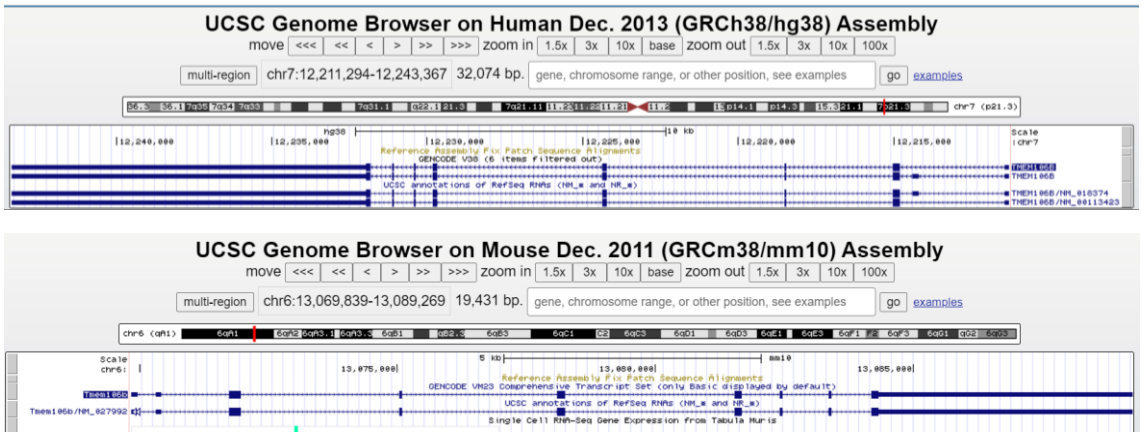
Exercici 1. Estratègies d'alineament

1. El programa CLUSTAL realitza alineaments globals de dos o més seqüències. Connecteu-vos al servidor implementat en l'EBI per comparar la seqüència CDS del gen *TMEM106B* obtinguda des de RefSeq (UCSC) per a humà i ratolí en la PAC1 anterior (hg38 i mm10, respectivament).

Per realitzar aquest exercici cal accedir a la web UCSC i cercar les dos seqüències esmentades:



Imatge 1 Seqüències CDS del gen TMEM106B obtinguda des de RefSeq (UCSC) per a humà i ratolí.

A partir d'això es necessari obtenir les dos seqüències CDS de les seqüències corresponents a partir de RefSeq:

```
>hg38_refGene_NM_018374 range=chr7:12214811-12231975 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ATGGGAAAGTCTCTTTCTCATTTGCCTTTGCATTCAAGCAAAGAAGATGC
TTATGATGGAGTCACATCTGAAAACATGAGGAATGGACTGGTTAATAGTG
AAGTCCATAATGAAGATGGAAGAAATGGAGATGTCTCTCAGTTTCCATAT
GTGGAATTTACAGGAAGAGATAGTGTCACCTGCCCTACTTGTCAGGGAAC
AGGAAGAATTCCTAGGGGGCAAGAAAACCAACTGGTGGCATTGATTCCAT
ATAGTGATCAGAGATTAAGGCCAAGAAGAACAAAGCTGTATGTGATGGCT
TCTGTGTTTGTCTGTCTACTCCTTTCTGGATTGGCTGTGTTTTTCCTTTT
CCCTCGCTCTATCGACGTGAAATACATTGGTGTAAAATCAGCCTATGTCA
GTTATGATGTTCAGAAGCGTACAATTTATTTAAATATCACAAACACACTA
AATATAACAAACAATAACTATTACTCTGTCGAAGTTGAAAACATCACTGC
CCAAGTTCAATTTTCAAAAACAGTTATTGGAAAGGCACGCTTAAACAACA
TAACCATTATTGGTCCACTTGATATGAAACAAATTGATTACACAGTACCT
ACCGTTATAGCAGAGGAAATGAGTTATATGTATGATTTCTGTACTCTGAT
ATCCATCAAAGTGCATAACATAGTACTCATGATGCAAGTTACTGTGACAA
CAACATACTTTGGCCACTCTGAACAGATATCCCAGGAGAGGTATCAGTAT
GTCGACTGTGGAAGAAACACAACTTATCAGTTGGGGCAGTCTGAATATTT
AAATGTACTTCAGCCACAACAGTAA
```

Imatge 2 Seqüència CDS del gen TMEM106B de hg38

```
>mm10_refGene_NM_027992 range=chr6:13071744-13084326 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ATGGGAAAGTCTCTTTCTCACTTACCTTTGCATTCAAATAAAGAAGATGG
CTATGATGGCGTTACATCGACAGACAATATGAGAAATGGATTGGTTAGCA
GTGAAGTGCACAACGAAGACGGAAGAAATGGAGATGTCTCTCAGTTCCCA
TATGTGGAATTTACTGGAAGAGATAGTGTCACTTGTCCCACTTGCCAAGG
AACAGGAAGAATTCCTAGGGGACAAGAAAACCAACTGGTGGCATTGATTC
CATATAGTGATCAGCGGTTACGGCCAAGAAGAACAAAGCTGTATGTGATG
GCGTCTGTGTTTGTCTGCCTGCTCCTGTCTGGATTGGCTGTGTTTTTTCT
TTTCCCTCGATCTATTGAGGTGAAGTACATTGGAGTAAAATCAGCCTATG
TCAGCTACGACGCTGAAAAGCGAACCATATATTTAAATATCACGAACACA
CTAAATATAACAAATAATAACTATTATTCTGTTGAAGTTGAAAACATCAC
TGCTCAAGTCCAGTTTTCAAAAACCGTGATTGGAAAGGCTCGTTTAAACA
ACATAACTAACATTGGCCCACTTGATATGAAGCAGATTGATTATACGGTA
CCCACAGTTATTGCAGAGGAAATGAGTTACATGTATGATTTCTGTACACT
GCTCTCCATCAAAGTGCACAACATAGTACTCATGATGCAAGTTACTGTAA
CAACAGCATACTTTGGACACTCTGAGCAGATATCTCAGGAAAGGTACCAG
TATGTCGACTGTGGAAGGAACACGACTTACCAGTTGGCCCAGTCTGAGTA
TCTAAATGTCCTTCAGCCACAACAATAA
```

Imatge 3 Seqüència CDS del gen TMEM106B de mm10

Per fer la comparació de seqüencies s'utilitzarà el programa Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) en el que s'introdueixen les dues seqüències CDS obtingudes.

En la següent imatge mostrem l'alineament global resultant en el que observem una similaritat molt alta 88,85%. Aquest fet es dedueix ja que es tracta de dos gens ortòlegs.

Imatge 4 Resultat de comparar les dos seqüències CDS de hg38_refGene_NM_018374 i mm10_refGene_NM_027992.

2. Repetiu aquest mateix alineament global, utilitzant ara les respectives proteïnes d'aquest gen en cada espècie (que prèviament heu de tornar a recuperar de l'entrada de RefSeq). Valoreu el grau d'homologia entre aquestes dues seqüències.

```
>NP_060844 length=274
MGKSLSHLPLHSSKEDAYDGVTSENMRNGLVNSEVHNEDGRNGDVSQFPY
VEFTGRDSVTCPTCQGTGRIPRGQENQLVALIPYSDQRLRPRRTKLYVMA
SVFVCLLLSGLAVFFLFPRSIDVKYIGVKSAYVSYDVQKRTIYLNITNTL
NITNNNYYSVEVENITAQVQFSKTVIGKARLNNITIIGPLDMKQIDYTVP
TVIAEEMSYMYDFCTLISIKVHNIVLMMQVTVTTTYFGHSEQISQERYQY
VDCGRNTTYQLGQSEYLNVLQPQQ
```

