

Problemas & soluciones de los temas 1 y 3

Tema 1:

Problema 1: BLAST y isoformas

Mediante el BLAST hemos analizado esta [secuencia](#) determinando que se trata de la isoforma 1 del gen BRCA1. Por lo tanto el alineamiento que podemos visualizar para la isoforma 2, en realidad corresponde a un alineamiento entre las dos isoformas. Si lo analizamos encontramos en el alineamiento un hueco que termina con un desemparejamiento:

```
Query 1441 DLRNPEQSTSEK-----AVLTSQKSSEYPISQNPEGLSADKFEV 1479
          DLRNPEQSTSEK                               +VLTSQKSSEYPISQNPEGLSADKFEV
Sbjct 1441 DLRNPEQSTSEKDSHIHGQRNNSMFSKRPREHISVLTSQKSSEYPISQNPEGLSADKFEV 1500
```

- a) Se trata de dos isoformas anotadas y por lo tanto el hueco debería reflejar cambios en el splicing. ¿El hueco se manifiesta también si alineamos las secuencias de mRNA y CDS?

La siguiente solución se elaboró con las secuencias que había en la base de datos en 2020

Si, lo podemos observar tanto usando las secuencias de CDS como los mRNA

mRNA:

```

NM_007300.4      4401  |||||
NM_007294.4      4451  ACAAAGCACATCAGAAAAAG----- 4470
NM_007300.4      4451  |||||
NM_007294.4      4471  -----CAGTATTAAC TTCACAG 4487
NM_007300.4      4501  |||||
NM_007300.4      4501  CCATGTTTTCTAAAAGGCCTAGAGAACATATATCAGTATTAAC TTCACAG 4550
```

CDS

NM_007294.4_c	4351	GAAAAAG-----	4357
NM_007300.4_c	4351	GAAAAAGATTCGCATATACATGGCCAAAGGAACAACCTCCATGTTTTCTAA	4400
NM_007294.4_c	4358	-----CAGTATTAACCTTCACAGAAAAGTAGTGAAT	4387
NM_007300.4_c	4401	AAGGCCTAGAGAACATATATCAGTATTAACCTTCACAGAAAAGTAGTGAAT	4450

b) Según el esquema de abajo, de qué 'tipo' de splicing se trata?

Para averiguarlo tenemos que fijarnos primero en posición del hueco, que es la 4471 (la última base es la 4470). Si buscamos en las anotaciones de las dos isoformas

(https://www.ncbi.nlm.nih.gov/nuccore/NM_007294.4

https://www.ncbi.nlm.nih.gov/nuccore/NM_007300.4)

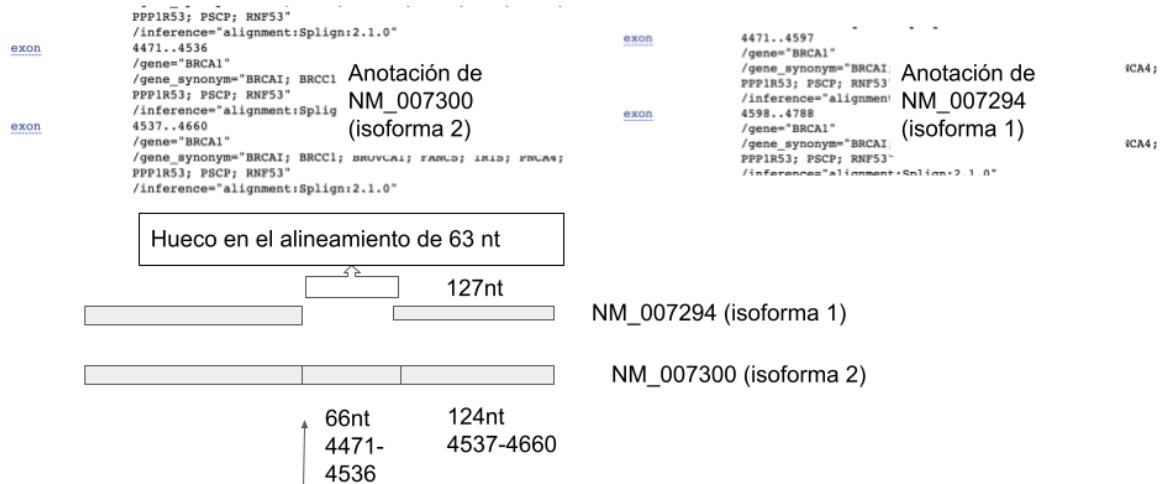
vemos que en esta posición empieza un exon en ambas isoformas. Sin embargo tienen longitudes diferentes. **4471-4536 (isoforma 1) y 4471-4597 (isoforma 2).**

```

exon
.....
4471..4536
/gene="BRCA1"
/gene_synonym="BRCAI; BRCC1; BROVCA1; FANCS; IRIS; PNCA4;
PPP1R53; PSCP; RNF53"
/inference="alignment:Splice:2.1.0"
-----

exon
.....
4471..4597
/gene="BRCA1"
/gene_synonym="BRCAI; BRCC1; BROVCA1; FANCS; IRIS; PNCA4;
PPP1R53; PSCP; RNF53"
/inference="alignment:Splice:2.1.0"

```

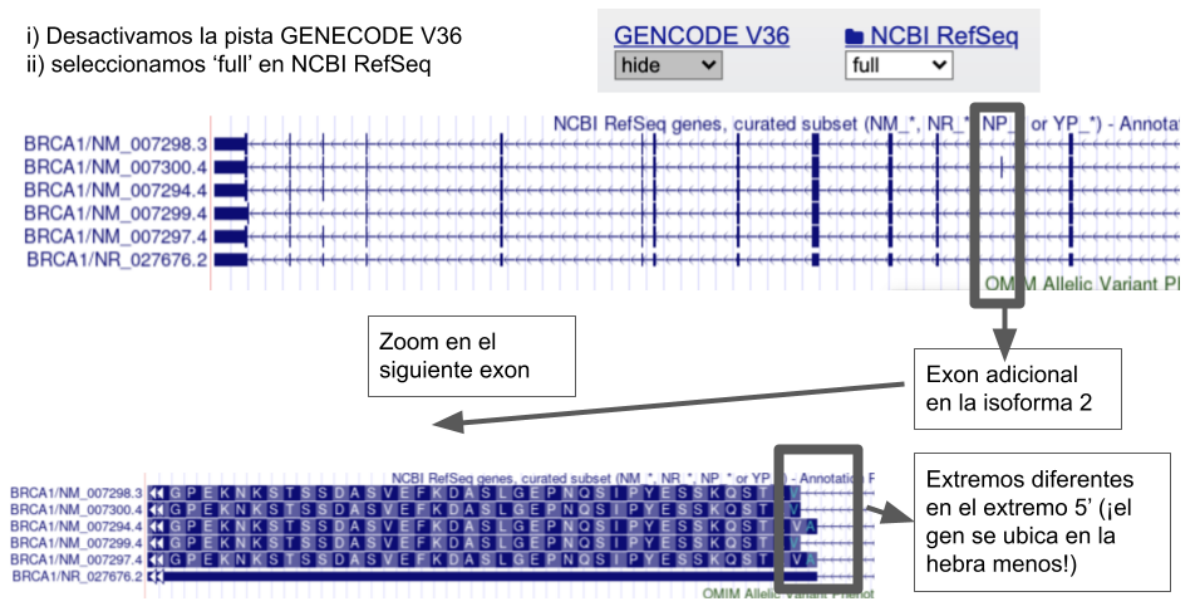


Dos eventos de splicing: i) exon skipping (uso de un exon adicional de 66 nt en la isoforma 2), ii) corte-empalme en 5' del siguiente exon

