Informe Global

TÉCNICAS Y MODELOS ALGORÍTMICOS

Pilar Piñeiro Pereda

2016

ÍNDICE

INTRODUCCIÓN

En este informe se analizan los temas, códigos y resultados vistos durante la asignatura de Técnicas y Modelos Algorítmicos.

Se han realizado unos programas para calcular los k-mers, los dinucleótidos y los codones de una secuencia en formato FASTA usando como parámetros la k (número de k-mers, que será 2 para dinucleótidos) y el nombre de archivo de secuencia.
También se ha realizado un código en C y en Python para calcular un diccionario a partir de la secuencia, y para comparar distintos diccionarios y establecer similitudes.

Se han usado estos resultados para sacar conclusiones acerca de las secuencias estudiadas.

EVALUACIÓN DE RENDIMIENTO

- TIEMPOS DE EJECUCIÓN SEGÚN LONGITUD Y K PARA EL ALGORITMO KMERS

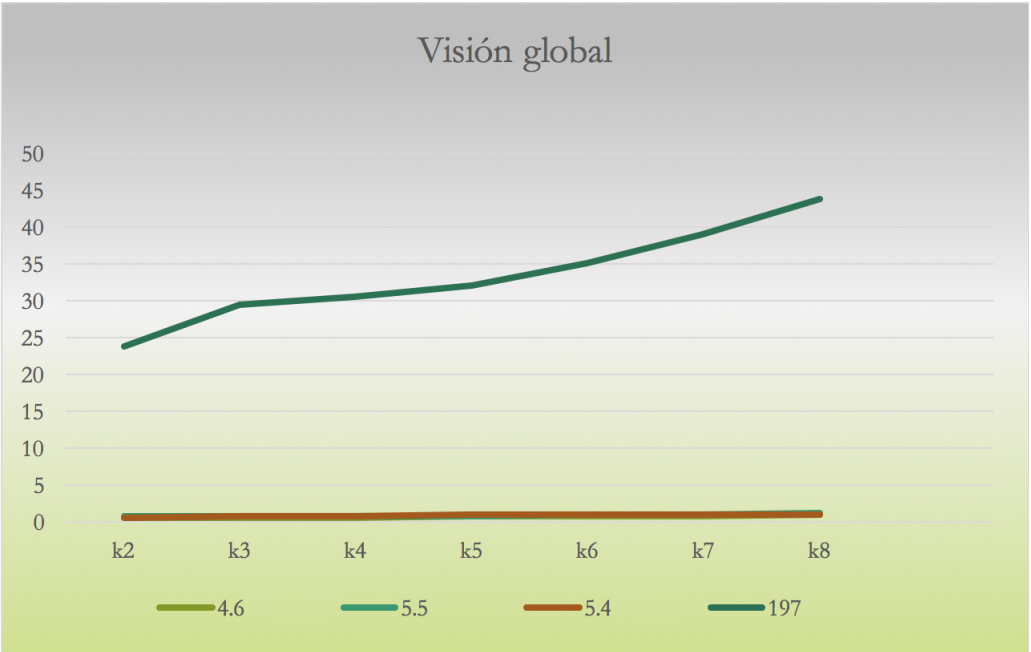
Para observar cómo varia el tiempo de ejecución según el tamaño del K-mer (K) y la longitud de la secuencia (MB), he ejecutado el algoritmo con el comando "time" y he tomado el "real time" para los valores de K 2, 3, 4, 5, 6, 7 y 8. He hecho esto para cuatro secuencias distintas de diferente longitud: Mus Musculus, E.Colik12, E.ColiO y Schizosaccharomyces Pombe .

En la tabla siguiente podemos ver los datos obtenidos tras las ejecuciones del algoritmo:

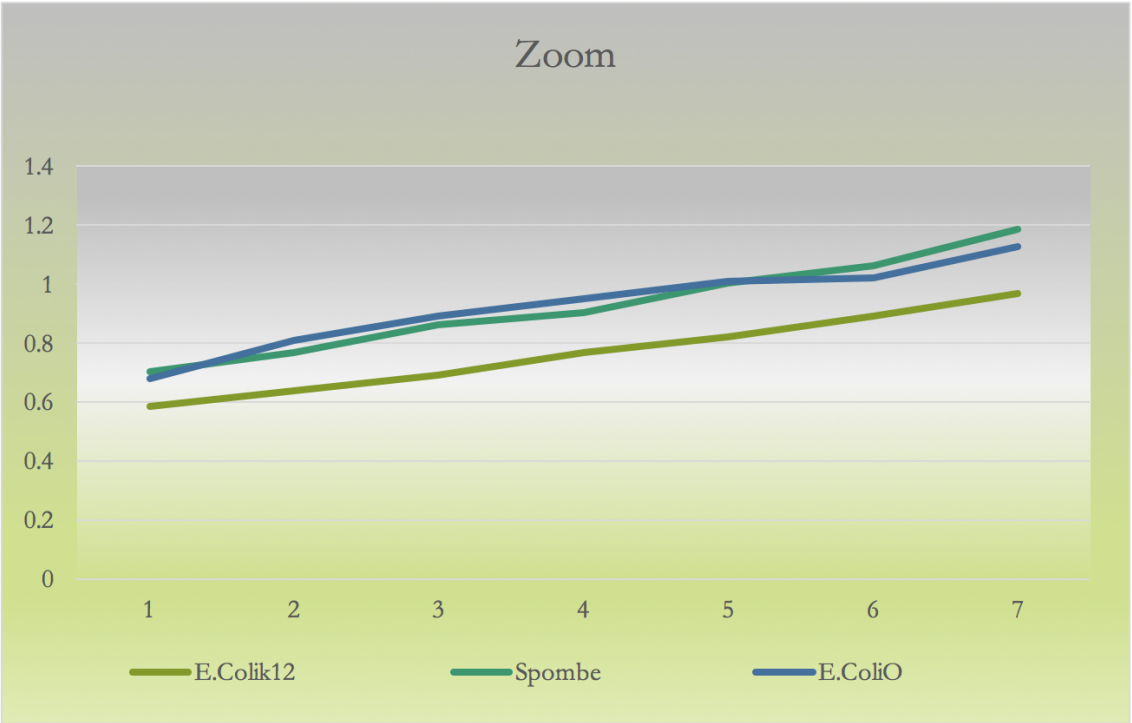
Organismo	**Longitud (MBp)**	**k=2**	**k=3**	**k=4**	**k=5**	**k=6**	**k=7**	**k=8**
E.Colik 12	4.6	0.58	0.63	0.69	0.76	0.82	0.89	0.96
Spombe	5.5	0.70	0.76	0.86	0.90	1.00	1.06	1.18
E.ColiO	5.4	0.68	0.80	0.89	0.95	1.01	1.02	1.12
MusMusculus	197	23.91	29.51	30.62	32.03	35.16	39.00	43.93

