Diego Vázquez Zambrano

Ejercicio 1. Estrategias de alineamiento [50%]

1. El programa CLUSTAL realiza alineamientos globales de dos o más secuencias. Conectaos al servidor implementado en el EBI para comparar la secuencia CDS del gen *TMEM106B* obtenida desde RefSeq (UCSC) para humano y ratón en la PEC1 anterior (hg38 y mm10, respectivamente).

http://www.ebi.ac.uk/Tools/msa/clustalo/

```
CCDS5358.1      ATGGGAAAGTCTCTTTCTCATTTGCCTTTGCATTCAAGCAAAGAAGATGCTTATGATGGA   60
CCDS19914.1     ATGGGAAAGTCTCTTTCTCACTTACCTTTGCATTCAAATAAAGAAGATGGCTATGATGGC   60
                ******************* ** ************  **********  ********

CCDS5358.1      GTCACATCT---GAAAACATGAGGAATGGACTGGTTAATAGTGAAGTCCATAATGAAGAT   117
CCDS19914.1     GTTACATCGACAGACAATATGAGAAATGGATTGGTTAGCAGTGAAGTGCACAACGAAGAC   120
                ** *****    ** ** ***** ****** ******  ******** ** ** *****

CCDS5358.1      GGAAGAAATGGAGATGTCTCTCAGTTTCCATATGTGGAATTTACAGGAAGAGATAGTGTC   177
CCDS19914.1     GGAAGAAATGGAGATGTCTCTCAGTTCCCATATGTGGAATTTACTGGAAGAGATAGTGTC   180
                ************************** ***************** ***************

CCDS5358.1      ACCTGCCCTACTTGTCAGGGAACAGGAAGAATTCCTAGGGGGCAAGAAAACCAACTGGTG   237
CCDS19914.1     ACTTGTCCCACTTGCCAAGGAACAGGAAGAATTCCTAGGGGACAAGAAAACCAACTGGTG   240
                ** ** ** ***** ** ********************** ******************

CCDS5358.1      GCATTGATTCCATATAGTGATCAGAGATTAAGGCCAAGAAGAACAAAGCTGTATGTGATG   297
CCDS19914.1     GCATTGATTCCATATAGTGATCAGCGGTTACGGCCAAGAAGAACAAAGCTGTATGTGATG   300
                ************************ * *** ******************************

CCDS5358.1      GCTTCTGTGTTTGTCTGTCTACTCCTTTCTGGATTGGCTGTGTTTTTCCTTTTCCCTCGC   357
CCDS19914.1     GCGTCTGTGTTTGTCTGCCTGCTCCTGTCTGGATTGGCTGTGTTTTTTCTTTTCCCTCGA   360
                ** ************** ** ***** ******************** ***********

CCDS5358.1      TCTATCGACGTGAAATACATTGGTGTAAAATCAGCCTATGTCAGTTATGATGTTCAGAAG   417
CCDS19914.1     TCTATTGAGGTGAAGTACATTGGAGTAAAATCAGCCTATGTCAGCTACGACGCTGAAAAG   420
                ***** ** ***** ******** ******************** ** ** * * * ***

CCDS5358.1      CGTACAATTTATTTAAATATCACAAACACACTAAATATAACAAACAATAACTATTACTCT   477
CCDS19914.1     CGAACCATATATTTAAATATCACGAACACACTAAATATAACAAATAATAACTATTATTCT   480
                ** ** ** ************** ******************** ********** ***

CCDS5358.1      GTCGAAGTTGAAAACATCACTGCCCAAGTTCAATTTTCAAAAACAGTTATTGGAAAGGCA   537
CCDS19914.1     GTTGAAGTTGAAAACATCACTGCTCAAGTCCAGTTTTCAAAAACCGTGATTGGAAAGGCT   540
                ** ******************** ***** ** ********** ** **********

CCDS5358.1      CGCTTAAACAACATAACCATTATTGGTCCACTTGATATGAAACAAATTGATTACACAGTA   597
CCDS19914.1     CGTTTAAACAACATAACTAACATTGGCCCACTTGATATGAAGCAGATTGATTATACGGTA   600
                ** ************** *  ***** ************** ** ******** ** ***

CCDS5358.1      CCTACCGTTATAGCAGAGGAAATGAGTTATATGTATGATTTCTGTACTCTGATATCCATC   657
CCDS19914.1     CCCACAGTTATTGCAGAGGAAATGAGTTACATGTATGATTTCTGTACACTGCTCTCCATC   660
                ** ** ***** ***************** ***************** *** * ******

CCDS5358.1      AAAGTGCATAACATAGTACTCATGATGCAAGTTACTGTGACAACAACATACTTTGGCCAC   717
CCDS19914.1     AAAGTGCACAACATAGTACTCATGATGCAAGTTACTGTAACAACAGCATACTTTGGACAC   720
                ******** ****************************** ****** ********** ***

CCDS5358.1      TCTGAACAGATATCCCAGGAGAGGTATCAGTATGTCGACTGTGGAAGAAACACAACTTAT   777
CCDS19914.1     TCTGAGCAGATATCTCAGGAAAGGTACCAGTATGTCGACTGTGGAAGGAACACGACTTAC   780
                ***** ******** ***** ***** ******************** ***** *****

CCDS5358.1      CAGTTGGGGCAGTCTGAATATTTAAATGTACTTCAGCCACAACAGTAA       825
CCDS19914.1     CAGTTGGCCCAGTCTGAGTATCTAAATGTCCTTCAGCCACAACAATAA       828
                *******  ******** *** ******* ************** ***
```

```
Percent Identity  Matrix - created by Clustal2.1


   1: CCDS5358.1   100.00   88.85
   2: CCDS19914.1   88.85  100.00
```

Ya que son dos especies muy cercanas, era de esperar un porcentaje de identidad tan alto

2. Repetid este mismo alineamiento global, utilizando ahora las respectivas proteínas de este gen en cada especie (que previamente debéis volver a recuperar de la entrada de RefSeq). Valorad el grado de homología entre estas dos secuencias.

```
CCDS5358.1_prot       MGKSLSHLPLHSSKEDAYDGVTS-ENMRNGLVNSEVHNEDGRNGDVSQFPYVEFTGRDSV      59
CCDS19914.1_prot      MGKSLSHLPLHSNKEDGYDGVTSTDNMRNGLVSSEVHNEDGRNGDVSQFPYVEFTGRDSV      60
                      ***********.***.****** :*******.*************************

CCDS5358.1_prot       TCPTCQGTGRIPRGQENQLVALIPYSDQRLRPRRTKLYVMASVFVCLLLSGLAVFFLFPR      119
CCDS19914.1_prot      TCPTCQGTGRIPRGQENQLVALIPYSDQRLRPRRTKLYVMASVFVCLLLSGLAVFFLFPR      120
                      ************************************************************

CCDS5358.1_prot       SIDVKYIGVKSAYVSYDVQKRTIYLNITNTLNITNNNYYSVEVENITAQVQFSKTVIGKA      179
CCDS19914.1_prot      SIEVKYIGVKSAYVSYDAEKRTIYLNITNTLNITNNNYYSVEVENITAQVQFSKTVIGKA      180
                      **:**************.:****************************************

CCDS5358.1_prot       RLNNITIIGPLDMKQIDYTVPTVIAEEMSYMYDFCTLISIKVHNIVLMMQVTVTTTYFGH      239
CCDS19914.1_prot      RLNNITNIGPLDMKQIDYTVPTVIAEEMSYMYDFCTLLSIKVHNIVLMMQVTVTTAYFGH      240
                      ****** ****************************:***************:****

CCDS5358.1_prot       SEQISQERYQYVDCGRNTTYQLGQSEYLNVLQPQQ       274
CCDS19914.1_prot      SEQISQERYQYVDCGRNTTYQLAQSEYLNVLQPQQ       275
                      ********************.************
```

```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#

     1: CCDS5358.1_prot   100.00   95.99
     2: CCDS19914.1_prot   95.99  100.00
```

El grado de homología de estas dos secuencias es del 95.99% como se observa en esta matriz.

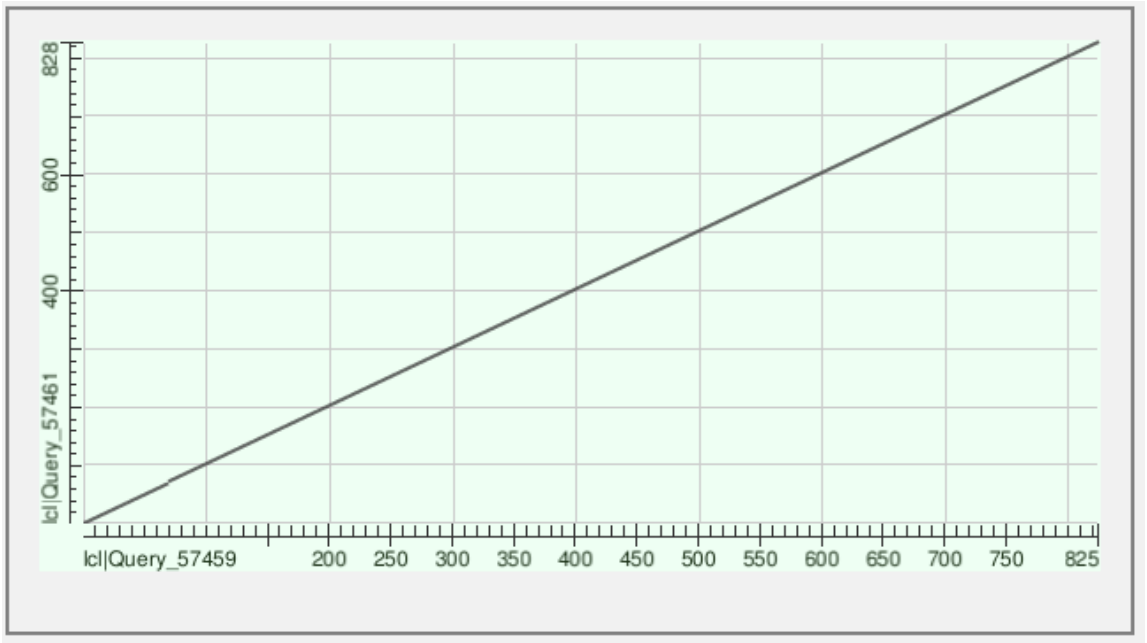
Una homología del 95.99% entre la secuencia hg38 (genoma humano versión 38) y la secuencia del mus musculus (ratón) indica que hay una alta similitud entre las dos secuencias. Esto sugiere que hay una relación evolutiva cercana entre los dos organismos, lo que no es sorprendente ya que los humanos y los ratones comparten un ancestro común. Además, esta alta homología sugiere que muchos de los genes y proteínas en los humanos y los ratones son similares o idénticos, lo que puede ser útil en la investigación biomédica y en la comprensión de las enfermedades humanas. Sin embargo, también es importante tener en cuenta que hay diferencias significativas entre los genomas de los humanos y los ratones, y que estas diferencias pueden tener implicaciones importantes en la investigación y el tratamiento de enfermedades.

