFFEDLADKPYTLEDYDVSFGSGVLTVKLGGDLGTYVINKQTPNKQIWLSPPSSGPKR	LAEFFEDLADKPYTLEDYDVSFGSGVLTVKLGGDLGTYVINKQTPNKQIWLSPPSSGPKR	LAEFFEDLADKPYTLEDYDVSFGSGVLTVKLGGDLGTYVINKQTPNKQIWLSPPSSGPKR	LAEFFEDLADKPYTLEDYDVSFGSGVLTVKLGGDLGTYVINKQTPNKQIWLSPPSSGPKR	LAEFFEDLADKPYTLEDYDVSFGSGVLTVKLGGDLGTYVINKQTPNKQIWLSPPSSGPKR
4	YDWTKGNWVYSHDGVLShELLArELTKALnTKLDSLSSLAGSGKgT-----	YDWTKGNWVYSHDGVLShELLArELTKALnTKLDSLSSLAGSGKgT-----	YDWTKGNWVYSHDGVLShELLAELTKALKLDSLSSLAGSGKgT-----						

The screenshot shows a sequence alignment interface with the following data:

	Raton	Rata	Caballo	Vaca	Perro	Rhesus	Macaco	Humano	Chimpance
1	MWAFGGAAVGL-----LPRTASrAsAWgn----PRwrEpivtCGRRGL-----hvtV	MWTFGRRAaAGL-----LPRTASrAsAvwRn---PRgrErigtCGRRGL-----hvtA	MiyrspaAIsGLgerdvwrerAaPclArgRai---PsvsrlppspGKRGivrsasaGk	MWTlGRRsVAsf-----LPRSALPgfAptRaga--PRpAkdlLsGlpGL-----R1gt	MWTlGRRAaAGL-----LPRSAPPgaaAgagTRgP--traApLhGgRGL-----Rvgt	MWTFGRRAVAGL-----L---ASPSpAqAQTLTRAPRlAELaQlCsRRGL-----RTGI	MWTFGRRAVAGL-----L---ASPSpAqAQTLTRAPRlAELaQlCsRRGL-----RTGI	MWTFGRRAVAGL-----L---ASPSpAqAQTLTrvPrpAElaPlCGRRL-----RTdI	MWTFGRRAVAGL-----L---ASPSpAqAQTLTrvPrpAElaPlCGRRL-----RTGI
2	NAGATRHa--hlNLhyL-QilNiKKQSVcvhLRnlGldnPsSLDETaYERLAEETLDS	NAdairhs--hlNLhyLgQilNiKKQSVcvhLRnSGTLGnPSSLDETtaYERLAEETLDA	gtkgReggaSSlshflsQilNVKKQSVcvMhLrtTGTlgdPGSLDETTYERLAEETLDS	akapaRsqs-SlsLrcLnQtlDVKKQSVciNLRatGTLGdaGtLdTTYERLAEETLDS	gAargphsA-nlsLhhlnQlvNVKKQSVCLMNmRtvGtvsPGSLDETTYERLAEETLDS	NATCtHht-SSNLRglNQIrNVkrQSVyLMNLRKSGTlGHPGSLddTTYERLAEETLDS	NATRtHht-SSNLRglNQIrNVkrQSVyLMNLRKSGTlGHPGSLddTTYERLAEETLDS	dATCtprrA-SSNqRglNQiwNVKKQSVyLMNLRKSGTlGHPGSLDETTYERLAEETLDS	dATCtprrA-SSNLRglNQiwNVKKQSVyLMNLRKSGTlGHPGllgsnpyERLAEETLdf
3	LAEFFEDLADKPYTLEDYDVSFGGGVLT1KLGGDLGTYVINKQTPNKQIWLSPPSSGPKR	LAEFFEDLADKPYTLDKYDVSFGGGVLT1KLGGDLGTYVINKQTPNKQIWLSPPSSGPKR	LAEFFEDLADKPYTLEDYDVSFGSGVLT1KmGGDLGTYVINKQTPNKQIWLSPPSSGPKR	LAEFFEDLADKPYTLEDYDVSFGSGVLTVKLGGDLGTYVINKQTPNKQIWLSPPSSGPKR	LAEFFEDLADKPYTLEDYDVSFGSGVLTVKLGGDLGTYVINKQTPNKQIWLSPPSSGPKR	LAEFFEDLADKPYTLEDYDVSFGSGVLTVKLGGDLGTYVINKQTPNKQIWLSPPSSGPKR	LAEFFEDLADKPYTLEDYDVSFGSGVLTVKLGGDLGTYVINKQTPNKQIWLSPPSSGPKR	LAEFFEDLADKPYTLEDYDVSFGSGVLTVKLGGDLGTYVINKQTPNKQIWLSPPSSGPKR	LAEFFEDLADKPYTLEDYDVSFGSGVLTVKLGGDLGTYVINKQTPNKQIWLSPPSSGPKR
4	YDWTKGNWVYSHDGVLShELLArELTKALnTKLDSLSSLAGSGKgT-----	YDWTKGNWVYSHDGVLShELLArELTKALnTKLDSLSSLAGSGKgT-----	YDWTKGNWVYSHDGVLShELLAELTKALKLDSLSSLAGSGKgT-----						

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

Utilizando el archivo de Illumina_MySeq.fastq:

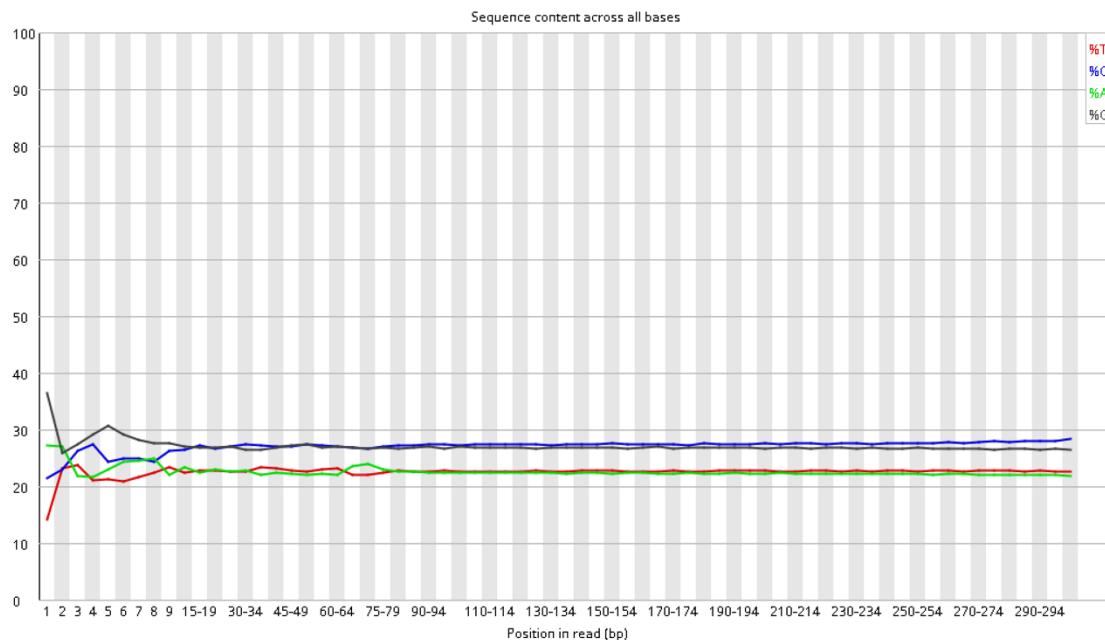
2.3.1 Realiza un primer análisis de calidad con la herramienta fastqc. Proporciona el comando utilizado y una captura de pantalla de la sección del informe que muestra “*Per base sequence content*”.

Comando: fastqc Illumina_MySeq.fastq

Captura:

```
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

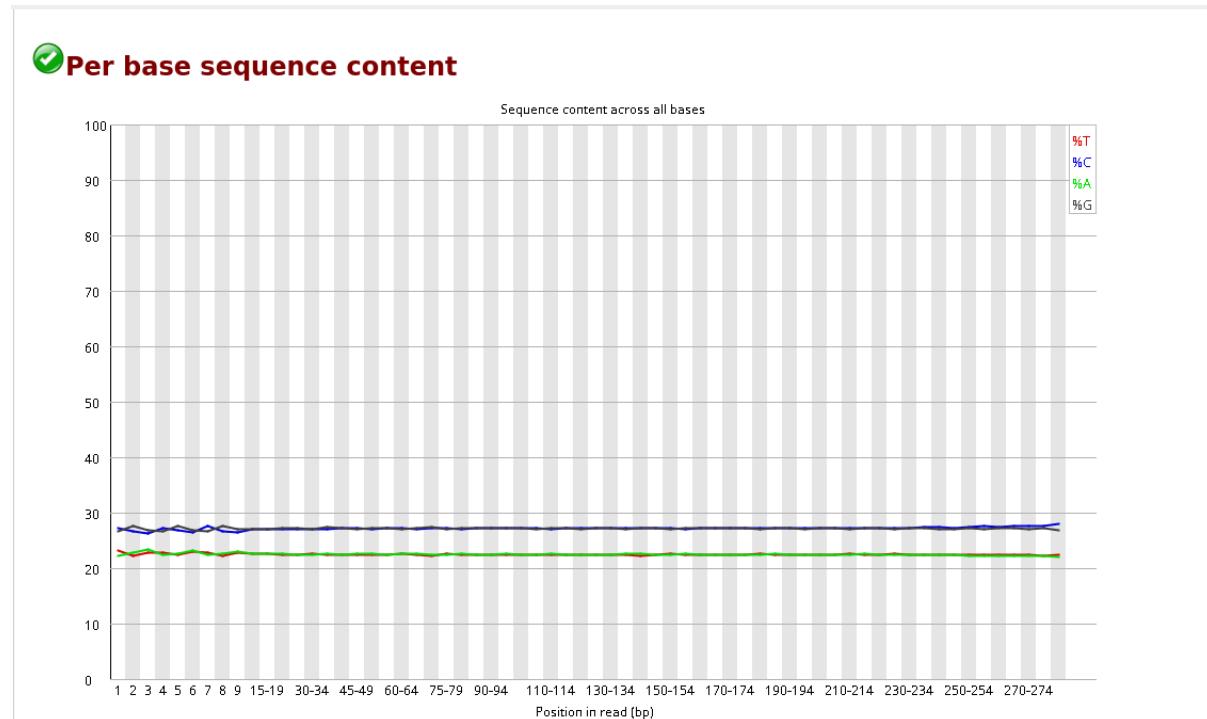
Y fastqc del nuevo archivo: fastqc Illumina_MySeq_prinseq_good_DHXy_prinseq_good_Kw5Q.fastq

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-1keohkce3uhb4 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-1keohkce3uhb4 Descargas]$
```

Per base sequence content

