

Salvatore Javier Balivo López

Tarea 1 Bioinformatica

Introducción

La alta complejidad de la cognición humana está asociada con el sistema nervioso central y su organización sináptica (Silberis et al.2016). Las sinapsis son ambientes complejos donde numerosos componentes como neurotransmisores, receptores de membrana, canales de iones y proteínas de señales de transducción, actúan juntos para crear una fuerte red de interacción. Los genes ortólogos que generan estas estructuras han sido encontrados también en especies sin neuronas (Sakarya et al. 2007; Burkhardt et al. 2011; Emes and Grant 2011, Emes and Grant 2012; Suga et al. 2013; Burkhardt 2015).

En este análisis se toma como referencia la proteína humana CHRM1 (Also known as M1; HM1; M1R) en el que se estudiará su distribución taxonómica.

Preparación del ambiente de computación

Para poder llevar a cabo los analisis he tenido que instalar en My Binder el programa Blast con bioconda y la base de datos ejecutando el escript proporcionado por la docencia `preparar_ambiente.sh`

```
In [1]: system2(command = './preparar_ambiente.sh', wait = TRUE)
```

Métodos

Se utilizará **blastp** (Altschul *et al.*, 1990) con diferentes umbrales de valor E, para determinar a qué ritmo aumenta la distribución taxonómica de las secuencias encontradas a medida que se relaja el grado de similitud exigido. Es decir, a medida que acepto como resultados del **blastp** secuencias con un valor E mayor. Se explorará la distribución taxonómica de la proteína humana CHRM1 describiendo las secuencias homólogas utilizando la base de datos Swissprot donde podemos encontrar secuencias revisadas a mano, significa que las secuencias que nos propondrá la base de datos están certificadas y no son un mero producto de predicción informática.

Resultados

BLASTP individual

Como primer paso busco las secuencias más parecidas a mi query (secuencia proporcionada por la docencia en el archivo CHRM1.fas), ejecuto una única búsqueda con blastp, usando un

umbral de valor E de 1.0e-50, que es muy exigente: sólo las secuencias extremadamente parecidas a la original aparecerán en los resultados.

In [2]:

```
#como query utilizo el archivo CHRM1.fas
BlastpOut01 <- system2(command = 'blastp',
                      args = c('-db', 'swissprot',
                                '-query', 'CHRM1.fas',
                                '-evalue', '1.0e-50',
                                '-outfmt', '"7 saccver pident length qstart qend sstart send evalue staxid ssciname sblastname"'),
                      stdout = TRUE)

# El resultado del comando anterior, guardado en "BlastOut01", es un archivo
# texto plano. Con las funciones textConnection() y read.table(), lo transformo
# en un "data frame", una tabla donde cada columna es una variable. La opción
# "col.names" en read.table() sirve para dar nombre a las columnas.
TablaOut01 <- read.table(textConnection(BlastpOut01),
                        sep = '\t',
                        col.names = c('saccver', 'pident', 'length', 'qstart', 'qend', 'sstart', 'send', 'evalue', 'staxid', 'ssciname', 'sblastname'))

# Dimensiones de la tabla:
dim(TablaOut01)
```

37 · 11

In [3]:

TablaOut01

A data.frame: 37 × 11

saccver	pident	length	qstart	qend	sstart	send	evalue	staxid	ssciname	sblastname
<dbl>	<dbl>	<int>	<int>	<int>	<int>	<int>	<dbl>	<int>	<fct>	<fct>
P11229.2	100.000	460	1	460	1	460	0.00e+00	9606	Homo sapiens	pr
P56489.1	99.565	460	1	460	1	460	0.00e+00	9544	Macaca mulatta	pr
Q5R949.1	99.565	460	1	460	1	460	0.00e+00	9601	Pongo abelii	pr
P04761.1	99.130	460	1	460	1	460	0.00e+00	9823	Sus scrofa	eve unq
P12657.2	98.913	460	1	460	1	460	0.00e+00	10090	Mus musculus	r
P08482.1	98.696	460	1	460	1	460	0.00e+00	10116	Rattus norvegicus	r
Q9N2A4.1	51.731	520	10	438	51	564	1.14e-170	9598	Pan troglodytes	pr
P20309.1	51.731	520	10	438	51	564	2.19e-170	9606	Homo sapiens	pr
Q9N2A3.1	51.737	518	10	436	51	562	1.08e-169	9595	Gorilla gorilla gorilla	pr
P11483.1	50.769	520	10	438	51	564	3.15e-169	9823	Sus scrofa	eve unq
Q9N2A2.1	51.670	509	19	438	62	564	3.59e-169	9600	Pongo pygmaeus	pr
P41984.1	50.577	520	10	438	51	564	3.75e-169	9913	Bos taurus	eve unq