3. El programa BLAST realiza alineamientos locales. Conectaos a BLAST, en el servidor principal del NCBI, para buscar qué versión de este programa debéis utilizar para alinear dos secuencias. Realizad ahora el alineamiento local de las dos regiones CDS del gen *TMEM106B*.

Tenemos que utilizar el programa BLASTN para llevar a cabo la alineación. Ello sin contrastar la base de datos de secuencias completas. Hay que indicarlo explícitamente ya que por defecto te pone otra opción.

http://www.ncbi.nlm.nih.gov/blast/

Se observa gran similitud entre ambas secuencias según el “Dot Plot”.



Se observa una identidad del 89%. La mayoría de la secuencia CDS es prácticamente idéntica.

Sequence ID: Query_57461 Length: 828 Number of Matches: 1

Range 1: 1 to 828 Graphics ▼ Next Match

Score	Expect	Identities	Gaps	Strand
1064 bits(1179)	0.0	733/828(89%)	3/828(0%)	Plus/Plus

```
Query  1    ATGGGAAAGTCTCTTTCTCATTTGCCTTTGCATTCAAGCAAAGAAGATGCTTATGATGGA  60
            ||||||||||||||||||| || ||||||||||||   ||||||||  ||||||||
Sbjct  1    ATGGGAAAGTCTCTTTCTCACTTACCTTTGCATTCAAATAAAGAAGATGGCTATGATGGC  60

Query  61   GTCACATCT---GAAAACATGAGGAATGGACTGGTTAATAGTGAAGTCCATAATGAAGAT  117
            || |||||    || || ||||| ||||||  ||||||   ||||||| || || |||||
Sbjct  61   GTTACATCGACAGACAATATGAGAAATGGATTGGTTAGCAGTGAAGTGCACAACGAAGAC  120

Query  118  GGAAGAAATGGAGATGTCTCTCAGTTTCCATATGTGGAATTTACAGGAAGAGATAGTGTC  177
            |||||||||||||||||||||||| ||||||||||||||| ||||||||||||||||
Sbjct  121  GGAAGAAATGGAGATGTCTCTCAGTTCCCATATGTGGAATTTACTGGAAGAGATAGTGTC  180

Query  178  ACCTGCCCTACTTGTCAGGGAACAGGAAGAATTCCTAGGGGGCAAGAAAACCAACTGGTG  237
            || || || ||||| || |||||||||||||||||||||| ||||||||||||||||||
Sbjct  181  ACTTGTCCCACTTGCCAAGGAACAGGAAGAATTCCTAGGGGACAAGAAAACCAACTGGTG  240

Query  238  GCATTGATTCCATATAGTGATCAGAGATTAAGGCCAAGAAGAACAAAGCTGTATGTGATG  297
            ||||||||||||||||||||||| | ||| ||||||||||||||||||||||||||||
Sbjct  241  GCATTGATTCCATATAGTGATCAGCGGTTACGGCCAAGAAGAACAAAGCTGTATGTGATG  300

Query  298  GCTTCTGTGTTTGTCTGTCTACTCCTTTCTGGATTGGCTGTGTTTTTCCTTTTCCCTCGC  357
            || |||||||||||||| || ||||| |||||||||||||||||||| ||||||||||||
Sbjct  301  GCGTCTGTGTTTGTCTGCCTGCTCCTGTCTGGATTGGCTGTGTTTTTTCTTTTCCCTCGA  360

Query  358  TCTATCGACGTGAAATACATTGGTGTAAAATCAGCCTATGTCAGTTATGATGTTCAGAAG  417
            ||||| || ||||| |||||||| |||||||||||||||||||| || || | | | |||
Sbjct  361  TCTATTGAGGTGAAGTACATTGGAGTAAAATCAGCCTATGTCAGCTACGACGCTGAAAAG  420
```

4. Ahora utilizad el servidor de CLUSTAL para alinear globalmente la secuencia *genomicA.txt* y la secuencia *genomicB.txt* que encontraréis adjuntas a este enunciado.

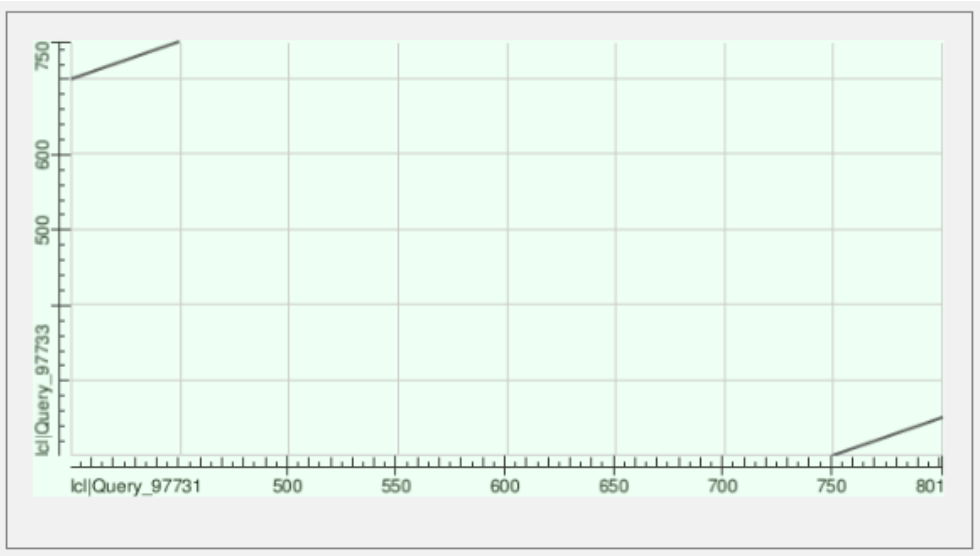
Estas dos secuencias presentan un bajo nivel de similitud (55.04% de identidad), con numerosos gaps entre bloques de conservación mínima.

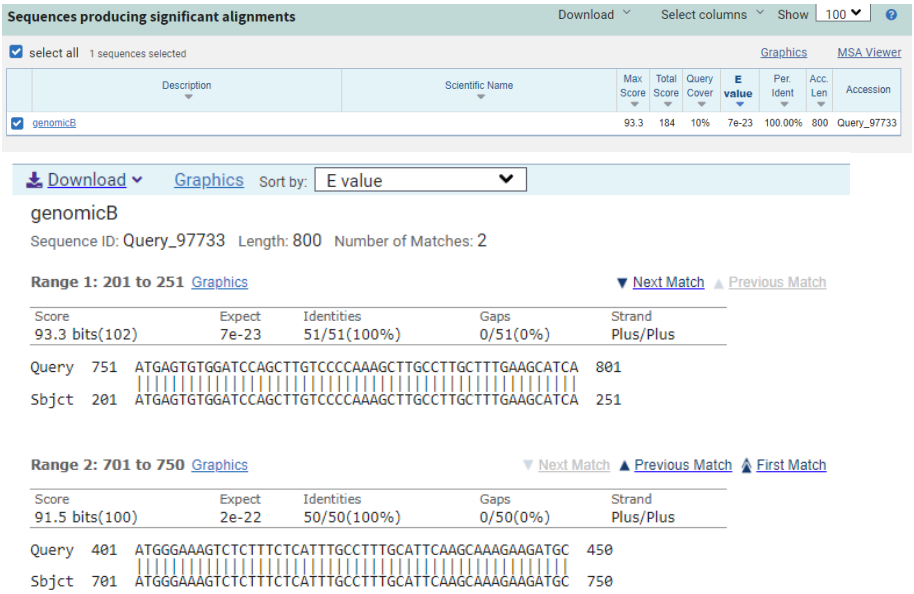
```
genomicA      cagaagaattgcttgaaccagggaggtggaggttgcagtgagcagaga----tcacgcca      56
genomicB      ----------------gctgggatgtgg--ggagcagtgttctgaggctgagcaggaca      41
                             * **** ****  *  ******  * ***      ** * **

genomicA      ctgcactcctgcttaagtgacagagtgagact-ccatctcaaaaaaaaaaaaaaaattcc      115
genomicB      gtga----------------ggccttgggcctggcctctgaaacca------ttttttcc      79
               **                    ** * **  * *** ***  *          ****

genomicA      tat------tatgtgcttgagtaataccacccactctggcaaatcttaaaaaagctcttg      169
genomicB      acctaggcctctgagcctgtgtcctataacttattg--------------caggctgtta      125
                    * ** ** ** **  **  **  * *               * *** **

genomicA      gccgggtgcagt--ggctcatgcctgtaatccccagaagaattgcttgaaccagggaggt      227
genomicB      gaagcaggcagactactttctggatgctttgctgcttagaatttttt------------      173
              *  *   ****      *  **  **   * *     ******  **

genomicA      ggaggttgcagtgagcagagatcacgccactgcactcctgcttaagtgacagagtgagac      287
genomicB      -----------ctgccagatatcctaggtcatcac-----------tctATGAGTGTGGA      211
                       **** ***      *  ***           *    ***** *

genomicA      tccatctcaaaaaaaaaaaaaaaattcctattatgtgcttgagtaataccacccactctg      347
genomicB      TCCAGCTTGTCCCCAA-----AGCTTGCCTT-----GC----------------------      239
              **** **       **     *  ** *  *     **

genomicA      gcaaatcttaaaaaagctcttggccgggtgcagtggctcatgcctgtaatcccATGGGAA      407
genomicB      -------TTTGA-AGCATCatggg-aggagctgtctctaagatctctaaagtgactttga      290
                    **  * *   ** ***   ** ** **  ** *   ** ***    *     *

genomicA      AGTCTCTTTCTCATTTGCCTTTGCATTCAAGCAAAGAAGATGCagttccccatttctgtc      467
genomicB      ggccttttgctcattgtctt--ggatattagccct-----t--ggcacccttttagtcac      341
               * ** ** ******  * *  * **   ***       *   *  ***  **  *  *
```

```
Percent Identity  Matrix - created by Clustal2.1


  1: genomicA    100.00   55.04
  2: genomicB     55.04  100.00
```

5. Proceded ahora a efectuar el alineamiento local con BLAST de la secuencia genómica *genomicA.txt* y la secuencia *genomicB.txt* adjuntadas con el enunciado.