A simple vista, la tabla muestra que el tiempo de ejecución de la secuencia de MusMusculus es mucho mayor, con lo que en la gráfica nos será difícil comparar todos los tiempos.

La gráfica muestra en el eje x los valores para K, en el eje Y los tiempos en segundos y cada color de línea corresponde a una secuencia distinta, con su correspondiente longitud.



Los tiempos de ejecución de las secuencias de E.Colik12, E.ColiO y Schizosaccharomyces Pombe quedan escondidas debajo de la de Mus Musculus, que tiene un tiempo de ejecución que aumenta linealmente. Parecería que el tiempo para las secuencias más cortas se mantiene constante, pero esto se debe a las distintas escalas. Veamos un zoom de las secuencias más cortas:



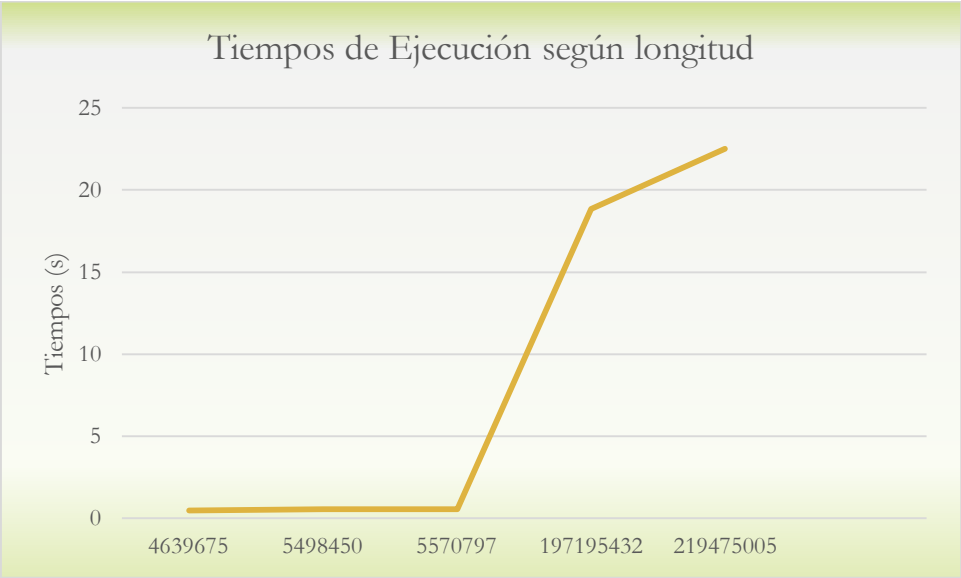
Aquí vemos más claro el incremento (también lineal) del tiempo en la ejecución de estas tres secuencias, que en la gráfica global parecía sin pendiente, debido a las diferencias de escala.

Vemos que el tiempo crece según aumenta la k, pero en mucha menor proporción que como aumenta con la secuencia más larga (la del ratón o Mus Musculus).

- TIEMPOS DE EJECUCIÓN SEGÚN LONGITUD PARA EL ALGORITMO CODONES

Para ver el crecimiento del tiempo de ejecución a medida que la longitud va aumentando según la secuencia, usamos las mismas secuencias que usamos para Kmers: Mus Musculus , Homo Sapiens, EColik12 , EColiO y Schizosaccharomyces Pombe. Ejecutamos con 'time' y tomamos el real time. Los datos recogidos se presentan a continuación:

Organismo	**Tiempo (s)**	**Longitud(MBp)**
E.colik12	0.47	4.6
E.coliO	0.55	5.5
MusMuscu	18.85	197.2
HomoSapiens	22.50	219.4
Spombe	0.54	5.5



El tiempo de ejecución aumenta sin una gran pendiente hasta llegar a la secuencia del ratón (Mus Musculus) , donde se incrementa linealmente con una gran pendiente. La diferencia entre el tiempo de ejecución para el humano y el ratón no es tan grande, con lo que entre ambos tiempos de ejecución la pendiente disminuye.

Si comparamos los tiempos de ejecución mas cortos para K-mers (k=2) y los de Codones para las mismas secuencias, vemos que el algoritmo Codones es más rápido en ejecución y sus tiempos escalan más lentamente al aumentar la longitud que en el algoritmo K-mers.

Organismo	**Longitud(MBp)**	**Tiempo K-mers k=2 (s)**	**Tiempo Codones (s)**
E.Colik12	4.6	0.58	0.47
Spombe	5.6	0.70	0.54
E.ColiO	5.5	0.68	0.55
MusMusculus	197.19	23.91	18.85
	Media->	1.61	1.27

