

GENÓMICA COMPUTACIONAL - PEC 2

por Diego Vallarino

EJERCICIO 1

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).

```

CLUSTAL O(1.2.4) multiple sequence alignment

human  ATGGGAAAGTCTCTTTCTCATTTCCTTTGCATTCAAGCAAAGAAGATGCTTATGATGGA 60
raton   ATGGGAAAGTCTCTTTCTCACTTACCTTTGCATTCAAATAAAGAAGATGGCTATGATGGC 60
*****

human  GTCACATCT--GAAACATGAGGAATGGACTGGTTAATAGTGAAGTCCATAATGAAGAT 117
raton   GTTACATCGACAGACAATATGAGAAATGGATTGGTTAGCAGTGAAGTGCACAACGAAGAC 120
*****

human  GGAAGAAATGGAGATGTCTCTCAGTTTCCATATGTGGAATTTACAGGAAGAGATAGTGTC 177
raton   GGAAGAAATGGAGATGTCTCTCAGTTCCATATGTGGAATTTACTGGAAGAGATAGTGTC 180
*****

human  ACCTGCCCTACTTGTGAGGGAACAGGAAGAATTCCTAGGGGGCAAGAAAACCAACTGGTG 237
raton   ACTTGTCCCACTTGCCAAGGAACAGGAAGAATTCCTAGGGGCAAGAAAACCAACTGGTG 240
*****

human  GCATTGATTCCATATAGTGATCAGAGATTAAGGCCAAGAAGAACAAAGCTGTATGTGATG 297
raton   GCATTGATTCCATATAGTGATCAGCGGTTACGGCCAAGAAGAACAAAGCTGTATGTGATG 300
*****

human  GCTTCTGTGTTTGTCTGTCTACTCCTTTCTGATTGGCTGTGTTTTTCTTTTCCCTCGC 357
raton   GCGTCTGTGTTTGTCTGCCTGCTCTGTCTGGATTGGCTGTGTTTTTCTTTTCCCTCGA 360
*****

human  TCTATCGACGTGAAATACATTGGTGTAAATCAGCCTATGTCAGTTATGATGTTTCAAG 417
raton   TCTATTGAGGTGAAGTACATTGGAGTAAATCAGCCTATGTCAGCTACGACGCTGAAAG 420
*****

human  CGTACAATTTATTTAAATATCACAAACACACTAAATATAACAAACAATAACTATTACTCT 477
raton   CGAACCATATATTTAAATATCACGAACACACTAAATATAACAAATAATAACTATTATTCT 480
*****

human  GTCGAAGTTGAAAACATCACTGCCCCAAGTTCAATTTTCAAAAACAGTTATTGGAAGGCA 537
raton   GTTGAAGTTGAAAACATCACTGCTCAAGTCCAGTTTCAAAAACCGTATTGGAAGGCT 540
*****

human  CGCTTAAACAACATAAACCATTATTGGTCCACTTGATATGAAACAAATTGATTACACAGTA 597
raton   CGTTTAAACAACATAAACAATTGGCCCACTTGATATGAAGCAGATTGATTATACGGTA 600
*****

human  CCTACCGTTATAGCAGAGGAAATGAGTTATATGTATGATTTCTGTACTCTGATATCCATC 657
raton   CCCACAGTTATTGCAGAGGAAATGAGTTACATGTATGATTTCTGTACTCTGCTCTCCATC 660
*****

human  AAAGTGCATAACATAGTACTCATGATGCAAGTTACTGTGACAACAACATACTTTGGCCAC 717
raton   AAAGTGCACAACATAGTACTCATGATGCAAGTTACTGTAAACAACAGCATACTTTGGACAC 720
*****

human  TCTGAACAGATATCCAGGAGAGGTATCAGTATGTCGACTGTGGAAGAAACACAACCTTAT 777
raton   TCTGAGCAGATATCTCAGGAAAGGTACCAAGTATGTCGACTGTGGAAGGAACACGACTTAC 780
*****

human  CAGTTGGGGCAGTCTGAATATTTAAATGTACTTCAGCCACAACAGTAA 825
raton   CAGTTGGCCAGTCTGAGTATCTAATGTCTTCAGCCACAACAATAA 828
*****

```

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.

```

CLUSTAL O(1.2.4) multiple sequence alignment

human      MGKSLSHLPLHSSKEDAYDGVTS-ENMRNGLVNSEVHNEDGRNGDVVSQFPYVEFTGRDSV  59
raton      MGKSLSHLPLHSNKEDGYDGVTS-ENMRNGLVNSEVHNEDGRNGDVVSQFPYVEFTGRDSV  60
*****
human      TCPTCQGTGRIPRGQENQLVALIPYSDQRLRPRTKLYVMASVFCVLLSGLAVFFLFPR  119
raton      TCPTCQGTGRIPRGQENQLVALIPYSDQRLRPRTKLYVMASVFCVLLSGLAVFFLFPR  120
*****
human      SIDVKYIGVKSAVSYDVQKRTIYLNITNTLNITNNNYYSVEVENITAQVQFSKTIGKA  179
raton      SIEVKYIGVKSAVSYDAEKRTIYLNITNTLNITNNNYYSVEVENITAQVQFSKTIGKA  180
*****
human      RLNNITIIIGPLDMKQIDYVPTVIAEEMSYMYDFCTLLSIKVHNIIVLMMQVTVTTTYFGH  239
raton      RLNNITIIIGPLDMKQIDYVPTVIAEEMSYMYDFCTLLSIKVHNIIVLMMQVTVTTTYFGH  240
*****
human      SEQISQERYQYVDCGRNTTYQLGQSEYLMVLQPQQ  274
raton      SEQISQERYQYVDCGRNTTYQLAQSEYLMVLQPQQ  275
*****

```

El grado de homología es cercano al 96%.

```

#
#
# Percent Identity Matrix - created by Clustal2.1
#
#
1: human      100.00   95.99
2: raton      95.99  100.00

```

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*.