En este caso comparten dos únicos fragmentos según el DotPlot. Tiene un valor E bastante bajo por lo que no parece un juego del azar. De estos dos fragmentos se podría realizar un estudio biológicamente interesante.





6. Comparad los resultados del alineamiento global y local en los dos casos anteriores (2 CDSs o las secuencias *genomicA.txt* y *genomicB.txt)*. Decidid cuál de los dos programas probados es más adecuado para cada caso en función de la estrategia empleada.

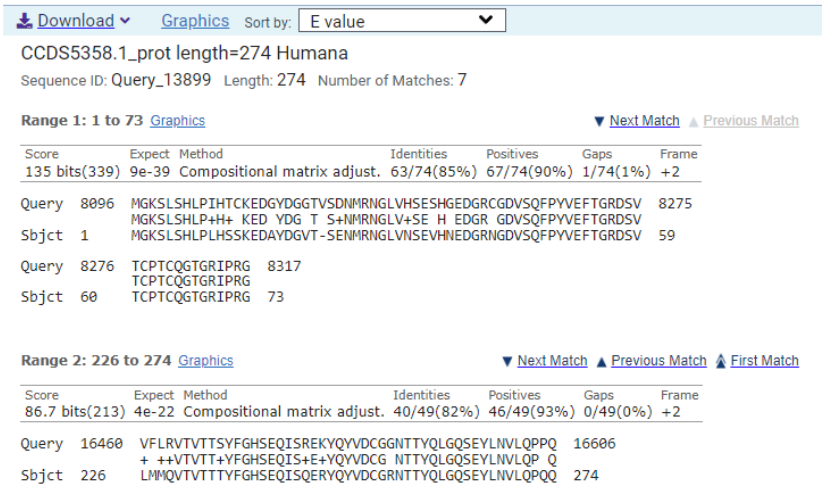
A partir de los resultados obtenidos, podemos concluir lo siguiente para cada caso:
En el caso de las CDSs, el alineamiento global es el más adecuado para posibles análisis de homología, ya que las dos secuencias son muy similares tanto en su tamaño como en su contenido. Si se realiza un alineamiento local, se obtendría un resultado similar, ya que no existen similitudes locales más fuertes que la tendencia general mostrada a lo largo de toda la secuencia. Por lo tanto, es lógico pensar que la comparación global es la más adecuada en este caso. Cuando se comparan regiones codificantes o proteínas homólogas, se debe emplear la estrategia global.
En el caso de las secuencias adjuntas, dado que un primer intento con la estrategia global no produce resultados significativos, se deben efectuar búsquedas locales para resaltar (en caso de que existan) aquellos motivos cortos conservados entre ambas secuencias. Con el programa BLASTN emergen claramente dos patrones de nucleótidos con posible relevancia biológica que no se pueden detectar utilizando una estrategia global. Por lo tanto, en este caso, es necesario utilizar una estrategia de búsqueda local para encontrar patrones cortos conservados entre las secuencias.

7. Unos investigadores que trabajan con el genoma del pollo (*chicken*) nos envían la secuencia adjunta *genomicC.txt*, pues sospechan que la forma ortóloga de nuestro gen *TMEM106B* está codificada en su interior. Decidid qué versión de BLAST debéis utilizar para validar esta hipótesis con la proteína humana (que tenéis de pasos previos), anotando su homóloga en esta región genómica de pollo. En caso de respuesta

afirmativa, interpretad el grado de homología resultante entre ambas proteínas.

Para relacionar un fragmento de ADN con una proteína, es necesario utilizar una herramienta llamada BLAST. Existen diferentes variantes de BLAST, y para este caso en particular, debe elegir la variante que permita traducir la secuencia de ADN a proteína, para luego comparar las dos proteínas resultantes. La variante adecuada para esta tarea es BLASTX. Sin embargo, si se está utilizando una proteína humana como consulta y una base de datos de secuencias genómicas, entonces la variante apropiada sería TBLASTN.

Se ha empleado BLASTX y hemos obtenido un 85% de identidad. Podemos, por tanto, decir que ambas son homólogas.



También se observa un porcentaje de similitud del 90%

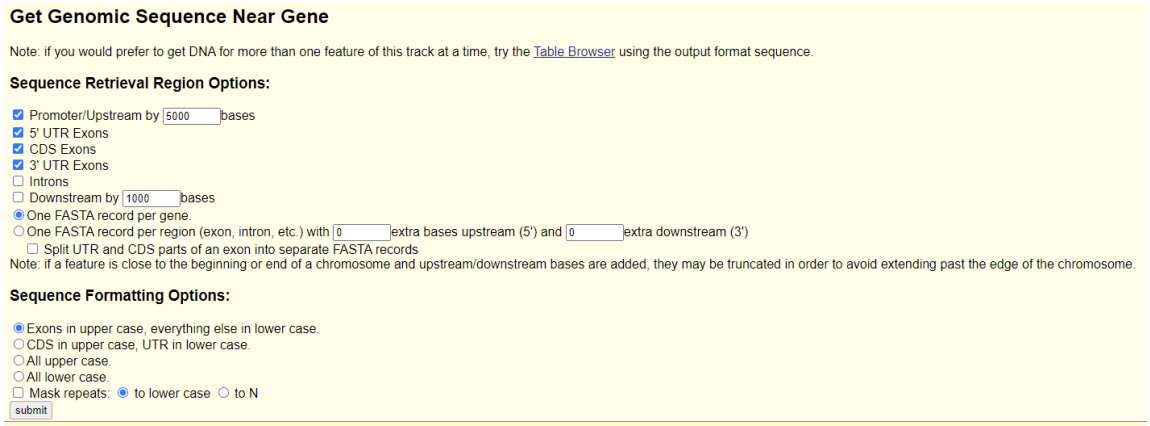
8. El programa MEME representa una familia alternativa de herramientas bioinformáticas para comparar secuencias. Definid en pocas palabras qué tipo de tarea realiza esta aplicación y cómo puede ser empleado dentro del área de estudio de la regulación génica mediante factores de transcripción:

http://meme-suite.org/

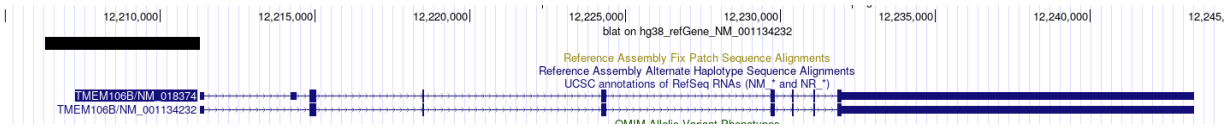
MEME es una herramienta bioinformática que se utiliza para identificar motivos de secuencia conservados en un conjunto de secuencias de ADN. Estos motivos pueden ser utilizados para predecir la presencia de factores de transcripción que se unen a ellos y regulan la expresión génica. En el área de estudio de la regulación génica mediante factores de transcripción, MEME puede ser empleado para identificar los motivos de secuencia que son reconocidos por los

factores de transcripción y, por lo tanto, ayudar a entender cómo se regula la expresión génica en diferentes condiciones y en diferentes tipos de células.

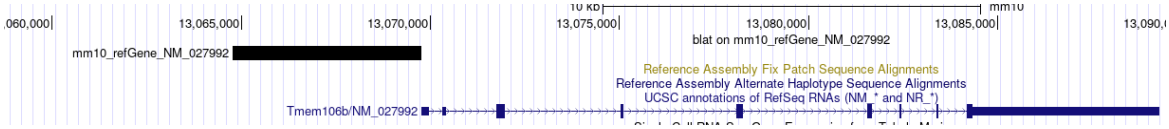
9. Vamos a estudiar la regulación transcripcional de nuestro gen *TMEM106B* a lo largo de la evolución. En primer lugar, empleando el navegador genómico de UCSC y las anotaciones de RefSeq, debéis extraer la región promotora del gen (seleccionad 5000 nucleótidos de longitud justo antes del inicio de transcripción del gen en cada especie) para estas especies: humano (hg38), ratón (mm10), rata (rn6) y pollo (galgal6).