saccver	pident	length	qstart	qend	sstart	send	evalue	staxid	ssciname	sblas
<fct>	<dbl>	<int>	<int>	<int>	<int>	<int>	<dbl>	<int>	<fct>	
P08483.1	51.670	509	19	438	61	563	3.87e-169	10116	Rattus norvegicus	r
P08912.2	51.935	491	23	441	28	518	4.60e-168	9606	Homo sapiens	pr
Q9ERZ3.1	50.586	512	19	438	61	563	3.14e-167	10090	Mus musculus	r
P56490.1	51.527	491	23	441	28	518	6.41e-167	9544	Macaca mulatta	pr
Q5IS53.1	51.324	491	23	441	28	518	9.87e-164	9598	Pan troglodytes	pr
P49578.1	48.837	516	10	438	99	613	1.15e-156	9031	Gallus gallus	
P17200.1	48.214	448	24	437	41	483	1.12e-141	9031	Gallus gallus	
P32211.1	48.198	444	24	437	30	472	1.56e-138	10090	Mus musculus	r
P10980.2	47.153	439	23	437	21	459	9.08e-138	10116	Rattus norvegicus	r
P08173.2	47.973	444	24	437	31	472	9.18e-138	9606	Homo sapiens	pr
P08485.1	47.973	444	24	437	30	471	4.95e-137	10116	Rattus norvegicus	r
Q9ERZ4.2	46.925	439	23	437	21	459	8.25e-137	10090	Mus musculus	r
P08172.1	46.697	439	23	437	21	459	3.31e-136	9606	Homo sapiens	pr
P41985.2	45.796	452	23	437	20	458	6.36e-136	9913	Bos taurus	eve unq
P06199.1	46.241	439	23	437	21	459	6.72e-135	9823	Sus scrofa	eve unq
Q9N2A7.1	46.759	432	30	437	2	433	3.76e-134	9598	Pan troglodytes	pr
P30372.1	44.444	459	2	437	3	459	2.76e-132	9031	Gallus gallus	
P30544.1	45.975	472	18	438	25	478	1.12e-129	8355	Xenopus laevis	fr toi
P08911.1	73.488	215	20	234	24	238	9.43e-111	10116	Rattus norvegicus	r
Q920H4.2	74.882	211	20	230	25	235	9.45e-110	10090	Mus musculus	r
P16395.2	53.061	245	12	251	87	323	3.30e-77	7227	Drosophila melanogaster	
Q9U7D5.2	50.000	208	24	230	64	271	1.08e-66	6239	Caenorhabditis elegans	nem
Q9JI35.2	33.333	408	39	433	49	427	4.11e-55	10141	Cavia porcellus	r
									Ascaris heligolandica	

La tabla de resultados incluye 37 secuencias encontradas, incluyendo la misma secuencia usada como *query*. Las columnas de la tabla són:

- **saccver**: número de acceso o identificador de la secuencia encontrada (*subject*).
- **pident**: porcentaje de identidad entre la *query* y el *subject* en su alineamiento local.

- **length**: longitud del alineamiento entre *query* y *subject*.
- **qstart**: primera posición de la *query* alineada.
- **qend**: última posición alineada en la *query*.
- **sstart**: primera posición de la *subject* alineada.
- **send**: última posición de la *subject* alineada.
- **eval**: valor E del alineamiento. Es decir, número de alineamientos de igual o mejor puntuación esperados por azar en una base de datos igual de grande.
- **staxid**: identificador de la especie a la que pertenece la *subject*, en la base de datos de taxonomía del NCBI.
- **ssciname**: nombre científico de la especie a la que pertenece la *subject*.
- **sblastname**: nombre de un grupo taxonómico de rango superior al que pertenece la *subject*, usado para facilitar la interpretación.

BLASTP en serie

Para ver cómo aumenta el número de secuencias a medida que aumenta el umbral de valor E, deberíamos repetir la búsqueda unas cuantas veces, con valores del parámetro `-eval` diferentes. Podemos hacerlo de forma automática usando la función `lapply()`. El código siguiente ejecutará una búsqueda de **blastp** para cada valor del vector