Confirmación con el Navegador Genómico

Verificación mediante el Navegador genómico: introducimos BRCA1

- Desactivamos la pista GENCODE V36
- seleccionamos 'full' en NCBI RefSeq



Eventos de splicing y sus frecuencias:

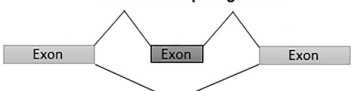
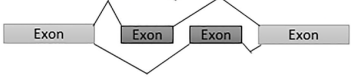
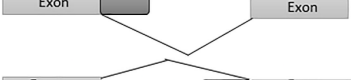
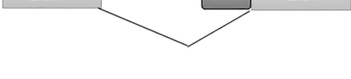
Alternative Splicing Events		Observed Frequency		Potential Outcomes
		Humans	Plants	
	(i) Exon Skipping (ES)	~40%	~8%	Diversifying proteome complexity
	(ii) Mutually Exclusive Exons (MXE)	~10.7%	NA	Diversifying proteome complexity
	(iii) Intron Retention (IR)	~5%	~60%	(i) Sequester in the nucleus and produce on demand, (ii) PTC+ degraded by NMD pathway, or (iii) may generate truncated protein
	(iv) Alternative 5' Splice Site (A5SS)	~7.9%	~7.5%	Diversifying proteome complexity
	(v) Alternative 3' Splice Site (A3SS)	~18.4%	~15.5%	Diversifying proteome complexity
	(vi) Exon Intron (EI)	~4%	~4%	(i) Generates complete protein domain, (ii) intrinsically disordered protein, or (iii) post-translational modifications to improve substrate specificity

Figura tomado de: <https://www.frontiersin.org/articles/10.3389/fpls.2019.00708/full>

c) ¿Por qué hay un desemparejamiento al final del hueco?

Si una región codificante sufre una inserción o delección que es múltiple de 3, entonces no cambia la pauta de lectura. Sin embargo, si los nucleótidos eliminados no corresponden a una ristra de codones, entonces veremos adicionalmente un desemparejamiento. Consideramos la siguiente secuencia de la eliminamos 3 nucleótidos (abajo con ---)

ATGAAGTTG**GAT**ATC

ATG---TTGGATATC → ATGTTGGATATC → se han eliminado en fase y por eso no hay desemparejamiento en el alineamiento

ATGAAGTTG**GAT**ATC

ATGA---TGGATATC → ATG**ATG**GATATC

→ no se han eliminado en fase y aparece un nuevo codón (en azul) que alineando las proteínas se nos manifiesta como desemparejamiento

Problema 2: Caracterizar una secuencia anónima

Tenemos la siguiente [secuencia anónima](#). Nos hacemos las siguientes preguntas:

- ¿De qué especie es la secuencia?
- ¿Corresponde a un gen conocido?

Mediante el blastn podemos observar que es el gen ATF3 (variante transcripcional 3) en humano.

select all 100 sequences selected		GenBank		Graphics		Distance tree of results	
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☒	Homo sapiens activating transcription factor 3 (ATF3), transcript variant 3, mRNA	3411	3411	100%	0.0	100.00%	NM_001030287.4
☒	PREDICTED: Pan troglodytes activating transcription factor 3 (ATF3), transcript variant X3, mRNA	3332	3332	100%	0.0	99.24%	XM_016938388.2
☒	Homo sapiens activating transcription factor 3 (ATF3), transcript variant 1, mRNA	3269	3269	95%	0.0	100.00%	NM_001674.4
☒	PREDICTED: Homo sapiens activating transcription factor 3 (ATF3), transcript variant X2, mRNA	3269	3269	95%	0.0	100.00%	XM_011509579.1
☒	PREDICTED: Homo sapiens activating transcription factor 3 (ATF3), transcript variant X1, mRNA	3269	3269	95%	0.0	99.94%	XM_005273146.1
☒	Homo sapiens activating transcription factor 3, mRNA (cDNA clone MGC:12746 IMAGE:4138076), complete cds	3264	3264	95%	0.0	99.94%	BC006322.2

- ¿Corresponde exactamente a un transcrito conocido?

Si, ya que tanto la secuencia de entrada como la referencia tienen longitud 1847 pb y además la Query Coverage y Similitud son del 100%

Homo sapiens activating transcription factor 3 (ATF3), transcript variant 3, mRNA

NCBI Reference Sequence: NM_001030287.4

[FASTA](#) [Graphics](#)

Go to: ☐

LOCUS NM_001030287 1847 bp mRNA linear PRI 17-SEP-2020
 DEFINITION Homo sapiens activating transcription factor 3 (ATF3), transcript variant 3, mRNA.

- ¿En qué cromosoma se ubica esta secuencia?

Tanto mediante la página del NCBI gene (<https://www.ncbi.nlm.nih.gov/gene/467>) como mediante el BLAT podemos ver que se ubica en el cromosoma 1

BLAT Search Results

Go back to [chr21:8207400-8208946](#) on the Genome Browser.

Custom track name:

Custom track description:

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser details	NM_001030287.4	1844	1	1847	1847	100.0%	chr1	+	212565407	212620775	55369
browser details	NM_001030287.4	25	1474	1505	1847	96.3%	chr1	-	71753904	71753938	35
browser details	NM_001030287.4	21	1710	1731	1847	100.0%	chr5	-	137642976	137642998	23
browser details	NM_001030287.4	20	1483	1502	1847	100.0%	chr1	-	91831502	91831521	20

- ¿Cuántos exones tendrá esta secuencia juzgando solamente por la salida del BLAT?

Haciendo click en details vemos que hay 4 bloques de alineamiento, que podrían corresponder a 4 exones

[NM_001030287.4](#)

[Human.chr1](#)

[block1](#)

[block2](#)

[block3](#)

[block4](#)

[together](#)

- ¿Cuántos exones codificantes tiene este transcrito? ¿Hay una forma de averiguarlo sin usar la [anotación](#)?