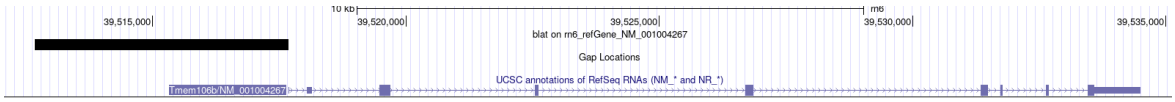
Humano



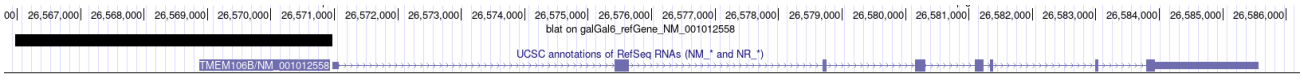
Ratón



Rata

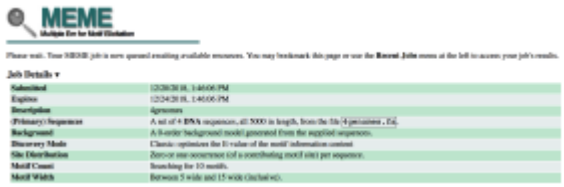


Pollo



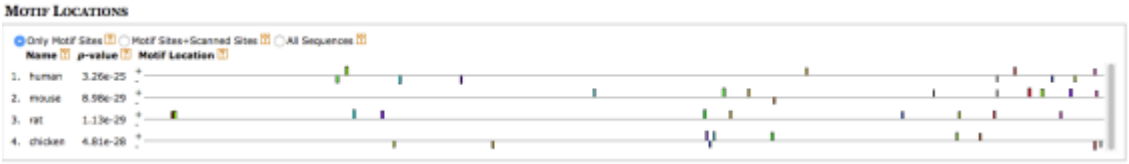
10. En segundo lugar, emplead el programa MEME para comparar esas cuatro secuencias ortólogas. Buscamos los 10 mejores motivos que posean una longitud entre 5 y 15 pares de bases. Explorad qué función puede jugar el programa TOMTOM integrado dentro de la *suite* de programas MEME y efectuad una prueba con alguno de los motivos identificados.

Efectuamos la comparación:

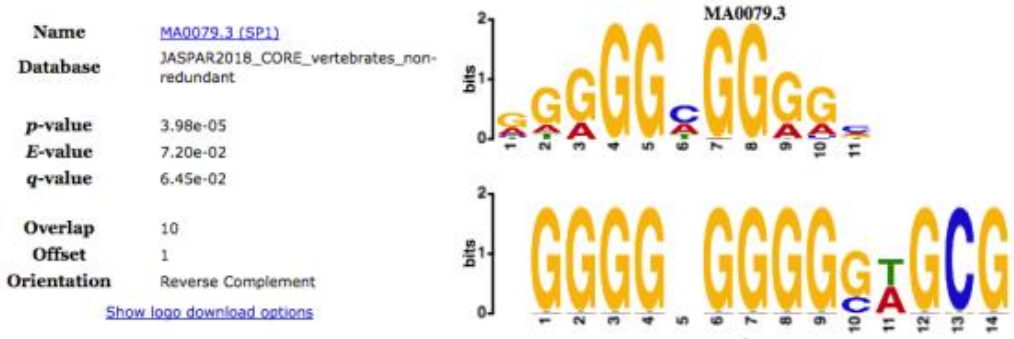


Obtenemos los resultados de los 10 mejores motivos de una longitud entre 5 y 10

Podemos explorar funciones como su distribución:



Gracias a TOMTOM podemos explorar las similitudes entre los motivos que hemos identificado y aquellos que son conocidos por su capacidad de unirse a factores de transcripción. Como ejemplo aquí está el primer motivo, el SP1



Ejercicio 2. Anotación computacional de genes [50%]

Estamos colaborando con un laboratorio de biología molecular que sospecha que la secuencia *anonima.fa* codifica un gen humano. Este fragmento genómico está representado en el formato FASTA habitual con una cabecera inicial y la secuencia a continuación (adjunto al enunciado):

```
>human
GCCGCGGCGCCTTTGTGACGCCATCAGCCCGCGCGCCGCCGCCGCCGCCT
TCTGTGCAGTCGCGGCCCGGGCGGACGGTGGCTGGCTGCTCCGCAGCGCT
CGGCTGGCTGCAGCGGCACCGCGGGTTGCGCGGCCGGGGATGCTCCAGCG
GGCGCGATGGCCCCCGCCATGCAGCCGGCCGAGATCCAATTTGCCCAGCG
```

1. Deseamos conocer las coordenadas de los exones que constituyen el gen codificado en esta secuencia. Como primer paso de nuestro protocolo de anotación, debéis utilizar el programa GENEID para recuperar el mejor gen identificado computacionalmente en esta región del genoma humano:

GENEID:
http://genome.crg.es/geneid.html

geneid predictions on sequence submitted from are:

```
## date Mon May 22 14:25:18 2023
## source-version: geneid v 1.2 -- geneid@imim.es
# Sequence human - Length = 37571 bps
# Optimal Gene Structure. 2 genes. Score = 31.87
# Gene 1 (Forward). 11 exons. 622 aa. Score = 31.58
   First      157      286       9.81   + 0 1    8.07    2.83   20.67    0.00   AA   1: 44 human_1
Internal    10376    10458       1.45   + 2 0    5.58    2.65    3.77    0.00   AA  44: 71 human_1
Internal    12800    12857       0.89   + 0 1    3.87    3.09    5.54    0.00   AA  72: 91 human_1
Internal    15504    15655      -0.00   + 2 0    0.91    4.65    5.41    0.00   AA  91:141 human_1
Internal    16764    16828       1.03   + 0 2    4.32    1.67    6.10    0.00   AA 142:163 human_1
Internal    17225    17406       5.73   + 1 1    3.69    3.72   15.71    0.00   AA 163:224 human_1
Internal    23771    23865      -1.35   + 2 0   -0.44    3.68    4.25    0.00   AA 224:255 human_1
Internal    25045    25142       2.96   + 0 2    3.54    0.05   14.52    0.00   AA 256:288 human_1
Internal    26262    26281       2.17   + 1 1    6.90    4.53    0.77    0.00   AA 288:295 human_1
Internal    27296    27427       2.70   + 2 1    0.45    5.26   10.70    0.00   AA 295:339 human_1
Terminal    28008    28858       6.20   + 2 0    4.56    0.00   21.14    0.00   AA 339:622 human_1

>human_1|geneid_v1.2_predicted_protein_1|622_AA
MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY
CMWVQDEPLLQEELANTIAQLVHAVNNSAAQACVWFFSRIKVFLDVLMKEVLCPESQSPN
GVRFHFIDIYLDELSKVGGKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVD
QSPFVPEETMEEQKTKVGDGDLSAEEIPENEVSLRRAVSKKKTALGKNHSRKDGLSDERG
RDDCGTFEDTGPLLQFDYKAVADRLLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQ
LSFAEDISADEDDQILSQGKHKKKGNKLLEKTNLEKEKGSRVFCVEEEDSESSLQKRRRK
KKKKHHLQPENPGPGGAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEH
PPAVPMHNKRKRPRKKSPRAHREMLESAVLPPEDMSQSGPSGSHPQGPRGSPTGGAQLLK
RKRKLGVVPVNGSGLSTPAWPPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLEL
CGLPSQKTASLKKRKKMRVMSNLVEHNGVLESEAGQPQALVRWEHPQASSPQRHSLASMG
LHCLLRGRVGAGGQASGLSSS*

# Gene 2 (Forward). 4 exons. 141 aa. Score = 0.28
   First    30518    30529      -2.92   + 0 0    1.42    1.23    1.21    0.00   AA   1:  4 human_2
Internal    30780    30932       0.68   + 0 0    2.81    3.31    5.01    0.00   AA   5: 55 human_2
Internal    31931    31994       2.93   + 0 1    4.88    4.80    5.31    0.00   AA  56: 77 human_2
Terminal    33682    33875      -0.40   + 2 0   -0.70    0.00   12.53    0.00   AA  77:141 human_2

>human_2|geneid_v1.2_predicted_protein_2|141_AA
MKIKGSSGTCSSLKKQKLRAESDFVKFDTPFLPKPLFFRRAKSSTATHPPGPAVQLNKTP
SSSKKVTFGLNRNMTAEFKKTDKSILVSPTGPSRVAFDPEQKPLHGVLKTPTSSPASSPL
VAKKPLTTTPRRRPRAMDFF*
```

Species:
Homo sapiens

Command:
`geneid -P /soft/GeneID/geneid_1.2/human.param /tmp/WebFiles/fastas/geneid25238.fasta`

Running time:
0.13 secs

2. Como segundo componente de nuestro *pipeline*, debéis emplear GENSCAN para recuperar el gen codificado internamente en esta secuencia humana:

GENSCAN:
http://hollywood.mit.edu/GENSCAN.html

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Init	+	157	286	130	0	1	107	80	324	0.752	33.81
1.02	Intr	+	10376	10458	83	0	2	94	92	26	0.829	2.96
1.03	Intr	+	12800	12857	58	1	1	97	99	62	0.963	6.66
1.04	Intr	+	14362	14447	86	2	2	47	95	49	0.678	1.04
1.05	Intr	+	15128	15189	62	1	2	53	86	51	0.694	-1.07
1.06	Intr	+	15526	15655	130	1	1	27	99	108	0.642	6.50
1.07	Intr	+	16764	16828	65	2	2	78	83	73	0.995	3.32
1.08	Intr	+	17225	17406	182	2	2	77	91	192	0.962	17.91
1.09	Intr	+	23771	23865	95	0	2	37	94	55	0.688	0.68
1.10	Intr	+	25045	25142	98	0	2	64	26	129	0.640	3.11
1.11	Intr	+	26262	26281	20	0	2	91	100	-1	0.600	-2.35
1.12	Intr	+	27296	27427	132	0	0	41	121	120	0.872	11.22
1.13	Intr	+	27663	27851	189	1	0	51	67	92	0.625	2.96
1.14	Intr	+	28008	28732	725	1	2	85	95	470	0.762	38.55
1.15	Intr	+	30236	30380	145	1	1	71	48	71	0.368	1.26
1.16	Intr	+	30589	30671	83	2	2	30	51	91	0.478	-1.04
1.17	Intr	+	30780	30932	153	2	0	100	101	109	0.999	13.67
1.18	Intr	+	31931	31994	64	1	1	114	131	52	0.996	10.39
1.19	Term	+	33682	33875	194	2	2	52	55	187	0.999	9.38