Valores E maximos. El resultado será una lista de tablas, como la anterior.

```
In [4]: Valores_E_maximos <- c('1.0e-50', '1.0e-40', '1.0e-30', '1.0e-20', '1.0e-10',
                              '1.0e-08', '1.0e-06', '1.0e-04', '1.0e-02', '1')

# La función lapply() aplicará una función (segundo argumento) sobre
# cada uno de los valores de la lista o vector indicados en el primer
# argumento (Valores_E_maximos, en este caso). La función que aplicamos
# la definimos entre llaves ("{}"). Incluye todos los pasos necesarios
# para obtener una tabla de resultados como la anterior.
Lista_de_Tablas <- lapply(Valores_E_maximos,
                          function(x) {
                            BlastpOut <- system2(
                              command = 'blastp',
                              args = c('-db', 'swissprot',
                                       '-query', 'CHRM1.fas',
                                       '-eval', x,
                                       '-outfmt',
                                       '"7 saccver pident length qstart
                                       qend sstart send eval
                                       ssciname sblastname"')
                              stdout = TRUE)
                            read.table(textConnection(BlastpOut),
                                       sep = '\t',
                                       col.names = c('saccver', 'pident',
                                                    'qend', 'sstart', 'send', 'eval',
                                                    'ssciname', 'sblastname'))
                          })
```

El objeto `Lista_de_Tablas` guarda las tablas de resultados de los diez **blastp** ejecutados. Podemos acceder a una tabla individual mediante los dobles corchetes:

```
In [6]: tail(Lista_de_Tablas[[10]])
```

A data.frame: 6 × 11

saccver	pident	length	qstart	qend	sstart	send	eval	staxid	ssciname	sblast
---------	--------	--------	--------	------	--------	------	------	--------	----------	--------

	<fct>	<dbl>	<int>	<int>	<int>	<int>	<int>	<dbl>	<int>	<fct>	
769	Q62795.1	28.477	151	22	170	34	181	1.32e-09	39089	Phoca groenlandica	carn
770	Q86917.1	25.758	198	42	233	108	289	1.32e-09	10269	Sheeppox virus KS-1	v
771	Q9P1P4.1	23.834	193	42	225	48	234	1.47e-09	9606	Homo sapiens	pri
772	Q9P1P4.1	32.143	84	350	433	241	324	4.38e-06	9606	Homo sapiens	pri
773	Q9Y5X5.2	27.338	139	22	160	144	280	1.58e-09	9606	Homo sapiens	pri
774	Q9Y5X5.2	30.380	79	359	430	370	447	2.83e-04	9606	Homo sapiens	pri

In [7]:

```
tail(Lista_de_Tablas[[6]])
```

A data.frame: 6 × 11

	saccver	pident	length	qstart	qend	sstart	send	evalue	staxid	ssciname	sblastna
	<fct>	<dbl>	<int>	<int>	<int>	<int>	<int>	<dbl>	<int>	<fct>	<f
572	Q86917.1	25.758	198	42	233	108	289	1.32e-09	10269	Sheeppox virus KS-1	virt
573	Q9P1P4.1	23.834	193	42	225	48	234	1.47e-09	9606	Homo sapiens	prim
574	Q9Y5X5.2	27.338	139	22	160	144	280	1.58e-09	9606	Homo sapiens	prim
575	Q6W3F4.1	30.709	127	39	165	21	145	1.87e-09	9615	Canis lupus familiaris	carniv
576	Q64077.1	28.221	163	2	163	9	169	1.88e-09	10141	Cavia porcellus	rod
577	P32302.1	27.835	194	17	204	43	234	1.99e-09	9606	Homo sapiens	prim

En estas tablas de homología generadas por blastp, como era de esperar se han encontrado secuencias pertenecientes a especies muy cercanas a *Homo sapiens* pero también encontramos una secuencia vírica con un valor de E muy bueno perteneciente a la especie *Sheeppox virus KS-1* (secuencia identificadora: Q86917.1). Este es un resultado inesperado dado que según los estudios de Visardi et al. 2021 en el que explican que la proteína que estamos utilizando como query aparece en el linaje del último ancestro común entre cordados y cnidarios. Es decir, no se conocen proteínas homólogas en ctenóforos, poríferos, placozoa, ni mucho menos en hongos ni en plantas. Las consideraciones sobre este resultado se encuentran en el apartado de 'Discusión'.