FRECUENCIAS RELATIVAS DE DINUCLEÓTIDOS PARA DISTINTAS ESPECIES

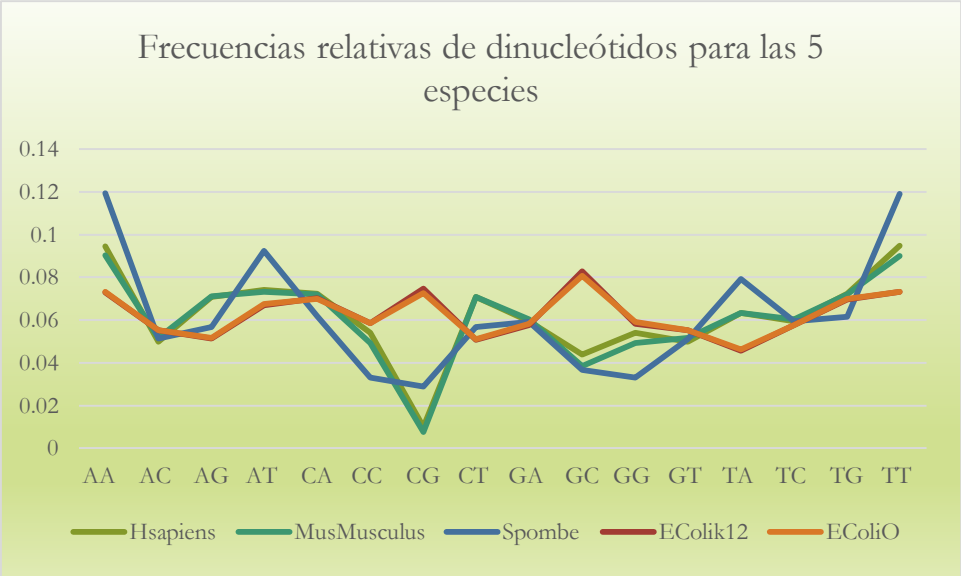
Con el algoritmo K-mers vamos a comparar las frecuencias de k-mers con K=2 (dinucleótidos) para distintas especies:
E.Colik12, E.ColiO, HomoSapiens, MusMusculus y Schizosaccharomyces Pombe.

El objetivo de este experimento es observar la semejanza que hay entre las frecuencias para cada dinucleótido entre las dos E.Coli y entre HomoSapiens y MusMusculus. Veremos que Schizosaccharomyces Pombe no se asemeja a ninguna de las secuencias de las especies anteriormente nombradas. He usado los archivos proporcionados en el Campus Virtual para este estudio.

La tabla siguiente muestra los resultados obtenidos:

Dinucleotidos	**Hsapiens**	**MusMusculus**	**Spombe**	**EColik12**	**EColiO**
AA	0.09	0.09	0.11	0.07	0.07
AC	0.04	0.05	0.05	0.05	0.05
AG	0.07	0.07	0.05	0.05	0.05
AT	0.07	0.07	0.09	0.06	0.06
CA	0.07	0.07	0.06	0.07	0.06
CC	0.05	0.04	0.03	0.05	0.05
CG	0.01	0.01	0.02	0.07	0.07
CT	0.07	0.07	0.05	0.05	0.05
GA	0.05	0.06	0.05	0.05	0.05
GC	0.04	0.03	0.03	0.08	0.08
GG	0.05	0.04	0.03	0.05	0.05
GT	0.04	0.05	0.05	0.05	0.05
TA	0.06	0.06	0.07	0.04	0.04
TC	0.05	0.06	0.05	0.05	0.05
TG	0.07	0.07	0.06	0.06	0.06
TT	0.09	0.08	0.11	0.07	0.07

Graficando estos resultados, obtenemos la siguiente gráfica:



La frecuencia de dinucleótidos en la secuencia de HomoSapiens se asemeja a la de Mus Musculus, y la de E.Colik12 se asemeja a la de E.ColiO.
Entre estas dos semejanzas se observa una diferencia: E.ColiO y E.Colik12 son más parecidas (cercanas) entre sí. Vamos a cuantificar esta diferencia.

Para ello, calculamos la correlación de Pearson entre las frecuencias de Mus Musculus y HomoSapiens, y repetimos el proceso para E.Colik12 y E.ColiO:

PearsonEColi	**PearsonMusHSap**
0.997	0.991

Como se podía ver en la gráfica, queda demostrado que la semejanza entre las secuencias de E.Coli es mayor.

USO PREFERENTE DE CODONES POR DISTINTAS ESPECIES

El uso preferente de codones se refiere a diferencias en la frecuencia de aparición de codones sinónimos en DNA codificante. Los códigos genéticos de distintos organismos frecuentemente están desviados hacia el uso de uno de los muchos codones que codifican el mismo aminoácido, es decir, hay una mayor frecuencia de uso de un codón sobre los otros. El uso preferente de codones es más fuerte en genes que se expresan mucho.

La siguiente tabla compara el uso preferente de codones entre distintas especies:

Codons that differ from those preferred in man and rat are highlighted

Amino acid	Human		Rat		*E. coli*		*S. cerevisiae*		*S. frugiperda*	
	Preferred codon	% use	Preferred codon	% use	Preferred codon	% use	Preferred codon	% use	Preferred codon	% use
Ala	GCC	41	GCC	41	GCC	34	GCT	38	GCT	37
Arg	CGG	21	AGG	21	CGC	38	AGA	48	AGA	24
Asn	AAC	55	AAC	60	AAC	54	AAT	59	AAC	63
Asp	GAC	54	GAC	58	GAT	63	GAT	65	GAC	58
Cys	TGC	56	TGC	56	TGC	55	TGT	63	TGC	58
Gln	CAG	75	CAG	76	CAG	66	CAA	69	CAG	51
Glu	GAG	59	GAG	62	GAA	68	GAA	71	GAG	52
Gly	GGC	35	GGC	35	GGC	39	GGT	47	GGA	32
His	CAC	59	CAC	62	CAT	57	CAT	64	CAC	60
Ile	ATC	50	ATC	55	ATT	50	ATT	46	ATC	47
Leu	CTG	41	CTG	42	CTG	49	TTG	29	CTG	31
Lys	AAG	58	AAG	64	AAA	75	AAA	58	AAG	58
Met	ATG	100	ATG	100	ATG	100	ATG	100	ATG	100
Phe	TTC	56	TTC	60	TTT	57	TTT	59	TTC	65
Pro	CCC	33	CCC	32	CCG	51	CCA	41	CCT/CCA	29
Ser	AGC	24	AGC	25	AGC	26	TCT	27	TCC	20
Thr	ACC	37	ACC	38	ACC	42	ACT	35	ACT	32
Trp	TGG	100	TGG	100	TGG	100	TGG	100	TGG	100
Tyr	TAC	57	TAC	61	TAT	58	TAT	56	TAC	67
Val	GTG	48	GTG	48	GTG	36	GTT	39	GTG	39
Trm	TGA	51	TGA	50	TAA	62	TAA	48	TAA	64