3. Finalmente, como tercer componente del proceso, utilizad el programa FGENESH para identificar también la predicción de este sistema:
FGENESH:
http://www.softberry.com/

```
FGENESH 2.6 Prediction of potential genes in Homo_sapiens genomic DNA
Time     :   Mon May 22 14:52:44 2023
Seq name: human
Length of sequence: 37571
Number of predicted genes 1: in +chain 1, in -chain 0.
Number of predicted exons 16: in +chain 16, in -chain 0.
Positions of predicted genes and exons: Variant   1 from   1, Score:115.993872
  G Str   Feature   Start        End    Score           ORF          Len
  1 +     1 CDSf       157 -       286   28.30       157 -       285     129
  1 +     2 CDSi     10376 -     10458    7.43     10378 -     10458      81
  1 +     3 CDSi     12800 -     12857    6.33     12800 -     12856      57
  1 +     4 CDSi     14362 -     14447    3.34     14364 -     14447      84
  1 +     5 CDSi     15128 -     15189    2.40     15128 -     15187      60
  1 +     6 CDSi     15526 -     15655    6.39     15527 -     15655     129
  1 +     7 CDSi     16764 -     16828    6.62     16764 -     16826      63
  1 +     8 CDSi     17225 -     17406   12.24     17226 -     17405     180
  1 +     9 CDSi     23771 -     23865    2.10     23773 -     23865      93
  1 +    10 CDSi     25045 -     25142    0.60     25045 -     25140      96
  1 +    11 CDSi     26262 -     26281   -1.21     26263 -     26280      18
  1 +    12 CDSi     27296 -     27427    8.29     27298 -     27426     129
  1 +    13 CDSi     28008 -     28732   33.29     28010 -     28732     723
  1 +    14 CDSi     30780 -     30932    9.89     30780 -     30932     153
  1 +    15 CDSi     31931 -     31994   13.04     31931 -     31993      63
  1 +    16 CDSl     33682 -     33875    1.90     33684 -     33875     192
  1 +       PolA     33923                -4.47
```

4. Para evaluar la coherencia de las predicciones obtenidas por cada programa, emplead CLUSTAL para comparar las proteínas reportadas por GENEID, GENSCAN y FGENESH. Realizad una primera interpretación de estos resultados en el contexto de este alineamiento global.

```
GENEID          MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY      60
GENSCAN         MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY      60
FGENESH:        MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY      60
                ************************************************************

GENEID          CMWVQDEPLLQEELANTIAQLVHAVNNSAAQAC--------------------------      93
GENSCAN         CMWVQDEPLLQEELANTIAQLVHAVNNSAAQHLFIQTFWQTMNREWKGIDRLRLDKYYML      120
FGENESH:        CMWVQDEPLLQEELANTIAQLVHAVNNSAAQHLFIQTFWQTMNREWKGIDRLRLDKYYML      120
                ********************************

GENEID          ---------------VWFFSRIKVFLDVLMKEVLCPESQSPNGVRFHFIDIYLDELSKVG      138
GENSCAN         IRLVLRQSFEVLKRNGWEESRIKVFLDVLMKEVLCPESQSPNGVRFHFIDIYLDELSKVG      180
FGENESH:        IRLVLRQSFEVLKRNGWEESRIKVFLDVLMKEVLCPESQSPNGVRFHFIDIYLDELSKVG      180
                        *  ****************************************

GENEID          GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVDQSPFVPEETMEEQKTKVG      198
GENSCAN         GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVDQSPFVPEETMEEQKTKVG      240
FGENESH:        GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVDQSPFVPEETMEEQKTKVG      240
                ************************************************************

GENEID          DGDLSAEEIPENEVSLRRAVSKKKTALGKNHSRKDGLSDERGRDDCGTFEDTGPLLQFDY      258
GENSCAN         DGDLSAEEIPENEVSLRRAVSKKKTALGKNHSRKDGLSDERGRDDCGTFEDTGPLLQFDY      300
FGENESH:        DGDLSAEEIPENEVSLRRAVSKKKTALGKNHSRKDGLSDERGRDDCGTFEDTGPLLQFDY      300
                ************************************************************

GENEID          KAVADRLLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQLSFAEDISADEDDQILSQ      318
GENSCAN         KAVADRLLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQLSFAEDISADEDDQILSQ      360
FGENESH:        KAVADRLLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQLSFAEDISADEDDQILSQ      360
                ************************************************************

GENEID          GKHKKKGNKLLEKTNLEKE-------------------------------------       337
GENSCAN         GKHKKKGNKLLEKTNLEKEKGKQELQGALGGGCLMTTRDLWFLPLSPKISGNGTISVPYV      420
FGENESH:        GKHKKKGNKLLEKTNLEKE-------------------------------------       379
                *******************

GENEID          ----------------------KGSRVFCVEEEDSESSLQKRRRKKKKKHHLQPENPGPG      375
GENSCAN         FINGQKEGFQSQLGMEEVGPDDKGSRVFCVEEEDSESSLQKRRRKKKKKHHLQPENPGPG      480
FGENESH:        ----------------------KGSRVFCVEEEDSESSLQKRRRKKKKKHHLQPENPGPG      417
                                       ************************************

GENEID          GAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK      435
GENSCAN         GAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK      540
FGENESH:        GAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK      477
                ************************************************************

GENEID          KSPRAHREMLESAVLPPEDMSQSGPSGSHPQGPRGSPTGGAQLLKRKRKLGVVPVNGSGL      495
GENSCAN         KSPRAHREMLESAVLPPEDMSQSGPSGSHPQGPRGSPTGGAQLLKRKRKLGVVPVNGSGL      600
FGENESH:        KSPRAHREMLESAVLPPEDMSQSGPSGSHPQGPRGSPTGGAQLLKRKRKLGVVPVNGSGL      537
                ************************************************************

GENEID          STPAWPPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRK      555
GENSCAN         STPAWPPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRK      660
FGENESH:        STPAWPPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRK      597
                ************************************************************

GENEID          KMRVMSNLVEHNGVLESEAGQPQALVRWEHPQA--------------SSPQRHSL-ASMG      600
GENSCAN         KMRVMSNLVEHNGVLESEAGQPQALAAHLNLPEPPVCRQRHWAAHTSESQVRDPVSLWVA      720
FGENESH:        KMRVMSNLVEHNGVLESEAGQPQAL-----------------------------------      622
                *************************

GENEID          LHCLLRGR--------------------------------VGAGGQASGLSSS-------      621
GENSCAN         VSCCTRNECPGPASVVLCVKPELCRMEGLSASAVRKTAGRRGSSGTCSSLKKQKLRAESD      780
FGENESH:        ----------------------------------------GSSGTCSSLKKQKLRAESD      641
                                                        *:.* .*.*...

GENEID          ------------------------------------------------------------      621
GENSCAN         FVKFDTPFLPKPLFFRRAKSSTATHPPGPAVQLNKTPSSSKKVTFGLNRNMTAEFKKTDK      840
FGENESH:        FVKFDTPFLPKPLFFRRAKSSTATHPPGPAVQLNKTPSSSKKVTFGLNRNMTAEFKKTDK      701

GENEID          ------------------------------------------------------- 621
GENSCAN         SILVSPTGPSRVAFDPEQKPLHGVLKTPTSSPASSPLVAKKPLTTTPRRRPRAMDFF 897
FGENESH:        SILVSPTGPSRVAFDPEQKPLHGVLKTPTSSPASSPLVAKKPLTTTPRRRPRAMDFF 758
```

Podemos obtener las proteínas anotadas por cada programa y guardarlas en un archivo en formato FASTA para alinearlas con el programa CLUSTAL. Si GENEID ha dividido la proteína en dos partes, podemos unirlas basándonos en las predicciones de los otros dos programas. Es importante identificar primero las regiones comunes detectadas por todos los programas y luego estudiar las áreas menos conservadas de la proteína resultante.
De los resultados obtenidos, podemos concluir que la mayoría de las predicciones son muy similares, aunque hay tres puntos de conflicto en la proteína. Es posible que GENSCAN haya recuperado un péptido más largo debido a un error de sobrepredicción. También observamos que la estrategia de combinar las dos proteínas reconocidas por GENEID funciona bien, aunque algunos exones pueden no haber sido detectados claramente debido a que el punto de unión no es el más adecuado desde un punto de vista lógico.

5. Finalmente, para comparar cuantitativamente los tres sistemas de predicción, rellenad la siguiente tabla con las coordenadas de todos los exones identificados dentro del mejor gen presentado por cada programa. Seleccionad dos de estos exones para realizar una búsqueda con BLASTP contra la base de datos completa de proteínas. Interpretad estos resultados para elaborar una primera anotación factible de este gen en función de estas predicciones:

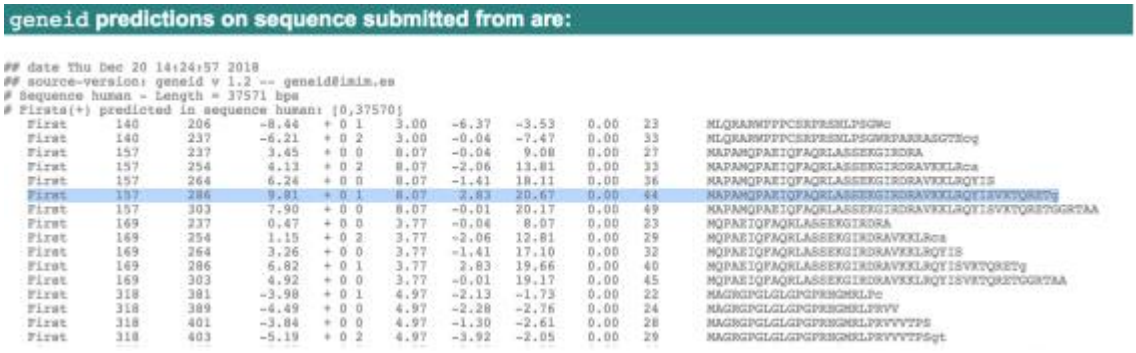
	GENEID	**GENSCAN**	**FGENESH**
Exón 157-286	**X**	**X**	**X**
Exón 10376-10458	**X**	**X**	**X**
Exón 12800-12857	**X**	**X**	**X**
Exón 14362-14447		**X**	**X**
Exón 15128-15189		**X**	**X**
Exón 15504-15655	**X**		
Exón 15526-15655		**X**	**X**
Exón 16724-16828	**X**	**X**	**X**
Exón 17225-17406	**X**	**X**	**X**
Exón 23771-23865	**X**	**X**	**X**
Exón 25045-25142	**X**	**X**	**X**
Exón 26262-26281	**X**	**X**	**X**
Exón 27296-27427	**X**	**X**	**X**
Exón 27663-27851		**X**	
Exón 28008-28858	**X**		
Exón 28008-28732		**X**	**X**