Usando la anotación podemos ver que el primer exon va de la posición 1 hasta 77, pero la CDS no empieza hasta la posición 82. Por lo tanto la región codificante empieza en el segundo exon, y por eso hay 3 exones codificantes (al menos una parte de su secuencia es parte de la región codificante).

```
exon          1..77
              /gene="ATF3"
              /inference="alignment:Splign:2.1.0"
misc_feature   49..51
              /gene="ATF3"
              /note="upstream in-frame stop codon"
exon          78..321
              /gene="ATF3"
              /inference="alignment:Splign:2.1.0"
CDS           82..627
```

Sin usar la anotación tendríamos que usar un programa como el [ORF finder](#) o [ORFfinder](#) para detectar el ORF (Open Reading Frame, Marco de lectura abierto) más largo. Vemos que el ORF más largo efectivamente va de 82-627 igual que la región CDS anotada. El primer bloque de alineamiento tiene una longitud de 77 igual que el primer exon anotado. De esta forma podríamos averiguar que el primer exon es no-codificante sin recurrir a la anotación

ORF Finder results

Results for 1847 residue sequence "NM_001030287.4 Homo sapiens activating transcription factor 3 (ATF3), transcript variant 3, mRNA" starting "GTGACAAGAA"

```
>ORF number 1 in reading frame 1 on the direct strand extends from base 82 to base 627.
ATGATGCTTCAACACCCAGGCCAGGTCTCTGCCTCGGAAGTGAGTGCTTCTGCCATCGTC
CCCTGCCTGTCCCTCTCTGGGTCACTGGTGTGAGGATTTTGCTAACCTGACGCCCTTT
GTCAAGGAAGAGCTGAGGTTTGCCATCCAGAACAGCACCTCTGCCACCGGATGTCCTCT
GCGCTGGAATCAGTCACTGTCTAGCGACAGACCCCTCGGGGTGTCCATCAGAAAAGCCGAG
GTAGCCCTGAAGAAGATGAAGGAAAAAGAGGCGACGAGAAAGAAATAAGATTGCAGCT
GCAAGTGCAGAAACAGAAGAGGAGAAGACGGAGTGCCTGCAGAAAGAGTCGGAGAAG
CTGGAAGTGTGAATGCTGAAGTGAAGGCTCAGATTGAGGAGCTCAAGAACGAGAAGCAG
CATTTGATATACATGCTCAACCTTCATCGGCCACGTGTATTGTCCGGGCTCAGAAATGGG
AGGACTCCAGAAGATGAGAGAACCTCTTTATCCAACAGATAAAGAAGGAACATTGCAG
AGCTAA
```

Problema 3: Buscar un transcrito en otra especie

Buscamos la secuencia del mRNA de SOX9 (NM_000346.4) dentro del genoma del chimpancé y ratón

- ¿Que observamos si nos fijamos en el número de posibles exones? ¿Qué explicación puede tener?.

Al igual que en humano, en el chimpancé vemos 3 bloques (o posibles exones). En el ratón hay 28 bloques. La distancia evolutiva con el chimpancé es mucho menor y por lo tanto es más probable que la estructura genética sea la misma que en humano.

- ¿Que observamos si en vez de alinear el mRNA entero, solamente alineamos la región codificante? ¿Qué explicación puede tener?

Observamos que ahora hay 6 bloques en el ratón comparado con los 3 exones codificantes en humano y chimpancé. Sabemos que la CDS se encuentra mucho más conservada comparada con las regiones UTR. En estas regiones no-traducidas puede haber muchas deleciones y inserciones lo que puede llevar a la fragmentación del alineamiento con BLAT (más bloques)

- ¿Que observamos si buscamos la secuencia en una especie de invertebrado?

No deberíamos obtener resultados ya que SOX9 no está en invertebrados.

Ahora descargamos la región codificante del gen SOX9 de ratón y la alineamos mediante el Smith-Waterman frente a la CDS de humano.

- ¿Qué podemos observar y qué explicación le podemos dar si comparamos con lo que observamos en el alineamiento de la CDS de humano frente al genoma del ratón?

Observamos que hay dos deleciones de longitud 3 en la CDS del ratón, pero en general existe un alto grado de conservación. El alineamiento mediante el BLAT se diferencia en dos aspectos: i) es frente a la secuencia genómica que contiene los intrones y ii) es un método ad hoc y no exhaustivo como el alineamiento de Smith-Waterman.

https://www.ncbi.nlm.nih.gov/nuccore/NM_011448.4

- a) ¿Qué diferencias observamos entre el alineamiento local y global?
- b) ¿Cuál sería la explicación evolutiva?
- a) Observamos que solamente una subsecuencia alinea (local). Gran parte de las dos secuencias no alinean, es decir que no hay una secuencia correspondiente en la otra especie (global).

Smith-Waterman (local)

```
#####
#
# Aligned_sequences: 2
# 1: mouse_SRY
# 2: human_SRY
# Matrix: EBL05UP62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 160
# Identity: 74/160 (46.2%)
# Similarity: 102/160 (63.8%)
# Gaps: 17/160 (10.6%)
# Score: 357.0
#
#####

mouse_SRY 1 HEGHVKRPYNAFVMSRGERHKLQGNPSHQNTEISKQLCRWKSLEAE 50
             ::..|||||::|::|::|::|::|::|::|::|::|::|::|::|::|
human_SRY 50 VQORVKRPYNAFVMSRQGRKHALENPRWRNSEISKQLQYQWILTEAE 105
             VQORVKRPYNAFVMSRQGRKHALENPRWRNSEISKQLQYQWILTEAE

mouse_SRY 51 KRPFQEAQRLKILHREKYPYNYQPHRAKVSQRSGILQPAVASTKLYN 100
             |.|||||::|::|::|::|::|::|::|::|::|::|::|::|::|::|::|
human_SRY 106 KRPFQEAQRLQAPHREKYPYNYQPHRAKVSQRSGILQPAVASTKLYN 155
             |.|||||::|::|::|::|::|::|::|::|::|::|::|::|::|::|::|

mouse_SRY 101 LLQMDRPHAITVRQDSRAHLYSKQGSFYQPVDIPTGHLQ--QQQQ 148
             :|::|.. .||::|::|::|::|::|::|::|::|::|::|::|::|
human_SRY 156 EVQLDNR---LYRDOCTKATHSRHEHQ-----LGHLPINAA 190
             :|::|.. .||::|::|::|::|::|::|::|::|::|::|::|::|

mouse_SRY 149 QQQQQQFHHH 158
             ::||::|::|
human_SRY 191 SPQQRCHYSH 200
             ::||::|::|

#####
#
#
```