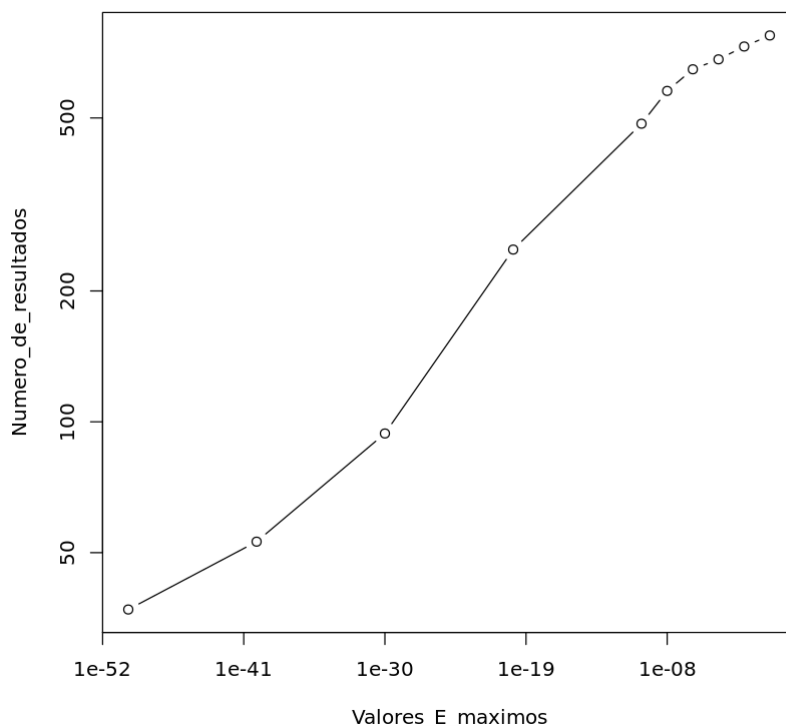
Para saber cuántas secuencias homólogas ha encontrado **blastp** en la base de datos con cada valor del parámetro `-eval`, aplico la función `dim()` que devuelve las dimensiones del objeto en cuestión (número de filas y de columnas) y me quedo con la primera dimensión

(número de filas):

```
In [8]: Numero_de_resultados <- sapply(Lista_de_Tablas, function(x) dim(x)[1])
Numero_de_resultados
```

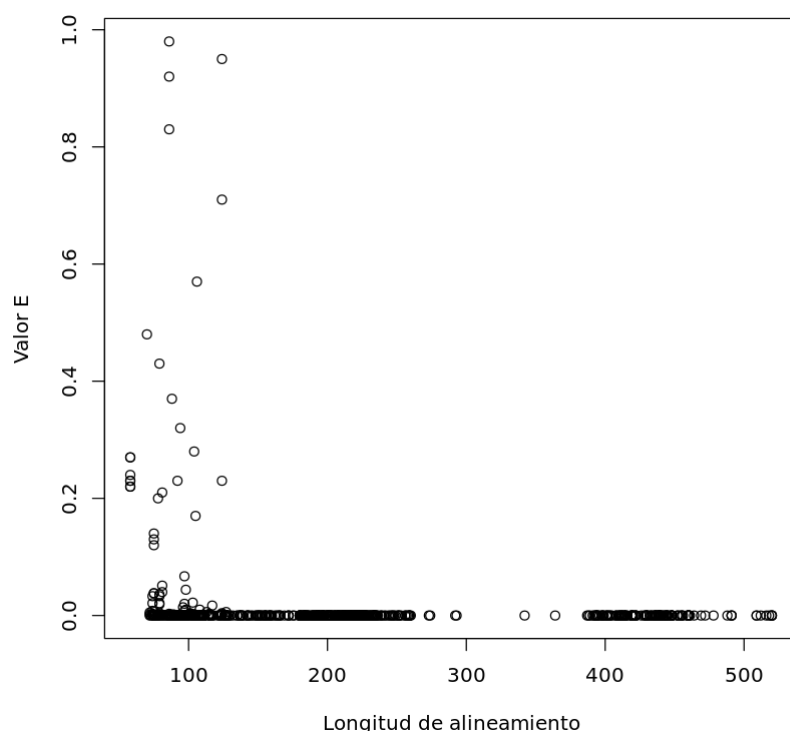
37 · 53 · 94 · 249 · 485 · 577 · 647 · 682 · 730 · 774

```
In [9]: # En la representación gráfica, se puede usar escala logarítmica solamente
# en un eje ("log='x'"), en los dos, o en ninguno.
plot(Valores_E_maximos, Numero_de_resultados, log = 'xy', type = 'b')
```



A continuación examino la relación entre la longitud del alineamiento y el valor E de los resultados guardados en la última tabla, la número 10, en la que he usado un valor E de 1 y que contiene un mayor número de resultados.

```
In [10]: # El símbolo "$" extrae de un "data frame" una columna. Si se ejecuta
# "Lista_de_Tablas[[3]]$length", se vera que el resultado es el vector
# de las longitudes de los alineamientos de la tercera tabla, por ejemplo.
plot(Lista_de_Tablas[[10]]$length, Lista_de_Tablas[[10]]$value,
      xlab = 'Longitud de alineamiento', ylab = 'Valor E')
```