**Los codones que difieren de los preferidos en Homo Sapiens y Mus Musculus están resaltados.*
Fuente: http://www.uky.edu/Pharmacy/ps/porter/CodonUsage/preferred_codons.htm

PROCEDIMIENTO

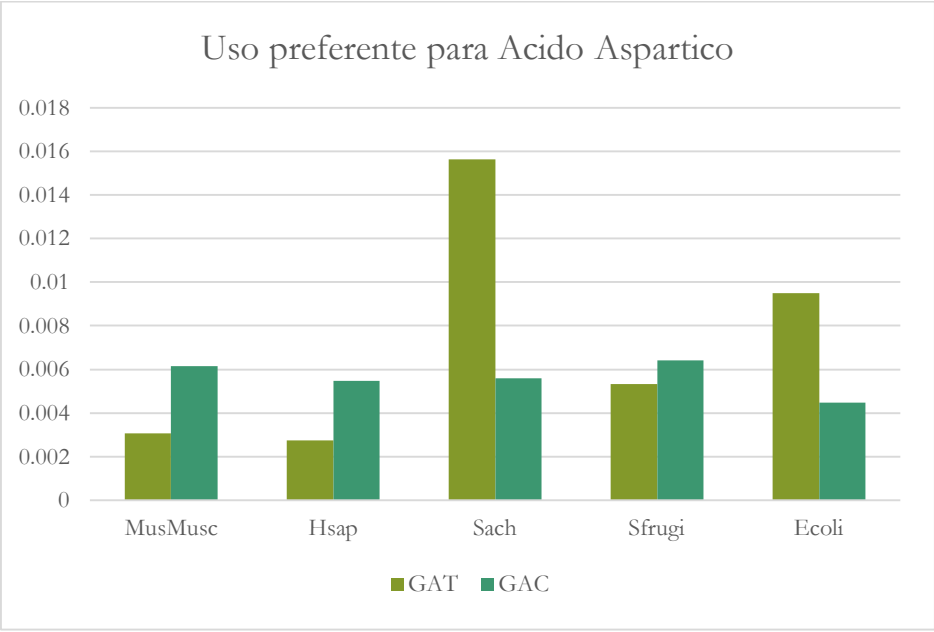
Es importante destacar que, para comparar uso de codones entre distintas especies, debemos comparar genes. Con genes obtenidos para cada una de las especies que aparecen en la tabla, he ejecutado el algoritmo Codones para obtener las frecuencias relativas de los codones usados por esos genes, y así poder observar si mis resultados se corresponden con los de la tabla anterior. Las secuencias de los genes de cada especie han sido obtenidas de http://www.ncbi.nlm.nih.gov/ .

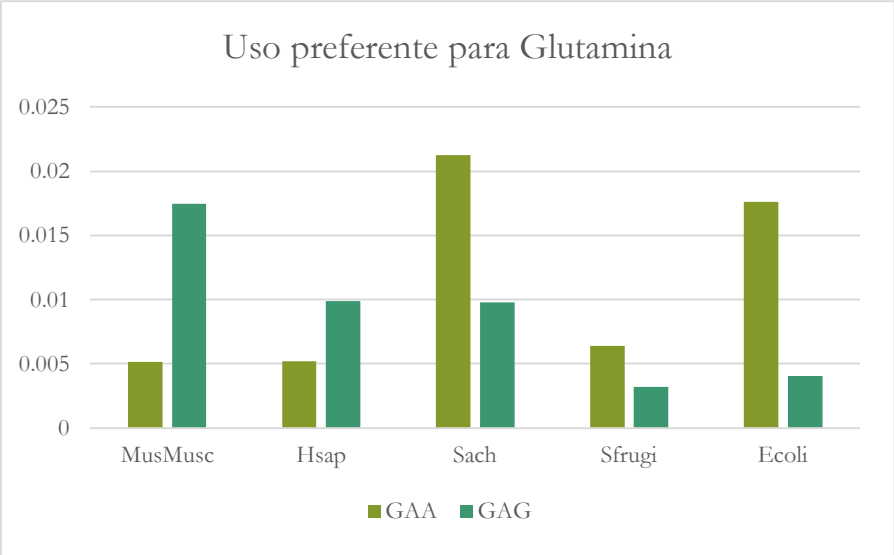
Codon	MusMusc	Hsap	Sach	Sfrugi	Ecoli	
AAA	0.0041	0.0010	0.0254	0.0063	0.0284	Lys
AAT	0.0020	0.0021	0.0116	0.0127	0.0189	Asn
AAC	0.0054	0.0021	0.0069	0.0010	0.0081	Asn
AAG	0.0095	0.0046	0.0116	0.0031	0.0040	Lys
ATA	0.0044	0.0010	0.0076	0.0159	0.0216	Ile
ATT	0.0034	0.0027	0.0142	0.0351	0.0081	Ile
ATC	0.0044	0.0021	0.0051	0.0001	0.0054	Ile
ATG	0.0054	0.0035	0.0065	0.0031	0.0013	Met
ACA	0.0102	0.0046	0.0069	0.0010	0.0149	Thr
ACT	0.0034	0.0038	0.0083	0.0085	0.0027	Thr
ACC	0.0054	0.0043	0.0025	0.0002	0.0027	Thr
ACG	0.0020	0.0016	0.0014	0.0003	0.0002	Thr
AGA	0.0047	0.0054	0.0081	0.0159	0.0108	Arg
AGT	0.0020	0.0021	0.0023	0.0021	0.0040	Ser
AGC	0.0044	0.0049	0.0032	0.0002	0.0067	Ser
AGG	0.0030	0.0052	0.0020	0.0012	0.0004	Arg
TAA	0.0020	0.0035	0.0010	0.0010	0.0013	stop
TAT	0.0013	0.0005	0.0083	0.0234	0.0203	Tyr
TAC	0.0041	0.0010	0.0055	0.0010	0.0067	Tyr
TAG	0.0023	0.0019	0.0011	0.0001	0.0015	stop
TTA	0.0041	0.0019	0.0097	0.0533	0.0081	Leu
TTT	0.0051	0.0035	0.0072	0.0351	0.0094	Phe
TTC	0.0065	0.0041	0.0048	0.0010	0.0040	Phe
TTG	0.0058	0.0054	0.0062	0.0031	0.0027	Leu
TCA	0.0017	0.0049	0.0039	0.0031	0.0040	Ser
TCT	0.0071	0.0082	0.0060	0.0106	0.0054	Ser
TCC	0.0061	0.0043	0.0030	0.0013	0.0021	Ser
TCG	0.0020	0.0016	0.0032	0.0011	0.0010	Ser
TGA	0.0037	0.0038	0.0002	0.0063	0.0002	Sec
TGT	0.0041	0.0032	0.0016	0.0031	0.0011	Cys
TGC	0.0027	0.0041	0.0004	0.0002	0.0013	Cys
TGG	0.0075	0.0090	0.0034	0.0021	0.0054	Trp
CAA	0.0058	0.0032	0.0090	0.0053	0.0081	Gln
CAT	0.0017	0.0032	0.0037	0.1112	0.0054	His
CAC	0.0054	0.0052	0.0016	0.0011	0.0013	His
CAG	0.0119	0.0082	0.0025	0.0011	0.0011	Gln
CTA	0.0041	0.0030	0.0034	0.0010	0.0027	Leu
CTT	0.0027	0.0046	0.0039	0.0002	0.0040	Leu
CTC	0.0078	0.0109	0.0004	0.0002	0.0013	Leu
CTG	0.0160	0.0120	0.0027	0.0002	0.0003	Leu
CCA	0.0051	0.0082	0.0046	0.0010	0.0081	Pro
CCT	0.0058	0.0120	0.0055	0.0074	0.0040	Pro
CCC	0.0061	0.0104	0.0014	0.0010	0.0002	Pro
CCG	0.0020	0.0063	0.0014	0.0002	0.0002	Pro