Needleman-Wunsch (global)

```
#####
#
# Aligned_sequences: 2
# 1: mouse_SRY
# 2: human_SRY
# Matrix: EBL05UP62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 452
# Identity: 74/452 (16.4%)
# Similarity: 103/452 (22.8%)
# Gaps: 305/452 (67.5%)
# Score: 353.0
#
#####

mouse_SRY 1 ..... 0
human_SRY 1 HQSYASAHLSVFNSSDYSPAVQENIPALRSSSFLCTESGSKYQCETGE 50
mouse_SRY 1 ----HEGHVKRPYNAFVMSRGERHKLQGNPSHQNTEISKQLCRWKS 45
             ::..|||||::|::|::|::|::|::|::|::|::|::|::|::|
human_SRY 51 NSKSNWQGRVKRPYNAFVMSRQGRKHALENPRWRNSEISKQLQYQWIL 100
mouse_SRY 46 LTEAEKRPFFQEAQRLKILHREKYPYNYQPHRAKVSQRSGILQPAVAS 95
             |||||::|::|::|::|::|::|::|::|::|::|::|::|::|::|
human_SRY 101 LTEAEKRPFFQEAQRLQAPHREKYPYNYQPHRAKVSQRSGILQPAVAS 150
mouse_SRY 96 TKLYNLLQMDRPHAITVRQDSRAHLYSKQGSFYQPVDIPTGHLQ- 144
             :|::|::|::|::|::|::|::|::|::|::|::|::|::|::|
human_SRY 151 SVLCSEVQLDNR---LYRDOCTKATHSRHEHQ-----LGHLP 185
mouse_SRY 145 -QQQQQQQQQFHHHQQQQQFYQPVDIPTGHLQ- 193
             ::||::|::|::|::|::|::|::|::|::|::|::|::|::|
human_SRY 186 INAAASSPQQRCHYSHTKL 204
mouse_SRY 194 DHHQQQQQFHHHQQQQQFYQPVDIPTGHLQ- 243
human_SRY 205 ..... 204
mouse_SRY 244 QQQQQQFHHHQQQQQFYQPVDIPTGHLQ- 293
human_SRY 205 ..... 204
```

- b) Después de analizar los alineamientos, la hipótesis sería que existe un región conservada (aproximadamente la que podemos observar con el alineamiento local)

Si buscamos en la literatura (evolution of SRY genes) podemos encontrar que efectivamente solamente esta región se encuentra conservada entre humano y ratón. Mutaciones en esta región pueden provocar reversión sexual.

<http://physrev.physiology.org/content/87/1/1/F12>

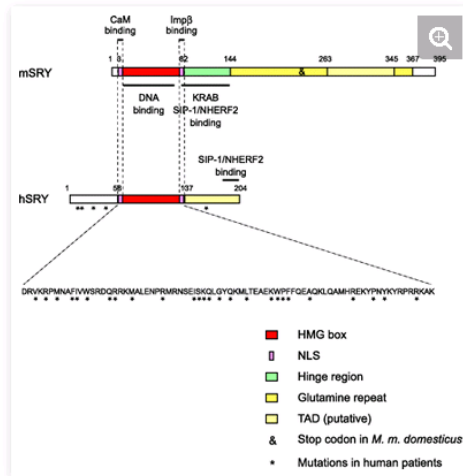


FIG. 12.

[Download figure](#) | [Open in new tab](#) | [Download powerpoint](#)

Structure of mouse and human SRY protein. The HMG DNA-binding domain is shown in red and the large glutamine-rich domain of the mouse SRY COOH terminus is in dark yellow. Nuclear translocation is mediated by one NLS (pink) at either end of the HMG domain. The NH₂-terminal NLS is recognized and bound by calmodulin (CaM), whereas the COOH-terminal acts via importin β . For both mouse and human SRY, a putative transactivation domain (TAD, light yellow) has been described. The hinge or bridge region (green) interacts with mouse SRY-interacting protein 1 (SIP-1/NHERF2) and the KRAB-only protein, whereas human SRY interacts with SIP-1/NHERF2 via its COOH terminus. Sex-reversing mutations in human SRY (marked by asterisks) leading to gonadal dysgenesis or hermaphroditism are mainly found in the HMG domain.

- Entre humanos y ratón, solo el dominio HMG (High Mobility Group) está conservado
- HMG es responsable para la unión al DNA
- Se ha observado que mutaciones en esta región causan reversión sexual.

Problema 5: La insulina

Comparar las secuencias de la insulina (pre-proinsulina) a nivel de ADN (solamente la región codificante) y proteína entre humano y el macaco (*Macaca fascicularis*)

- ¿En qué alineamiento (ADN o proteína) podemos observar más cambios evolutivos?
- ¿Por qué?

Observamos más cambios a nivel de ADN (véase abajo). Se debe a que el código genético es degenerado, es decir que podemos tener cambios a nivel de ADN que no cambian la secuencia de la proteína.

- Analizamos las posiciones con desemparejamientos en el alineamiento de las CDS.
- ¿Las posiciones se distribuyen aleatoriamente o podemos detectar algún patrón?

No se distribuyen aleatoriamente. Todos los cambios menos dos se producen en la tercera posición de los codones.

```

#=====
#
# Aligned_sequences: 2
# 1: INS_human
# 2: Macaca_INS
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 110
# Identity:      108/110 (98.2%)
# Similarity:    108/110 (98.2%)
# Gaps:          0/110 ( 0.0%)
# Score: 577.0
#
#
#=====

INS_human      1 MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFY    50
                |||
Macaca_INS      1 MALWMRLLPLLALLALWGPDPAPAFVNQHLCGSHLVEALYLVCGERGFFY    50

INS_human      51 TPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIC    100
                |||
Macaca_INS      51 TPKTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIC    100

INS_human      101 SLYQLENYCN      110
                |||
Macaca_INS      101 SLYQLENYCN      110

#-----
#-----