Imatge 5 Seqüencia de proteïnes de hg38

```
>NP_082268 length=275
MGKSLSHLPLHSNKEDGYDGVTSTDNMRNGLVSSEVHNEDGRNGDVSQFP
YVEFTGRDSVTCPTCQGTGRIPRGQENQLVALIPYSDQRLRPRRTKLYVM
ASVFVCLLLSGLAVFFLFPRSIEVKYIGVKSAYVSYDAEKRTIYLNITNT
LNITNNNYYSVEVENITAQVQFSKTVIGKARLNNITNIGPLDMKQIDYTV
PTVIAEEMSYMYDFCTLLSIKVHNIVLMMQVTVTTAYFGHSEQISQERYQ
YVDCGRNTTYQLAQSEYLNVLQPQQ
```

Imatge 6 Seqüencia de proteïnes de mm10

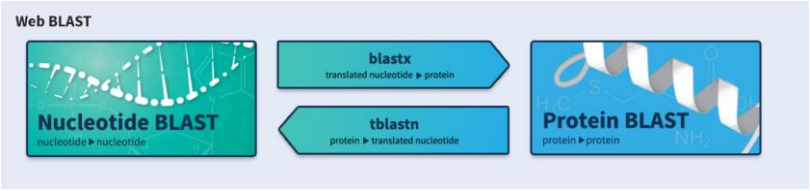
```
NP_060844      MGKSLSHLPLHSSKEDAYDGVTS-ENMRNGLVNSEVHNEDGRNGDVSQFPYVEFTGRDSV    59
NP_082268      MGKSLSHLPLHSNKEDGYDGVTSTDNMRNGLVSSEVHNEDGRNGDVSQFPYVEFTGRDSV    60
               ***********.***.****** :*******.*************************

NP_060844      TCPTCQGTGRIPRGQENQLVALIPYSDQRLRPRRTKLYVMASVFVCLLLSGLAVFFLFPR    119
NP_082268      TCPTCQGTGRIPRGQENQLVALIPYSDQRLRPRRTKLYVMASVFVCLLLSGLAVFFLFPR    120
               ************************************************************

NP_060844      SIDVKYIGVKSAYVSYDVQKRTIYLNITNTLNITNNNYYSVEVENITAQVQFSKTVIGKA    179
NP_082268      SIEVKYIGVKSAYVSYDAEKRTIYLNITNTLNITNNNYYSVEVENITAQVQFSKTVIGKA    180
               **:**************.:****************************************

NP_060844      RLNNITIIGPLDMKQIDYTVPTVIAEEMSYMYDFCTLISIKVHNIVLMMQVTVTTTYFGH    239
NP_082268      RLNNITNIGPLDMKQIDYTVPTVIAEEMSYMYDFCTLLSIKVHNIVLMMQVTVTTAYFGH    240
               ****** ****************************:*****************:****

NP_060844      SEQISQERYQYVDCGRNTTYQLGQSEYLNVLQPQQ      274
NP_082268      SEQISQERYQYVDCGRNTTYQLAQSEYLNVLQPQQ      275
               ********************.************
```

Imatge 7 Alineament global entre les dos seqüencaies de proteïnes.

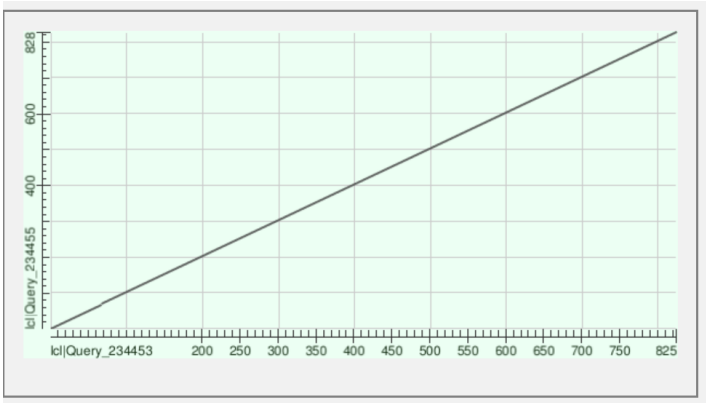
En aquest alienament global s'observa un percetantge de identificació encara més elevat, del 95,99%.

3. El programa BLAST realitza alineaments locals. Connecteu-vos a BLAST, en el servidor principal del NCBI, per buscar quina versió d'aquest programa heu d'utilitzar per alinear dues seqüències. Realitzeu ara l'alineament local de les dues regions CDS del gen *TMEM106B*.

Per alinear dues seqüències de les regions CDS del gen TMEM106 de humà i ratolí caldrà utilitzar la versió blastn.



Imatge 8 Versions del programa BLAST



Imatge 9 Matriu de punts resultants de l'alineament local de les seqüències de hg38 (eix X) i mm10 (eix Y)

En aquesta matriu resultant cal apreciar el resultat de la alta similaritat entre les dues seqüencies.

Score	Expect	Identities	Gaps	Strand
1000 bits(541)	0.0	733/828(89%)	3/828(0%)	Plus/Plus

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

Query  61   GTCACAT---CTGAAAACATGAGGAATGGACTGGTTAATAGTGAAGTCCATAATGAAGAT  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

Query  418  CGTACAATTTATTTAAATATCACAAACACACTAAATATAACAAACAATAACTATTACTCT  477
            || || || |||||||||||||| |||||||||||||||||| |||||||||| |||
Sbjct  421  CGAACCATATATTTAAATATCACGAACACACTAAATATAACAAATAATAACTATTATTCT  480
```

Imatge 10 Resultat de l'alineament.

En aquest alineament es pot veure l'alta identificació entre les dues seqüències 733/828 (89%).

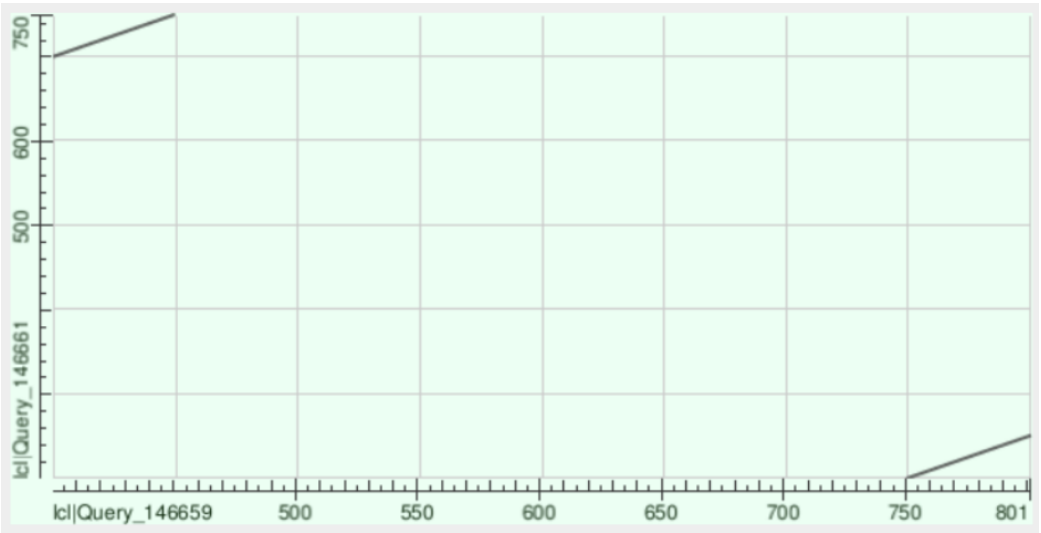
4. Ara utilitzeu el servidor de CLUSTAL per alinear globalment la seqüència *genomicA.txt* i la seqüència *genomicB.txt* que trobareu adjuntes a aquest enunciat.