CGA	0.0027	0.0043	0.0011	0.0021	0.0027	Arg
CGT	0.0013	0.0016	0.0016	0.0042	0.0013	Arg
CGC	0.0078	0.0049	0.0004	0.0001	0.0002	Arg
CGG	0.0030	0.0054	0.0004	0.0010	0.0002	Arg
GAA	0.0051	0.0052	0.0212	0.0063	0.0176	Glu
GAT	0.0030	0.0027	0.0156	0.0053	0.0094	Asp
GAC	0.0061	0.0054	0.0055	0.0064	0.0044	Asp
GAG	0.0174	0.0098	0.0097	0.0031	0.0040	Glu
GTA	0.0027	0.0013	0.0025	0.0031	0.0067	Val
GTT	0.0017	0.0041	0.0065	0.0117	0.0067	Val
GTC	0.0023	0.0046	0.0027	0.0014	0.0011	Val
GTG	0.0102	0.0060	0.0023	0.0002	0.0010	Val
GCA	0.0041	0.0090	0.0039	0.0021	0.0067	Ala
GCT	0.0071	0.0087	0.0060	0.0106	0.0013	Ala
GCC	0.0102	0.0074	0.0030	0.0010	0.0040	Ala
GCG	0.0023	0.0074	0.0020	0.0012	0.0011	Ala
GGA	0.0078	0.0123	0.0027	0.0074	0.0135	Gly
GGT	0.0065	0.0071	0.0104	0.0095	0.0027	Gly
GGC	0.0065	0.0112	0.0023	0.0010	0.0013	Gly
GGG	0.0082	0.0126	0.0025	0.0042	0.0027	Gly

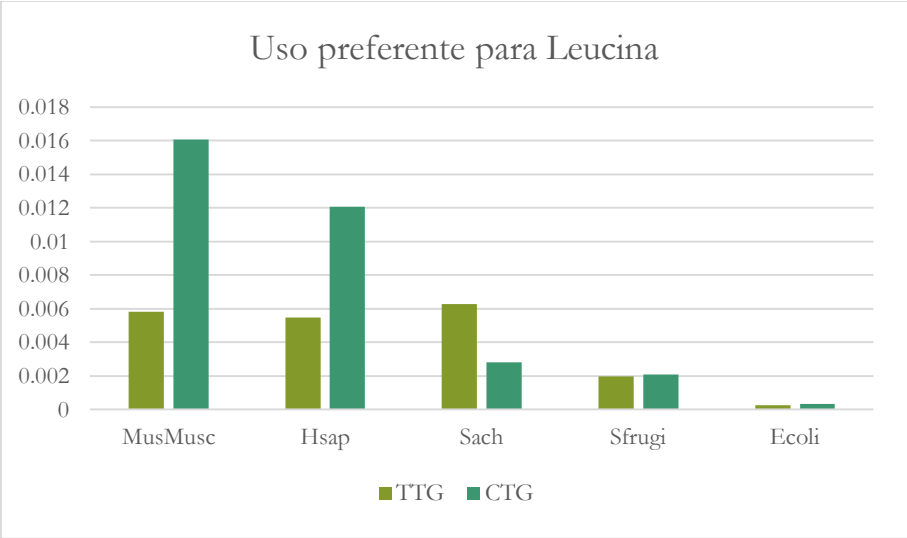
Aunque hemos obtenido las frecuencias para todos los codones, vamos a observar los codones que nos interesan y que son algunos de los que están destacados en la tabla anteriormente presentada: son los codones preferidos por las especies E coli, S.cervesiae y S.frugipeda, que difieren de los codones que son normalmente mas usados por HomoSapiens y MusMusculus para la codificación de un mismo aminoácido.