```

```

..
# Length: 332
# Identity:      316/332 (95.2%)
# Similarity:    316/332 (95.2%)
# Gaps:          0/332 ( 0.0%)
# Score: 1516.0
#
#
#=====

Macaca_cds      1 ATGGCCCTGTGGATGCGCCTCCTGCCCTGCTGGCGCTGCTGGCCCTCTG      50
|||||
EMBOSS_001      1 ATGGCCCTGTGGATGCGCCTCCTGCCCTGCTGGCGCTGCTGGCCCTCTG      50

Macaca_cds     51 GGGACCTGACCCGGCCCGGCCTTTGTGAACCAGCACCTGTGCGGCTCCC     100
|||||
EMBOSS_001     51 GGGACCTGACCCAGCCGCGAGCCTTTGTGAACCAACACCTGTGCGGCTCAC     100

Macaca_cds     101 ACCTAGTGGAAGCTCTCTACCTGGTGTGCGGGGAACGAGGCTTCTTCTAC     150
|||||
EMBOSS_001     101 ACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAACGAGGCTTCTTCTAC     150

Macaca_cds     151 ACACCCAAGACCCGCCGGGAGGCAGAGGACCTCAGGTGGGGCAGGTGGA     200
|||||
EMBOSS_001     151 ACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGTGGGGCAGGTGGA     200

Macaca_cds     201 GCTGGGCGGGGGCCCTGGCGCAGGCAGCCTGCAGCCCTTGGCGCTGGAGG     250
|||||
EMBOSS_001     201 GCTGGGCGGGGGCCCTGGTGCAGGCAGCCTGCAGCCCTTGGCCCTGGAGG     250

Macaca_cds     251 GGTCCCTGCAGAAGCGCGGCATCGTGGAGCAGTGTGCACCAGCATCTGC     300
|||||
EMBOSS_001     251 GGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGTACCAGCATCTGC     300

Macaca_cds     301 TCCCTCTACCAGCTGGAGAACTACTGCAACTA      332
|||||
EMBOSS_001     301 TCCCTCTACCAGCTGGAGAACTACTGCAACTA      332

```

Problema 6: El gen XIST

Descargamos la secuencia del transcrito maduro del gen XIST:

https://www.ncbi.nlm.nih.gov/nuccore/NR_001564.2

- ¿En qué cromosoma se localiza el gen?

BLAT Search Results

Go back to [chrX:15,560,138-15,602,945](#) on the Genome Browser.

Custom track name:

Custom track description:

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser details	NR_001564.2	19275	1	19280	19296	100.0%	chrX	-	73820651	73852753	32103

- ¿Cuántos exones tiene?

Tiene 6 bloques de alineamientos lo que coincide con los 6 exones anotados en genbank.

Alignment of NR_001564.2

[NR_001564.2](#)

[Human.chrX](#)

[block1](#)

[block2](#)

[block3](#)

[block4](#)

[block5](#)

[block6](#)

[together](#)

- ¿Cuántos exones codificantes tiene?

No tiene exones codificantes ya que se trata de un gen no-codificante.

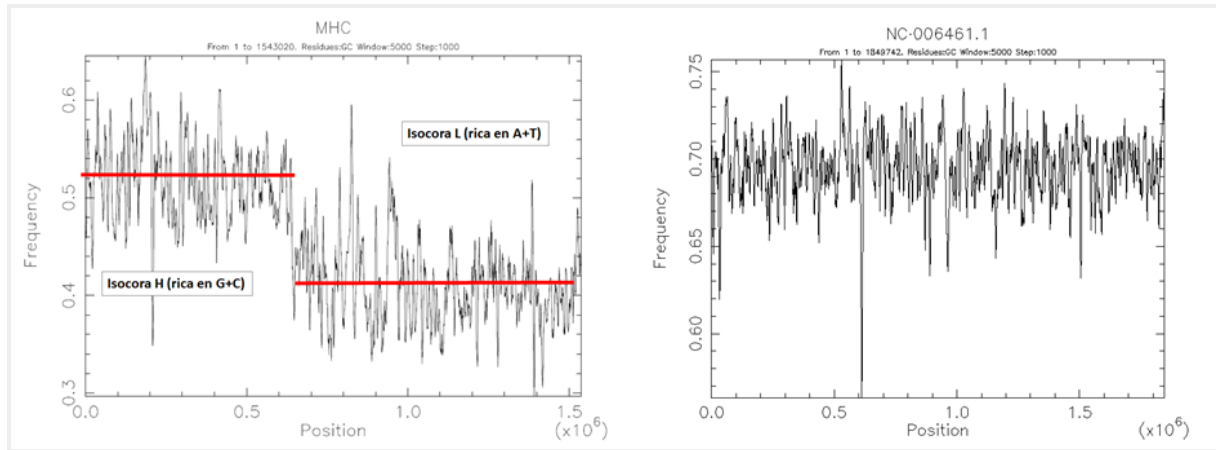
Tema 3:

1. Comparar la secuencia del MHC con una bacteria

Comparar la fluctuación del %G+C de la secuencia MHC con la de NC_006461.1

¿Qué diferencias observamos si comparamos la distribución espacial del G+C entre las dos secuencias?

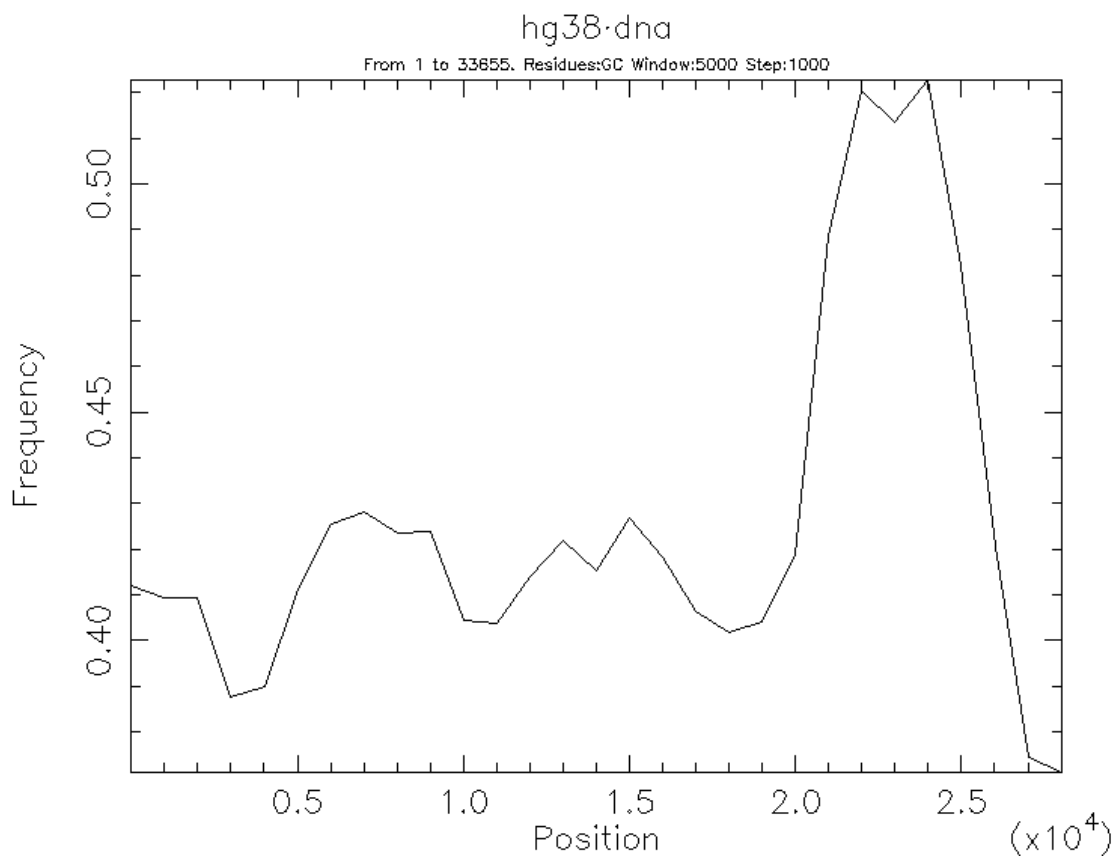
Primero vamos a seleccionar los mismos parámetros del **freak** para poder comparar los resultados. En este caso, ventana de 5000 bp y 1000 bp de salto. En la bacteria *Thermus thermophilus* observamos una distribución espacial del %G+C mucho más homogénea a lo largo del cromosoma. Vemos solamente una región claramente más pobre en %G+C (que llega a aprox. 56%). Se puede tratar de una [transferencia horizontal](#) de una bacteria más rica en A+T o un gen/operón rico en A+T.