```
genomicA    cagaagaattgcttgaaccagggaggtggaggttgcagtgagcagagatcacgccactgc      60
genomicB    -----------------gctgggatg--tggggagcagtgttctgaggctgagcag-gac      40
                            * **** *    **  ******  * ***     **     *

genomicA    actcctgcttaagtgacagagtgagactccatctcaaaaaaaaaaaaaaaattcctatta      120
genomicB    agtgaggccttgggcctggcct----------ctgaaaccatttttttccacctaggcctc      90
            * *   ** *  *     *  *          ** ***  *        *  *     *

genomicA    tgtgcttgagtaataccacccactctggcaaatcttaaaaaagctcttggccgggtgcag      180
genomicB    tgagcctgtgtcctataacttattgcaggctgttagaagc------------aggcagac      138
            ** ** ** **  **  **  * *   *    *   **              **   *

genomicA    tggctcatgcctgtaatccccagaagaattgcttgaaccagggaggtggaggttgcagtg      240
genomicB    tactttctggatgctttgctgcttagaatttttttctgcca-------------------      179
            *   *  **  **   * *     ******  **    *

genomicA    agcagagatcacgccactgcactcctgcttaagtgacagagtgagactccatctcaaaaa      300
genomicB    ----ga-----------tatcctaggtcatcactctATGAGTGTGGATCCAGCTTGT---      221
                **           *   **    * * * *    ***** *  **** **