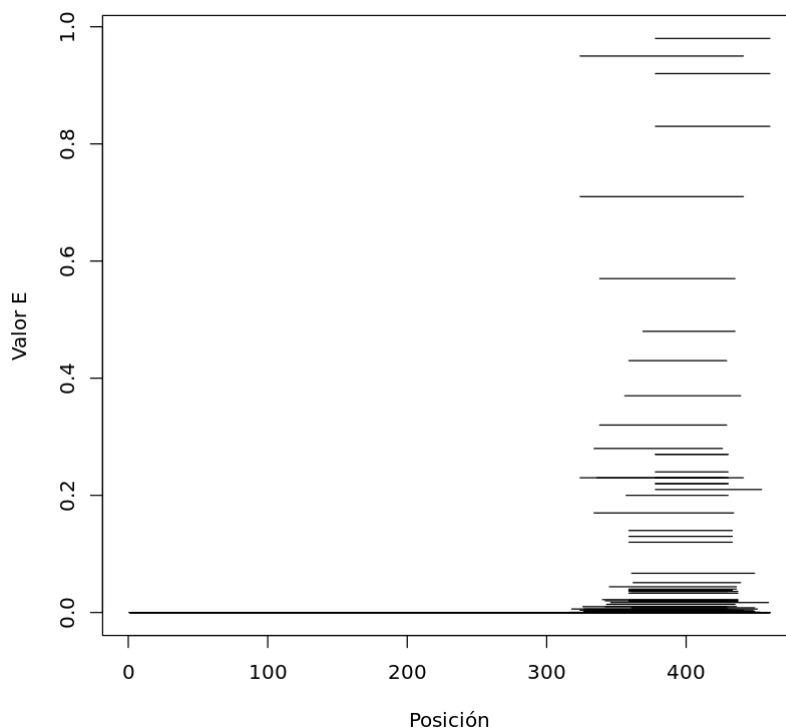
En este gráfico se muestran las coincidencias en los alineamientos y la posición de los círculos indica la longitud de ese alineamiento. Considerando que la proteína que estamos analizando CHRM1 es de 460 aminoácidos y que la mayoría de secuencias homólogas están entre 50 / 260 y entre 400 / 520 podemos deducir que existen zonas muy conservadas de esta proteína. Los valores de E también son muy parecidos, solo 24 superan el 0.1. Este resultado refuerza la idea de que en esta proteína existen zonas muy bien conservadas (aunque parciales respecto a CHRM1) que llevan a cabo una función muy específica que la naturaleza no ha querido perder.

Cabe preguntar qué posiciones de CHRM1 son esas que participan en los alineamientos parciales.

In [11]:

```
inicios <- Lista_de_Tablas[[10]]$qstart
finales <- Lista_de_Tablas[[10]]$qend
valoresE <- Lista_de_Tablas[[10]]$evalue

# La función segments() añade segmentos a un gráfico previo. Por eso, repre
# antes un gráfico con los rangos adecuados, pero vacío (type='n'). Para de
# el rango del eje horizontal, determino primero el valor máximo de "qend".
AlineamientoMaximo <- max(Lista_de_Tablas[[10]]$qend)
plot(c(0, AlineamientoMaximo), range(valoresE), type='n', xlab='Posición',
segments(inicios, valoresE, finales, valoresE))
```



Este gráfico muestra que la homología encontrada con valores E mayores no está distribuida aleatoriamente a lo largo de la secuencia de CHRM1, se concentra entre las posiciones 320 y 460, aproximadamente. Se puede asegurar que la región C-terminal es la conservada y es la que tiene fundamental importancia para el buen funcionamiento de la proteína.

Por último, voy a explorar la distribución taxonómica de las secuencias encontradas. La columna `sblastname` ofrece una clasificación fácilmente interpretable de las secuencias. Una posibilidad es contar en cada tabla las secuencias encontradas de cada grupo taxonómico. Para ello, aplico la función `table()` a esa columna de cada tabla en `Lista_de_Tablas`:

In [12]:

```
lapply(Lista_de_Tablas, function(x) table(x$sblastname))
```

[[1]]

	birds	crustaceans	even-toed ungulates	f
lies	3	1	5	
1				
	frogs & toads	nematodes	primates	rod
ents	1	1	13	
12				