2. Analizar el G+C a lo largo de esta **secuencia**.

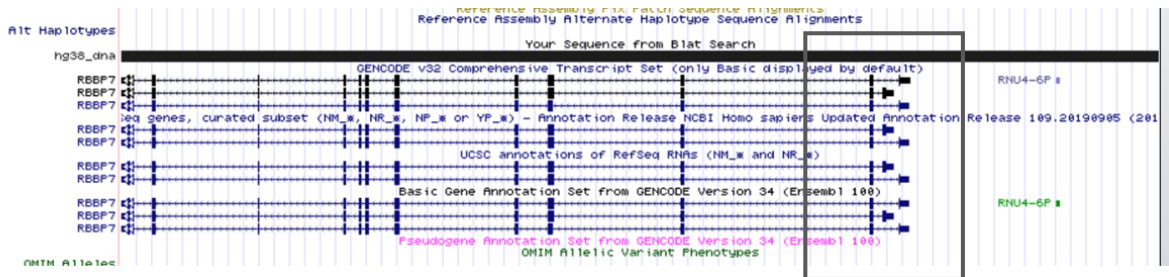
- ¿Que observamos?

Se puede observar un pico en G+C en la posición 23.000 aproximada.



- ¿Qué explicación tiene?

Si buscamos la secuencia en el genoma mediante el BLAT, vemos que alinea perfectamente con el cromosoma X del genoma humano. Mediante el enlace 'Browser' podemos visualizar la región (véase abajo). Vemos que en esta región se encuentra el gen RBBP7 que se ubica en la hebra menos indicado por las flechas que apuntan hacia la izquierda. Por lo tanto, el pico en G+C coincide con la región promotora de este gen.



3. Comparar la composición de dinucleótidos

Comparar la composición de la [sec29](#) con una secuencia obtenida de *C. elegans* ([¿Como?](#))

- ¿Qué diferencias observamos?

En la secuencia 29 (humano) vemos una ratio entre las frecuencias observadas y esperadas de los dinucleótidos CG de 0.24, es decir, solamente encontramos la cuarta parte de lo que hubiéramos esperado. En la secuencia de *E. elegans* que tiene un tamaño comparable no observamos esta falta de CGs.

OUTPUT FILE [outfile](#)

```
#
# Output from 'compseq'
#
# The Expected frequencies are calculated from the observed single
# base or residue frequencies in these sequences
#
# The input sequences are:
#   sec29

Word size      2
Total count    2271170

#
# Word  Obs Count  Obs Frequency  Exp Frequency  Obs/Exp Frequency
#
AA  261729  0.1148856  0.1052765  1.0912755
AC  120224  0.0527722  0.0597962  0.8825342
AG  148411  0.0651448  0.0584558  1.1144280
AT  208819  0.0916609  0.1009349  0.9081189
CA  162156  0.0711782  0.0597962  1.1903465
CC  91813  0.0403012  0.0339638  1.1865946
CG  18114  0.0079511  0.0332025  0.2394738
CT  147766  0.0648617  0.0573302  1.1313710
GA  131518  0.0577297  0.0584558  0.9875773
GC  77790  0.0341458  0.0332025  1.0284129
GG  90893  0.0398974  0.0324582  1.2291919
GT  110238  0.0483888  0.0560451  0.8633908
TA  183779  0.0806696  0.1009349  0.7992242
TC  130023  0.0570734  0.0573302  0.9955217
TG  153021  0.0671684  0.0560451  1.1984698
TT  241876  0.1061712  0.0967723  1.0971238

Other  0  0.0000000  0.0000000  1000000000.0000000
```

OUTPUT FILE [outfile](#)

```
#
# Output from 'compseq'
#
# The Expected frequencies are calculated from the observed single
# base or residue frequencies in these sequences
#
# The input sequences are:
#   cell_dna

Word size      2
Total count    2149999

#
# Word  Obs Count  Obs Frequency  Exp Frequency  Obs/Exp Frequency
#
AA  304608  0.1416782  0.1012391  1.3994415
AC  96203  0.0447456  0.0584472  0.7655734
AG  104024  0.0483833  0.0577262  0.8381517
AT  179254  0.0833740  0.1007685  0.8273816
CA  126238  0.0587154  0.0584472  1.0045888
CC  84901  0.0394889  0.0337426  1.1702963
CG  79401  0.0369307  0.0332264  1.1081533
CT  104397  0.0485568  0.0581755  0.8346604
GA  129620  0.0602884  0.0577262  1.0443861
GC  82821  0.0385214  0.0332264  1.1558842
GG  82621  0.0384284  0.0329152  1.1674953
GT  95003  0.0441875  0.0574578  0.7690418
TA  123622  0.0574986  0.1007685  0.5706013
TC  131012  0.0609358  0.0581755  1.0474489
TG  124019  0.0576833  0.0574578  1.0039240
TT  302255  0.1405838  0.1003001  1.4016320

Other  0  0.0000000  0.0000000  1000000000.0000000
```

- ¿Qué explicación puede tener?

En *C. elegans* no hay una metiltransferasa que metila las citocinas en el contexto CG. Y por lo tanto no hay pérdida evolutiva de CGs por desaminación espontánea de la citosina metilada a timina.

4. Comparar la composición en dinucleótidos con una secuencia aleatorizada

Aleatorizar ([suffleseq](#)) la [sec29](#) y determinar el uso de dinucleótidos.

- ¿Qué diferencias hay en el uso de dinucleótidos entre la sec29 y su secuencia randomizada? ¿Por qué?