Exón 30236-30380		**X**	
Exón 30518-30529	**X**		
Exón 30780-30932	**X**	**X**	**X**
Exón 31931-31994	**X**	**X**	**X**
Exón 33682-33875	**X**	**X**	**X**

Cada fila de esta primera tabla representa un exón. En amarillo se encuentran los exones con cierto grado de solapamiento. Ahora podemos escoger uno de los exones identificados por los tres programas o selecciones aquellos encontrado por dos programas por lo menos

	GENEID	**GENSCAN**	**FGENESH**
Exón 157-286	**X**	**X**	**X**
Exón 10376-10458	**X**	**X**	**X**
Exón 12800-12857	**X**	**X**	**X**
Exón 14362-14447		**X**	**X**
Exón 15128-15189		**X**	**X**
Exón 15526-15655		**X**	**X**
Exón 16724-16828	**X**	**X**	**X**
Exón 17225-17406	**X**	**X**	**X**
Exón 23771-23865	**X**	**X**	**X**
Exón 25045-25142	**X**	**X**	**X**
Exón 26262-26281	**X**	**X**	**X**
Exón 27296-27427	**X**	**X**	**X**
Exón 28008-28732		**X**	**X**
Exón 30780-30932	**X**	**X**	**X**
Exón 31931-31994	**X**	**X**	**X**
Exón 33682-33875	**X**	**X**	**X**

Ahora podemos utilizar GENEID para buscar los exones



Subrayado en azul se encuentra el primer exón.
Ahora vamos a BLASTP e identificamos su gen:

RRP1B protein, partial [Homo sapiens]

Sequence ID: AAH14005.1 Length: 408 Number of Matches: 1

Range 1: 1 to 43 GenPept Graphics ▼ Next Match ▲ Previous Match

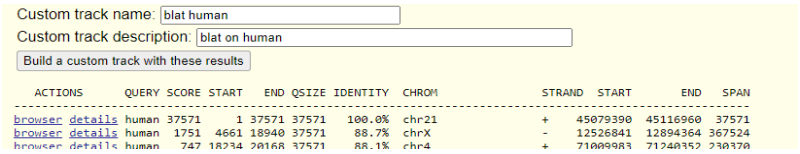
Score	Expect	Method	Identities	Positives	Gaps
88.6 bits(218)	2e-19	Compositional matrix adjust.	43/43(100%)	43/43(100%)	0/43(0%)

```
Query  1   MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRET  43
           MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRET
Sbjct  1   MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRET  43
```

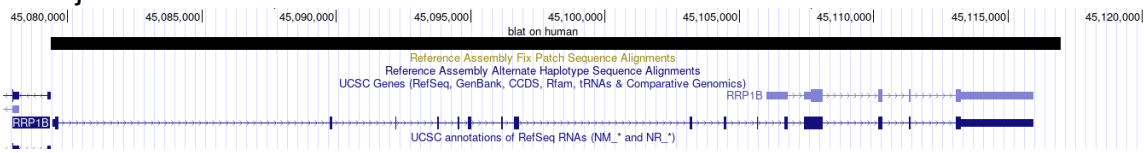
6. Aprovechad BLAT para identificar en qué parte del genoma humano se encuentra *anonima.fa* (cromosoma, inicio, final, hebra). Verificad visualmente que el inicio y el final de nuestra secuencia encajan con la región correcta.

BLAT: http://genome.ucsc.edu
http://genome.ucsc.edu/cgi-bin/hgBlat?command=start

Mostramos los primeros:



Encajan:



7. Convertid manualmente nuestras predicciones de GENEID, GENSCAN y FGENESH en formato GFF para visualizarlas como Custom tracks en UCSC (será necesario adaptar las coordenadas de los exones para trasladarlos sobre el cromosoma 21):

Información para crear una Custom track:
http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#CustomTracks

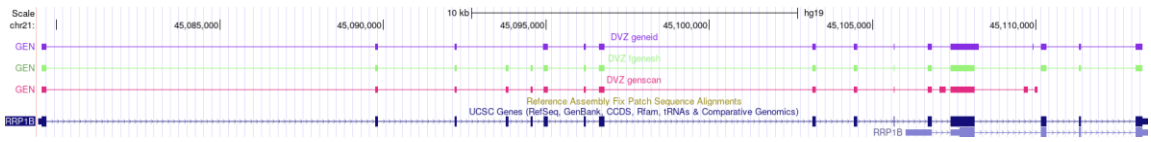
Información sobre el formato GFF:
http://genome.ucsc.edu/FAQ/FAQformat.html#format3

Debéis lograr un resultado similar a la siguiente pantalla (en vuestro caso, cambiad EBG por las iniciales de vuestro nombre y apellidos):

track	name="DVZ	geneid"	visibility=2	color=138,43,226				
chr21	geneid_v1.2	gen	45079546	45079675	9.81	.	.	GEN
chr21	geneid_v1.2	gen	45089765	45089847	1.45	.	.	GEN
chr21	geneid_v1.2	gen	45092189	45092246	0.89	.	.	GEN
chr21	geneid_v1.2	gen	45094893	45095044	-0.00	.	.	GEN
chr21	geneid_v1.2	gen	45096153	45096217	1.03	.	.	GEN
chr21	geneid_v1.2	gen	45096614	45096795	5.73	.	.	GEN
chr21	geneid_v1.2	gen	45103160	45103254	-1.35	.	.	GEN
chr21	geneid_v1.2	gen	45104434	45104531	2.96	.	.	GEN
chr21	geneid_v1.2	gen	45105651	45105670	2.17	.	.	GEN
chr21	geneid_v1.2	gen	45106685	45106816	2.70	.	.	GEN
chr21	geneid_v1.2	gen	45107397	45108247	6.20	.	.	GEN
chr21	geneid_v1.2	gen	45109907	45109918	-2.92	.	.	GEN
chr21	geneid_v1.2	gen	45110169	45110321	0.68	.	.	GEN
chr21	geneid_v1.2	gen	45111320	45111383	2.93	.	.	GEN
chr21	geneid_v1.2	gen	45113071	45113264	-0.40	.	.	GEN

track	name="DVZ	fgenesh"	visibility=2	color=138,243,126				
chr21	fgenesh	gen	45079546	45079675	28.30	.	.	GEN
chr21	fgenesh	gen	45089765	45089847	7.43	.	.	GEN
chr21	fgenesh	gen	45092189	45092246	6.33	.	.	GEN
chr21	fgenesh	gen	45093751	45093836	3.34	.	.	GEN
chr21	fgenesh	gen	45094517	45094578	2.40	.	.	GEN
chr21	fgenesh	gen	45094915	45095044	6.39	.	.	GEN
chr21	fgenesh	gen	45096153	45096217	6.62	.	.	GEN
chr21	fgenesh	gen	45096614	45096795	12.24	.	.	GEN
chr21	fgenesh	gen	45103160	45103254	2.10	.	.	GEN
chr21	fgenesh	gen	45104434	45104531	0.60	.	.	GEN
chr21	fgenesh	gen	45105651	45105670	-1.21	.	.	GEN
chr21	fgenesh	gen	45106685	45106816	8.29	.	.	GEN
chr21	fgenesh	gen	45107397	45108121	33.29	.	.	GEN
chr21	fgenesh	gen	45110169	45110321	9.89	.	.	GEN
chr21	fgenesh	gen	45111320	45111383	13.04	.	.	GEN
chr21	fgenesh	gen	45113071	45113264	1.90	.	.	GEN

track	name="DVZ	genscan"	visibility=2	color=238,43,126				
chr21	genscan	gen	45079546	45079675	33.81	.	.	GEN
chr21	genscan	gen	45089765	45089847	2.96	.	.	GEN
chr21	genscan	gen	45092189	45092246	6.66	.	.	GEN
chr21	genscan	gen	45093751	45093836	1.04	.	.	GEN
chr21	genscan	gen	45094517	45094578	-1.07	.	.	GEN
chr21	genscan	gen	45094915	45095044	6.50	.	.	GEN
chr21	genscan	gen	45096153	45096217	3.32	.	.	GEN
chr21	genscan	gen	45096614	45096795	17.91	.	.	GEN
chr21	genscan	gen	45103160	45103254	0.68	.	.	GEN
chr21	genscan	gen	45104434	45104531	3.11	.	.	GEN
chr21	genscan	gen	45105651	45105670	-2.35	.	.	GEN
chr21	genscan	gen	45106685	45106816	11.22	.	.	GEN
chr21	genscan	gen	45107052	45107240	2.96	.	.	GEN
chr21	genscan	gen	45107397	45108121	38.55	.	.	GEN
chr21	genscan	gen	45109625	45109769	1.26	.	.	GEN
chr21	genscan	gen	45109978	45110060	-1.04	.	.	GEN



Con esta imagen podemos observar que es idéntica gráficamente a la anotación real de RefSeq

8. Emplead el Table Browser de UCSC para calcular la correlación, dentro de la región genómica delimitada por la secuencia *anonima.fa*, entre las predicciones de (a) GENEID y GENSCAN, (b) GENEID y FGENESH, (c) GENSCAN y FGENESH. A continuación, repetid el mismo procedimiento para calcular la correlación entre cada predicción individual y el gen anotado por el consorcio RefSeq.