genomicA    aaaaaaaaaaattcctattatgtgcttgagtaataccacccactctggcaaatcttaaaa      360
genomicB    -------------------------------------CCCCAAAGCTTGCCTTGCTTTGAA      245
                                                 * ** *  ** **    ***  **
```

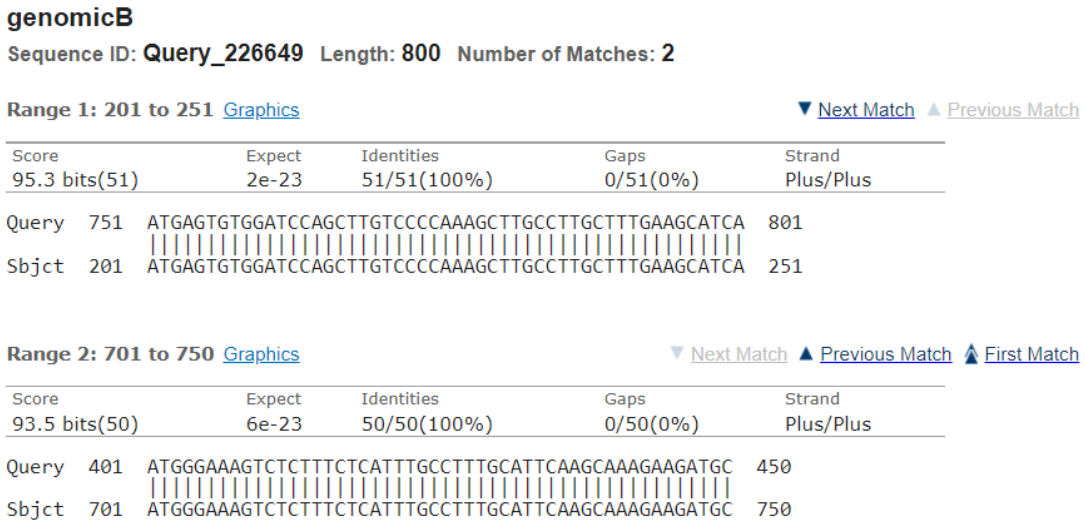
Imatge 11 Alineament global de les seqüències genomicA.txt i genomicB.txt

Aquestes dues seqüencies presenten un baix nivell de similaritat del 44% de identitat.

5. Procediu ara a efectuar l'alineament local amb BLAST de la seqüencia genòmica *genomicA.txt* i la seqüència *genomicB.txt* adjuntades amb l'enunciat.



Imatge 12 Matriu resultant de les seqüencies genomicA.txt i genomicB.txt



Imatge 13 Fragments conservats entre genomicA.txt i genomicB.txt

Fet l'alineament local observem dos fragments conservats entre genomicA.txt i genomicB.txt en diferents ubicacions.

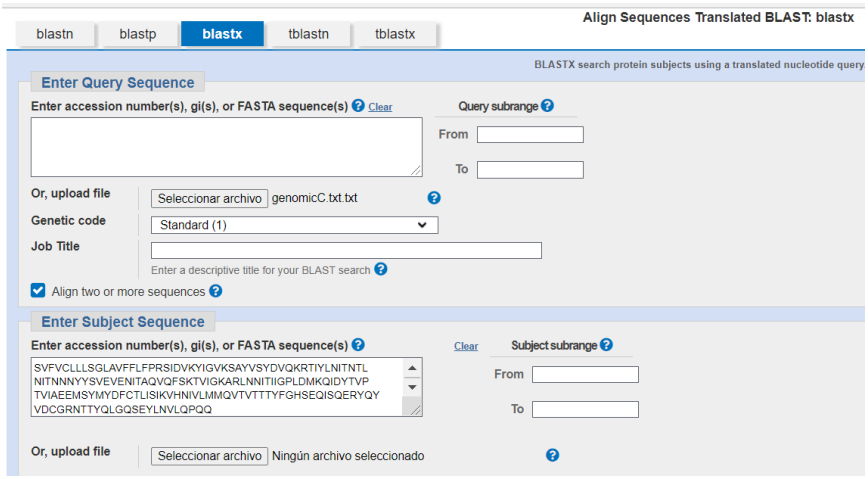
6. Compareu els resultats de l'alineament global i local en els dos casos anteriors (2 CDSs o les seqüències *genomicA.txt* i *genomicB.txt)*. Decidiu quin dels dos programes provats és més adequat per a cada cas en funció de l'estratègia emprada.

En el cas de l'alineament per comparar la seqüència CDS del gen TMEM106B obtinguda des de RefSeq (UCSC) per a humà i ratolí, **l'alineament global** és el més adequat, ja que aquestes dos seqüències codifiquen elements amb un funció biològica semblant amb una estructura i tamany gairebé similar, hg38_refGene_NM_018374 (825 nucleòtids) i mm10_refGene_NM_027992 (828 nucleòtids). En funció de la seva distancia evolutiva entre aquestes dues espècies comparades, la semblança final serà més o menys accentuada. Per tant per fer aquestes comparacions, és precís calcular un alineament global que recobreixi la totalitat de les seqüències corresponents, permeten associar als dos gens a cada residu el seu equivalent en l'altre organisme. Amb aquest alineament també ens permetrà observar el grau d'homologia entre les dos seqüències.

En canvi en el cas de l'alineament de les dos seqüències adjuntes (genomaA.txt i genomaB.txt), la comparació més adequada vindria donada per un **alineament local** realitzat amb BLASTN, ja que en aquesta situació pretenem distingir aquells fragments més similars de la resta intentant obtenir un fragment parcialment conservat entre les dos seqüències, realitzant exclusivament la correspondència entre aquells fragments que tinguin una coincidència màxima de caràcters, descartant la resta de regions al llarg de les seqüències que no presentin una mínima similaritat.

7. Uns investigadors que treballen amb el genoma del pollastre (*chicken*) ens envien la seqüència adjunta *genomicC.txt*, doncs sospiten que la forma ortòloga del nostre gen *TMEM106B* està codificada en el seu interior. Decidiu quina versió de BLAST heu d'utilitzar per validar aquesta hipòtesi amb la proteïna humana (que teniu de passos previs), anotant la seva homòloga en aquesta regió genòmica de pollastre. En cas de resposta afirmativa, interpreteu el grau d'homologia resultant entre ambdues proteïnes.

En aquest cas es vol relacionar un fragment genòmic i una proteïna, per tant caldrà utilitzar la versió del programa que realitzi la traducció de la seqüència genòmica a proteïna per seguidament comparar aquestes dues proteïnes. Per tant utilitzarem la versió BLASTX.



Imatge 14 Query: genomicC.txt comparat amb la proteïna humana.

	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
	135	497	3%	9e-39	85.14%	274	Query_28977

Imatge 15 Grau d'homologia de les seqüencies estudiades: 85,14%.

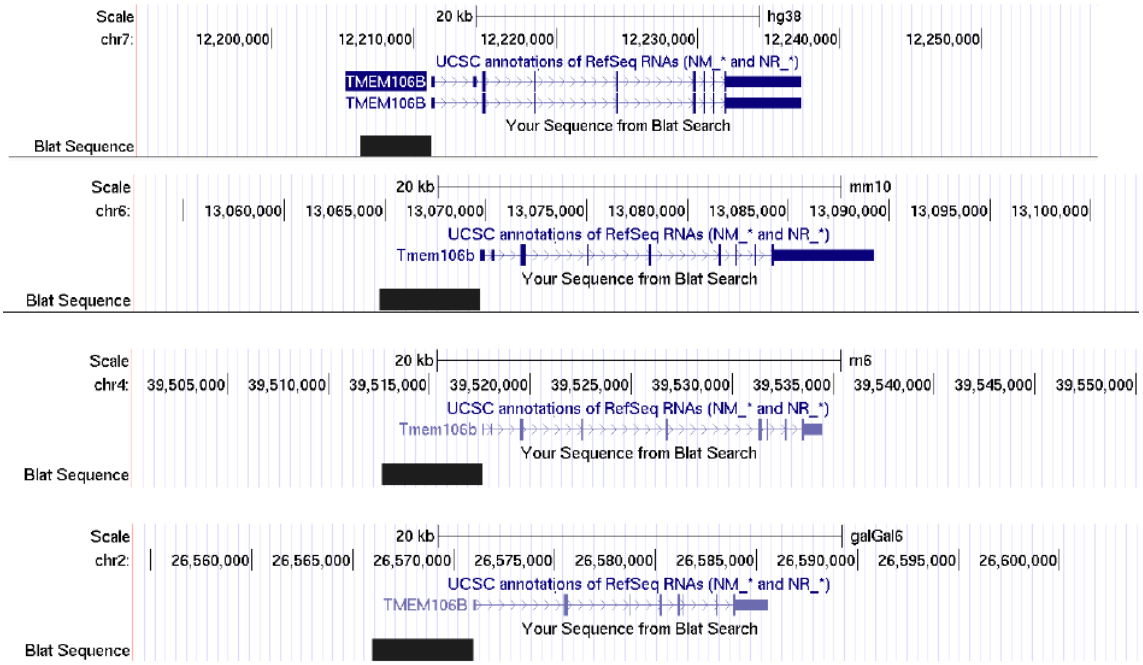
Observant el percentatge de 85% de identitat, podem afirmar que aquestes dues proteïnes són homòlogues (mateix gen, espècies diferents).

8. El programa MEME representa una família alternativa d'eines bioinformàtiques per comparar seqüències. Definiu en poques paraules quins tipus de tasca realitza aquesta aplicació i com pot ser emprat dins de l'àrea d'estudi de la regulació gènica mitjançant factors de transcripció:

MEME es tracta d'una eina bioinformàtica utilitzada per capturar motius conservats en diferents graus, situats en conjunts de seqüències, sense que aquests hagin d'estar situats en les mateixes posicions. Compara conjunts de seqüències que hipotèticament comparteixin algun mecanisme de regulació. En aquest programa podem observar un mapa interactiu del pipeline de treball.

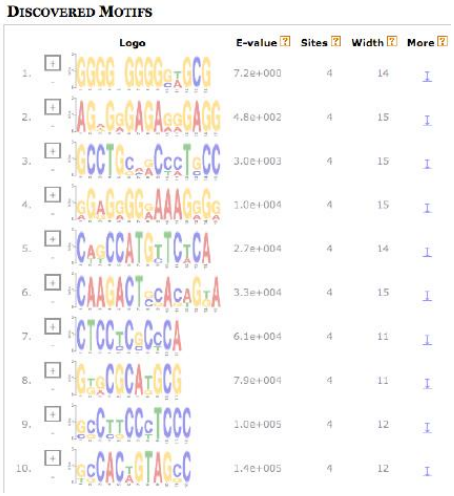
9. Anem a estudiar la regulació transcripcional del nostre gen *TMEM106B* al llarg de l'evolució. En primer lloc, emprant el navegador genòmic d'UCSC i les anotacions de RefSeq, heu d'extreure la regió promotora del gen (seleccioneu 5000 nucleòtids de longitud just abans de l'inici de transcripció del gen en cada espècie) per a aquestes espècies: humà (hg38), ratolí (mm10), rata (rn6) i pollastre (galgal6).

Utilitzarem la interfície de RefSeq per seleccionar 500 nucleòtids previs abans de l'inici de la transcripció del gen TSS. Seguidament podem utilitzar BLAT per tal de comprovar que la seqüència que hem seleccionat sigui exactament la mateixa amb aquella zona.



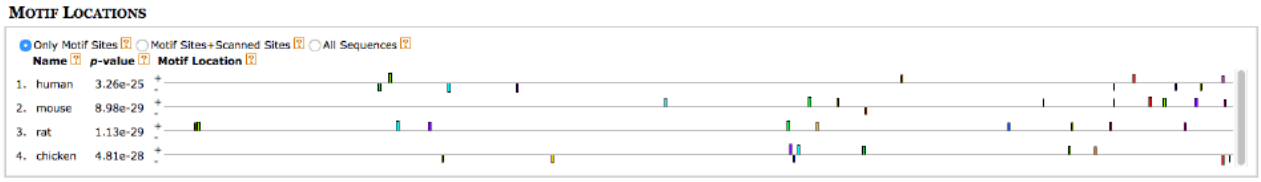
Imatge 16 hg38, mm10, rn6, galGal6 (respectivament)

10. En segon lloc, empreu el programa MEME per comparar aquestes quatre seqüències ortòlogues. Busquem els 10 millors motius que posseeixin una longitud entre 5 i 15 parells de bases. Exploreu quina funció pot jugar el programa TOMTOM integrat dins de la *suite* de programes MEME i efectueu una prova amb algun dels motius identificats.



Imatge 17 Llista MEME dels 10 motius resultants

Cap d'aquest 10 motius resulten significatius.

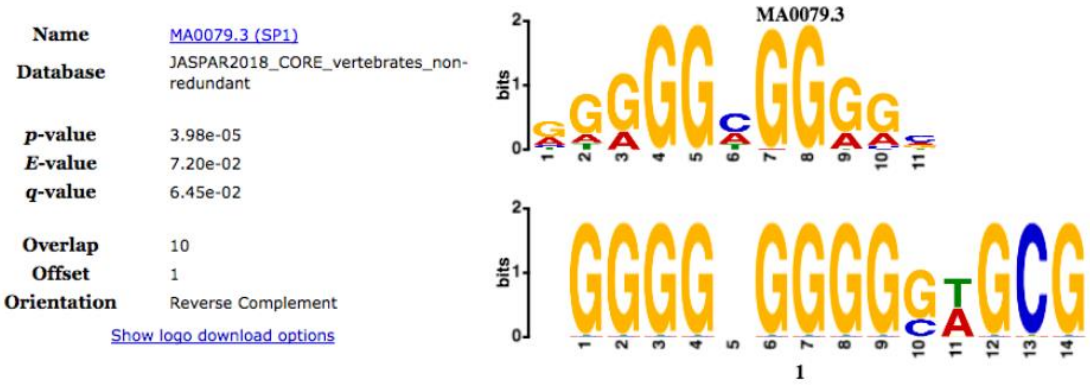


Imatge 18 Distribució de les seqüències promotores.

Amb el programa TOMTOM integrat dins de MEME podem cercar motius coneguts de unió a factor de transcripció coincidint amb motius ja identificats, sense la necessitat de utilitzar catàlegs de models de predicció.

Per exemple:

Imatge 19 Primer motiu del llistat.



Imatge 20 Motiu de TOMTOM i el nostre primer motiu.

Exercici 2. Anotació computacional de gens

Estem col·laborant amb un laboratori de biologia molecular que sospita que la seqüència *anonima.fa* codifica un gen humà. Aquest fragment genòmic està representat en el format FASTA habitual amb una capçalera inicial i la seqüència a continuació (adjunt a l'enunciat):

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

1. Desitgem conèixer les coordenades dels exons que constitueixen el gen codificat en aquesta seqüència. Com a primer pas del nostre protocol d'anotació, heu d'utilitzar el programa GENEID per recuperar el millor gen identificat computacionalment en aquesta regió del genoma humà:

```
## 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
```

Imatge 21 Coordenades del gen codificat en anonima.fa

2. Com a segon component del nostre *pipeline*, heu d'emprar GENSCAN per recuperar el gen codificat internament en aquesta seqüència humana:

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. Finalment, com a tercer component del procés, utilitzeu el programa FGENESH per identificar també la predicció d'aquest sistema:

```
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
```

Imatge 22 Prediction of potential genes in Homo_sapiens

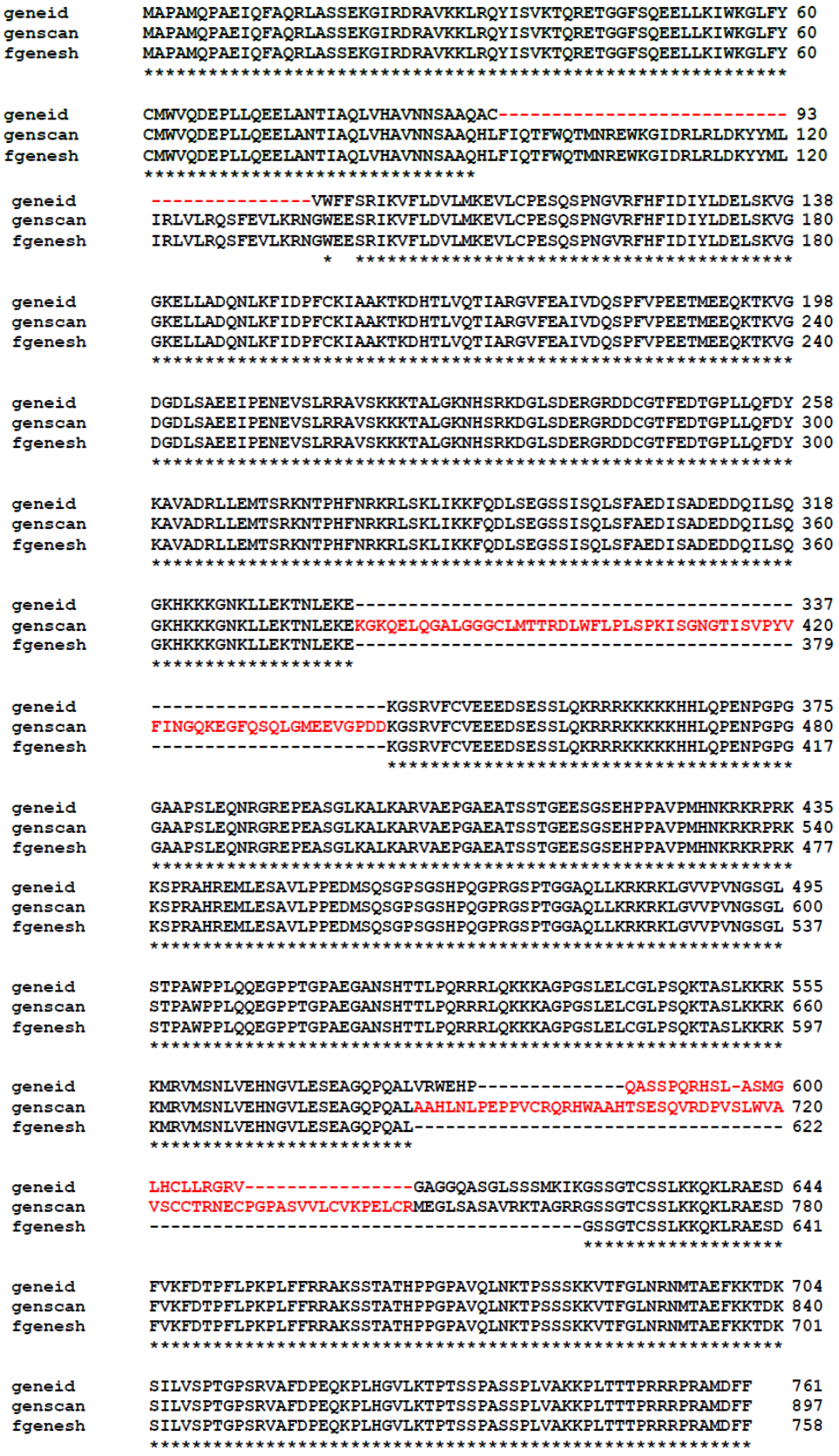
4. Per avaluar la coherència de les prediccions obtingudes per cada programa, empreu CLUSTAL per comparar les proteïnes reportades per GENEID, GENSCAN i FGENESH. Realitzeu una primera interpretació d'aquests resultats en el context d'aquest alineament global.

De cada programa podem utilitzar les proteïnes anotades. El resultat és un fitxer de format FASTA per ser alineat globalment amb CLUSTAL.

GENEID→ ja que aquest programa ha dividit la proteïna, procedim a combinar-la en funció de les prediccions generades als altres dos programes restants.

Dels resultats podem concloure que un percentatge alt de les prediccions és molt similar, tot i trobar-nos amb tres punts de conflicte al llarg de la proteïna.

El mecanisme de fusionar les dos proteïnes reconegudes per GENEID funciona bé, tot i detectar que alguns exons no s'hagin detectat completament.



Imatge 23 Multiple sequence alignment (geneid, genscan, fgenesh)

5. Finalment, per comparar quantitativament els tres sistemes de predicció, empleneu la següent taula amb les coordenades de tots els exons identificats dins del millor gen presentat per cada programa. Seleccioneu dos d'aquests exons per realitzar una recerca amb BLASTP contra la base de dades completa de proteïnes. Interpreteu aquests resultats per elaborar una primera anotació factible d'aquest gen en funció d'aquestes prediccions:

	GENEID	GENSCAN	FGENESH
Exon 157-286	●	●	●
Exon 10376 - 10458	●	●	●
Exon 12800 - 12857	●	●	●
Exon 14362 - 14447		●	●
Exon 15128 - 15189		●	●
Exon 15504 - 15655	●		
Exon 15526 - 15655		●	●
Exon 16764 - 16828	●	●	●
Exon 17225 - 17406	●	●	●
Exon 23771 - 23865	●	●	●
Exon 25045 - 25142	●	●	●
Exon 26262 - 26281	●	●	●
Exon 27296 - 27427	●	●	●
Exon 27663 - 27851		●	
Exon 28008 - 28858	●		
Exon 28008 - 28732		●	●
Exon 30236 - 30380		●	
Exon 30518 - 30529	●		
Exon 30780 - 30932	●	●	●
Exon 31931 - 31994	●	●	●
Exon 33682 - 33875	●	●	●

Imatge 24 Taula comparativa dels sistemes de predicció GENEID GENSCAN i FGENESH

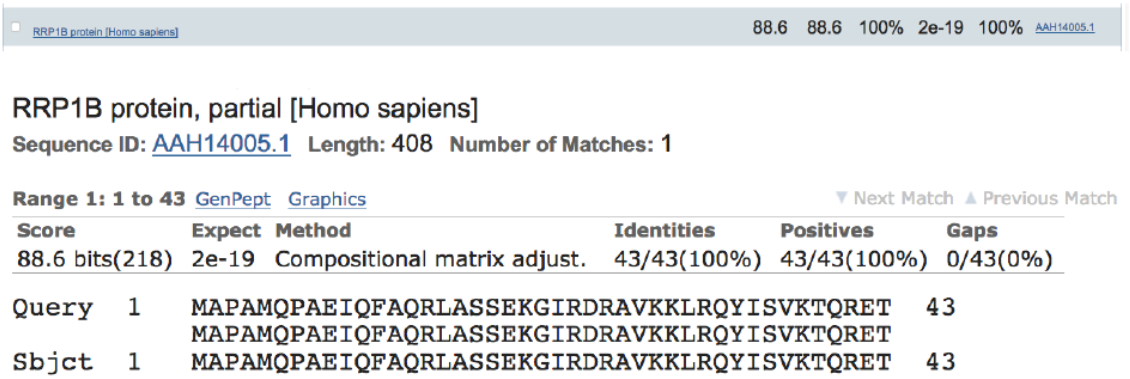
En la taula trobem representant els exons (línies) identificats per un o més programes. Subratllats en gris estan marcats els exons equivalents, existint cert solapament entre ells.

Fet això, ja es pot utilitzar GENEIsD per buscar dos exons i obtenir la seva seqüència d'aminoàcids.