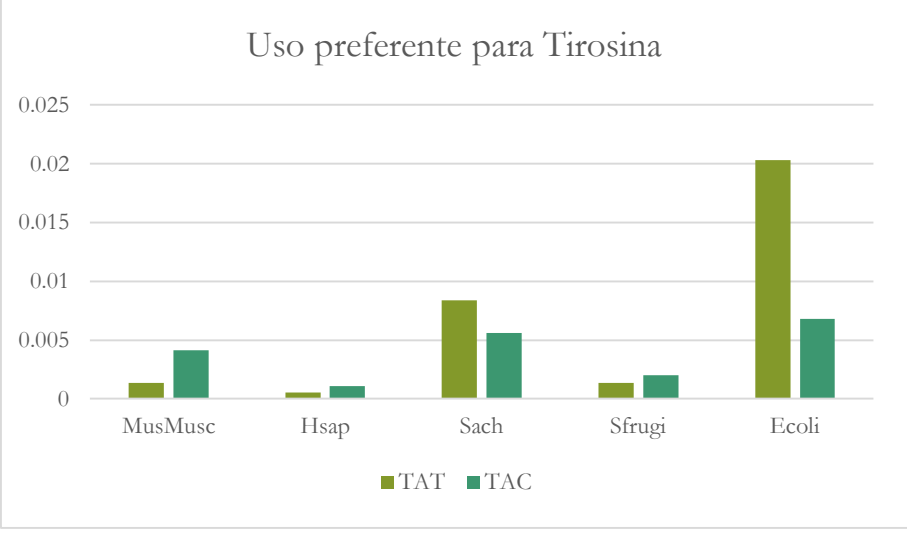
Veamos las tablas siguientes:





- Aunque en la tabla se indica que S frugi prefiere el codón GAG, para el gen que obtuvimos la frecuencia de GAA es mayor. Esto es normal, ya que contamos con un solo gen de cada especie.





Se han hecho estudios en los cuales se eliminan copias del codón más usado, resultando en un decrecimiento en la eficiencia de la traducción solo en algunos genes , que son los más expresados. Esto puede traducirse en una correlación alta entre el uso preferente de codones y la eficiencia de la traducción en genes muy expresados y por tanto altamente traducidos.

Otra razón para el uso preferente de codones en el caso de procariotas es:

> "En el caso ciertos genomas procariotas existe una presión para que los genes más expresados utilicen los codones sinónimos que se corresponden con los tRNAs más abundantes."
> *Grupo de Genómica Evolutiva. Departamento de Bioquímica y Biotecnología, Universidad Rovira i Virgili (URV), c/ Marcel·li Domingo s/n, Campus Sescelades, 43007 Tarragona*

Se conoce que el uso preferente de codones refleja el efecto de mutaciones y selección natural para la optimización del proceso de traducción (una traducción más rápida y más exacta).

LLACTOCOCUS LLACTIS Y TRANSFERENCIA HORIZONTAL DE GENES

Haciendo uso del fichero "Llactis_Genes" proporcionado en el Campus Virtual, vamos a analizar los genes de Llactococus Llactis: este fichero es multifasta, es decir, contiene muchos genes a la vez.

Para ello, será necesario adaptar el código de Kmers, con el objetivo de poder leer y procesar todos los genes del mismo fichero.

• BÚSQUEDA DE GENES OBTENIDOS POR TRANSFERENCIA HORIZONTAL.

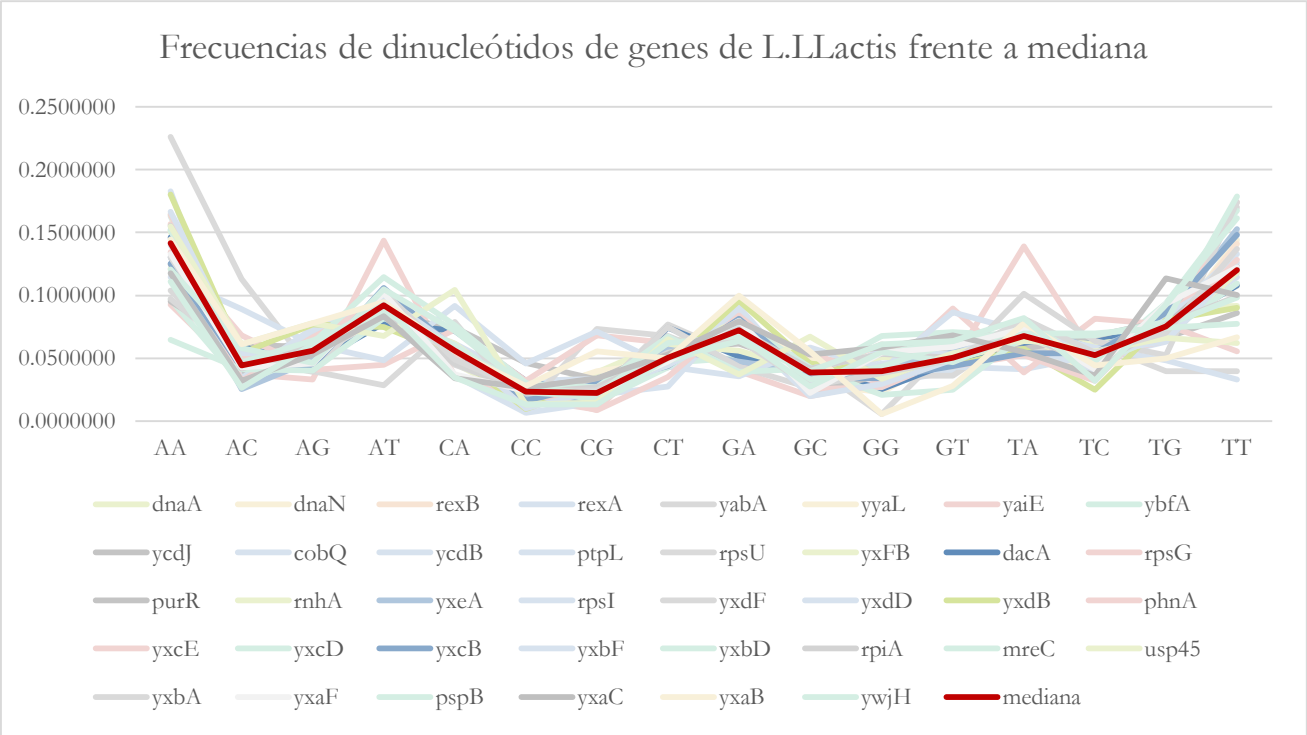
Acerca de la transferencia horizontal de genes:

La transferencia horizontal de genes o HGT en inglés (Horizontal Gene Transfer) consiste en la transmisión del genoma o parte de éste de un organismo a otro que no es su descendiente. El tipo de transferencia habitual, llamado transferencia vertical de genes, es el que se da desde un organismo a su descendencia, como ocurre en la reproducción sexual. La transferencia horizontal tiene un papel muy importante en la evolución.