En la secuencia aleatorizada vemos que todas las ratios están muy cercas de 1 confirmando que el valor 1 corresponde a lo que esperaríamos en una secuencia totalmente aleatoria.

OUTPUT FILE [outfile](#)

Output from 'compseq'

The Expected frequencies are calculated from the observed single
base or residue frequencies in these sequences

The input sequences are:
sec29

Word size 2
Total count 2278170

Word Obs Count Obs Frequency Exp Frequency Obs/Exp Frequency
#

AA 261729 0.1148856 0.1052765 1.0912755
AC 120224 0.0527722 0.0597962 0.8825342
AG 148411 0.0651448 0.0584558 1.1144280
AT 208819 0.0916609 0.1009349 0.9081189
CA 162156 0.0711782 0.0597962 1.1903465
CC 91813 0.0403012 0.0339638 1.1865946
CG 18114 0.0079511 0.0332025 0.2394738
CT 147766 0.0648617 0.0573302 1.1313710
GA 131518 0.0577297 0.0584558 0.9875773
GC 77790 0.0341458 0.0332025 1.0284129
GG 90893 0.0398974 0.0324582 1.2291919
GT 110238 0.0483888 0.0560451 0.8633908
TA 183779 0.0806696 0.1009349 0.7992242
TC 130023 0.0570734 0.0573302 0.9955217
TG 153021 0.0671684 0.0560451 1.1984698
TT 241876 0.1061712 0.0967723 1.0971238

Other 0 0.0000000 0.0000000 1.000000000.0000000

OUTPUT FILE [outfile](#)

Output from 'compseq'

The Expected frequencies are calculated from the observed single
base or residue frequencies in these sequences

The input sequences are:
sec29_random

Word size 2
Total count 2278170

Word Obs Count Obs Frequency Exp Frequency Obs/Exp Frequency
#

AA 239918 0.1053117 0.1052765 1.0003349
AC 136691 0.0600004 0.0597962 1.0034143
AG 132857 0.0583174 0.0584558 0.9976320
AT 229716 0.1008336 0.1009349 0.9989965
CA 136420 0.0598814 0.0597962 1.0014250
CC 76985 0.0337925 0.0339638 0.9949570
CG 75835 0.0332877 0.0332025 1.0025670
CT 130610 0.0573311 0.0573302 1.0000160
GA 133404 0.0585575 0.0584558 1.0017394
GC 75516 0.0331477 0.0332025 0.9983497
GG 73972 0.0324699 0.0324582 1.0003607
GT 127547 0.0559866 0.0560451 0.9989559
TA 229440 0.1007124 0.1009349 0.9977962
TC 130658 0.0573522 0.0573302 1.0003835
TG 127775 0.0560867 0.0560451 1.0007416
TT 220826 0.0969313 0.0967723 1.0016433

Other 0 0.0000000 0.0000000 1.000000000.0000000

- Existen otros dinucleótidos en la sec29 cuya ratio obs/esp se desvía de 1. ¿A qué se podría deber?

Resaltan los dinucleótidos TG y CA con 1.1985 y 1.1903 respectivamente. TG es el producto de desaminación de CG en la hebra más y CA en la hebra menos y por eso detectamos más de lo que se esperarí por azar.

5. Determinar el uso de codones con compseq

El programa **compseq**, por defecto determina la frecuencia de N-meros mediante ventanas solapantes. Sin embargo existe un parámetro que nos permite definir un 'salto' (como en el programa freak).

- ¿como se llama el parámetro y que valor tenemos que especificar?

Mediante el parámetro -frame podemos analizar los trinucleótidos en un marco de lectura dado. En este caso tiene que tomar el valor 3.

```
compseq -frame 3 NM_007295_CDS.txt
```

```
#
# Output from 'compseq'
#
# Only words in frame 3 will be counted.
# The Expected frequencies are calculated on the (false) assumption that every
# word has equal frequency.
#
```

The input sequences are:

NM_007294.4_cds_NP_009225.1_1

Word size 3

Total count 1863

#	# WordObs	Count	Obs Frequency	Exp Frequency	Obs/Exp Frequency
#					
AAA	79		0.0424047	0.0156250	2.7139023
AAC	38		0.0203972	0.0156250	1.3054214
AAG	58		0.0311326	0.0156250	1.9924852
AAT	83		0.0445518	0.0156250	2.8513151
ACA	38		0.0203972	0.0156250	1.3054214
ACC	24		0.0128824	0.0156250	0.8244767
ACG	7		0.0037574	0.0156250	0.2404724
ACT	42		0.0225443	0.0156250	1.4428341
AGA	34		0.0182501	0.0156250	1.1680086
AGC	44		0.0236178	0.0156250	1.5115405
AGG	24		0.0128824	0.0156250	0.8244767
AGT	61		0.0327429	0.0156250	2.0955448
ATA	27		0.0144928	0.0156250	0.9275362
ATC	21		0.0112721	0.0156250	0.7214171
ATG	29		0.0155663	0.0156250	0.9962426
ATT	29		0.0155663	0.0156250	0.9962426
CAA	40		0.0214707	0.0156250	1.3741278
CAC	16		0.0085883	0.0156250	0.5496511
CAG	57		0.0305958	0.0156250	1.9581320
CAT	33		0.0177134	0.0156250	1.1336554
CCA	37		0.0198604	0.0156250	1.2710682
CCC	16		0.0085883	0.0156250	0.5496511
CCG	1		0.0005368	0.0156250	0.0343532
CCT	42		0.0225443	0.0156250	1.4428341
CGA	4		0.0021471	0.0156250	0.1374128
CGC	2		0.0010735	0.0156250	0.0687064
CGG	5		0.0026838	0.0156250	0.1717660
CGT	7		0.0037574	0.0156250	0.2404724
CTA	22		0.0118089	0.0156250	0.7557703
CTC	16		0.0085883	0.0156250	0.5496511
CTG	27		0.0144928	0.0156250	0.9275362
CTT	23		0.0123457	0.0156250	0.7901235
GAA	135		0.0724638	0.0156250	4.6376812
GAC	26		0.0139560	0.0156250	0.8931830
GAG	63		0.0338164	0.0156250	2.1642512
GAT	59		0.0316694	0.0156250	2.0268384