```
# Firsts(+) predicted in sequence human: [0,37570]
   First      140      206      -8.44   + 0 1    3.00   -6.37   -3.53    0.00   23     MLQRARWPPPCSRPRSNLPSGWc
   First      140      237      -6.21   + 0 2    3.00   -0.04   -7.47    0.00   33     MLQRARWPPPCSRPRSNLPSGWRPARRASGTEcg
   First      157      237       3.45   + 0 0    8.07   -0.04    9.08    0.00   27     MAPAMQPAEIQFAQRLASSEKGIRDRA
   First      157      254       4.13   + 0 2    8.07   -2.06   13.81    0.00   33     MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRca
   First      157      264       6.24   + 0 0    8.07   -1.41   18.11    0.00   36     MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYIS
   First      157      286       9.81   + 0 1    8.07    2.83   20.67    0.00   44     MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETg
   First      157      303       7.90   + 0 0    8.07   -0.01   20.17    0.00   49     MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGRTAA
   First      169      237       0.47   + 0 0    3.77   -0.04    8.07    0.00   23     MQPAEIQFAQRLASSEKGIRDRA
   First      169      254       1.15   + 0 2    3.77   -2.06   12.81    0.00   29     MQPAEIQFAQRLASSEKGIRDRAVKKLRca
   First      169      264       3.26   + 0 0    3.77   -1.41   17.10    0.00   32     MQPAEIQFAQRLASSEKGIRDRAVKKLRQYIS
   First      169      286       6.82   + 0 1    3.77    2.83   19.66    0.00   40     MQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETg
   First      169      303       4.92   + 0 0    3.77   -0.01   19.17    0.00   45     MQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGRTAA
   First      318      381      -3.98   + 0 1    4.97   -2.13   -1.73    0.00   22     MAGRGPGLGLGPGPRHGMRLPc
   First      318      389      -4.49   + 0 0    4.97   -2.28   -2.76    0.00   24     MAGRGPGLGLGPGPRHGMRLPRVV
   First      318      401      -3.84   + 0 0    4.97   -1.30   -2.61    0.00   28     MAGRGPGLGLGPGPRHGMRLPRVVVTPS
   First      318      403      -5.19   + 0 2    4.97   -3.92   -2.05    0.00   29     MAGRGPGLGLGPGPRHGMRLPRVVVTPSgt
```

Imatge 25 Seqüència d'aminoàcids del primer exó.

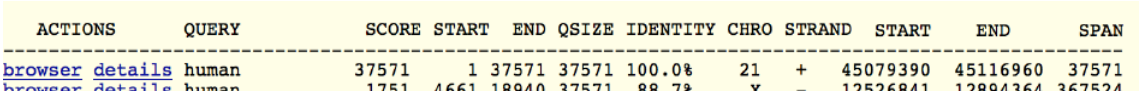
Segons BLASTP, l'exó correspon la gen RRP1B:



Imatge 26 Gen RRP1B

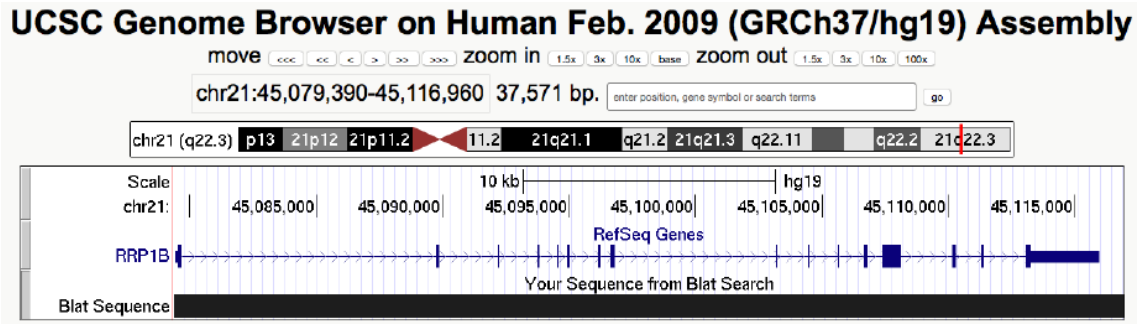
6. Aprofiteu BLAT per identificar en quina part del genoma humà es troba *anonima.fa* (cromosoma, inici, final, bri). Verifiqueu visualment que l'inici i el final de la nostra seqüència encaixen amb la regió correcta.

Feta la cerca, es recuperen diversos hits ordenats de més a menys probable:



Imatge 27 Resultats de la cerca

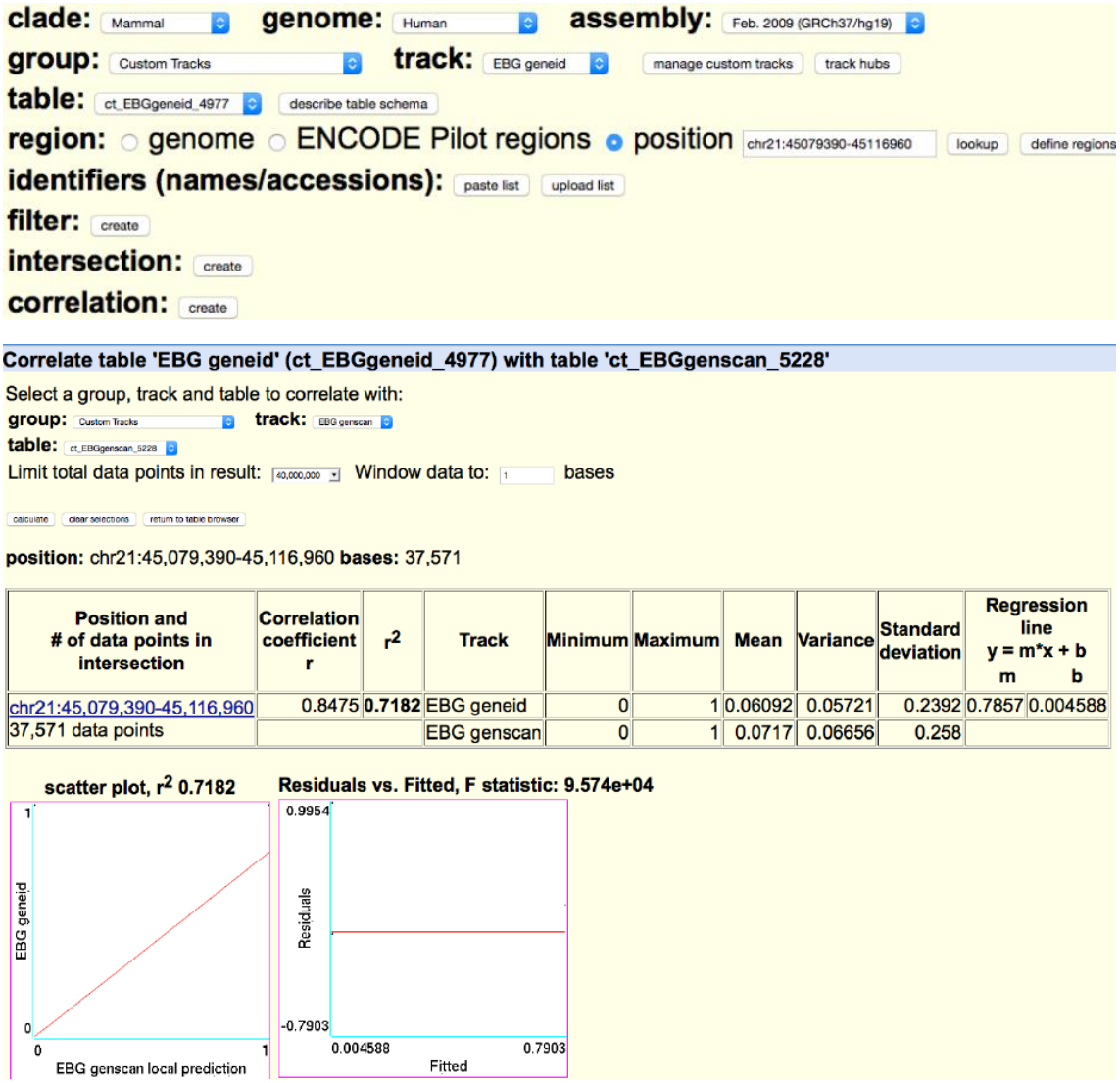
El primer resultat, utilitzant la versió hg19, coincideix amb totes les posicions de la seqüència en qüestió.



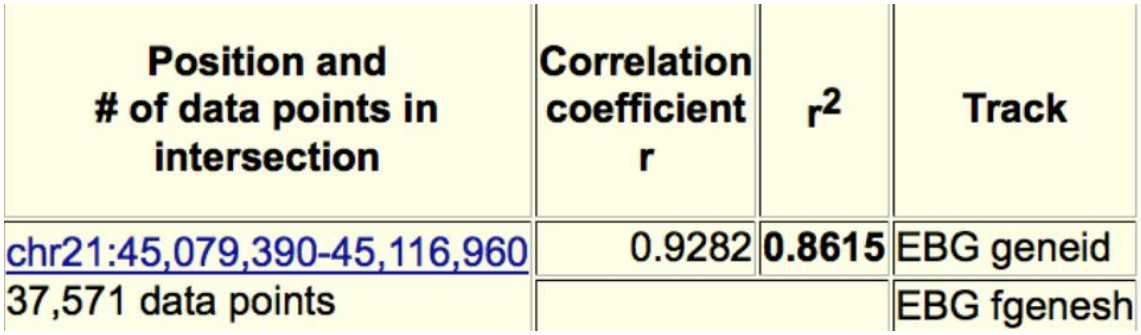
Imatge 28 Resultat BLAT

7. Convertiu manualment les nostres prediccions de GENEID, GENSCAN i FGENESH a format GFF per visualitzar-les com Custom tracks en UCSC (serà necessari adaptar les coordenades dels exons per traslladar-los sobre el cromosoma 21):

8. Empreu el *Table Browser* d'UCSC per calcular la correlació, dins de la regió genòmica delimitada per la seqüència *anonima.fa*, entre les prediccions de (a) GENEID i GENSCAN, (b) GENEID i FGENESH, (c) GENSCAN i FGENESH. A continuació, repetiu el mateix procediment per calcular la correlació entre cada predicció individual i el gen anotat pel consorci RefSeq.



Imatge 29 GENEID – GENSCAN: correlación de 0.85



Imatge 30 GENEID – FGENESH: correlación de 0.93

Position and # of data points in intersection	Correlation coefficient r	r^2	Track
chr21:45,079,390-45,116,960	0.9139	**0.8352**	EBG genscan
37,571 data points			EBG fgenesh

Imatge 31 GENSCAN – FGENESH: correlación de 0.91

Position and # of data points in intersection	Correlation coefficient r	r^2	Track
chr21:45,079,390-45,116,960	0.5916	**0.35**	EBG geneid
37,571 data points			RefSeq Genes

Imatge 32 GENEID – REFSEQ: correlación de 0.59

Position and # of data points in intersection	Correlation coefficient r	r^2	Track
chr21:45,079,390-45,116,960	0.5766	**0.3324**	EBG genscan
37,571 data points			RefSeq Genes

Imatge 33 GENSCAN – REFSEQ: correlación de 0.58

Position and # of data points in intersection	Correlation coefficient r	r^2	Track
chr21:45,079,390-45,116,960	0.6419	**0.412**	EBG fgenesh
37,571 data points			RefSeq Genes

Imatge 34 FGENESH – REFSEQ: correlación de 0.64

9. Per acabar, efectueu amb CLUSTAL l'alineament múltiple global de les tres proteïnes predites per cada programa juntament amb la proteïna real RRP1B. Analitzeu acuradament cada secció de la proteïna a la recerca de les millors prediccions en aquest fragment. Amb totes aquestes informacions, decidiu quin programa ha efectuat la millor predicció.