Para estudiar la distribución de frecuencias en los genes de Llactococcus Llactis he ejecutado el código Kmers adaptado para un archivo multifasta. Posteriormente, grafiqué muchos de los genes obtenidos, de manera que se pudiera apreciar cada gen y su tendencia. Calculé la media para comparar estos genes con dicha media.

NOTA IMPORTANTE: para no graficar cada uno de los 100 genes manualmente, he escogido un número grande de genes de manera aleatoria para que la distribución de los datos tienda a una distribución normal. De esta forma, se consigue estudiar de forma general el organismo sin utilizar cada uno de los genes. Al ser genes escogidos de forma aleatoria y ser un número elevado de genes, podemos generalizar los resultados.

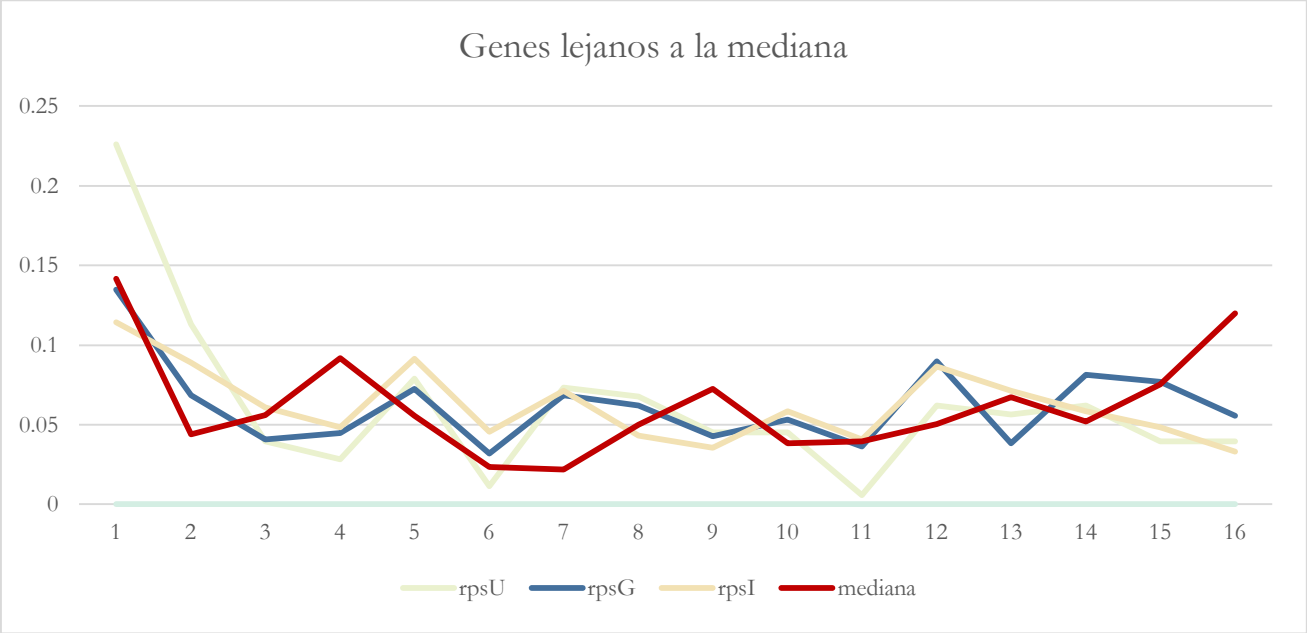
En la siguiente gráfica veremos que los genes tienen unas frecuencias similares entre sí, pero que hay algunos genes que se comportan de manera distinta a la mediana (rojo):



Correlación de Pearson con la mediana para cada gen:

Pearson	Genes
0.961386	dnaA
0.981898	dnaN
0.988788	rexB
0.980133	rexA
0.840747	yabA
0.950934	yyaL
0.754836	yaiE
0.827429	ybfA
0.920073	ycdJ
0.973152	cobQ
0.944669	ycdB
0.951884	ptpL
0.500322	rpsU
0.949996	yxFB
0.918056	dacA
0.463311	rpsG
0.956521	purR
0.97056	rnhA
0.943099	yxeA
0.158653	rpsI
0.789324	yxdF
0.952278	yxdD
0.889058	yxdB
0.953629	phnA
0.984073	yxcE
0.942267	yxcD
0.952817	yxcB
0.97882	yxbF
0.659482	yxbD
0.964616	rpiA
0.967047	mreC
0.698823	usp45
0.968658	yxbA
0.951864	yxaF
0.872031	pspB
0.842298	yxaC
0.739041	yxaB
0.879842	ywjH

Viendo los valores, podemos decir que el gen rpsI tiene una posible procedencia por transferencia horizontal, pero no es el único: rpsU y rpsG también pueden proceder de otro organismo. Vamos a graficar los genes con un valor menor a 0.6 en correlación frente a la mediana.



Saber de qué organismo proceden estos genes será de gran interés para estudiar la evolución de esta bacteria. La inferencia filogenética nos permite saber de qué modo evolucionaron los organismos, y este gen es un ejemplo de herramienta para ir rastreando la huella evolutiva de un ser vivo.

PROCESO DE CREACIÓN DE UN DICCIONARIO PARA UNA SECUENCIA DADA

Un diccionario de una secuencia es un archivo de texto en el que podemos ver las 'palabras' o fragmentos de la secuencia con la posición que ocupan dentro de ésta.