Tomamos el Blast Nucleotide y tomamos la frecuencia FASTA de cada una de las dos secuencias. A continuación, se muestra en la siguiente figura:

blastn

blastp

blastx

tblastn

tblastx

Align

BLASTN programs search

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

>human

ATGGGAAAGTCTCTTTCTCATTGCTTTGCATTCAAGCAAAGAAGATGCT

TATGATGGAGTCACATCTG

AAAACATGAGGAATGGACTGGTTAATAGTGAAGTCCATAATGAAGATGGAA

Query subrange [?](#)

From

To

Or, upload file

Seleccionar archivo

Ningún archivo seleccionado [?](#)

Job Title

human

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

>raton

ATGGGAAAGTCTCTTTCTCATTACCTTTGCATTCAAATAAGAAGATGGC

TATGATGGCGTTACATCGA

CAGACAATATGAGAAATGGATTGGTTAGCAGTGAAGTGCACAACGAAGAC

Subject subrange [?](#)

From

To

Or, upload file

Seleccionar archivo

Ningún archivo seleccionado [?](#)

Program Selection

Optimize for

☒ Highly similar sequences (megablast)
 ☐ More dissimilar sequences (discontiguous megablast)
 ☐ Somewhat similar sequences (blastn)

Choose a BLAST algorithm [?](#)

BLAST

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

☐ Show results in a new window

Range 1: 1 to 828 [Graphics](#) [▼ Next Match](#)

Score	Expect	Identities	Gaps	Strand
1000 bits(541)	0.0	733/828(89%)	3/828(0%)	Plus/Plus
Query 1	ATGGGAAAGTCTCTTTCTCATTGCTTTGCATTCAAGCAAAGAAGATGCTTATGATGGA	60		
Sbjct 1	ATGGGAAAGTCTCTTTCTCATTGCTTTGCATTCAAATAAGAAGATGGCTATGATGGC	60		
Query 61	GTACAT ---CTGAAACATGAGGAATGGACTGGTTAATAGTGAAGTCCATAATGAAGAT	117		
Sbjct 61	GTTACATCGACAGACAATATGAGAAATGGATTGGTTAGCAGTGAAGTGCACAACGAAGAC	120		
Query 118	GGAAGAAATGGAGATGCTCTCAGTTTCCATATGTGGAATTACAGGAAGAGATAGTGTC	177		
Sbjct 121	GGAAGAAATGGAGATGCTCTCAGTTCCCATATGTGGAATTTACTGGAAGAGATAGTGTC	180		
Query 178	ACCTGCCCTACTTGTACAGGAACAGGAAGAATTCCTAGGGGGCAAGAAAACCACTGGTG	237		
Sbjct 181	ACTTGTCCTACTTGTCAAGGAACAGGAAGAATTCCTAGGGGGCAAGAAAACCACTGGTG	240		
Query 238	GCATTGATTCATATAGTGATCAGAGATTAAGGCCAAGAAGAACAAAGCTGTATGTGATG	297		
Sbjct 241	GCATTGATTCATATAGTGATCAGCGGTACGGCCAAGAAGAACAAAGCTGTATGTGATG	300		
Query 298	GCTTCGTGTTTGTCTGTCTACTCCTTCTGGATTGGCTGTGTTTTCTTTTCCCTCGC	357		
Sbjct 301	GCGTCGTGTTTGTCTGCCTGCTCCTGTCTGGATTGGCTGTGTTTTCTTTTCCCTCGA	360		
Query 358	TCTATCGACGTGAAATACATTGGTGTAATAACAGCCTATGTGAGTTATGATGTTGAGAAG	417		
Sbjct 361	TCTATTGAGGTGAAGTACATTGGAGTAAATCAGCCTATGTGAGTACGACGTGAAAAG	420		
Query 418	CGTACAATTTATTTAAATATCACAAACACACTAAATATAACAAACAATAACTATTACTCT	477		
Sbjct 421	CGAACCATATATTTAAATATCACGAACACACTAAATATAACAAATAATACTATTATTCT	480		
Query 478	GTCGAAGTTGAAAACATCACTGCCAAGTTCAATTTTCAAAAACAGTTATTGGAAAGGCA	537		
Sbjct 481	GTTGAAGTTGAAAACATCACTGCTCAAGTCCAGTTTTCAAAACCGTGATTGGAAAGGCT	540		
Query 538	CGCTTAAACAACATAACCATTATTGGTCCACTTGATGAAACAAATTGATTACACAGTA	597		
Sbjct 541	CGTTTAAACAACATAACTAACATTGGCCCACTTGATGAAGCAGATTGATTATACGGTA	600		
Query 598	CCTACCGTTATAGCAGAGGAAATGAGTTATATGTATGATTTCTGTACTCTGATATCCATC	657		
Sbjct 601	CCCACAGTTATTGCAGAGGAAATGAGTTACATGTATGATTTCTGTACACTGCTCTCCATC	660		
Query 658	AAAGTGCATAACATAGTACTCATGATGCAAGTTACTGTGACAACAACATACTTTGGCCAC	717		
Sbjct 661	AAAGTGCACAACATAGTACTCATGATGCAAGTTACTGTAAACAACAGCATACTTTGGACAC	720		
Query 718	TCTGAACAGATATCCAGGAGAGGTATCAGTATGTCGACTGTGGAAGAAACACAACCTAT	777		
Sbjct 721	TCTGAGCAGATATCTCAGGAAAGGTACAGTATGTCGACTGTGGAAGGAACACGACTTAC	780		
Query 778	CAGTTGGGCGAGTCTGAATATTTAAATGTACTTCAGCCACAACAGTAA	825		
Sbjct 781	CAGTTGGCCAGTCTGAGTATCTAAATGTCCTTCAGCCACAACAATAA	828		

Sequence ID: Query_492119 Length: 828 Number of Matches: 1

Range 1: 1 to 828 [Graphics](#)

▼ [Next Match](#)

Score	Expect	Identities	Gaps	Strand
1000 bits(541)	0.0	733/828(89%)	3/828(0%)	Plus/Plus
Query Sbjct	1 1	ATGGGAAAGTCTCTTTCTCATTTGCCTTTGCATTCAAGCAAGAAGATGCTTATGATGGA		60 60
	C.A.....AT.....GC.....C		
Query Sbjct	61 61	GTCACAT---CTGAAAAACATGAGGAATGGACTGGTTAATAGTGAAGTCCATAATGAAGAT		117 120
		..T...CGA.A..C.T...A...T...GC.....G..C..C....C		
Query Sbjct	118 121	GGAAGAAATGGAGATGTCTCTCAGTTTCCATATGTGGAATTACAGGAAGAGATAGTGTC		177 180
	C.....T.....		
Query Sbjct	178 181	ACCTGCCCTACTTGTGACGGGAACAGGAAGAATTCCTAGGGGGCAAGAAACCAACTGGTG		237 240
		..T..T..C.....C..A.....A.....		
Query Sbjct	238 241	GCATTGATTCCATATAGTGTATCAGAGATTAAAGGCCAAGAAAGCAAAAGCTGTATGTGATG		297 300
	C.G..C.....		
Query Sbjct	298 301	GCTTCTGTGTTTGTCTGTCTACTCCTTCTCGATTGGCTGTGTTTTTCCTTTTCCCTCGC		357 360
		..G.....C..G....G.....T.....A		
Query Sbjct	358 361	TCTATCGACGTGAAATACATTGGTGTAATAACAGCCTATGTCAGTTATGATGTTCAAGAG		417 420
	T..G.....G.....A.....C.....C..C.G.A..		
Query Sbjct	418 421	CGTACAATTATTATTAATATCAACAACACACTAAATATAACAACAATAACTATTACTCT		477 480
		..A..C..A.....G.....T.....		
Query Sbjct	478 481	GTCGAAGTTGAAAAACATCACTGCCCAAGTTCAATTTTCAAAAACAGTTATTGGAAGGCA		537 540
		..T.....T...C..G.....C..G.....T		
Query Sbjct	538 541	CGCTTAAACAACATAACCATTATTGGTCCACTTGATATGAACAATAATTGATTACACAGTA		597 600
	T.AC.....G..G.....T..G..		
Query Sbjct	598 601	CCTACCGTTATAGCAGAGGAAATGAGTTATATGTATGATTTCGTACTCTGATATCCATC		657 660
		..C..A...T.....C.....A..C..C.....		
Query Sbjct	658 661	AAAGTGCATAACATAGTACTCATGATGCAAGTTACTGTGACAAACATACTTTGGCCAC		717 720
	C.....A.....A.....		
Query Sbjct	718 721	TCTGAACAGATATCCAGGAGAGGTATCAGTATGTCGACTGTGGAAGAAACACAACCTTAT		777 780
	G.....T...A...C.....G.....G.....C		
Query Sbjct	778 781	CAGTTGGGGCAGTCTGAATATTTAAATGTACTTCAAGCCACAACAGTAA	825 828	
	CC.....G...C.....A.....		