```
geneid      MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY
genscan     MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY
fgenesh     MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY
RRP1B       MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY
            ************************************************************

geneid      CMWVQDEPLLQEELANTIAQLVHAVNNSAAQAC--------------------------
genscan     CMWVQDEPLLQEELANTIAQLVHAVNNSAAQHLFIQTFWQTMNREWKGIDRLRLDKYYML
fgenesh     CMWVQDEPLLQEELANTIAQLVHAVNNSAAQHLFIQTFWQTMNREWKGIDRLRLDKYYML
RRP1B       CMWVQDEPLLQEELANTIAQLVHAVNNSAAQHLFIQTFWQTMNREWKGIDRLRLDKYYML
            *********************************

geneid      ---------------VWFFSRIKVFLDVLMKEVLCPESQSPNGVRFHFIDIYLDELSKVG
genscan     IRLVLRQSFEVLKRNGWEESRIKVFLDVLMKEVLCPESQSPNGVRFHFIDIYLDELSKVG
fgenesh     IRLVLRQSFEVLKRNGWEESRIKVFLDVLMKEVLCPESQSPNGVRFHFIDIYLDELSKVG
RRP1B       IRLVLRQSFEVLKRNGWEESRIKVFLDVLMKEVLCPESQSPNGVRFHFIDIYLDELSKVG
                         *  ****************************************

geneid      GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVDQSPFVPEETMEEQKTKVG
genscan     GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVDQSPFVPEETMEEQKTKVG
fgenesh     GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVDQSPFVPEETMEEQKTKVG
RRP1B       GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVDQSPFVPEETMEEQKTKVG
            ************************************************************

geneid      DGDLSAEEIPENEVSLRRAVSKKKTALGKNHSRKDGLSDERGRDDCGTFEDTGPLLQFDY
genscan     DGDLSAEEIPENEVSLRRAVSKKKTALGKNHSRKDGLSDERGRDDCGTFEDTGPLLQFDY
fgenesh     DGDLSAEEIPENEVSLRRAVSKKKTALGKNHSRKDGLSDERGRDDCGTFEDTGPLLQFDY
RRP1B       DGDLSAEEIPENEVSLRRAVSKKKTALGKNHSRKDGLSDERGRDDCGTFEDTGPLLQFDY
            ************************************************************

geneid      KAVADRLLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQLSFAEDISADEDDQILSQ
genscan     KAVADRLLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQLSFAEDISADEDDQILSQ
fgenesh     KAVADRLLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQLSFAEDISADEDDQILSQ
RRP1B       KAVADRLLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQLSFAEDISADEDDQILSQ
            ************************************************************

geneid      GKHKKKGNKLLEKTNLEKE-------------------------------------------
genscan     GKHKKKGNKLLEKTNLEKEKGKQELQGALGGGCLMTTRDLWFLPLSPKISGNGTISVPYV
fgenesh     GKHKKKGNKLLEKTNLEKE-------------------------------------------
RRP1B       GKHKKKGNKLLEKTNLEKE-------------------------------------------
            ******************

geneid      ----------------------KGSRVFCVEEEDSESSLQKRRRKKKKKHHLQPENPGPG
genscan     FINGQKEGFQSQLGMEEVGPDDKGSRVFCVEEEDSESSLQKRRRKKKKKHHLQPENPGPG
fgenesh     ----------------------KGSRVFCVEEEDSESSLQKRRRKKKKKHHLQPENPGPG
RRP1B       ----------------------KGSRVFCVEEEDSESSLQKRRRKKKKKHHLQPENPGPG
                                   ************************************

geneid      GAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK
genscan     GAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK
fgenesh     GAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK
```