Estas 'palabras' son los k-mers (palabras de longitud k). El formato en que están mostrados los k-mers es una lista con cada k-mer sin repetirse , y para cada k-mer todas las posiciones en las que aparece:

```
seq : TCAGACGATT GAAGAATCAT n=20
pos : 0123456789 0123456789
```

A	2, 4, 7, 11, 12, 14, 15, 18	AA	11, 14	AAG	11
C	1, 5, 17	AC	4	AAT	14
G	3, 6, 10, 13	AG	2, 12	ACG	4
T	0, 8, 9, 16, 19	AT	7, 15, 18	AGA	2, 12
		CA	1, 17	ATC	15
		CC		ATT	7
		CG	5	CAG	1
		CT		CAT	17
		GA	3, 6, 10, 13	CGA	5
		GC		GAA	10, 13
		GG		GAC	3
		GT		GAT	6
		TA		TCA	0, 16
		TC	0, 16	TGA	9
		TG	9	TTG	8
		TT	8		

Tabla obtenida de una diapositiva del campus virtual

El diccionario es de una gran utilidad para calcular la semejanza que hay entre dos secuencias: si ambas presentan el mismo k-mer en la misma posición, podemos considerar que hay una 'coincidencia' y sumar las coincidencias para sacar conclusiones.

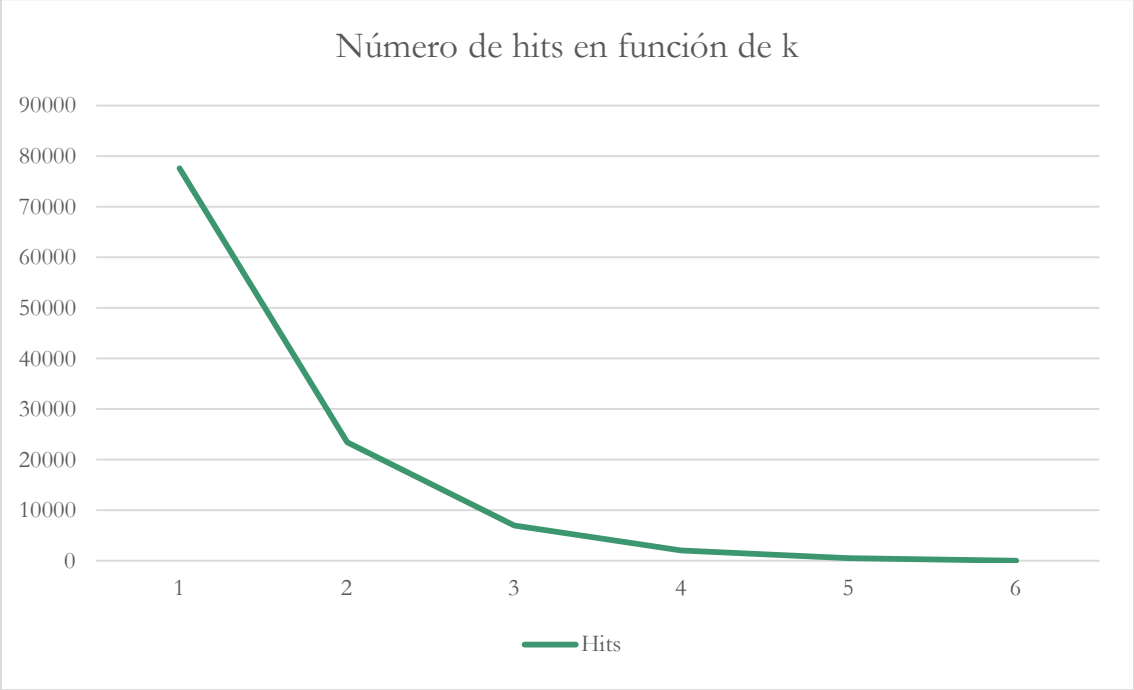
PROCEDIMIENTO EN C:

- Cogemos los k-mers de la secuencia **con solapamiento** y guardamos la posición en la que está cada k-mer. Escribimos el k-mer y su posición en un fichero de salida.
- A partir de ese fichero de salida, leemos los k-mers encontrados con la posición guardando estos datos en una estructura.
- Para poder agrupar los k-mers, nos conviene ordenar la estructura para que cada k-mer esté al lado de los mismos k-mers.
- Una vex ordenados, vamos guardando el k-mer con todas las posiciones que hay para ese k-mer en otra estructura. Este será el diccionario, que podrá imprimirse en otro fichero de salida.
-

DICCIONARIO EN PYTHON Y CÁLCULO DE HITS:

Para manejarme con mayor soltura, he creado un diccionario en Python (sirviéndome de Biopython para leer los archivos fasta) y además he hecho un programa para calcular los hits entre dos diccionarios previamente creados.

Aquí expongo un ejemplo hecho con dos archivos fasta: Myco01.fasta y Myco04.fasta



k	Hits
2	77642
3	23418
4	7065
5	2092
6	612
7	No resultado

Al aumentar la k, es más difícil que el k-mer coincida muchas veces. Podemos decir que el número de hits va descendiendo casi exponencialmente.

Vamos a comparar otras dos secuencias de Mycoplasma para ver cuales son mas parecidas entre sí:

Organismo		Hits (k=2)
Myco01vsMyco06		79193
Myco04vsMyco06		84416
Myco01vsMyco04		77642

Mycoplasma04 y el Mycoplasma06 son más parecidos, ya que tienen un mayor número de hits.
¿Por qué serán más cercanos el Mycoplasma04 y el Mycoplasma06?
Vamos a llamar a cada Mycoplasma por su nombre verdadero:

Myco01-> Mycoplasma agalactiae

Myco04-> Mycoplasma bovis PG45

Myco06-> Mycoplasma bovis Hubei-1

04 y 06 son más cercanos porque son dos cepas de un Mycoplasma que afecta a los bovinos, en cambio Myco01 es otro Mycoplasma que afecta a las ovejas. Es más lejano.