[[2]]

	birds	bony fishes	crustaceans	even-toed ungul
ates	3	4	1	
7				
	flies	frogs & toads	moths	nemat
odes	1	1	1	
1				

	primates 15	rodents 19		
[[3]]				
	birds	bony fishes	carnivores	crustac
eans	4	6	2	
1				
even-toed ungulates		flies	frogs & toads	gastro
1	8	2	1	
1	grasshoppers	marsupials	moths	nemat
odes	2	1	3	
2				
odd-toed ungulates		primates	rabbits & hares	rod
ents	2	26	1	
32				
[[4]]				
	birds	bony fishes	carnivores	crustac
eans	6	18	16	
2				
even-toed ungulates		flies	frogs & toads	gastro
1	23	12	8	
2	grasshoppers	insectivores	marsupials	m
oths	2	2	4	
5				
nematodes	odd-toed ungulates		placentals	prim
ates	7	3	7	
52				
rabbits & hares		rodents		
	5	75		
[[5]]				
	birds	bivalves	bony fishes	carniv
ores	11	1	25	
29				
cephalopods		crustaceans	even-toed ungulates	f
lies	5	3	44	
23				
frogs & toads		gastropods	grasshoppers	insectiv
ores	13	6	2	
2				
lancelets		lizards	marsupials	m
oths	2	1	4	
7				
nematodes	odd-toed ungulates		placentals	prim
ates	14	4	7	
100				
rabbits & hares		rodents	viruses	
	11	170	1	
[[6]]				

	birds	bivalves	bony fishes	carniv
ores	11	1	27	
36	cephalopods	crustaceans	even-toed ungulates	f
lies	6	3	49	
28	frogs & toads	gastropods	grasshoppers	insectiv
ores	15	6	2	
2	lancelets	lizards	marsupials	m
oths	2	1	6	
7	nematodes	odd-toed ungulates	placentals	prim
ates	14	4	7	
117	rabbits & hares	rodents	viruses	
	13	218	2	

[[7]]

	birds	bivalves	bony fishes	carniv
ores	12	1	32	
43	cephalopods	crustaceans	even-toed ungulates	f
lies	6	3	58	
29	frogs & toads	gastropods	grasshoppers	insectiv
ores	18	7	2	
2	lancelets	lizards	marsupials	m
oths	2	1	6	
8	nematodes	odd-toed ungulates	placentals	prim
ates	14	5	7	
132	rabbits & hares	rodents	viruses	
	13	244	2	

[[8]]

	birds	bivalves	bony fishes	carniv
ores	12	1	32	
44	cephalopods	crustaceans	even-toed ungulates	f
lies	6	3	61	
29	frogs & toads	gastropods	grasshoppers	insectiv
ores	20	9	2	
2	lancelets	lizards	marsupials	m
oths	2	1	7	
8	nematodes	odd-toed ungulates	placentals	prim
ates	14	5	7	
140				

```

        rabbits & hares          rodents          viruses
              14                  261                  2

[[9]]

        birds          bivalves          bony fishes          carniv
ores              14                  1                  34
47
        cephalopods          crustaceans even-toed ungulates          f
lies              7                  3                  67
30
        frogs & toads          gastropods          grasshoppers          insectiv
ores              21                  9                  2
2
        lancelets          lizards          marsupials          m
oths              2                  1                  7
8
        nematodes odd-toed ungulates          placentals          prim
ates              15                  5                  7
150
        rabbits & hares          rodents          viruses
              17                  279                  2

[[10]]

        birds          bivalves          bony fishes          carniv
ores              15                  2                  35
50
        cephalopods          crustaceans even-toed ungulates          f
lies              10                  3                  73
32
        frogs & toads          gastropods          grasshoppers          insectiv
ores              22                  9                  2
2
        lancelets          lizards          marsupials          m
oths              3                  1                  7
9
        nematodes odd-toed ungulates          placentals          prim
ates              15                  5                  7
162
        rabbits & hares          rodents          viruses

```

El resultado del código anterior es una lista de recuentos de los diferentes grupos taxonómicos presentes en cada tabla de resultados de blastp. Se observa, que esta proteína se encuentra en muchos taxones distintos: insectos, gasterópodos, pájaros, ranas y crustáceos pero sobretodo en roedores y primates. Para ver la relación entre la longitud del alineamiento y la distribución taxonómica, a continuación repito el recuento de grupos taxonómicos pero contando solamente las secuencias con un alineamiento de al menos 380 aminoácidos. Es un alineamiento muy largo respecto a la longitud de mi secuencia query que tiene 460 aminoácidos, así que debería encontrar homologías solo en taxones cercanos a primates y roedores.