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.

He codificado en RStudio para poder acceder a cada uno de los txt. Este fue el código escrito:

```
untar("D:/Master en BioEstadística/Materias/2.Genómica
computacional/PEC2/input.tar.gz", exdir = "D:/Master en
BioEstadística/Materias/2.Genómica computacional/PEC2")
```

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

DNA

sequences in any supported format:

```

ctttttatcttcacacatgctgcgcacaaagtttccaaactgtat
gcttttttccctttaaagtaagattcagctttatagctattcttgc
atggggagacagatgaatcatatggtagagaggaagacacagagagagac
tggatgtagtgacagacacttaagcaatcaatccacagatgaactc
gaa-aaagaaagcttcttccacaaagggatgtaactctcagagaggaac
tgcaccatgatcaatcaactcccccacaggaataacatcatgtttccaa
ATGGGAAAGTCTCTTTTCATTTGCTTTTCATTCAAGCAAAAGAGATGC
tgcacatgatcaatcaactcccccacaggaataacatcatgtttccaa

```

Or, upload a file |

Seleccionar archivo

Ningún archivo seleccionado

Use a example sequence |

Clear sequence |

See more example inputs

STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

The default settings will fulfil the needs of most users.

More options...

(Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

genomicA	cagaagaattgcttgaaccaggagggtggagggttcagtgagcagagatcacgccactgc	60
genomicB	-----gctgggatg--tggggagcagtgcttctgaggctgagcag-gac	40
	* * * * *	
genomicA	actcctgcttaagtacagagtgagactccatctcaaaaaaaaaaaaaattcctatta	120
genomicB	agtgaggccttgggcctggcct-----ctgaaaccatttttccacctaggcctc	90
	* * * * *	
genomicA	tgtgcttgagtaataccaccactctggcaaatcttaaaaaagctcttggccgggtgcag	180
genomicB	tgagcctgtgtcctataacttattgcaggctgttagaagc-----aggcagac	138
	* * * * *	
genomicA	tggctcatgcctgtaatcccagaagaattgcttgaaccaggagggtggagggttcagtg	240
genomicB	tactttctggatgcttctgctttagaattttttctgcca-----	179
	* * * * *	
genomicA	agcagagatcacgccactgcactcctgcttaagtacagagtgagactccatctcaaaaa	300
genomicB	---ga-----tattctaggtcatcactctATGAGTGTGGATCCAGCTTGT---	221
	* * * * *	
genomicA	aaaaaaaaaattcctattatgtgcttgagtaataccaccactctggcaaatcttaaaa	360
genomicB	-----CCCCAAGCTTGCTTGCCTTGAA	245
	* * * * *	
genomicA	aagctcttggccgggtgcagtggtcatgcctgtaatccCATGGGAAAGTCTTTCTCA	420
genomicB	GCATCat--gggaggagctgtctctaagatctctaaagtgactttgaggccttttctca	303
	* * * * *	
genomicA	TTTGCTTTGCATTCAAGCAAGAAGATGCagttcccatcttctgtgccacacctctga	480
genomicB	ttgtcttgatattagccctt-----ggcacccttttagtcagcctaattccccta	354
	* * * * *	
genomicA	gatgggtgcctgtgtctgtcattgtttcttgaatcaatctagacctcagttctaaagaacc	540
genomicB	gcaagtgggtgtccacagcctgtttat-----attcctctctc--aataatgc	401
	* * * * *	
genomicA	ctaaaaactctgtccgtgaatcttgggggaagggaagtcaatgtaaaatacttcata	600
genomicB	tttttattctctgccacatgg--ctggctacaggttttccaaacttgta---tgcttt	455
	* * * * *	
genomicA	ttgtatttctaagatgtctatttcccttt--gtgattattttgactgcaagtgtccgtg	658
genomicB	gtttcccttttaaatgtaagtttcagctttaagtcatttcttgcaggggagcagatga	515
	* * * * *	
genomicA	aatcttgggggaagggaagtcaatgtaaaatacttcataattgtatttctaagatgtc	718
genomicB	atcatatggtgagagagggaagtcacagagagagactaggatgtggtaccagactttaag	575
	* * * * *	
genomicA	tatttccctttgtgattattttgactgcaagATGAGTGTGGATCCAGCTTGTCCTCAAA	778
genomicB	caatcaaatctcacgtgaactaactgagcaagaa-----gtgacttatcaccag	625
	* * * * *	
genomicA	GCTTGCTTTGCTTTGAAGCATCagttcccatatt-----ctgtcggcacacctc	827
genomicB	gggtgttaacattcatgaggatctgcccacatgatccaatcacctcccaccaggaaat	685
	* * * * *	
genomicA	tgagatggtgcctgtgtctgtcattgtttcttgaatcaatctagacctcagttctaaaga	887
genomicB	cacattggtttccaaATGGGAAAGTCTCTTTCTATTG-----CCTTTGCATTCA	736
	* * * * *	
genomicA	accctaaaaactcagttcccatcttctgtgccacacctctgagatggtgcctgtgtctg	947
genomicB	AGCAAAAGAGATGCTgcccacatgatccaatcacctcccaccaggaaatcacattgggaa	796
	* * * * *	
genomicA	tcattgtttcttgaatcaatctagacctcagttctaaagaaccctaaaaactc	1000
genomicB	tcac-----	800

```

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