```
geneid      GAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK
genscan     GAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK
fgenesh     GAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK
RRP1B       GAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK
            ************************************************************

geneid      KSPRAHREMLESAVLPPEDMSQSGPSGSHPQGPRGSPTGGAQLLKRKRKLGVVPVNGSGL
genscan     KSPRAHREMLESAVLPPEDMSQSGPSGSHPQGPRGSPTGGAQLLKRKRKLGVVPVNGSGL
fgenesh     KSPRAHREMLESAVLPPEDMSQSGPSGSHPQGPRGSPTGGAQLLKRKRKLGVVPVNGSGL
RRP1B       KSPRAHREMLESAVLPPEDMSQSGPSGSHPQGPRGSPTGGAQLLKRKRKLGVVPVNGSGL
            ************************************************************

geneid      STPAWPPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRK
genscan     STPAWPPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRK
fgenesh     STPAWPPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRK
RRP1B       STPAWPPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRK
            ************************************************************

geneid      KMRVMSNLVEHNGVLESEAGQPQALVRWEHP--------------QASSPQRHSL-ASMG
genscan     KMRVMSNLVEHNGVLESEAGQPQALAAHLNLPEPPVCRQRHWAAHTSESQVRDPVSLWVA
fgenesh     KMRVMSNLVEHNGVLESEAGQPQAL-----------------------------------
RRP1B       KMRVMSNLVEHNGVLESEAGQPQAL-----------------------------------
            **************************

geneid      LHCLLRGRV----------------GAGGQASGLSSSMKIKGSSGTCSSLKKQKLRAESD
genscan     VSCCTRNECPGPASVVLCVKPELCRMEGLSASAVRKTAGRRGSSGTCSSLKKQKLRAESD
fgenesh     ----------------------------------------GSSGTCSSLKKQKLRAESD
RRP1B       ----------------------------------------GSSGTCSSLKKQKLRAESD
                                                     ********************

geneid      FVKFDTPFLPKPLFFRRAKSSTATHPPGPAVQLNKTPSSSKKVTFGLNRNMTAEFKKTDK
genscan     FVKFDTPFLPKPLFFRRAKSSTATHPPGPAVQLNKTPSSSKKVTFGLNRNMTAEFKKTDK
fgenesh     FVKFDTPFLPKPLFFRRAKSSTATHPPGPAVQLNKTPSSSKKVTFGLNRNMTAEFKKTDK
RRP1B       FVKFDTPFLPKPLFFRRAKSSTATHPPGPAVQLNKTPSSSKKVTFGLNRNMTAEFKKTDK
            ************************************************************

geneid      SILVSPTGPSRVAFDPEQKPLHGVLKTPTSSPASSPLVAKKPLTTTPRRRPRAMDFF
genscan     SILVSPTGPSRVAFDPEQKPLHGVLKTPTSSPASSPLVAKKPLTTTPRRRPRAMDFF
fgenesh     SILVSPTGPSRVAFDPEQKPLHGVLKTPTSSPASSPLVAKKPLTTTPRRRPRAMDFF
RRP1B       SILVSPTGPSRVAFDPEQKPLHGVLKTPTSSPASSPLVAKKPLTTTPRRRPRAMDFF
            ************************************************************
```

Imatge 35 Alineament múltiple global de les tres proteïne predite.

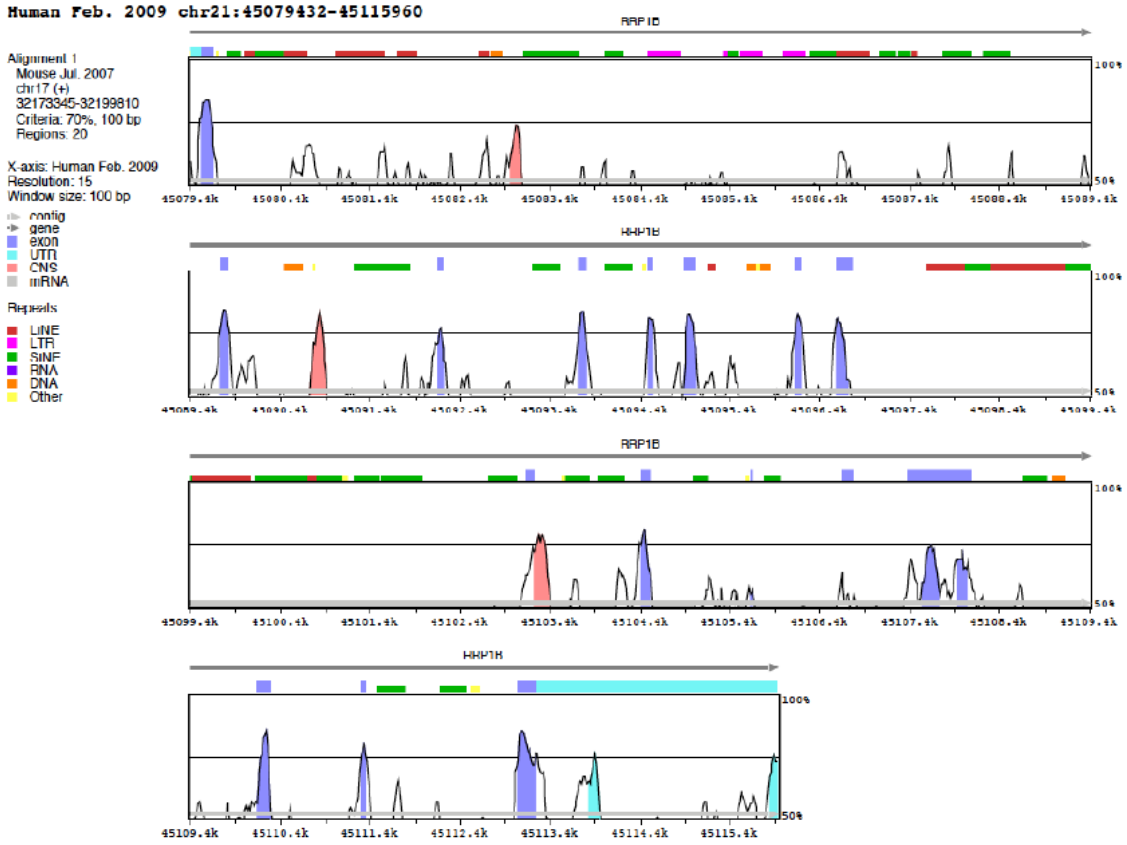
10. El navegador genòmic VISTA permet observar la conservació entre diversos genomes. Analitzeu la documentació existent sobre aquesta aplicació i esbrineu el significat que tenen les gràfiques i els colors emprats sobre cada alineament entre dos genomes. Posteriorment, seleccioneu el nostre gen d'estudi per analitzar el grau de conservació que posseeixen els exons d'aquest. Raoneu breument sobre com podríem millorar les prediccions inicials servides per GENEID, GENSCAN i FGENESH utilitzant aquesta informació sobre la conservació de seqüència en regions funcionals.

El programa VISTA respresenta de manera gràfica els alineaments de múltiples genomes amb la finalitat de identificar aquelles regions funcionals amb un alt grau de conservació al llarg dels anys.

Un dels trets més significatius d'aquests gràfics són aquelles àrees conservades dins de certes regions genòmiques reben una coloració diferent fàcilment identificable. En aquest cas tenim que el vermell són les regions no codificants (introns), blau clar regions UTR del gen estudiat i blau fosc per les regions codificants (CDS).

Podem apreciar que a mesura que ens allunyem (evolutivament parlant) la senyal de conservació disminueix.

Per millorar aquestes prediccions prodriem afegir valors numèrics a aquells exons predits que encaixin sobre regions conservades.



Imatge 36 Resultat interpretat genòmic VISTA