In [15]:

```
lapply(Lista_de_Tablas, function(x) {
  filtro <- x$length >= 380
  table(x[filtro, 'sblastname'])
})
```

[[1]]

	birds	crustaceans	even-toed ungulates	f
lies	3	1	5	
0				
	frogs & toads	nematodes	primates	rod
ents	1	0	13	
10				

[[2]]

	birds	bony fishes	crustaceans	even-toed ungul
ates	3	4	1	
7				
	flies	frogs & toads	moths	nemat
odes	0	1	1	
0				
	primates	rodents		
	15	15		

[[3]]

	birds	bony fishes	carnivores	crustac
eans	4	6	1	
1				
even-toed ungulates		flies	frogs & toads	gastro
pod	8	1	1	
1				
	grasshoppers	marsupials	moths	nemat
odes	0	0	3	
0				
odd-toed ungulates		primates	rabbits & hares	rod
ents	2	25	1	
26				

[[4]]

	birds	bony fishes	carnivores	crustac
eans	4	10	4	
1				
even-toed ungulates		flies	frogs & toads	gastro
pod	10	1	3	
1				
	grasshoppers	insectivores	marsupials	m
oths	0	2	1	
3				
	nematodes	odd-toed ungulates	placentals	prim
ates	3	3	7	
32				

	rabbits & hares		rodents		
	2		33		
[[5]]					
	birds		bivalves	bony fishes	carniv
ores	4		0	10	
4	cephalopods		crustaceans	even-toed ungulates	f
lies	0		1	13	
1	frogs & toads		gastropods	grasshoppers	insectiv
ores	3		1	0	
2	lancelets		lizards	marsupials	m
oths	0		0	1	
3	nematodes	odd-toed ungulates		placentals	prim
ates	3		3	7	
36	rabbits & hares		rodents	viruses	
	2		43	0	

[[6]]					
	birds		bivalves	bony fishes	carniv
ores	4		0	10	
4	cephalopods		crustaceans	even-toed ungulates	f
lies	0		1	14	
2	frogs & toads		gastropods	grasshoppers	insectiv
ores	3		1	0	
2	lancelets		lizards	marsupials	m
oths	0		0	1	
3	nematodes	odd-toed ungulates		placentals	prim
ates	3		3	7	
37	rabbits & hares		rodents	viruses	
	2		46	0	

[[7]]					
	birds		bivalves	bony fishes	carniv
ores	4		0	10	
4	cephalopods		crustaceans	even-toed ungulates	f
lies	0		1	15	
2	frogs & toads		gastropods	grasshoppers	insectiv
ores	3		1	0	
2	lancelets		lizards	marsupials	m
oths	0		0	1	

3					
ates	nematodes	odd-toed ungulates	placentals		prim
38	3	3	7		
	rabbits & hares	rodents	viruses		
	2	46	0		
[[8]]					
ores	birds	bivalves	bony fishes		carniv
4	4	0	10		
lies	cephalopods	crustaceans	even-toed ungulates		f
2	0	1	15		
ores	frogs & toads	gastropods	grasshoppers		insectiv
2	3	1	0		
oths	lancelets	lizards	marsupials		m
3	0	0	1		
ates	nematodes	odd-toed ungulates	placentals		prim
38	3	3	7		
	rabbits & hares	rodents	viruses		
	2	46	0		
[[9]]					
ores	birds	bivalves	bony fishes		carniv
4	4	0	10		
lies	cephalopods	crustaceans	even-toed ungulates		f
2	0	1	15		
ores	frogs & toads	gastropods	grasshoppers		insectiv
2	3	1	0		
oths	lancelets	lizards	marsupials		m
3	0	0	1		
ates	nematodes	odd-toed ungulates	placentals		prim
38	3	3	7		
	rabbits & hares	rodents	viruses		
	2	46	0		
[[10]]					
ores	birds	bivalves	bony fishes		carniv
4	4	0	10		
lies	cephalopods	crustaceans	even-toed ungulates		f
2	0	1	15		
ores	frogs & toads	gastropods	grasshoppers		insectiv

	3	1	0	
2				
	lancelets	lizards	marsupials	m
oths				
	0	0	1	
3				
	nematodes	odd-toed ungulates	placentals	prim
ates				
	3	3	7	
38				
	rabbits & hares	rodents	viruses	

A pesar del valor de homología (alto) de 380 aminoácidos siguen encontrándose homologías en la mayoría de taxones: roedores, pájaros, peces, crustáceos, suinos, moscas, conejos y gasterópodos.

Discusión

Utilizando la secuencia de la proteína CHRM1 de *Homo sapiens* como consulta (*query*) en búsquedas de blastp, encontramos proteínas homólogas en multitud de especies entre los animales superiores, incluyendo insectos, nematodos, gasterópodos y virus. El hecho de que esté en taxones muy diferentes testimonia el hecho de que esta proteína aparece en un ancestro común a los taxones expuestos anteriormente. Esta conclusión hace que estudiar el origen de esta proteína pueda ser útil para la diferenciación filogenética en animales. La distribución de los alineamientos a lo largo de la secuencia de CHRM1 (de 460 aminoácidos) muestra que gran parte de la homología encontrada se limita a una región cerca de 100 aminoácidos, aproximadamente entre los residuos 320 y 450. Esta región constituye un dominio conservado.