Correlate table 'DVZ genscan' (ct_DVZgenscan_5145) with table 'ct_DVZfgenesh_9282'

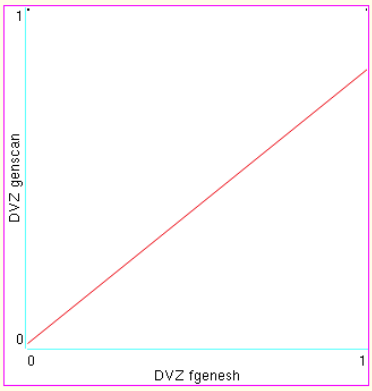
Select a group, track and table to correlate with:
group: Custom Tracks **track:** DVZ fgenesh
table: ct_DVZfgenesh_9282
Limit total data points in result: 40,000,000 Window data to: 1 bases

calculate | clear selections | return to table browser

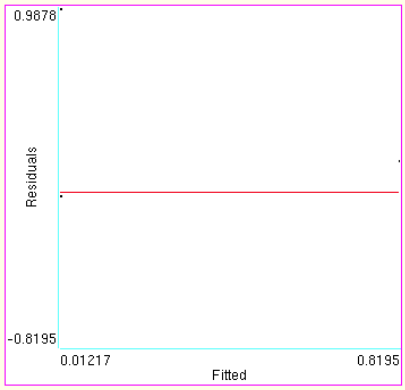
position: chr21:45,079,432-45,115,960 **bases:** 36,529

Position and # of data points in intersection	Correlation coefficient r	r^2	Track	Minimum	Maximum	Mean	Variance	Standard deviation	Regression line y = m*x + b m	b
chr21:45,079,432-45,115,960	0.8063	**0.6502**	DVZ genscan	0	1	0.0625	0.05859	0.2421	0.8073	0.01217
36,529 data points			DVZ fgenesh	0	1	0.06233	0.05845	0.2418		

scatter plot, r^2 0.6502



Residuals vs. Fitted, F statistic: 6.789e+04



Correlate table 'DVZ genscan' (ct_DVZgenscan_5145) with table 'ct_DVZgeneid_1704'

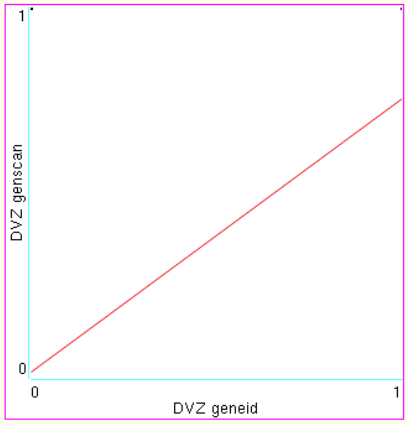
Select a group, track and table to correlate with:
group: Custom Tracks **track:** DVZ geneid
table: ct_DVZgeneid_1704
Limit total data points in result: 40,000,000 Window data to: 1 bases

calculate | clear selections | return to table browser

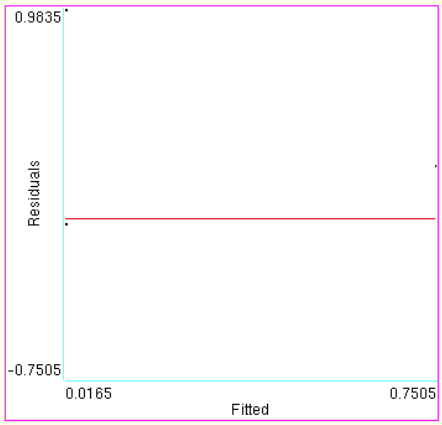
position: chr21:45,079,432-45,115,960 **bases:** 36,529

Position and # of data points in intersection	Correlation coefficient r	r^2	Track	Minimum	Maximum	Mean	Variance	Standard deviation	Regression line y = m*x + b m	b
chr21:45,079,432-45,115,960	0.7349	**0.5401**	DVZ genscan	0	1	0.0625	0.05859	0.2421	0.734	0.0165
36,529 data points			DVZ geneid	0	1	0.06266	0.05874	0.2424		

scatter plot, r^2 0.5401



Residuals vs. Fitted, F statistic: 4.29e+04



Correlate table 'DVZ fgenesh' (ct_DVZfgenesh_9282) with table 'ct_DVZgeneid_1704'

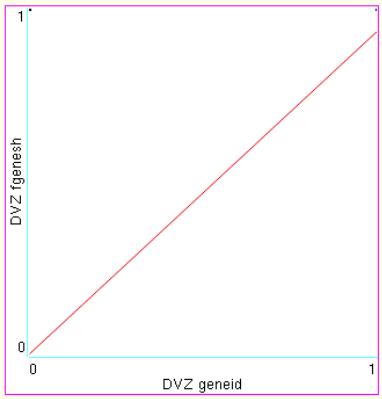
Select a group, track and table to correlate with:
group: Custom Tracks **track:** DVZ geneid
table: ct_DVZgeneid_1704
Limit total data points in result: 40,000,000 Window data to: 1 bases

calculate | clear selections | return to table browser

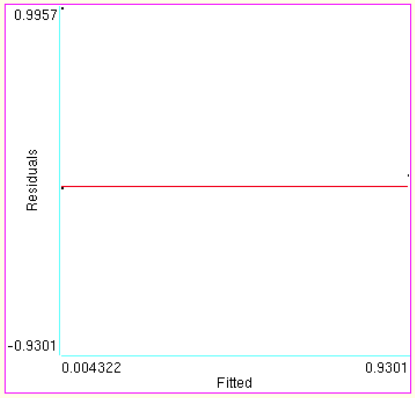
position: chr21:45,079,432-45,115,960 **bases:** 36,529

Position and # of data points in intersection	Correlation coefficient r	r^2	Track	Minimum	Maximum	Mean	Variance	Standard deviation	Regression line y = m*x + b m	b
chr21:45,079,432-45,115,960	0.9281	**0.8613**	DVZ fgenesh	0	1	0.06233	0.05845	0.2418	0.9258	0.004322
36,529 data points			DVZ geneid	0	1	0.06266	0.05874	0.2424		

scatter plot, r^2 0.8613



Residuals vs. Fitted, F statistic: 2.268e+05



Correlate table 'DVZ fgenesh' (ct_DVZfgenesh_9282) with table 'refGene'

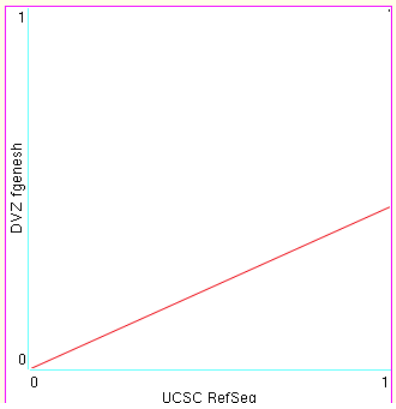
Select a group, track and table to correlate with:
group: Genes and Gene Predictions **track:** NCBI RefSeq
table: UCSC RefSeq (refGene)
Limit total data points in result: 40,000,000 Window data to: 1 bases

calculate | clear selections | return to table browser

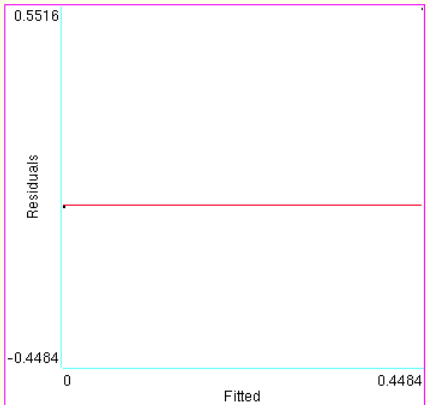
position: chr21:45,079,432-45,115,960 **bases:** 36,529

Position and # of data points in intersection	Correlation coefficient r	r^2	Track	Minimum	Maximum	Mean	Variance	Standard deviation	Regression line y = m*x + b m	b
chr21:45,079,432-45,115,960	0.6417	**0.4117**	DVZ fgenesh	0	1	0.06233	0.05845	0.2418	0.4484	0
36,529 data points			UCSC RefSeq	0	1	0.139	0.1197	0.346		

scatter plot, r^2 0.4117



Residuals vs. Fitted, F statistic: 2.557e+04



Correlate table 'DVZ genscan' (ct_DVZgenscan_5145) with table 'refGene'

Select a group, track and table to correlate with:

group: Genes and Gene Predictions **track:** NCBI RefSeq

table: UCSC RefSeq (refGene)

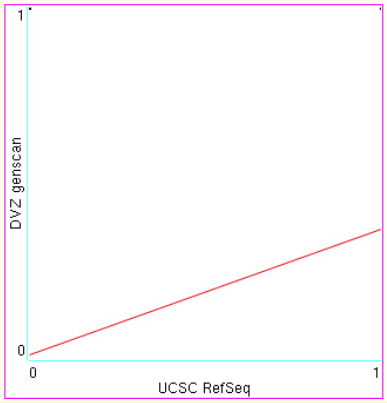
Limit total data points in result: 40,000,000 Window data to: 1 bases

calculate | clear selections | return to table browser

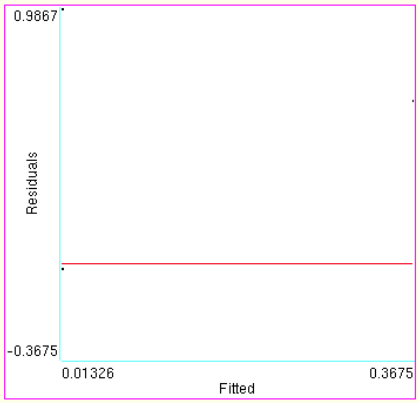
position: chr21:45,079,432-45,115,960 **bases:** 36,529

Position and # of data points in intersection	Correlation coefficient r	r^2	Track	Minimum	Maximum	Mean	Variance	Standard deviation	Regression line y = m*x + b m	b
chr21:45,079,432-45,115,960	0.5062	**0.2563**	DVZ genscan	0	1	0.0625	0.05859	0.2421	0.3542	0.01326
36,529 data points			UCSC RefSeq	0	1	0.139	0.1197	0.346		

scatter plot, r^2 0.2563



Residuals vs. Fitted, F statistic: 1.259e+04



Correlate table 'DVZ geneid' (ct_DVZgeneid_1704) with table 'refGene'

Select a group, track and table to correlate with:

group: Genes and Gene Predictions **track:** NCBI RefSeq

table: UCSC RefSeq (refGene)