GCA	35	0.0187869	0.0156250	1.2023618
GCC	14	0.0075148	0.0156250	0.4809447
GCG	2	0.0010735	0.0156250	0.0687064
GCT	33	0.0177134	0.0156250	1.1336554
GGA	30	0.0161031	0.0156250	1.0305958
GGC	16	0.0085883	0.0156250	0.5496511
GGG	17	0.0091251	0.0156250	0.5840043
GGT	24	0.0128824	0.0156250	0.8244767
GTA	21	0.0112721	0.0156250	0.7214171
GTC	15	0.0080515	0.0156250	0.5152979
GTG	27	0.0144928	0.0156250	0.9275362
GTT	38	0.0203972	0.0156250	1.3054214
TAA	0	0.0000000	0.0156250	0.0000000
TAC	14	0.0075148	0.0156250	0.4809447
TAG	0	0.0000000	0.0156250	0.0000000
TAT	17	0.0091251	0.0156250	0.5840043
TCA	37	0.0198604	0.0156250	1.2710682
TCC	16	0.0085883	0.0156250	0.5496511
TCG	1	0.0005368	0.0156250	0.0343532
TCT	65	0.0348900	0.0156250	2.2329576
TGA	1	0.0005368	0.0156250	0.0343532
TGC	12	0.0064412	0.0156250	0.4122383
TGG	10	0.0053677	0.0156250	0.3435319
TGT	32	0.0171766	0.0156250	1.0993022
TTA	34	0.0182501	0.0156250	1.1680086
TTC	16	0.0085883	0.0156250	0.5496511
TTG	34	0.0182501	0.0156250	1.1680086
TTT	33	0.0177134	0.0156250	1.1336554
Other	0	0.0000000	0.0000000	10000000000.0000000

- ¿Coinciden las frecuencias de codones?

Si analizamos la misma secuencia mediante **CUSP**, podemos observar que las frecuencias de trinucleótidos detectados con el **compseq** corresponden con las frecuencias de los codones del **CUSP**.

Word size 3
Total count 1863

```
#
# Word Obs Count Obs Frequency Exp Frequency Obs/Exp
Frequency
#
GAA 135 0.0724638 0.0156250 4.6376812
GAC 26 0.0139560 0.0156250 0.8931830
GAG 63 0.0338164 0.0156250 2.1642512
GAT 59 0.0316694 0.0156250 2.0268384
GCA 35 0.0187869 0.0156250 1.2023618
GCC 14 0.0075148 0.0156250 0.4809447
GCG 2 0.0010735 0.0156250 0.0687064
GCT 33 0.0177134 0.0156250 1.1336554
GGA 30 0.0161031 0.0156250 1.0305958
GGC 16 0.0085883 0.0156250 0.5496511
GGG 17 0.0091251 0.0156250 0.5840043
GGT 24 0.0128824 0.0156250 0.8244767
GTA 21 0.0112721 0.0156250 0.7214171
GTC 15 0.0080515 0.0156250 0.5152979
GTG 27 0.0144928 0.0156250 0.9275362
GTT 38 0.0203972 0.0156250 1.3054214
TAA 0 0.0000000 0.0156250 0.0000000
TAC 14 0.0075148 0.0156250 0.4809447
TAG 0 0.0000000 0.0156250 0.0000000
TAT 17 0.0091251 0.0156250 0.5840043
TCA 37 0.0198604 0.0156250 1.2710682
TCC 16 0.0085883 0.0156250 0.5496511
TCG 1 0.0005368 0.0156250 0.0343532
TCT 65 0.0348900 0.0156250 2.232957
```

```
#CdsCount: 1
#Coding GC 41.22%
#1st letter GC 48.44%
#2nd letter GC 39.32%
#3rd letter GC 35.89%

#Codon AA Fraction Frequency Number
GCA A 0.417 18.777 35
GCC A 0.167 7.511 14
GCG A 0.024 1.073 2
GCT A 0.393 17.704 33
TGC C 0.273 6.438 12
TGT C 0.727 17.167 32
GAC D 0.306 13.948 26
GAT D 0.694 31.652 59
GAA E 0.682 72.425 135
GAG E 0.318 33.798 63
TTC F 0.327 8.584 16
TTT F 0.673 17.704 33
```

6. Uso de codones en una bacteria

Analizar el uso de codones de la bacteria [Thermus thermophilus](#)

- ¿Qué observamos si comparamos el %G+C en las diferentes posiciones del codón?

Después de obtener todas las CDS de la bacteria *Thermus thermophilus* y lanzar el CUSP obtenemos la siguiente salida:

OUTPUT FILE [outfile](#)

```
#CdsCount: 1957
#Coding GC 69.67%
#1st letter GC 72.13%
#2nd letter GC 44.83%
#3rd letter GC 92.05%
```

Vemos que hay 1957 secuencias de entrada. Si todas son CDS, deberíamos obtener el mismo número de codones de parada.

TAA	*	0.192	0.638	375
TAG	*	0.379	1.262	742
TGA	*	0.429	1.429	840

Lo podemos confirmar sumando los 3 codones de parada.

Si analizamos el G+C en las 3 posiciones de los codones vemos que existen diferencias muy notables con 72% en primera posición, 44.8% en segunda y 92% en tercera. Es decir que el 92% de los nucleótidos en tercera posición de los codones corresponden o a G o a C.

- ¿Qué explicación puede tener

Lo que observamos es un caso extremo del [sesgo de uso de codones](#). Por ejemplo, el 96.5% de todos los codones usados para codificar alanina son los que terminan en G (GCG) o C (GCC)

#Codon	AA	Fraction	Frequency	Number
GCA	A	0.014	1.597	939
GCC	A	0.739	84.877	49905
GCG	A	0.226	25.974	15272
GCT	A	0.021	2.453	1442

Esta bacteria vive en condiciones extremas (> 65 grados Celsius) lo que explica que su genoma es extremadamente rico en G+C (69.5%) ya que los enlaces entre G y C con 3 puentes de hidrógenos son termodinámicamente más estables que A/T con 2 enlaces entre ellos. Existe sesgo mutacional en bacterias. Un factor en el sesgo de mutaciones es la disponibilidad de bases en la replicación del ADN. En este caso se puede suponer que hay más nucleótidos C/G y por lo tanto los errores en la replicación serán más bien hacia G/C. Muchos aminoácidos están codificados por 4 codones que solamente cambian en la tercera posición. Por lo tanto, frecuentemente un cambio en la tercera posición no cambia el aminoácido codificado y por lo tanto estos cambios son más probables de ser aceptados evolutivamente al no sufrir selección negativa (ya que no cambian la secuencia y por lo tanto la función de la proteína). Es decir, en estas posiciones 'los genes pueden adaptar' su G+C a el del genoma. En la segunda posición, cualquier cambio significa un cambio de aminoácido y por lo tanto es más probable que se elimine por selección negativa. Eso puede explicar parte del sesgo tan extremo. Otro componente importante es la optimización de la traducción mediante el uso de codones y abundancia del anticodón.

Este ejemplo es para llamaros la atención sin entrar en más detalles. Más información podéis encontrar en:

https://en.wikipedia.org/wiki/Codon_usage_bias

<https://pubmed.ncbi.nlm.nih.gov/22889422/>

<https://www.nature.com/articles/nrm.2017.91>

<https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/codon-usage-bias>