```

```

1: genomicA    100.00   44.09
2: genomicB     44.09  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.

BLAST® » blastn suite

blastn blastp blastx tblastn tblastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

Query subrange [?](#)

From

To

Or, upload file genomicA.txt.txt [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☒ Align two or more sequences [?](#)

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

Subject subrange [?](#)

From

To

Or, upload file genomicB.txt.txt [?](#)

Program Selection

Optimize for

☒ Highly similar sequences (megablast)

☐ More dissimilar sequences (discontiguous megablast)

☐ Somewhat similar sequences (blastn)

Choose a BLAST algorithm [?](#)

BLAST Search nucleotide sequence using Megablast (Optimize for highly similar sequences)

genomicB

Sequence ID: Query_32727 Length: 800 Number of Matches: 2

Range 1: 201 to 251 [Graphics](#)

[▼ Next Match](#) [▲ Previous Match](#)

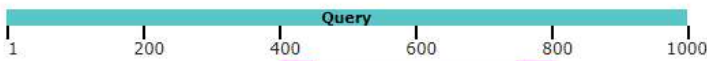
Score	Expect	Identities	Gaps	Strand
95.3 bits(51)	2e-23	51/51(100%)	0/51(0%)	Plus/Plus
Query 751	ATGAGTGTGGATCCAGCTTGTCCCAAAGCTTGCCTTGCTTTGAAGCATCA	801		
Sbjct 201	ATGAGTGTGGATCCAGCTTGTCCCAAAGCTTGCCTTGCTTTGAAGCATCA	251		

Range 2: 701 to 750 [Graphics](#)

[▼ Next Match](#) [▲ Previous Match](#) [▲ First Match](#)

Score	Expect	Identities	Gaps	Strand
93.5 bits(50)	6e-23	50/50(100%)	0/50(0%)	Plus/Plus
Query 401	ATGGGAAAGTCTCTTTCTCATTTGCCTTTGCATTCAAGCAAAGAAGATGC	450		
Sbjct 701	ATGGGAAAGTCTCTTTCTCATTTGCCTTTGCATTCAAGCAAAGAAGATGC	750		

Distribution of the top 2 Blast Hits on 1 subject sequences



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.

Un **alineamiento global** efectúa la correspondencia entre las secuencias completas, maximizando el número total de caracteres coincidentes a lo largo de las cadenas.

Un **alineamiento local** realiza exclusivamente la correspondencia entre aquellos fragmentos de las secuencias que poseen una coincidencia máxima de caracteres, descartando el resto de regiones a lo largo de dichas secuencias que no presentan una mínima similitud.

En este caso, a nivel de globalidad, el alineamiento era del 44%, pero cuando se analiza localmente, se encuentran 2 fragmentos en donde la coincidencia es muy alta. Como no se tiene toda la secuencia completa, parecería mejor utilizar el segundo caso, con un alineamiento local para poder tener mayor comprensión de esas dos regiones.

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.

The image shows the NCBI BLASTX search interface. At the top, there are tabs for different BLAST programs: blastn, blastp, **blastx**, tblastn, and tblastx. The **blastx** tab is selected. The interface is divided into two main sections: "Enter Query Sequence" and "Enter Subject Sequence".

Enter Query Sequence:

- Enter accession number(s), gi(s), or FASTA sequence(s): A text area containing a DNA sequence starting with ">genomicC".
- Query subrange: Fields for "From" and "To" positions.
- Or, upload file: A button labeled "Seleccionar archivo" and a status "Ningún archivo seleccionado".
- Genetic code: A dropdown menu set to "Standard (1)".
- Job Title: A text field containing "genomicC".
- Align two or more sequences: A checked checkbox.

Enter Subject Sequence:

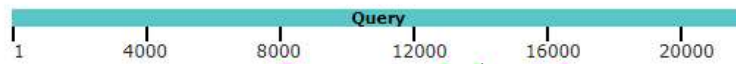
- Enter accession number(s), gi(s), or FASTA sequence(s): A text area containing a protein sequence starting with ">human".
- Subject subrange: Fields for "From" and "To" positions.
- Or, upload file: A button labeled "Seleccionar archivo" and a status "Ningún archivo seleccionado".

At the bottom, there is a large blue "BLAST" button and a checkbox for "Show results in a new window".

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> human		135	497	3%	9e-39	85.14%	274	Query_26769

El grado de homologia es relativamente alto, de aproximadamente 85%.

Distribution of the top 7 Blast Hits on 1 subject sequences



human

Sequence ID: **Query_26769** Length: **274** Number of Matches: **7**

Range 1: 1 to 73 [Graphics](#)

[Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
135 bits(339)	9e-39	Compositional matrix adjust.	63/74(85%)	67/74(90%)	1/74(1%)	+2
Query 8096	MGKSLSHLPIHTCKEDGYDGGTVSDNMRNGLVHSESHGEDGRCGDVSQFPYVEFTGRDSV	8275				
Sbjct 1	MGKSLSHLP+H+ KED YDG T S+NMRNGLV+SE H EDGR GDVSQFPYVEFTGRDSV	59				
Query 8276	TCPTCQGTGRIPRG	8317				
Sbjct 60	TCPTCQGTGRIPRG	73				

Range 2: 226 to 274 [Graphics](#)

[Next Match](#) [Previous Match](#) [First Match](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
86.7 bits(213)	4e-22	Compositional matrix adjust.	40/49(82%)	46/49(93%)	0/49(0%)	+2
Query 16460	VFLRVTVTTSYFGHSEQISREKYQYVDCGNTTYQLGQSEYLNVLQPPQ	16606				
Sbjct 226	+ ++VTVT+YFGHSEQIS+E+YQYVDCG NTTYQLGQSEYLNVLQPPQ	274				

Range 3: 154 to 198 [Graphics](#)

[Next Match](#) [Previous Match](#) [First Match](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
76.3 bits(186)	1e-20	Compositional matrix adjust.	40/45(89%)	42/45(93%)	0/45(0%)	+3
Query 13785	NNNYYSVEVANITAQVQFSKTIVIGKARLNNITNIGPLDMKQVNRT	13919				
Sbjct 154	NNNYYSVEV NITAQVQFSKTIVIGKARLNNIT IGPLDMKQ++ T	198				

Range 4: 195 to 210 [Graphics](#)

[Next Match](#) [Previous Match](#) [First Match](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
31.2 bits(69)	1e-20	Compositional matrix adjust.	13/16(81%)	14/16(87%)	0/16(0%)	+1
Query 14014	IDYMPVTVIQDEMSYM	14061				
Sbjct 195	IDYMPVTVIAEEMSYM	210				

Range 5: 95 to 143 [Graphics](#)

▼ [Next Match](#) ▲ [Previous Match](#) ▲ [First Match](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
79.0 bits(193)	1e-19	Compositional matrix adjust.	38/49(78%)	43/49(87%)	0/49(0%)	+1
Query 12835	KLYVTASVIVCLLLSGLAVFFLFPRSDVEYIGVKS VYVNYEQSRRRIY				12981	
Sbjct 95	KLYV ASV VCLLLSGLAVFFLFPRSDV+YIGVKS YV+Y+ +R IY				143	

Range 6: 73 to 94 [Graphics](#)

▼ [Next Match](#) ▲ [Previous Match](#) ▲ [First Match](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
48.9 bits(115)	1e-09	Compositional matrix adjust.	22/22(100%)	22/22(100%)	0/22(0%)	+3
Query 11376	GQENQLVALIPYSDQRLRPRT				11441	
Sbjct 73	GQENQLVALIPYSDQRLRPRT				94	

Range 7: 207 to 229 [Graphics](#)

▼ [Next Match](#) ▲ [Previous Match](#) ▲ [First Match](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
40.8 bits(94)	6e-07	Compositional matrix adjust.	17/23(74%)	19/23(82%)	0/23(0%)	+2
Query 15650	FSFFSDFCTLASIKVHNIIVMMQ				15718	
Sbjct 207	S+ DFCTL SIKVHNIIV+MMQ				229	

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:

El servidor web de MEME Suite proporciona un portal unificado para el descubrimiento y análisis en línea de motivos (*motifs*) de secuencia que representan características tales como sitios de unión de ADN y dominios de interacción de proteínas.

El popular algoritmo de descubrimiento de motivos MEME ahora se complementa con el algoritmo GLAM2 que permite el descubrimiento de motivos que contienen huecos. Tres algoritmos de escaneo de secuencias, MAST, FIMO y GLAM2SCAN, permiten escanear numerosas bases de datos de secuencias de ADN y proteínas en busca de motivos descubiertos por MEME y GLAM2. Los motivos de los factores de transcripción (incluidos los descubiertos mediante MEME) se pueden comparar con motivos en muchas bases de datos de motivos populares utilizando el algoritmo de exploración de la base de datos de patrón TOMTOM.

Los motivos de los factores de transcripción se pueden analizar adicionalmente para determinar la función putativa mediante la asociación con términos de Ontología Genética (GO) utilizando la herramienta de asociación de términos motivo-GO GOMO. La salida de MEME ahora contiene LOGOS de secuencia para cada motivo descubierto, así como botones para permitir que los motivos se envíen convenientemente a los algoritmos de escaneo de la base de datos de secuencias y patrón (MAST, FIMO y TOMTOM), o a GOMO, para su posterior análisis. La salida de GLAM2 contiene de manera similar botones para un análisis más detallado usando GLAM2SCAN y para volver a ejecutar GLAM2 con diferentes parámetros.

MEME Suite 5.3.3

► Motif Discovery

► Motif Enrichment

► Motif Scanning

► Motif Comparison

► Gene Regulation

► Manual

► Guides & Tutorials

► Sample Outputs

► File Format Reference

► Databases

► Download & Install


► Help

► Alternate Servers

► Authors & Citing

► Recent Jobs

Previous version 5.3.2



Tomtom

Motif Comparison Tool

Version 5.3.3

Tomtom compares one or more motifs against a database of known motifs (e.g., JASPAR). Tomtom will rank the motifs in the database and produce an alignment for each significant match (sample output for motif and JASPAR CORE 2014 database). See this Manual for more information.

Data Submission Form

Search one or more motifs against a motif database.

Input query motifs

Enter the motif(s) to compare to known motifs. ?

Type in motifs

Custom

Seleccionar archivo

Ningún archivo seleccionado ?

GGAACAGGAAGAATTCTTAGGGGCAAGAAACCAACTGGTGGCATTGAT

Select target motifs

Select a motif database or provide motifs to compare with. ?

Eukaryote DNA

DNA ?

Vertebrates (in vivo and in silico)

?

☐ Allow alphabet expansion ?

Run immediately

☐ Search with one motif (faster queue) ?

Input job details

(Optional) Enter your email address. ?

(Optional) Enter a job description. ?

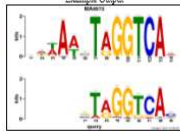
► Advanced options

Note: if the combined form inputs exceed 80MB the job will be rejected.

Start Search

Clear Input

Example Output



Version 5.3.3

Please send comments and questions to: memesuite@um.edu

Powered by Opal

[Home](#)
[Documentation](#)
[Downloads](#)
[Authors](#)
[Citing](#)

MEME version 4

ALPHABET= ACGT

strands: + -

Background letter frequencies (from unknown source):
A 0.250 C 0.250 G 0.250 T 0.250

MOTIF 1 GGAACAGGAAGAATTCTTAGGGGCAAGAAACCAACTGGTGGCATTGAT

letter-probability matrix: alength= 4 w= 50 nsites= 1 E= 0e+0

0.000000	0.000000	1.000000	0.000000
0.000000	0.000000	1.000000	0.000000
1.000000	0.000000	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000
0.000000	1.000000	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000
0.000000	0.000000	1.000000	0.000000
0.000000	0.000000	1.000000	0.000000
1.000000	0.000000	0.000000	0.000000
0.000000	0.000000	1.000000	0.000000
1.000000	0.000000	0.000000	0.000000
0.000000	0.000000	1.000000	0.000000
1.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	1.000000
0.000000	0.000000	0.000000	1.000000
0.000000	1.000000	0.000000	0.000000
0.000000	1.000000	0.000000	0.000000
0.000000	0.000000	0.000000	1.000000
1.000000	0.000000	0.000000	0.000000
0.000000	0.000000	1.000000	0.000000
0.000000	0.000000	1.000000	0.000000
0.000000	0.000000	1.000000	0.000000
0.000000	0.000000	1.000000	0.000000
0.000000	0.000000	1.000000	0.000000
0.000000	0.000000	1.000000	0.000000
0.000000	0.000000	1.000000	0.000000
0.000000	0.000000	1.000000	0.000000
0.000000	0.000000	1.000000	0.000000
0.000000	1.000000	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000

.....

EJERCICIO 2

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 predictions on sequence submitted from are:

```
## gff-version 2
## date Tue May 4 13:25:15 2021
## 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
human geneid_v1.2 First 157 286 9.81 + 0 human_1
human geneid_v1.2 Internal 10376 10458 1.45 + 2 human_1
human geneid_v1.2 Internal 12800 12857 0.89 + 0 human_1
human geneid_v1.2 Internal 15504 15655 -0.00 + 2 human_1
human geneid_v1.2 Internal 16764 16828 1.03 + 0 human_1
human geneid_v1.2 Internal 17225 17406 5.73 + 1 human_1
human geneid_v1.2 Internal 23771 23865 -1.35 + 2 human_1
human geneid_v1.2 Internal 25045 25142 2.96 + 0 human_1
human geneid_v1.2 Internal 26262 26281 2.17 + 1 human_1
human geneid_v1.2 Internal 27296 27427 2.70 + 2 human_1
human geneid_v1.2 Terminal 28008 28858 6.20 + 2 human_1
# Gene 2 (Forward). 4 exons. 141 aa. Score = 0.28
human geneid_v1.2 First 30518 30529 -2.92 + 0 human_2
human geneid_v1.2 Internal 30780 30932 0.68 + 0 human_2
human geneid_v1.2 Internal 31931 31994 2.93 + 0 human_2
human geneid_v1.2 Terminal 33682 33875 -0.40 + 2 human_2
Species:
Homo sapiens
Command:
geneid -P /soft/GeneID/geneid_1.2/human.param -G
/tmp/WebFiles/fastas/geneid29899.fasta
Running time:
0.12 secs
```

```
>human_1|geneid_v1.2_predicted_protein_1|622_AA
MAPAMQPAEIQFAQLASSEK GIRDRAVKKLRQYISVKTQRETGGFSQEELKIWKGLFY
CMWVQDEPLLQEELANTIAQLVHAVNNSAAQACVWFFSRIKVFLDVLMKEVLCPEQSPPN
GVRFFHFDIYLDLDEL SKVGGKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVD
QSPFVPEETMEEQKTKVGDGDL SAEIPENEVSLRRRAVSKKK TALGKNHSRKDGLSDERG
RDDCGTFEDTGPLLQFDYKAVADRLLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQ
LSFAEDISADEDDQILSQGKHKKKGK NLLKTNLEKEKGSRVFCVEEEDSESSLQKRRRK
KKKKHHLQPENPGPGAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEH
PPAVPMHNKRKRPRKKS PRAHREMLES AVLPPEDMSSQSGPSGSHPQGRGSPGTGGAQLLK
RKRKLGVVPVNGSLSTPAWPPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLEL
CGLPSQKTASLKKRKKMRVMSNLVEHNGVLESEAGQPQALVRWEHPQASSPQRHSLASMG
LHCLLRGRVGAGGQASGLSSS
>human_2|geneid_v1.2_predicted_protein_2|141_AA
MKIKGSSGTCSSLKKQKLRAESDFVKFDTPLPKPLFFRRAKSSTATHPPGPAVQLNKTP
SSSKKVTFLNRNMTAEFKKTDK SILVSPTGPSRVAFDPEQKPLHGV LKTPTSSPASSPL
VAKKPLTTTPRRRPRAMDF
```

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

```
>/tmp/05_04_21-09:27:48.fasta|GENSCAN_predicted_peptide_1|897_aa
MAPAMQPAEIQFAQLASSEK GIRDRAVKKLRQYISVKTQRETGGFSQEELKIWKGLFY
CMWVQDEPLLQEELANTIAQLVHAVNNSAAQH LFIQTFWQTMNREWKGIDRLRLDKYYML
IRLVLRQSFEVLKRNGWEESRIKVFLDVLMKEVLCPEQSPPNGVRFFHFDIYLDLDEL SKVG
GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVDQSPFVPEETMEEQKTKVG
DGDL SAEIPENEVSLRRRAVSKKK TALGKNHSRKDGLSDERG RDDCGTFEDTGPLLQFDY
```

KAVADRLLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQLSFAEDISADEDDQILSQ
 GKHKKKGNKLEKTNLEKEKGKQELQGALGGGCLMTTRDLWFLPLSPKISGNGTISVPYV
 FINGQKEGFQSQLGMEEVGPDDKGSRVFCVEEEDSESSLQKRRRKKKKKHHLQPENPGPG
 GAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK
 KSPRAHREMLES AVLPPEDMSQSGPSGSHPOGPRGSPTGGAQLLKRKRKLGVVPVNGSGL
 STPAWPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRK
 KMRVMSNLVEHNGVLESEAGQPQALAAHLNLPEPPVCRQRHWAHTSESQVRDPVSLWVA
 VSCCTRNECPGPASVVLCKPELCRMEGLSASAVRKTAGRRGSSGTCSSLKKQKLRAESD
 FVKFDTPLPKPLFFRAKSSTATHPPGPAVQLNKTPSSSKKVTFTGLNRNMTAEFKKTDK
 SILVSPTGPSRVAFDPEQKPLHGVLTPTTSSPASSPLVAKKPLTTTPRRRPRAMDFF

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

Predicted protein(s):

>FGENESH: 1 16 exon (s) 157 - 33875 758 aa, chain +
 MAPAMQPAEIQFAQLASSEKGI RDRVVKLRQYISVKTQRETGGFSQEELKIWKGLFY
 CMWVQDEPLLQEELANTIAQLVHAVNNSAAQHLLFIQTFWQTMNREWKGIDRLRLDKYYML
 IRLVLRQSFEVLKRNGWEESRIKVFLDVLMKEVLCPEQSQSPNGVRFHFIDYLDLSKVG
 GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVDQSPFVPEETMEEQTKVG
 DGDLSAEEIPENEVSLRRAVSKKKTALGKNHSRKDGLSDERGRDDCGTFEDTGPLLQFDY
 KAVADRLLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQLSFAEDISADEDDQILSQ
 GKHKKKGNKLEKTNLEKEKSRVFCVEEEDSESSLQKRRRKKKKKHHLQPENPGPGGAA
 PSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRKKSP
 RAHREMLES AVLPPEDMSQSGPSGSHPOGPRGSPTGGAQLLKRKRKLGVVPVNGSGLSTP
 AWPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRKKMR
 VMSNLVEHNGVLESEAGQPQALGSSGTCSSLKKQKLRAESDFVKFDTPLPKPLFFRAK
 SSTATHPPGPAVQLNKTPSSSKKVTFTGLNRNMTAEFKKTDKSILVSPTGPSRVAFDPEQK
 PLHGVLTPTTSSPASSPLVAKKPLTTTPRRRPRAMDFF

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.

```
# Percent Identity Matrix - created by Clustal2.1
#
#
1: geneid      100.00   94.35   99.30
2: Genscan     94.35  100.00  100.00
3: Fgenesh     99.30  100.00  100.00
```

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

```
# 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

# 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
```

FGENESH

G Str	Feature	Start	End	Score	ORF	Len
1 +	1 CDSf	157 -	286	28.30	157 -	285
1 +	2 CDSi	10376 -	10458	7.43	10378 -	81
1 +	3 CDSi	12800 -	12857	6.33	12800 -	57
1 +	4 CDSi	14362 -	14447	3.34	14364 -	84
1 +	5 CDSi	15128 -	15189	2.40	15128 -	60
1 +	6 CDSi	15526 -	15655	6.39	15527 -	129
1 +	7 CDSi	16764 -	16828	6.62	16764 -	63
1 +	8 CDSi	17225 -	17406	12.24	17226 -	180
1 +	9 CDSi	23771 -	23865	2.10	23773 -	93
1 +	10 CDSi	25045 -	25142	0.60	25045 -	96
1 +	11 CDSi	26262 -	26281	-1.21	26263 -	18
1 +	12 CDSi	27296 -	27427	8.29	27298 -	129
1 +	13 CDSi	28008 -	28732	33.29	28010 -	723
1 +	14 CDSi	30780 -	30932	9.89	30780 -	153
1 +	15 CDSi	31931 -	31994	13.04	31931 -	63
1 +	16 CDSi	33682 -	33875	1.90	33684 -	192
1 +	PoLA	33923		-4.47		

GENSCAN

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Init	+	162	291	130	2	1	107	80	324	0.752	33.81
1.02	Intr	+	10381	10463	83	2	2	94	92	26	0.829	2.96
1.03	Intr	+	12805	12862	58	0	1	97	99	62	0.963	6.66
1.04	Intr	+	14367	14452	86	1	2	47	95	49	0.678	1.04
1.05	Intr	+	15133	15194	62	0	2	53	86	51	0.694	-1.07
1.06	Intr	+	15531	15660	130	0	1	27	99	108	0.642	6.50
1.07	Intr	+	16769	16833	65	1	2	78	83	73	0.995	3.32
1.08	Intr	+	17230	17411	182	1	2	77	91	192	0.962	17.91
1.09	Intr	+	23776	23870	95	2	2	37	94	55	0.688	0.68
1.10	Intr	+	25050	25147	98	2	2	64	26	129	0.640	3.11
1.11	Intr	+	26267	26286	20	2	2	91	100	-1	0.600	-2.35
1.12	Intr	+	27301	27432	132	2	0	41	121	120	0.872	11.22
1.13	Intr	+	27668	27856	189	0	0	51	67	92	0.625	2.96
1.14	Intr	+	28013	28737	725	0	2	85	95	470	0.762	38.55
1.15	Intr	+	30241	30385	145	0	1	71	48	71	0.368	1.26
1.16	Intr	+	30594	30676	83	1	2	30	51	91	0.478	-1.04
1.17	Intr	+	30785	30937	153	1	0	100	101	109	0.999	13.67
1.18	Intr	+	31936	31999	64	0	1	114	131	52	0.996	10.39
1.19	Term	+	33687	33880	194	1	2	52	55	187	0.999	9.38
1.20	PlyA	+	35505	35510	6							-0.45

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.

```

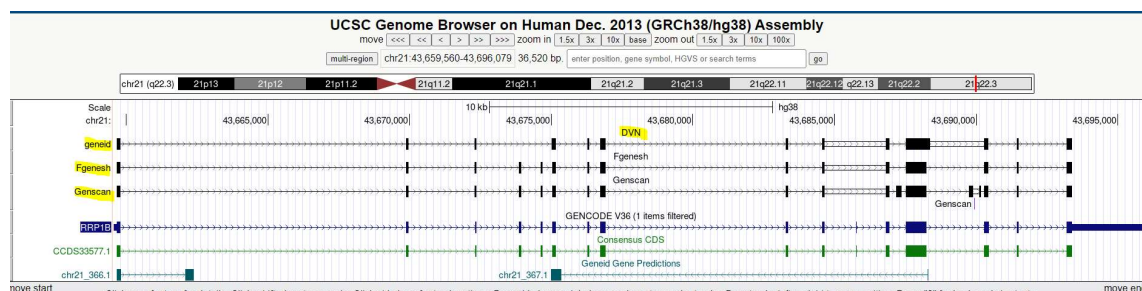
00000001 gccgcggcgccctttgtgacgccatcagcccgcgcgcgcgcgcgcgcct 00000050
>>>>>>> |||||  >>>>>>>
43659509 gccgcggcgccctttgtgacgccatcagcccgcgcgcgcgcgcgcgcct 43659558
.
.
.
00037451 ggtgacaaaatgagaccctgtctttaaaaaaaaaaaaaagccctagagg 00037500
>>>>>>> |||||  >>>>>>>
43696959 ggtgacaaaatgagaccctgtctttaaaaaaaaaaaaaagccctagagg 43697008

00037501 aagaaggaaatctgttgtaatgtattattttaaaatgtccagttttcaaca 00037550
>>>>>>> |||||  >>>>>>>
43697009 aagaaggaaatctgttgtaatgtattattttaaaatgtccagttttcaaca 43697058

00037551 aaaacaaggaagacattcaa 00037571
>>>>>>> |||||  >>>>>>>
43697059 aaaacaaggaagacattcaa 43697079

```

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):



9. Para acabar, efectuaad 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.

```

# Percent Identity Matrix - created by Clustal2.1
#
#
1: geneid      100.00  94.35  99.30  99.30
2: Genscan    94.35  100.00  100.00  100.00
3: Fgenesh    99.30  100.00  100.00  100.00
4: RRP1B      99.30  100.00  100.00  100.00

```

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.

La página web <http://www-gsd.lbl.gov/vista/> sirve como portal para acceder al conjunto de herramientas de VISTA.

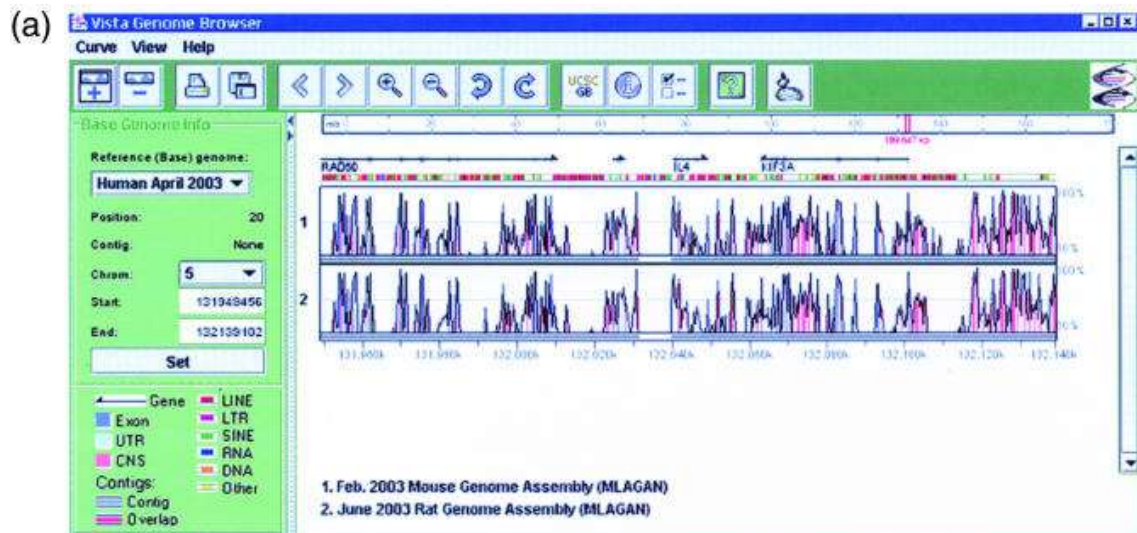
Uno de ellos es VISTA Browser, que permite al usuario ver alineaciones del genoma completo precalculadas de muchas especies. Hay tres servidores VISTA, GenomeVISTA, mVISTA y rVISTA, que permiten al usuario enviar secuencias de ADN para su análisis. Para GenomeVISTA, el usuario envía una única secuencia (borrador o terminada) que se compara con conjuntos completos de genoma completo disponibles públicamente. mVISTA es el programa original, diseñado para la comparación de secuencias ortólogas de diferentes especies.

rVISTA combina una búsqueda en la base de datos de sitios de unión de factores de transcripción con un análisis de secuencia comparativo. El programa Phylo-VISTA, un nuevo miembro de la familia de herramientas VISTA, permite al usuario visualizar múltiples datos de alineación de secuencias enviados mientras se tienen en cuenta las relaciones filogenéticas entre secuencias. El sitio web de VISTA también proporciona acceso a los análisis comparativos del conjunto de genes cardiovasculares, estudiado por el Programa de Berkeley para Aplicaciones Genómicas (PGA).

Las páginas de VISTA brindan una gran ayuda para seleccionar un tipo de análisis, encontrar los parámetros óptimos para un proyecto en particular y navegar por el sitio web.

Los gráficos de visualización muestran secuencias conservadas entre humanos y ratones (panel superior) y humanos y ratas (panel inferior) basadas en la alineación múltiple de tres genomas usando MLAGAN. El nivel de conservación (eje vertical) se muestra en las coordenadas de la secuencia humana (eje horizontal). Las regiones conservadas por encima del nivel de 70% / 100 pb se resaltan debajo de la curva, donde el rojo indica una región no codificante conservada, el azul, un exón conservado y el turquesa, una región no traducida. Los detalles de la pantalla se dan en la leyenda en el lado izquierdo del gráfico. El botón 'UCSC' abre otra ventana que contiene la vista del navegador UCSC reflejada del mismo intervalo con pistas VISTA integradas.

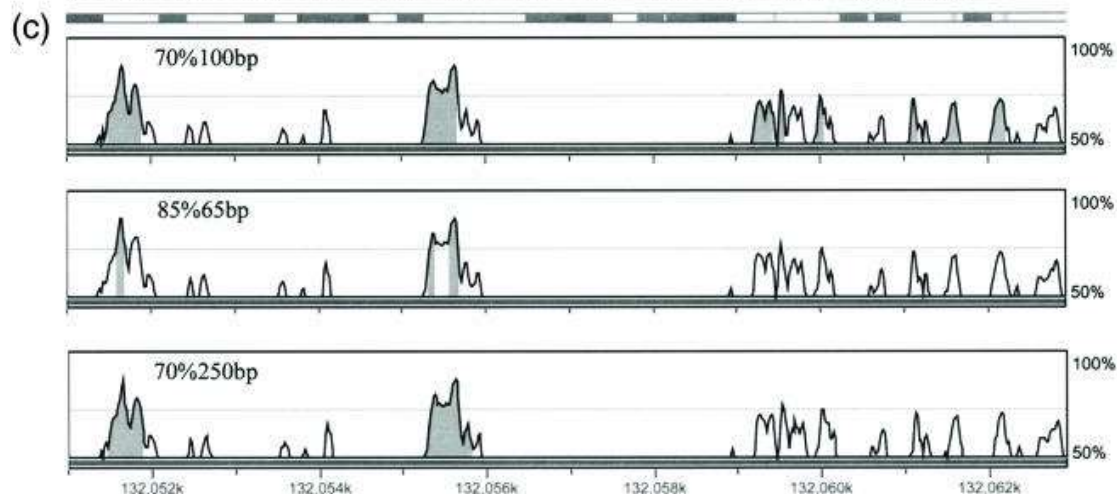
El navegador cuenta con una amplia ayuda en línea. **(b)** VISTA Browser generó una lista de elementos conservados de humanos / ratones en la región *KIF3A* con sus coordenadas en la secuencia humana (números sin corchetes) y del ratón (números entre corchetes), longitudes e identidades porcentuales y anotación funcional. Se muestran los elementos del comienzo del intervalo de 180 kb en *RAD50*. **(c)** Fragmento genómico corriente arriba del gen *KIF3A* que contiene múltiples elementos no codificantes conservados. El número de elementos conservados (coloreados) depende de la identidad porcentual seleccionada por el usuario y los límites de longitud que se muestran arriba de cada gráfico.



(b) Criteria: 70% identity over 100 bp

***** Conserved Regions - Human (Mouse) *****

131952851	(54292441)	to	131953108	(54292210)	=	258bp	at	69.4%	noncoding
131954117	(54291314)	to	131954245	(54291186)	=	129bp	at	89.9%	exon
131954246	(54291185)	to	131954339	(54291091)	=	98bp	at	71.4%	noncoding
131954479	(54290969)	to	131954644	(54290804)	=	166bp	at	87.3%	exon
131954759	(54289473)	to	131954891	(54289341)	=	135bp	at	71.1%	noncoding
131955242	(54288804)	to	131955435	(54288611)	=	194bp	at	89.7%	exon
131956186	(54288222)	to	131956392	(54288016)	=	207bp	at	73.4%	exon
131957525	(54284506)	to	131957654	(54284379)	=	130bp	at	70.0%	noncoding
131957779	(54284180)	to	131957961	(54283998)	=	183bp	at	85.2%	exon



Con respecto a cómo nos ayudaría contar con esta información, podríamos decir que para reducir sustancialmente el ruido en estas representaciones podemos emplear la comparación con otras secuencias relacionadas, aportando nueva información sobre la conservación regulatoria.

Existen diversas fuentes de conocimiento que nos permiten identificar regiones de genes que hipotéticamente comparten una colección de sitios de unión:

1. Genes que desempeñan funciones similares en el organismo.
2. Genes que en experimentos de expresión a gran escala poseen patrones parecidos de activación.
3. Genes ortólogos pertenecientes a múltiples especies.

En todos los casos, asumiendo que funciones similares deben ser implementadas mediante combinatorias de motivos comunes, la comparación de secuencias nos ayuda a focalizar nuestro interés únicamente sobre ciertas regiones conservadas para reforzar aquellas predicciones obtenidas con matrices de pesos. De todas estas alternativas, el análisis de regiones reguladoras que incluye información sobre conservación filogenética es el método que arroja resultados más prometedores.