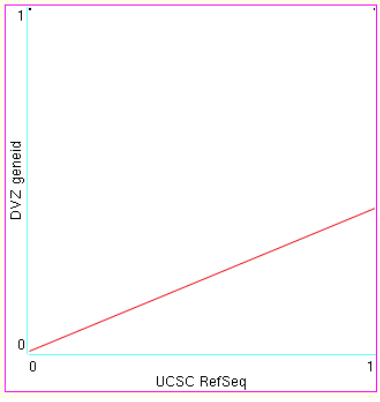
Limit total data points in result: 40,000,000 Window data to: 1 bases

calculate | clear selections | return to table browser

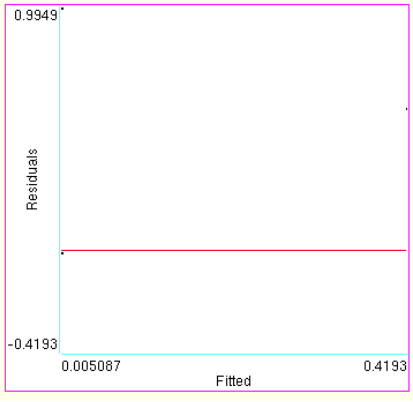
position: chr21:45,079,432-45,115,960 **bases:** 36,529

Position and # of data points in intersection	Correlation coefficient r	r^2	Track	Minimum	Maximum	Mean	Variance	Standard deviation	Regression line y = m*x + b m	b
chr21:45,079,432-45,115,960	0.5912	**0.3496**	DVZ geneid	0	1	0.06266	0.05874	0.2424	0.4142	0.005087
36,529 data points			UCSC RefSeq	0	1	0.139	0.1197	0.346		

scatter plot, r^2 0.3496



Residuals vs. Fitted, F statistic: 1.963e+04



Obtenemos que el que tiene mayor solapamiento con RefSeq es fgenesh.

9. Para acabar, efectuad con CLUSTAL el alineamiento múltiple global de las tres proteínas predichas por cada programa junto con la proteína real RRP1B. Analizad cuidadosamente cada sección de la proteína en busca de las mejores predicciones en ese fragmento. Con todas estas informaciones, decidid qué programa ha efectuado la mejor predicción.

```
geneid          MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY  60
FGENESH         MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY  60
Proteína        MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY  60
genscan         MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY  60
                ************************************************************

geneid          CMWVQDEPLLQEELANTIAQLVHAVNNSAAQAC--------------------------  93
FGENESH         CMWVQDEPLLQEELANTIAQLVHAVNNSAAQHLFIQTFWQTMNREWKGIDRLRLDKYYML  120
Proteína        CMWVQDEPLLQEELANTIAQLVHAVNNSAAQHLFIQTFWQTMNREWKGIDRLRLDKYYML  120
genscan         CMWVQDEPLLQEELANTIAQLVHAVNNSAAQHLFIQTFWQTMNREWKGIDRLRLDKYYML  120
                ******************************

geneid          ---------------VWFFSRIKVFLDVLMKEVLCPESQSPNGVRFHFIDIYLDELSKVG  138
FGENESH         IRLVLRQSFEVLKRNGWEESRIKVFLDVLMKEVLCPESQSPNGVRFHFIDIYLDELSKVG  180
Proteína        IRLVLRQSFEVLKRNGWEESRIKVFLDVLMKEVLCPESQSPNGVRFHFIDIYLDELSKVG  180
genscan         IRLVLRQSFEVLKRNGWEESRIKVFLDVLMKEVLCPESQSPNGVRFHFIDIYLDELSKVG  180
                      *  ****************************************

geneid          GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVDQSPFVPEETMEEQKTKVG  198
FGENESH         GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVDQSPFVPEETMEEQKTKVG  240
Proteína        GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVDQSPFVPEETMEEQKTKVG  240
genscan         GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVDQSPFVPEETMEEQKTKVG  240
                ************************************************************

geneid          DGDLSAEEIPENEVSLRRAVSKKKTALGKNHSRKDGLSDERGRDDCGTFEDTGPLLQFDY  258
FGENESH         DGDLSAEEIPENEVSLRRAVSKKKTALGKNHSRKDGLSDERGRDDCGTFEDTGPLLQFDY  300
Proteína        DGDLSAEEIPENEVSLRRAVSKKKTALGKNHSRKDGLSDERGRDDCGTFEDTGPLLQFDY  300
genscan         DGDLSAEEIPENEVSLRRAVSKKKTALGKNHSRKDGLSDERGRDDCGTFEDTGPLLQFDY  300
                ************************************************************

geneid          KAVADRLLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQLSFAEDISADEDDQILSQ  318
FGENESH         KAVADRLLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQLSFAEDISADEDDQILSQ  360
Proteína        KAVADRLLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQLSFAEDISADEDDQILSQ  360
genscan         KAVADRLLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQLSFAEDISADEDDQILSQ  360
                ************************************************************

geneid          GKHKKKGNKLLEKTNLEKE------------------------------------------  337
FGENESH         GKHKKKGNKLLEKTNLEKE------------------------------------------  379
Proteína        GKHKKKGNKLLEKTNLEKE------------------------------------------  379
genscan         GKHKKKGNKLLEKTNLEKEKGKQELQGALGGGCLMTTRDLWFLPLSPKISGNGTISVPYV  420
                ******************

geneid          ----------------------KGSRVFCVEEEDSESSLQKRRRKKKKKHHLQPENPGPG  375
FGENESH         ----------------------KGSRVFCVEEEDSESSLQKRRRKKKKKHHLQPENPGPG  417
Proteína        ----------------------KGSRVFCVEEEDSESSLQKRRRKKKKKHHLQPENPGPG  417
genscan         FINGQKEGFQSQLGMEEVGPDDKGSRVFCVEEEDSESSLQKRRRKKKKKHHLQPENPGPG  480
                                       ************************************

geneid          GAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK  435
FGENESH         GAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK  477
Proteína        GAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK  477
genscan         GAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK  540
                ************************************************************

geneid          KSPRAHREMLESAVLPPEDMSQSGPSGSHPQGPRGSPTGGAQLLKRKRKLGVVPVNGSGL  495
FGENESH         KSPRAHREMLESAVLPPEDMSQSGPSGSHPQGPRGSPTGGAQLLKRKRKLGVVPVNGSGL  537
Proteína        KSPRAHREMLESAVLPPEDMSQSGPSGSHPQGPRGSPTGGAQLLKRKRKLGVVPVNGSGL  537
genscan         KSPRAHREMLESAVLPPEDMSQSGPSGSHPQGPRGSPTGGAQLLKRKRKLGVVPVNGSGL  600
                ************************************************************

geneid          STPAWPPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRK  555
FGENESH         STPAWPPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRK  597
Proteína        STPAWPPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRK  597
genscan         STPAWPPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRK  660
                ************************************************************

geneid          KMRVMSNLVEHNGVLESEAGQPQALVRWEHPQAS--------------SPQRHSL-ASMG  600
FGENESH         KMRVMSNLVEHNGVLESEAGQPQAL-----------------------------------  622
Proteína        KMRVMSNLVEHNGVLESEAGQPQAL-----------------------------------  622
genscan         KMRVMSNLVEHNGVLESEAGQPQALAAHLNLPEPPVCRQRHWAAHTSESQVRDPVSLWVA  720
                *************************

geneid          LHCLLRG--------------------------------RVGAGGQASGLSSS-------  621
FGENESH         ----------------------------------------GSSGTCSSLKKQKLRAESD  641
Proteína        ----------------------------------------GSSGTCSSLKKQKLRAESD  641
```

```
genscan       VSCCTRNECPGPASVVLCVKPELCRMEGLSASAVRKTAGRRGSSGTCSSLKKQKLRAESD   780
                                                 *:.* .*.*...

geneid        ------------------------------------------------------------   621
FGENESH       FVKFDTPFLPKPLFFRRAKSSTATHPPGPAVQLNKTPSSSKKVTFGLNRNMTAEFKKTDK   701
Proteína      FVKFDTPFLPKPLFFRRAKSSTATHPPGPAVQLNKTPSSSKKVTFGLNRNMTAEFKKTDK   701
genscan       FVKFDTPFLPKPLFFRRAKSSTATHPPGPAVQLNKTPSSSKKVTFGLNRNMTAEFKKTDK   840

geneid        ------------------------------------------------------------   621
FGENESH       SILVSPTGPSRVAFDPEQKPLHGVLKTPTSSPASSPLVAKKPLTTTPRRRPRAMDFF     758
Proteína      SILVSPTGPSRVAFDPEQKPLHGVLKTPTSSPASSPLVAKKPLTTTPRRRPRAMDFF     758
genscan       SILVSPTGPSRVAFDPEQKPLHGVLKTPTSSPASSPLVAKKPLTTTPRRRPRAMDFF     897
```

En un compendio general vemos que FGENESH es la más acertada, aunque los tres programas recuperan bastante bien parte de la proteína. Utilizando la pista CDS vemos que es la que más idéntica y la cual tiene mejor correlación.

10. El navegador genómico VISTA permite observar la conservación entre diversos genomas. Analizad la documentación existente sobre esta aplicación y averiguad el significado que tienen las gráficas y los colores empleados sobre cada alineamiento entre dos genomas. Posteriormente, seleccionad nuestro gen de estudio para analizar el grado de conservación que poseen los exones de éste. Razonad brevemente sobre cómo podríamos mejorar las predicciones iniciales servidas por GENEID, GENSCAN y FGENESH utilizando esta información sobre la conservación de secuencia en regiones funcionales.

El software VISTA es una herramienta que visualiza la comparación de múltiples genomas para identificar áreas genómicas que han sido conservadas a lo largo de la evolución. Estas áreas conservadas se destacan en los gráficos de conservación con diferentes colores, como rojo para regiones no codificantes, azul claro para regiones UTR y azul oscuro para regiones codificantes. En nuestro caso, podemos observar que los exones tienen la mayor conservación a lo largo del árbol de especies. A medida que nos alejamos evolutivamente, la señal conservación se debilita, pero aún coincide con los elementos funcionales del gen. Para mejorar nuestras predicciones, podemos asignar un valor numérico a los exones predichos que se ajustan exactamente a las regiones conservadas en el gráfico. Esto aumentaría la evidencia de su posible existencia y nos ayudaría a descartar las regiones predichas que no están conservadas.