En dos tablas el blastp además de las secuencias esperadas en taxones cercanos a *Homo sapiens* y mamíferos, se ha encontrado una secuencia de virus de la especie *Sheeppox virus KS-1* (secuencia identificadora: Q86917.1) que muestra valores de homología muy parecidos a los que muestra *Homo sapiens* identificada con Q9P1P4.1. Una explicación que me atrevo dar, es que este virus haya transferido esta proteína al primate y los únicos cambios relevantes se han dado en la primera y última posición de la secuencia alineada del virus y del *Homo sapiens* (subject). Observando los valores de longitud de alineamiento: 198 para la secuencia de virus y 193 para la de *Homo sapiens* y que la primera posición de la query es la 42 en las dos lecturas, está claro que las únicas diferencias en los valores de primera y última posición alineadas de subject, se deben a una pérdida de aproximadamente 60 nucleótidos en la transferencia de virus a primate.

No he encontrado ningún artículo o indicio que pruebe esta teoría en el NCBI (link en referencias). Si mi explicación se revelara cierta estaríamos ante una secuencia aminoacídica de virus que ha tenido un rol importante en la evolución del sistema nervioso humano. Se podría añadir al pool secuencias con estas características que ya conocemos. Otra posibilidad sería que el virus haya adquirido el gen directamente del *Homo sapiens* recientemente y siga mostrando mucha homología. Una tercera respuesta puede residir en la posibilidad de que este virus haya transferido esta proteína al organismo ancestro común de los grupos que hemos encontrado. Para poder testear estas hipótesis se deben hacer varios estudios bioinformáticos: un alineamiento Muscle, usando MEGA por ejemplo, y otro estudio parecido al presente para ver en qué otros posibles taxones podemos encontrar evidencias de su

antigüedad y así poder descubrir el origen de esta proteína.

Referencias

- Stephen F. Altschul¹Warren Gish¹Webb Miller²Eugene W. Myers³David J. Lipman¹. Basic local alignment search tool. *JMB* Volume 215, Issue 3, 5 October 1990, Pages 403-410
- Pawel Burkhardt , Stegmann CM, Cooper B, Kloepper TH, Imig C, Varoqueaux F, Wahl MC, Fasshauer D. Primordial neurosecretory apparatus identified in the choanoflagellate *Monosiga brevicollis*. *PMID* 29 Aug 2011, 108(37):15264-15269 ,PMID: 21876177 PMCID: PMC3174607 [doi: 10.1073/pnas.1106189108]
- Pawel Burkhardt. The origin and evolution of synaptic proteins - choanoflagellates lead the way. *PabMed* 2015 Feb 15;218(Pt 4):506-14. [doi: 10.1242/jeb.110247]
- Richard David Emes, Seth G N Grant. The human postsynaptic density shares conserved elements with proteomes of unicellular eukaryotes and prokaryotes *PubMed* 2011 Mar 31;5:44. [doi: 10.3389/fnins.2011.00044]
- Richard D Emes, Seth G N Grant. Evolution of synapse complexity and diversity *PubMed* 2012;35:111-31. [doi: 10.1146/annurev-neuro-062111-150433.]
- Onur Sakarya, Kathryn A. Armstrong, Maja Adamska, Marcin Adamski, I-Fan Wang, Bruce Tidor, Bernard M. Degnan, Todd H. Oakley, Kenneth S. Kosik. A Post-Synaptic Scaffold at the Origin of the Animal Kingdom *PlosOne* Published: June 6, 2007 [<https://doi.org/10.1371/journal.pone.0000506>]
- John C Silbereis, Sirisha Pochareddy, Ying Zhu, Mingfeng Li, Nenad Sestan. The Cellular and Molecular Landscapes of the Developing Human Central Nervous System *PubMed* 2016 Jan 20;89(2):248-68. [doi: 10.1016/j.neuron.2015.12.008]
- Hiroshi Suga, Zehua Chen, Alex de Mendoza et al. The Capsaspora genome reveals a complex unicellular prehistory of animals *PubMed* 2013;4:2325. [doi: 10.1038/ncomms3325]
- Lucas Henriques Viscardi, Danilo Oliveira Imparato, Maria Cátira Bortolini, Rodrigo Juliani Siqueira Dalmolin. Ionotropic Receptors as a Driving Force behind Human Synapse Establishment *Molecular Biology and Evolution*, Volume 38, Issue 3, March 2021, Pages 735–744 [<https://doi.org/10.1093/molbev/msaa252>]
- secuencia proteica de la proteína virica con nombre Q86917.1 de Sheeppox virus KS-1 <https://www.ncbi.nlm.nih.gov/protein/2495049/>

In []: